Travel Mode Choice Prediction Using Imbalanced Machine Learning

Huanfa Chen and Yan Cheng

***Abstract*— Travel mode choice prediction is critical for travel demand prediction, which influences transport resource allocation and transport policies. Travel modes are often characterised by severe class imbalance and inequality, which leads to the inferior predictive performance of minority modes and bias in travel demand prediction. In existing studies, the class imbalance in travel mode prediction has not been addressed with a general approach. Basic resampling methods were adopted without much investigation, and the performance was assessed by commonly used metrics (e.g., accuracy), which is not suitable for predicting highly imbalanced modes. To this end, this paper proposes an evaluation framework to systematically investigate the combination of six over/undersampling techniques and three prediction methods. In a case study using the London Passenger Mode Choice dataset, results show that applying over/undersampling techniques on travel mode substantially improves the F1 score (i.e., the harmonic mean of precision and recall) of minority classes, without considerably downgrading the overall prediction performance or model interpretation. These findings suggest that combining over/undersampling techniques and statistical/machine-learning methods is appropriate for predicting travel mode, which effectively mitigates the influence of class imbalance while achieving high predictive accuracy and model interpretation. In addition, the combination of over/undersampling techniques and prediction methods enriches the model options for predicting mode choice, which would better support transport planning.**

***Index Terms*— Class imbalance, machine learning, oversampling, undersampling, travel mode choice.**

I. Introduction

TRAVEL mode choice prediction is an essential step of travel demand prediction. It affects not only resource allocation in transport planning and operation, but also transport policy-making with goals such as improving mobility and decarbonisation. Travel mode choice prediction aims for accurate aggregate prediction (i.e., prediction of market shares) as well as precise disaggregate prediction (i.e., prediction of modes of individual trips). Traditionally, travel mode prediction is approached using discrete choice models (DCM), including the multinomial logit model and its variants [1], [2]. Recently, there is a growing interest in using machine learning methods for modelling travel mode choice, including support vector machine (SVM), deep neural network (DNN), and extreme gradient boosting (XGB). It is reported that XGB and DNN methods have higher predictive power than discrete choice models in predicting travel mode [3], [4], [5], [6].

In travel mode prediction, the class imbalance between modes has become a common and prominent issue, which leads to the underestimation of the minority class. Due to different levels of transport service provision and transport policies, travel mode choice data are often highly imbalanced [7], i.e., some modes are used much more frequently than others. Table I shows the class imbalance of travel mode choice in literature. The degree of class imbalance is measured by the ratio of the number of trips of the minority mode to the majority mode [8]. In most datasets, this degree is less than 0.1, which indicates a high level of imbalance. The class imbalance would severely compromise the model estimation and predictive performance, as the model tends to focus on the majority class whilst ignoring the minority class [9]. Nevertheless, to the best of our knowledge, class imbalance in travel mode choice prediction has not received adequate attention and has not been well tackled. This study aims to deepen the understanding of whether and how mode choice imbalance can be tackled.

The rest of the paper is structured as follows. Section II starts by surveying the methods used in previous studies to tackle class imbalance in travel mode prediction and the evaluation metrics for assessing the performance of travel mode prediction; Section III firstly specifies the workflow of this paper, then briefly introduces the selected travel mode prediction models and over/undersampling (OUS) techniques to be combined, followed with a comprehensive evaluation framework proposed to assess the prediction using highly imbalanced dataset; Section IV describes the London Passenger Mode Choice dataset and variables used for prediction, as well as the setup of experiments; Section V evaluates the model performance using the framework proposed in Section II and discuss the suitability of various combinations. Finally, Section VI concludes this paper and proposes future research directions.

II. Literature Review

A. Methods Used to Tackle Class Imbalance in Travel Mode Prediction

Although the issue of class imbalance is one common challenge of predicting travel mode choices, which may cause

Manuscript received 26 August 2021; revised 25 March 2022 and 8 September 2022; accepted 29 December 2022. The Associate Editor for this article was T. Q. Dinh. *(Corresponding author: Yan Cheng.)*

Huanfa Chen is with the Centre for Advanced Spatial Analysis, University College London, W1T 4TJ London, U.K.

Yan Cheng is with the Key Laboratory of Road and Traffic Engineering of the Ministry of Education, Tongji University, Shanghai 200092, China, also with the Shanghai Key Laboratory of Rail Infrastructure Durability and System Safety, Tongji University, Shanghai 201804, China, and also with the Centre for Transport Studies, University College London, WCE1 6BT London, U.K. (e-mail: yan_cheng@tongji.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TITS.2023.3237681, provided by the authors.

Digital Object Identifier 10.1109/TITS.2023.3237681

TABLE I
CLASS IMBALANCE OF TRAVEL MODE CHOICE IN LITERATURE

Literature	Modes [a]	Majority mode	Max mode share (%)	Minority mode	Min mode share (%)	Standard deviation	Degree of class imbalance
Richards and Zill [5]	W, B, PT, C	C	80.00	B	1.50	31.95	0.019
Wang and Ross [6]	W, B, PT, C	C	83.21	B	0.98	33.78	0.012
Hagenauer and Helbich [10]	W, B, PT, C	C	52.28	PT	2.32	17.86	0.044
Kim [12]	W, B, PT, C	W	43.70	B	2.5	15.85	0.057
Zhao et al. [36]	W, B, PT, C	PT	35.28	C	14.89	7.76	0.422
Lee, Derrible and Pereira [57]	W, B, PT, C	PT	51.86	B	4.03	18.64	0.078
Cheng *et al.* [58]	W, B, PT, C, EM	EM	25.78	B	14.54	4.06	0.564
Zhang and Xie [59]	W, B, PT, C1, C2, C3	C1	72.30	B	1.00	25.12	0.014
Omrani *et al.* [60]	W&B, PT, C	C	76.83	W&B	6.7	31.01	0.087
Sekhar, Minal and Madhu [61]	W, B, Two-wheeler, Bus, Metro, C, Auto Rickshaw	C	36.00	B	0.60	12.74	0.017
Chang *et al.* [62]	W, B, Motorcycle, C(D), C(P), Bus, LRT, Rail, Other	C	32.35	Other	0.67	10.83	0.021

[a]W = Walking, B = Bike/Cycling, EM = E-motorcycle, PT = Public transport/Transit, C = Car/Driving, C(D) = Car driving, C(P) =Car passenger, C1 = Driving alone, C2 = Shared ride with two people, C3 = Shared ride with three people or more.

inferior performance for predicting modes with smaller shares [6], [7], [10], [11], [12], [13], limited efforts have been made to solve this problem. Hagenauer and Helbich [10] tried to deal with the class imbalance in travel modes by randomly oversampling the minority class and undersampling the majority class when pre-processing data. However, whether data over/undersampling improves the prediction is poorly understood because the prediction performance on the original and processed datasets was not compared. Pirra and Diana [11] adopted a modified SVM method that assigns different weights to different classes in the decision function of SVM, and found out this method outperformed the plain SVM. Qian et al. [14] introduced adjusting kernel scaling in developing an SVM model and found it improved the accuracy of the minority class classification in some cases. However, both methods are specifically designed for SVM and do not generalise to other machine learning models. Kim [12] used a class-specific weighting scheme in which each instance is assigned weights that are inversely proportional to the frequency distribution of classes. However, this approach treats all instances of a class as equally important to the classification. None of them proposed a general approach for addressing mode class imbalance, and there is a lack of systematic investigation into how and to what extent class imbalance can be tackled.

B. Evaluation Metrics to Assess the Performance of Travel Mode Prediction

Most studies on travel mode prediction adopt only one or multiple overall performance metrics, such as accuracy [15], [16], [17], recall (or sensitivity) [10], and log-loss [18]. These metrics are insufficient for highly imbalanced mode distribution because they ignore class-specific performance. Specifically, when the data is highly imbalanced, these overall metrics can be achieved by a trivial classifier that always predicts the most likely class. Furthermore, most studies only use metrics based on discretising the classification by assigning each prediction to the class with the highest probability. This is inadequate for imbalanced data because it is highly likely to result in non-representative mode shares. Therefore, an evaluation framework that includes metrics representing overall and mode-specific, aggregate and disaggregate performance of travel mode prediction is imperative. Rezaei et al. [19] tried to evaluate the impact of resampling techniques on the performance of logit models. However, machine learning models were not considered, and the research only investigated the sign and magnitude of the coefficients when carrying out behavioural analysis.

In summary, the tackling of the class imbalance in travel mode prediction remains unexplored. In the machine learning community, several techniques have been proposed to tackle class imbalance: over/undersampling the original dataset [20], [21], [22], [23], cost-sensitive learning [24], [25], [26], [27], [28], active learning [29], [30], [31], [32], and kernel-based methods [33], [34]. Among them, over/undersampling (OUS) is a straightforward and effective method for the imbalance problem and can be applied to a wide range of classifiers. This study investigates whether OUS techniques can enhance travel mode choice prediction by testing and comparing various combinations of OUS techniques and statistical or machine-learning methods with a comprehensive evaluation framework.

This study contributes to the literature on travel mode prediction as follows. Firstly, it proposes a comprehensive and multifaceted evaluation framework for travel mode prediction, which entails overall model performance, mode-specific performance, and model interpretation. Secondly, it presents a systematic investigation of over/undersampling techniques for tackling class imbalance in travel mode prediction. Thirdly, it verifies that it is viable and efficient to combine over/undersampling techniques and statistical/machine-learning models for predicting travel mode, which mitigates the influence of mode imbalance while achieving high predictive accuracy and model interpretation. This approach can inform transport planning and effectively avoid bias in travel demand prediction.

III. METHODOLOGY

This paper aims to investigate the impact of over/undersampling techniques on travel mode prediction. We firstly introduced three prediction methods and six

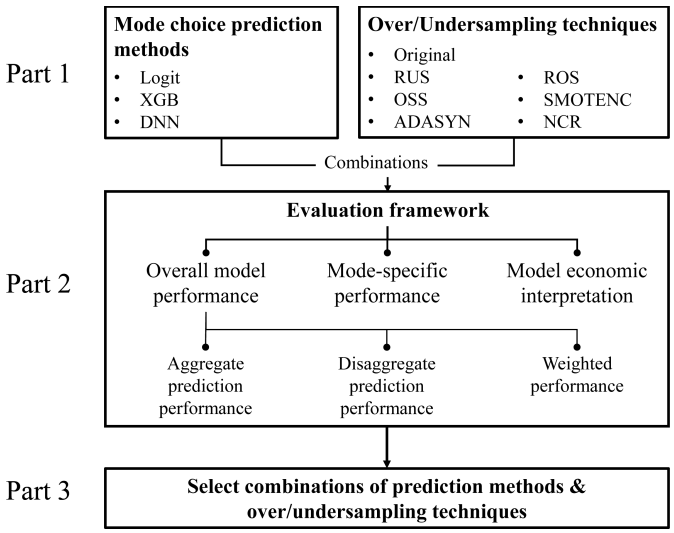


Fig. 1. Workflow of this study.

TABLE II
LIST OF ABBREVIATIONS AND ACRONYMS

Abbreviation	Full form
MADMS	Mean absolute deviation of market share
DCM	Discrete Choice Model
MNL	Multinomial Logit
XGB	Extreme Gradient Boosting
DNN	Deep Neural Network
OUS	Over/Undersampling
RUS	Random Undersampling
OSS	One-sided selection
NCR	Neighbourhood Cleaning Rule
ROS	Random Oversampling
SMOTENC	Synthetic Minority Over-sampling Technique – Nominal and Continuous
ADASYN	Adaptive Synthetic Sampling approach

over/undersampling (OUS) techniques to be investigated. The prediction methods include one traditional discrete choice model and two advanced machine learning models. Then we proposed a comprehensive evaluation framework for assessing the model performance of travel mode prediction on highly imbalanced travel datasets. Different combinations of prediction models and OUS techniques were evaluated and the best-performing combinations were selected and discussed, as shown in Fig. 1. Table II presents the list of abbreviations and acronyms used in this paper.

A. Travel Mode Choice Models

1) Logit Models: Logit models assume that passengers would choose a mode from a set of alternatives to maximise their utility. Under the random utility theory [35], logit models assume that each mode has a certain level of utility that consists of two components: a component representing the effects of observed explanatory variables (e.g., travel time, cost) and a random error reflecting the effects of unobserved variables. The utility of choosing mode i is:

$$U_{ni} = V_{ni} + \varepsilon_{ni} \ = \beta' \mathbf{x_{ni}} + \varepsilon_{ni} \quad (n \in N; i \in M_n) \quad (1)$$

where M_n is the set of available modes for trip n; N is the total number of trips; U_{ni} is the utility of alternative travel mode i for trip n; V_{ni} is the representative utility of alternative travel mode i for trip n; $\mathbf{x_{ni}}$ is a $1\times$ K vector of explanatory variables of alternative mode i for trip n; $\boldsymbol{\beta}$ is a $K\times1$ vector of coefficients of variables representing the weights attached to explanatory variables for trip i; and ε_{ni} is the random error of travel mode i for trip n. Different types of logit models are developed by specifying different types of random errors and choices of coefficients of explanatory variables. Most notably, the MNL model is formed when the error term is independently, identically Gumbel-distributed. In the MNL, the probability of trip n to choose travel mode i is given by (2). The coefficients of the MNL can be estimated using the maximum likelihood method.

$$\widehat{P}_{ni} = \frac{e^{V_{ni}}}{\sum\limits_{j \in M_n} e^{V_{nj}}} \, (i, j \in M_n; n \in N) \quad (2)$$

2) Machine Learning Models: Machine learning models consider mode choice prediction as a classification problem, i.e., given input variables, predicting the most likely mode and/or the probability of all alternatives. The objective is to learn a target function that maps input variables to the output target. A range of machine learning models have been used to predict travel mode choice, which include tree-based models, Naïve Bayes, support vector machine, and neural network [36]. Notably, the tree-based ensemble model (represented by extreme gradient boosting) and DNN have been attracting interest because of their high predictive power and capability of estimating choice probability [3], [4], [5].

Extreme gradient boosting (XGBoost) [37] is an efficient and scalable ensemble approach that uses decision trees as base predictors. The XGBoost is trained in an additive manner by starting from a low-accuracy decision tree and iteratively building trees to minimise a loss function. In each iteration, the instances that are misclassified by existing trees are given more weight. The final prediction of XGBoost is based on the weighted votes of base predictors, where the weight of a predictor is proportional to its predictive accuracy. XGBoost has proved suitable for mode prediction, due to the high predictive accuracy, robustness, interpretability, and ability to derive well-calibrated choice probabilities [4], [38].

Deep neural network (DNN) is an Artificial Neural Network (ANN) with multiple layers between the input and output layers. The DNN can model complex non-linear relationships between variables as the data goes through the weighted connections between DNN layers. The output of the DNN consists of k units corresponding to k classes of mode choice. Moreover, DNNs can reveal utility functions and behavioural patterns when applied to mode choice analysis [39]. Because of their extraordinary predictive power and satisfactory interpretability, DNNs have been adopted in transportation studies, including predicting travel mode, route choice, and automobile ownership.

B. Oversampling and Undersampling Techniques

Oversampling and undersampling techniques adjust the class distribution by replicating or synthesising samples in minority classes or by removing samples in majority classes. These techniques can be combined with various prediction

methods and are likely to tackle imbalance in travel mode prediction. Note that in prediction tasks involving a training and testing set, a good practice is to apply oversampling and undersampling to only the training set, not the testing set. This guarantees fair and unbiased model evaluation on the testing set. It is noteworthy that oversampling and undersampling fundamentally differ from sampling or resampling in statistics. Statistical sampling refers to extracting a subset of individuals from the population to infer characteristics of the whole population, and the extracted sample is expected to follow the distribution of the population.

In this study, six OUS techniques were selected and compared, as these methods represent the state-of-the-art sampling-based solutions for imbalanced data [40], [41]. These methods include two basic methods (RUS and ROS) and advanced methods because of their good performance in existing studies.

1) Undersampling Methods: Random undersampling method (RUS) works by randomly removing instances in major classes until the predefined class balance is achieved. This method is straightforward and efficient, with no assumptions about the data distribution. However, its major drawback is that potentially useful instances can be removed. In order to tackle this problem, new undersampling techniques have been proposed that identify and remove redundant, noisy and/or borderline instances from majority classes. Specifically, redundant instances are points that add little information about the majority classes, while noisy instances represent randomness in the data. Borderline instances are close to the boundary between classes and are unreliable as small changes to borderline instances' attributes would lead to considerable shifting of the decision boundary [42].

As one of the advanced undersampling approaches, One-Sided Selection (OSS) [20] combines Condensed Nearest Neighbour (CNN) (for removing redundant instances) and Tomek Links (for removing borderline/noisy instances). In Step 1, let S be the original data, a subset C is generated that contains all instances of the minor classes and a randomly selected majority instance. Then, for each instance in S, it is classified using its nearest neighbour in *C.* The misclassified instances are added into C. In this way, C does not contain redundant instances that are correctly classified by its nearest neighbours. In Step 2, minority class instances that belong to Tomek Links are removed from C. Tomek Links [43] can be briefly explained as follows: a pair of instances a and b is a Tomek Links if three criteria are met: (i) a and b belong to different classes, (ii) a's nearest neighbour is *b,* and (iii) b's nearest neighbour is *a.* By definition, instances that belong to Tomek Links are either noisy or boundary instances. The resulting set C is the output of OSS.

Another undersampling approach is the Neighbourhood Cleaning Rule [44], which adopts the rule of Wilson's Edited Nearest Neighbours (ENN) [45] to eliminate noisy/borderline major class instances. In NCR, the three nearest neighbours of each instance a are computed and used to classify a. If a is from the majority classes and is misclassified by the three nearest neighbours, a is removed as it is considered as a borderline/noisy instance. If a is a minor class instance and is misclassified by its nearest neighbours, then the majority class instances within a's nearest neighbours are removed.

2) Oversampling Methods: In random oversampling, minority class instances are randomly selected and repeated in the data until a balanced class distribution is obtained. It is subject to overfitting on the training data and thus fails to generalise to the unseen dataset. To avoid overfitting, more advanced oversampling approaches have been proposed that smartly create synthetic instances of minority classes. Among these approaches are Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic Sampling approach (ADASYN).

SMOTE [46] works by firstly selecting a minority class instance a at random and finding its k-nearest minority class instances called S. An instance b is randomly picked up from *S.* Then, a synthetic instance is generated as a weighted combination of a and b. This approach is plausible as new synthetic instances are generated from pairs of minority class instances that are sufficiently close.

Although SMOTE proves effective for continuous data, it is not applicable for data consisting of continuous and nominal variables. One such example is the travel survey data which include continuous variables of travel duration and cost, as well as nominal variables of gender, trip purpose, and trip mode. To deal with a mixture of continuous and nominal attributes, we use a variant of SMOTE, SMOTE-Nominal Continuous (SMOTE-NC for short). It differs from the SMOTE in two aspects. Firstly, the distance between two instances a and b consists of two components, namely the difference of continuous variables and the penalty term of differing nominal variables. The penalty term is defined as the median of standard deviations of all continuous variables across the minority class. Secondly, in the generation of synthetic instances, while the continuous variables are interpolated using the same procedure as SMOTE, each nominal variable is given the most frequent value in the k-nearest neighbours.

Alternatively, ADASYN [8] adaptively generates minority class instances based on the difficulty level of classifying original minority class instances. Specifically, it tends to generate synthetic instances close to the original instances that are incorrectly classified by a k-nearest neighbours classifier. It uses the same procedure as SMOTE-NC to generate synthetic instances by interpolation. Through this strategy, ADASYN increases the density of hard-to-classify minority class instances close to the borderline and then improves the classification performance.

C. Evaluation Framework

We proposed a comprehensive evaluation framework to systematically assess the performance of mode choice prediction from three aspects: overall model performance, mode-specific performance, and model economic interpretation, as shown in Part 2 of Fig. 1. The metrics of this framework not only assess how well the market share is predicted, but also how accurately the modes of individual trips are predicted. Furthermore, the prediction performance of each mode was discussed to explicitly show the impact of imbalance and how OUS techniques tackle this problem. Using economic interpretation metrics, we validated that applying OUS techniques will not distort travellers' behaviour patterns in mode choice prediction.

This framework is applicable to both logit models and machine learning models.

1) Overall Model Performance: The overall performance refers to the model's predictive power for the entire dataset that consists of multiple travel modes. Specifically, the overall performance includes three aspects, namely aggregate prediction performance, disaggregate prediction performance, and weighted performance.

Aggregate prediction performance of a model concerns the model's capability to reproduce and predict the aggregate choice distribution of each mode, i.e., the market mode share. This performance can be assessed using the mean absolute deviations of market share (MADMS), which is defined as:

$$MADMS = \frac{1}{|M|} \sum_{i \in M} \left| \frac{1}{|N_i|} \sum_{n \in N_i} \left(\widehat{P}_{ni} \right) - P_i \right| \tag{3}$$

where M is the set of travel modes; N_i is the set of trips that choose travel mode i; $|\cdot|$ is the cardinality function that outputs the number of elements in a set; $\widehat{P}_{ni}$ is the predicted choice probability of travel mode i for trip n; P_i is the actual market share of travel mode i. The MADMS metric is similar to the L1-norm error for mode share prediction [36], which is defined as the sum of the absolute differences between the predicted and actual market share predictions.

On the other hand, disaggregate prediction performance concerns the model's ability to accurately predict the mode of each trip record. In literature, this performance has been evaluated by a range of metrics, including accuracy, precision, recall, and F1 score. Herein, we mainly use accuracy and macro-average F1 score to evaluate the disaggregate prediction performance.

The metrics of accuracy, F1 score, and others are based on a summary of individual predictions. Given the predicted and actual labels, the prediction of each class can be summarised by a confusion matrix (see Fig. 2), where columns and rows represent predicted and actual labels, respectively. In this matrix, True Positive (TP) and True Negative (TN) are the numbers of positive and negative examples that are correctly classified, respectively, while False Positive (FP) and False Negative (FN) are the numbers of actually negative and positive examples that are incorrectly classified, respectively.

The accuracy of travel mode prediction is defined as the proportion of accurately predicted trip records to the total number of records, as below:

$$ACC = \frac{1}{|N|} \sum_{i \in M} (TP_i + TN_i) \tag{4}$$

where N is the set of all trips; M is the set of travel modes; TP_i and TN_i are the frequency of True Positive and True Negative instances of travel mode i, respectively.

Precision and recall are two common metrics to measure the predictive performance of each class, and both range from 0 to 1. Specifically, precision is the proportion of true positive predictions in the total positive predictions, while recall is the proportion of positive predictions that are correctly identified. They are defined as:

	Predicted Class *i*	**Predicted Not Class *i***
Actual Class *i*	True Positive (TP)	False Negative (FN)
Actual Not Class *i*	False Positive (FP)	True Negative (TN)

Fig. 2. Confusion Matrix.

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \tag{5}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{6}$$

where FP_i and FN_i are the frequency of False Positive and False Negative instances of travel mode i, respectively.

There is often a trade-off between precision and recall, meaning that improving one metric would lead to the reduction of the other. For this reason, F1 score has been proposed to reconcile both metrics, which is defined as the harmonic mean of precision and recall. The F1 score of travel mode i is expressed as:

$$F1_i = \frac{2}{Precision_i^{-1} + Recall_i^{-1}} = \frac{TP_i}{TP_i + \frac{1}{2}(FP_i + FN_i)} \tag{7}$$

In this study, F1 score is used in two ways. First, we use the F1 score of each mode to describe the mode-specific predictive performance. A higher F1 score indicates a better predictive performance of the mode. Second, we use the macro-average F1 score to describe the overall disaggregate predictive performance, which is the average F1 score across all modes, as shown in (8). Likewise, the higher the macro F1 score, the better the overall predictive performance.

$$MacroF1 = \frac{1}{M} \sum_{i \in M} F1_i \tag{8}$$

A main challenge in model evaluation and comparison is the conflict between different metrics. In other words, it is often impossible to simultaneously achieve the best performance in all metrics. Therefore, we propose a weighted method to combine the metrics into an overall score. All three metrics are firstly standardised by min-max scalar to keep the same scale, then each metric is multiplied with a weight. The sum of three weights equals one, with each of them ranging in value between zero and one. The overall score is defined as the sum of weighted standardised metrics in Equation (9):

$$\begin{aligned} E = &-W_1 \times S(MADMS) + W_2 \times S(ACC) \\ &+ W_3 \times S(MacroF1) \end{aligned} \tag{9}$$

where $S(\cdot)$ represents the standardisation procedure.

There are different approaches to determining the weights. First, the weights can be selected based on expert knowledge or the need for real-world applications. Second, if there is limited prior knowledge of the weights, it is recommended to try a range of weight values in a sensitivity test.

2) Mode-Specific Performance: The mode-specific performance refers to the model's predictive power for each travel mode. The mode-specific performance may vary significantly across modes, especially when the mode frequency is highly imbalanced. Many studies paid much attention to the overall model performance while ignoring the impact of imbalance on each travel mode. Here, we measured the mode-specific performance by the F1 score of each mode (as discussed above), which provides a detailed understanding of the model's capability and would reveal whether OUS improves the prediction of each mode.

3) Model Economic Interpretation: A well-performed prediction model for travel mode should have not only high predictive power but also accurate and reliable economic information regarding travel behaviours. In this context, the economic information includes the marginal effect and elasticity of travel modes regarding input variables, the value-of-time in different modes, and the substitution pattern of alternatives [39]. Interpreting the economic information of logit models is straightforward. Recent studies show that machine learning models can readily provide as reliable economic information as logit models [39].

In this study, the focus of the model interpretation is whether using OUS on the data would alter the economic information in the prediction models. To achieve this, we calculate the average elasticities of four travel modes with regard to travel duration or cost. Elasticity (also known as the standardised derivative) measures the per cent changes in the choice probability of a mode as a result of one per cent change in an input variable. Mathematically, it is defined as:

$$E_{ik} = \frac{1}{|N|} \sum_{n \in N} \frac{\partial \widehat{P}_{ni}}{\partial x_{nik}} \frac{x_{nik}}{\widehat{P}_{ni}} \tag{10}$$

where E_{ik} is the average elasticity of travel mode i with regard to the k_{th} variable; N is the total number of trips; $\widehat{P}_{ni}$ is the predicted choice probability of trip n of choosing travel mode i; x_{nik} is the k_{th} variable of travel mode i for trip n.

A positive elasticity means that an increase in the input variable leads to an increase in the choice probability of the given mode, while a negative value means an increase in the variable causes a decrease in the choice probability. We note that there are other metrics for economic information in travel behaviours. A comprehensive discussion of behaviour analysis in mode choice is available in Wang et al. [39].

IV. Data and Setup of Experiments

A. Data

The dataset of London Passenger Mode Choice (LPMC) from April 2012 to March 2015 [4] was used in this study. This dataset was derived from the London Travel Demand Survey (LTDS), an annual survey that captures a detailed snapshot of journeys made by every over-five-year-old member of the selected household on a selected day. The key steps that generate the LPMC dataset from LTDS include: (1) removing the trips that had the same postcode in origin and destination; (2) assigning each trip to one of the four travel modes; (3) simplifying the trip purposes to five main purposes; (4) adding travel time and cost information of four modes to LTDS by utilising Google Map API and Oyster cards. The resultant LPMC dataset contains 81,086 trips generated by 31,954 individuals across 17,616 households. The four main travel modes accounting for 99.5% of trips are walking, cycling, public transport, and driving (including car passenger, taxi, van and motorbike). Table III indicates that the mode shares are considerably stable between 2012 and 2015. Driving accounts for more than 40% of total trips, followed by public transport accounting for 35% of trips. In contrast, less than 3% of trips use cycling. The large difference between the major and minor modes reveals the severe class imbalance in mode choice, which is consistent with the mode choice data mentioned above.

LPMC dataset contains a wide range of variables about the household (e.g., household members, car ownership), individual (e.g., gender, age, ticket types) and trip (e.g., trip purpose, departure time, travel mode, trip duration and cost of alternatives). Compared with LTDS, which provides only the trip information of the chosen mode, one big improvement of LPMC is that it provides the trip cost and duration of all four alternative modes, which is estimated using an online directions service. We selected 14 variables for this study (see Table IV), which are in line with Wang et al. [39]. As the cost of walking and cycling is zero for all trips, they were not included in the list.

The duration and cost of travel modes are used differently in discrete choice and machine learning models. In discrete choice models, the duration and cost of a mode are only used in the corresponding utility function. In contrast, in machine learning, the duration and cost of all modes are fed into the algorithm, and then the algorithm automatically determines the variables for building models.

B. Setup of Experiments

To gain insight into the impact of class imbalance and different OUS techniques, we tested 18 combinations of three mode prediction models and six OUS techniques, as mentioned in Section 2. These combinations were compared with the models using the original dataset. The computation was conducted on a Windows 10 desktop (Intel i7 CPU, 3.1 GHz with 15 GBytes memory). The logit and machine-learning models were constructed and trained in Python using the packages listed in Table V.

There are two sources of randomness in the mode prediction: first, applying OUS techniques introduces randomness; second, the model training of DNN and XGB involves randomness, as the model training may identify local minima rather than global minima, which is called model non-identification challenge of machine learning. Therefore, each combination of models and OUS techniques is assessed 100 times and

TABLE III
MODE SHARES OF SURVEY YEAR 2012/13-2014/15 IN LPMC

Travel mode	Survey year 2012/13		Survey year 2013/14		Survey year 2014/15		Total	
	Trips	%	Trips	%	Trips	%	Trips	%
Walking	4969	17.60	4615	18.08	4684	16.91	14268	17.80
Cycling	757	2.97	787	2.75	861	2.88	2405	3.27
Public transport	9669	35.28	9435	35.19	9501	34.58	28605	36.10
Driving	12083	44.16	12451	43.97	11274	45.63	35808	42.83
Total	27478	100.00	27288	100.00	26320	100.00	81086	100.00

TABLE IV
EXPLANATORY VARIABLES FOR TRAVEL MODE CHOICE MODELLING

Category	Attribute name	Description	Variable types and values
Household-related	car_ownership	Car ownership status of household	Nominal: 0 - no cars in household; 1 - less than one car per adult; 2 - one or more cars per adult.
Individual-related	gender	Whether the person is female	Nominal: 0 - Male, 1 - Female.
	age	Age of the person making the trip in years	Continuous
	driving_license	Whether the person has a driving licence	Nominal: 0 - No, 1 - Yes.
Trip-related	distance	Straight line distance between trip origin and destination in metres	Continuous
	dur_walking	Predicted duration of walking route in hours	Continuous
	dur_cycling	Predicted duration of cycling route in hours	Continuous
	dur_pt_access	Predicted total access and egress time for public transport route in hours	Continuous
	dur_pt_inv	In-vehicle time for public transport route in hours	Continuous
	dur_pt_int_total	Interchange time on public transport route in hours	Continuous
	dur_driving	Predicted duration of driving route in hours	Continuous
	pt_n_interchanges	Number of interchanges on public transport route	Integer
	cost_pt	Estimated cost of public transport route in GBP	Continuous
	cost_driving	Estimated cost of driving route in GBP	Continuous

TABLE V
PYTHON LIBRARIES USED FOR BUILDING MACHINE LEARNING MODELS

Python package	Ver.	Use
biogeme [63]	3.2.6	MNL model estimation and validation
TensorFlow [64]	1.13	DNN model building
scikit-learn [65]	0.22	Machine learning model training and validation
hyperopt [66]	0.2.4	Model selection and hyperparameter tuning of XGB
imbalanced-learn [67]	0.5.0	Implementing undersampling and oversampling

the average metric is used. Specifically, the OUS is applied ten times, which generates ten datasets; for each dataset, the prediction model is repeated ten times.

We use holdout sample testing in order to emulate the real-world application of predicting future trips and to avoid data leakage. The dataset is split into a training set (April 2012–March 2014, totalling 54,766 instances) and a hold-out testing set (April 2014–March 2015, totalling 26,230 instances), which is consistent with the data splitting in Hillel et al. [4]. While the training set is used for model optimisation and final model training, the testing set provides an unbiased performance evaluation of final models.

Regarding model optimisation, we used the optimum hyper-parameters of the Opt-DNNs in Wang et al. [34] without further tuning. This is reasonable as both studies use the London dataset provided by Hillel et al. [4]. On the other hand, we tuned the hyperparameters of XGB using the sequential model-based optimisation algorithm (also known as the Bayesian optimisation) via the hyperopt library (using 100 iterations). The XGB hyperparameters were optimised on the original dataset without OUS techniques. The optimal hyperparameters of DNN and XGB are shown in Table VI.

V. RESULTS AND DISCUSSION

A. Overall Model Performance

Figs. 3-5 show how the three metrics (MADMS, accuracy and Macro F1 score) varied for each combination of travel mode prediction models and datasets. The details of the three metrics are available in Appendices A and B. Overall, machine learning models outperform the MNL models, with the DNN models showing the best aggregate predictive performance while XGB models have the best disaggregate predictive performance.

When aggregate predictive performance is considered, all models achieved better performance on the original dataset than on the resampled data. The DNN model had the lowest MADMS (0.0040), followed by the MNL and XGB models with comparable performance. The advanced undersampling techniques (i.e., OSS and NCR) could keep the MADMS at a low level. On the contrary, RUS and all the oversam-

TABLE VI
THE HYPERPARAMETERS OF RANDOM FOREST AND EXTREME GRADIENT BOOSTING

Model	Hyperparameter	Range	Optimum value
DNN	depth	[1,2,3,4,5,6,7,8,9,10]	6
	width	[25,50,100,150,200]	200
	L_1 penalty constants	[0.1, 10^{-2}, 10^{-3}, 10^{-5}, 10^{-10}, 10^{-20}]	10^{-5}
	L_2 penalty constants	[0.1, 10^{-2}, 10^{-3}, 10^{-5}, 10^{-10}, 10^{-20}]	10^{-3}
	Dropout rates	[0.01, 10^{-5}]	0.01
XGB	max_depth	[1,2,3,4,5,6,7,8,9,10,11]	10
	min_child_weight	[1,2,3,4,5,6,7,8,9,10,11]	1
	gamma	[0.0,5.0] [a]	0.5580
	eta	[0.0,1.0] [a]	0.0087
	n_estimators	[100, 150,200,250,300,350,400,450,500]	300

[a] The hyperparameter is randomly drawn from a continuous distribution.

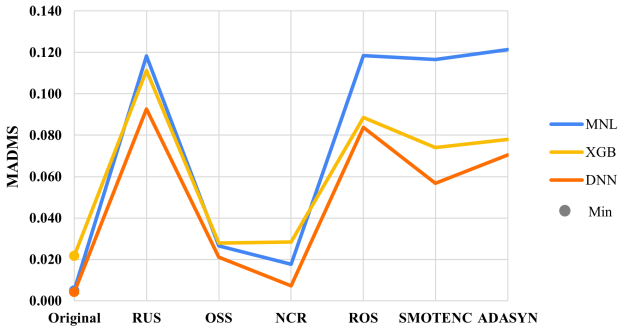


Fig. 3. Mean absolute deviations of market share of different methods and datasets.

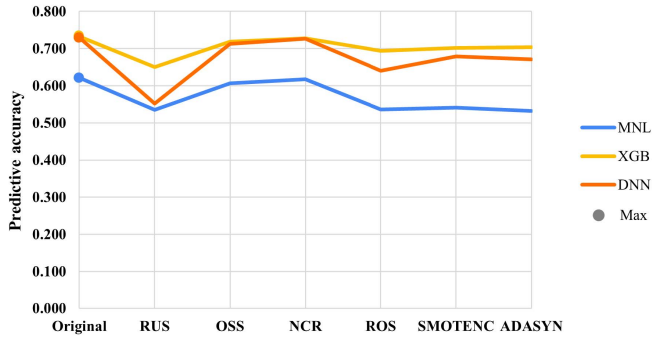


Fig. 4. Predictive accuracy of different methods and datasets.

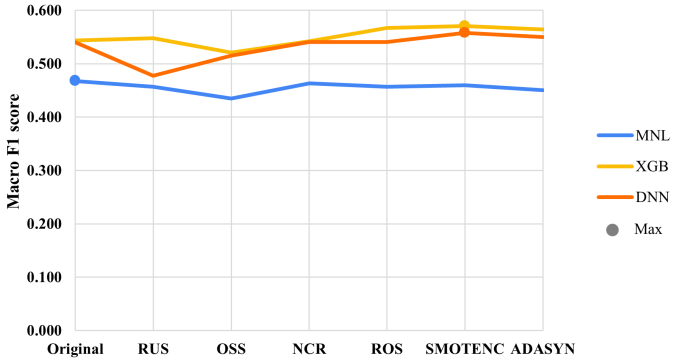


Fig. 5. Macro F1 score of different methods and datasets.

pling methods distort the results of market share. Particularly, the combination of MNL models and RUS or oversampling techniques should not be used if MADMS is the only criterion of mode prediction.

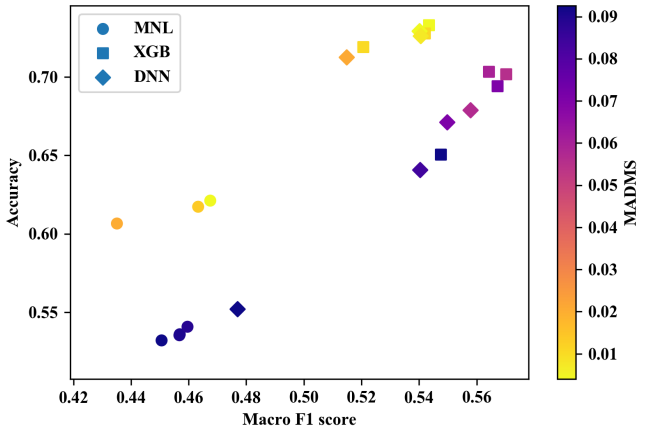


Fig. 6. Overall predictive power.

In terms of the metrics indicating disaggregate prediction performance, the accuracy showed a similar pattern as MADMS. All the models achieved their highest accuracy when the original dataset was used. In contrast, the Macro F1 score demonstrates a different trend. Although the MNL models still performed best when the original dataset was used, both machine learning methods achieved their best Macro F1 score when SMOTENC was used, followed by ADASYN. The highest macro F1 score was 0.5703 when the XGB method and SMOTENC-oversampled dataset were used. Thus, the oversampling techniques showed the capability of improving the Macro F1 score for XGB and DNN models. The trade-off between the three metrics is illustrated in Fig. 6.

The optimal combination of the prediction model and OUS technique depends on the relative importance of these metrics. To this end, we designed ten scenarios with differing weights and relative importance in the three metrics. Table VII presents the optimal and sub-optimal combinations in each scenario. Obviously, the machine learning methods achieve a more balanced performance with the three metrics as 19 out of 20 optimal or sub-optimal combinations were XGB or DNN models. The MNL model was the sub-optimal model only when accuracy and Macro F1-score were neglected in the evaluation (i.e., $W_1 = 1.0$). In addition, the original dataset was among the optimal or sub-optimal combinations when

TABLE VII
THE RECOMMENDED COMBINATIONS IN DIFFERENT WEIGHT SCENARIOS

W1	W2	W3	Highest score	Best combination	Second highest score	Second best combination
1/3	1/3	1/3	0.5862	DNN + Original	0.5730	DNN + NCR
1	0	0	0.0000	DNN + Original	-0.0082	MNL + Original
0.8	0.1	0.1	0.1759	DNN + Original	0.1527	DNN + NCR
0.5	0.25	0.25	0.4396	DNN + Original	0.4229	DNN + NCR
0	1	0	1.0000	XGB + Original	0.9815	DNN + Original
0.1	0.8	0.1	0.8652	XGB + Original	0.8629	DNN + Original
0.25	0.5	0.25	0.6850	DNN + Original	0.6712	DNN + NCR
0	0	1	1.0000	XGB + SMOTENC	0.9769	XGB + ROS
0.1	0.1	0.8	0.8249	XGB + SMOTENC	0.7903	XGB + ROS
0.25	0.25	0.5	0.6339	DNN + Original	0.6249	DNN + NCR

the weights were similar or when the dominant metric is MADMS or Macro F1-score. XGB models combined with SMOTENC or ROS datasets were the optimal or suboptimal combinations when Macro F1 score was the major metric (i.e., $W_3 \geq 0.8$). These results indicate that XGB models with oversampling techniques achieved better disaggregate prediction performance at the cost of inferior performance in the MADMS. Meanwhile, the models with the original dataset had high accuracy and lower MADMS but do not perform well on the Macro F1 score.

The OUS techniques add more flexibility to model selection for predicting travel mode. While machine learning models combined with the original dataset had the best overall performance in most scenarios, the combinations of XGB and oversampling techniques are the best choices if Macro F1 score is the focus of the prediction task.

B. Mode-Specific Prediction

The mode-specific F1 scores provide a better understanding of how different OUS techniques improve mode prediction. Fig. 7 shows that the mode-specific F1 scores of all the travel modes ranged from 0.50 to 0.78, except for cycling. Both machine learning models had higher F1 scores compared with the MNL model, especially for public transport and driving. XGB models performed best not only in Macro F1 score, but also in mode-specific F1 scores.

Notably, the F1 score of cycling is (nearly) zero for the models using the original dataset and the OSS and NCR datasets. This is because very few or no cycling records are correctly predicted. Given that cycling accounts for 3% of the total trips, the severe underprediction of cycling is problematic and unacceptable. This implies that we should be cautious about the overall performance (e.g., Macro F1 score), which might hide the underprediction of minority modes. Therefore, evaluating the prediction performance only at the overall level may be misleading. To avoid this misleading, it is essential to add mode-specific performance into our evaluation framework to enable a deep look into the impact of imbalance on each mode and how OUS techniques tackle this issue.

Another thing to note is that RUS exhibited good prediction performance for cycling, which is different from the other undersampling methods (OSS and OCR). This is because RUS reduces the modes with a higher share and leads to a dataset with equal share of different modes (or with no class imbalance). Similarly, the oversampling techniques (i.e., ROS, SMOTENC, and ADASYN) could mitigate the imbalance in the original dataset. Thus, the issue of underprediction for the minority class was substantially alleviated by using RUS and oversampling techniques, in which the F1 score of cycling is markedly improved in comparison with the original dataset. The implication is that using appropriate OUS techniques could lead to better predictive performance of the minority class (i.e., cycling in this study) without degrading the predictive performance of the other classes.

C. Model Interpretation

This section interprets the behavioural pattern and economic information in the constructed models by computing the elasticities of four travel modes regarding input variables. Specifically, for the seven models that achieved a high overall score (as recommended in Table VII), we calculated the elasticities for trip-related variables, including mode-specific duration, cost, and the number of interchanges in transit, as shown in Table VIII. Figures presenting the elasticities of each mode for recommended combinations are available in Appendix C.

In each panel, each entry represents the average elasticities of all respondents in the testing set, which indicates how much per cent changes in the choice probability of a mode would happen as a result of one per cent change of the corresponding variable. The elasticities of mode choices regarding their mode-specific variables are highlighted in Table VIII. It can be found the average elasticities in the models selected were largely reasonable in terms of signs. The highlighted entries in Table VIII were mostly negative, which is aligned with common sense as the higher travel cost and duration will reduce the probability of selecting the corresponding mode. However, a few exceptions of highlighted positive values did exist. For example, the elasticities of the duration of cycling were positive for the mode of cycling in Panels 2, 5 and 6. This can be attributed to the local irregularity of DNN or model non-identification of XGB and DNN models [39]. Local irregularity refers to that DNN models have locally irregular patterns (i.e., exploding gradients, the lack of monotonicity)

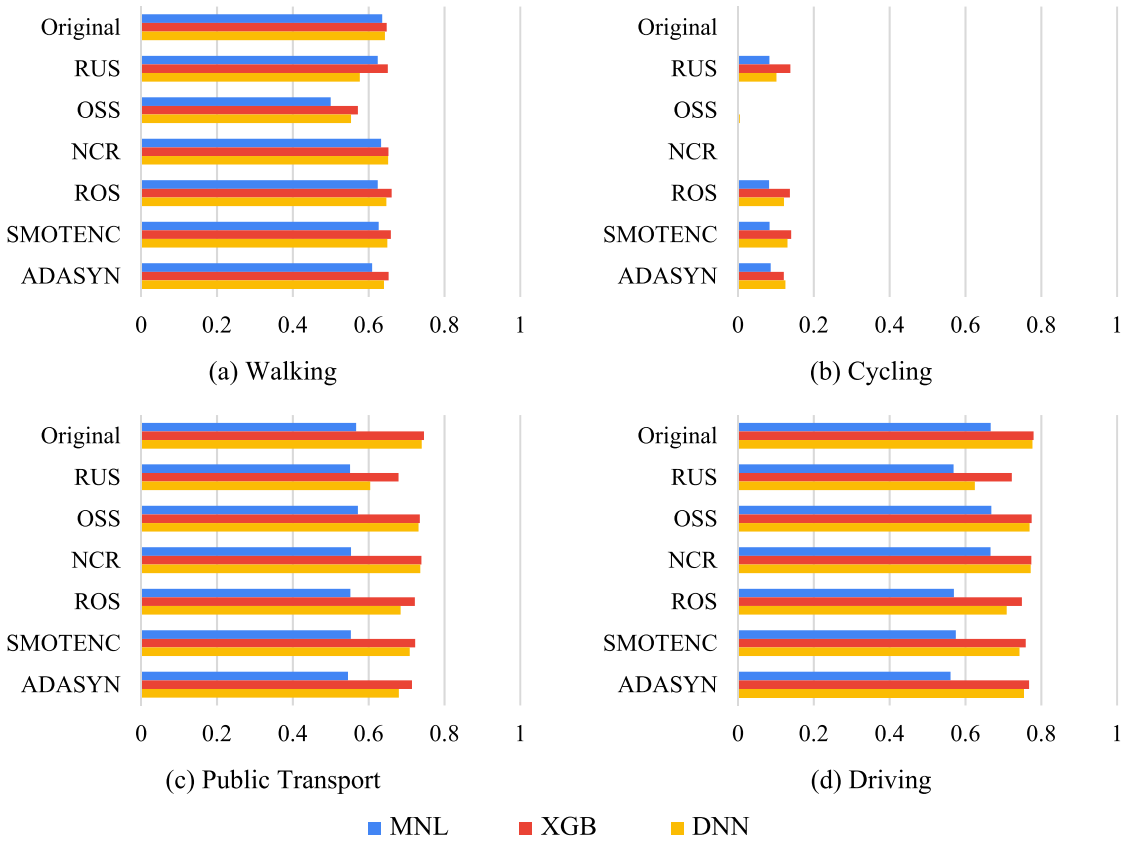


Fig. 7. Mode-specific F1 scores.

such that certain choice behaviours revealed by DNNs are not realistic. On the other hand, the model non-identification of machine-learning models refers to that the objective function of XGB or DNN is not globally convex and that the optimisation of XGB or DNN models may identify local minima or saddle points rather than global minima. In addition, it is worth noting that all highlighted entries in Panel 1 were negative except the number of interchanges in public transit (represented by pt_n_interchanges), which indicates the MNL model had a good performance in travel behaviour analysis.

The magnitudes of average elasticities in the models were mostly valid and consistent with existing studies. Wang et al. [39] reported the elasticities of travel modes in the same dataset using DNN and MNL models, which are very similar to Panels 1 and 5 in Table VIII. Notably, the elasticities of XGBs were much smaller in magnitude than MNLs and DNNs, although the relative magnitudes of the elasticity coefficients in XGBs were similar to those of MNLs and DNNs. For example, Panel 5 indicates that a 1% increase in accessing time, in-vehicle travel time and interchange time of public transit leads to a decrease of 0.55%, 0.34%, and 0.13% probabilities in using public transit, while in Panel 2, the corresponding probability decreases are 0.18%, 0.08%, and 0.05%. Although it is challenging to assess the validity of these results due to a lack of the ground truth of elasticity coefficients, it is indicated that these results were reasonable in relative magnitudes. This implies a need for further machine-learning-based mode choice studies that focus on validating the behavioural outputs [16].

The average elasticities in the models with OUS techniques were consistent with those using original datasets. If we compare the results of XGB models in Panels 2, 3, and 4, the corresponding average elasticities were quite close. Moreover, the elasticities of DNN models in Panels 5 and 6 were largely aligned with those reported by Wang et al. [39]. We can conclude that a combination of OUS techniques and machine learning models leads to models with valid and intuitive travel behaviours and economic information.

D. Limitation

In the LPMC dataset, each trip is labelled as one of the four modes: walking, cycling, public transport, and driving (which includes car passenger, taxi, van, and motorbike). For journeys consisting of multiple modes, the assigned mode is the one that covers the longest distance. This leads to the bias of the travel modes. The mixed mode prediction can be formulated as a multi-label classification [47], which predicts one or multiple labels (from a given label set) for unseen journeys. Another approach is to create more label classes by combining the current four modes (e.g., 'walking-public-transport'); however, the combination would result in 16 classes of travel modes, which is challenging for classification. We expect that the class imbalance issue will exist for both approaches, and therefore the OUS techniques are likely valid for the mixed mode prediction.

TABLE VIII
AVERAGE ELASTICITIES OF FOUR TRAVEL MODES WITH RESPECT TO INPUT VARIABLES

Panel 1: MNL + Original	**Walking**	**Cycling**	**Public Transport**	**Driving**
dur_walking	**-9.6348**	0.3942	0.3942	0.3942
dur_cycling	0.0460	**-1.9530**	0.0460	0.0460
dur_pt_access	0.3123	0.3123	**-0.5617**	0.3123
dur_pt_inv	0.2936	0.2936	**-0.3581**	0.2936
dur_pt_int_total	0.1163	0.1163	**-0.1381**	0.1163
pt_n_interchanges	-0.0079	-0.0079	**0.0087**	-0.0079
cost_pt	0.1230	0.1230	**-0.1631**	0.1230
dur_driving	0.7133	0.7133	0.7133	**-1.0541**
cost_driving	0.0678	0.0678	0.0678	**-0.2105**
Panel 2: XGB + Original	**Walking**	**Cycling**	**Public Transport**	**Driving**
dur_walking	**-0.0342**	-0.0037	-0.0073	0.1610
dur_cycling	0.0541	**0.0813**	0.0648	0.0133
dur_pt_access	0.0507	0.0685	**-0.1846**	0.1882
dur_pt_inv	0.0555	0.0652	**-0.0836**	0.1416
dur_pt_int_total	0.0074	0.0120	**-0.0476**	0.0582
pt_n_interchanges	0.0000	0.0000	**0.0000**	0.0000
cost_pt	0.0786	0.1187	**0.1056**	-0.0589
dur_driving	0.1188	0.1996	0.6441	**-0.2958**
cost_driving	0.0382	0.0450	0.0391	**0.0200**
Panel 3: XGB + ROS	**Walking**	**Cycling**	**Public Transport**	**Driving**
dur_walking	**-0.0248**	-0.1783	0.0616	0.2259
dur_cycling	0.0329	**-0.0291**	0.0913	0.0218
dur_pt_access	0.0468	0.0393	**-0.1533**	0.1711
dur_pt_inv	0.0548	0.0592	**-0.0808**	0.1372
dur_pt_int_total	0.0025	-0.0044	**-0.0398**	0.0601
pt_n_interchanges	0.0000	0.0000	**0.0000**	0.0000
cost_pt	0.0471	0.0348	**0.0643**	-0.0371
dur_driving	0.0614	0.0987	0.4729	**-0.2903**
cost_driving	0.0073	-0.0892	0.1072	**0.0268**
Panel 4: XGB + SMOTENC	**Walking**	**Cycling**	**Public Transport**	**Driving**
dur_walking	**-0.0460**	-0.1252	0.0496	0.1986
dur_cycling	0.0211	**-0.0311**	0.0893	0.0281
dur_pt_access	0.0333	0.0232	**-0.1535**	0.1858
dur_pt_inv	0.0569	0.0670	**-0.0691**	0.1217
dur_pt_int_total	0.0027	-0.0017	**-0.0341**	0.0471
pt_n_interchanges	0.0000	0.0000	**0.0000**	0.0000
cost_pt	0.1234	0.2061	**0.1027**	-0.1061
dur_driving	0.0543	0.0672	0.4539	**-0.2390**
cost_driving	0.0553	0.1554	0.0730	**-0.0294**
Panel 5: DNN + Original	**Walking**	**Cycling**	**Public Transport**	**Driving**
dur_walking	**-1.5014**	-0.9390	-0.3197	0.5960
dur_cycling	-0.1556	**0.0804**	0.3444	-0.1268
dur_pt_access	0.3143	0.1743	**-0.5515**	0.2855
dur_pt_inv	0.1353	0.1211	**-0.3438**	0.2621
dur_pt_int_total	-0.0152	-0.0622	**-0.1285**	0.0810
pt_n_interchanges	0.0162	0.0167	**0.0022**	0.0280
cost_pt	0.2562	0.1708	**-0.1211**	-0.0666
dur_driving	0.4069	0.6187	1.0191	**-0.8120**
cost_driving	0.0404	0.0104	0.0462	**-0.1256**
Panel 6: DNN + NCR	**Walking**	**Cycling**	**Public Transport**	**Driving**
dur_walking	**-1.7818**	-1.2356	-0.1525	0.6833
dur_cycling	-0.0445	**0.1085**	0.2706	-0.0855
dur_pt_access	0.3770	0.2787	**-0.5079**	0.2707
dur_pt_inv	0.2646	0.2858	**-0.2803**	0.2275
dur_pt_int_total	0.0135	-0.0434	**-0.1085**	0.0895
pt_n_interchanges	-0.0059	-0.0090	**-0.0104**	0.0449
cost_pt	0.3011	0.2352	**-0.1576**	-0.0842
dur_driving	0.4822	0.7817	0.9664	**-0.8859**
cost_driving	0.0220	0.0142	0.0399	**-0.1299**

VI. CONCLUSION

Class imbalance is a common and prominent problem in travel mode data, which leads to the underprediction of the minority class in travel mode prediction and causes biases in transport planning and policy-making. Although machine learning methods have obtained a high predictive accuracy in predicting travel modes, the problem of class imbalance has not been adequately discussed and addressed. This paper fills this research gap by proposing an evaluation framework for assessing the performance of travel mode prediction methods and OUS techniques. The contribution of the framework consists of at least two aspects: first, it examines not only the

overall performance of prediction with both aggregate and disaggregate metrics, but also the mode-specific performance that highlights the potential underprediction of minority modes. This framework also incorporates economic interpretation that examines whether the prediction provides valid travel behaviours. Second, because of the conflict between the aggregate and disaggregate metrics, we propose the overall score (i.e., the weighted sum of these metrics) that enables the performance comparison of travel mode prediction in different scenarios.

Using this framework, we conducted a systematic investigation of the combinations of statistical/machine-learning methods (i.e., MNL, DNN, and XGB) and six OUS techniques. It is found that although prediction models with the original dataset had better aggregate prediction performance, most OUS techniques could help improve the disaggregate prediction performance of machine learning models. RUS and oversampling techniques substantially improve the prediction of minority modes whilst keeping the overall prediction performance and model interpretation. On the other hand, the undersampling techniques of OSS and NCR fail to accurately predict the minority mode. Researchers should be careful about the selection of OUS techniques based on the purpose of travel mode choice prediction.

This research suggests that combining OUS techniques and statistical/machine-learning methods is appropriate for predicting travel mode, because it can effectively mitigate the influence of class imbalance while achieving high predictive accuracy and model interpretation. This methodology can effectively avoid bias in travel demand prediction and inform transport policy. For example, cycling is a healthy travel mode and has been advocated by many countries to improve micro-mobility and reduce carbon emissions [48], [49], [50], [51]. Since the outbreak of COVID-19, cycling has become more popular in many countries by substituting public transport in short and medium-distance journeys while keeping social distancing. However, as cycling is much less popular than driving or buses, the travel demand for cycling is often underestimated, which causes further problems in transport resource allocation and policy. While the minority class differs from area to area, a general principle is that no mode should be disadvantaged in prediction because each transport mode benefits some population groups while excluding others [52]. The methodology proposed in this research makes it possible to mitigate class imbalance in travel mode prediction. Moreover, this methodology enriches the model options for predicting mode choice, thereby providing greater flexibility of models for decision-making in transport planning.

The proposed methodology is generalisable to other classification-based transport studies that are subject to class imbalance. Some examples are driving safety risk prediction and driver sleepiness detection [53], [54], [55], where the frequency of incidents and sleepiness is very low and the data distribution is highly imbalanced. The proposed combination of OUS techniques and prediction methods is likely to mitigate class imbalance and improve the prediction for the minority classes. The evaluation framework proposed in this paper can serve to assess whether the class imbalance is addressed.

This research sheds light on several topics that are worth further investigation to improve mode choice prediction. One is the prediction of mixed-mode journeys, as this application is realistic and relevant. Another topic is preference heterogeneity in machine-learning mode choice prediction. While the DNN and XGBoost models in this paper are based on the average effects of mode choice, it would be interesting to look beyond the average effects in order to create models with better performance. Another topic is combining machine learning and causal inference in travel mode choice. While most machine learning models are based on associational relations between variables (e.g., Random Forest and XGBoost), they are subject to spurious correlation and might have limitations in model generalisation. Emerging methods that integrate machine learning with causal inference (e.g., causal forest) [56] might lead to an accurate and robust model for travel mode prediction, which is yet to be developed.

APPENDIX

A. Further Details of Mean Absolute Deviations of Market Share

See Table IX shown in the Supplementary Material.

B. Details of Accuracy and Macro F1 Score

See Tables X and XI shown in the Supplementary Material.

C. Elasticities of Each Mode for Recommended Combinations

See Fig. 8 shown in the Supplementary Material.

ACKNOWLEDGMENT

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

REFERENCES

[1] D. Brownstone, "Discrete choice modeling for transportation," in *Travel Behaviour Research: The Leading Edge*, D. Hensher, Ed. Amsterdam, The Netherlands: Pergamon, 2001, pp. 97–124. [Online]. Available: https://trid.trb.org/view/788683 and http://www.economics.uci.edu/~dbrownst/

[2] J. D. de Ortuúzar and L. G. Willumsen, *Modelling Transport*, 4th ed. Chichester, U.K.: Wiley, 2011.

[3] S. Wang, B. Mo, S. Hess, and J. Zhao, "Comparing hundreds of machine learning classifiers and discrete choice models in predicting travel behavior: An empirical benchmark," Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep., 2021. [Online]. Available: https://arxiv.org/abs/2102.01130

[4] T. Hillel, M. Z. E. B. Elshafie, and Y. Jin, "Recreating passenger mode choice-sets for transport simulation: A case study of London, U.K.," *Proc. Inst. Civil Engineers Smart Infrastruct. Construct.*, vol. 171, no. 1, pp. 29–42, Mar. 2018.

[5] M. J. Richards and J. C. Zill, "Modelling mode choice with machine learning algorithms," in *Proc. Australas. Transp. Res. Forum (ATRF)*, 2019, pp. 1–15.

[6] F. Wang and C. L. Ross, "Machine learning travel mode choices: Comparing the performance of an extreme gradient boosting model with a multinomial logit model," *Transp. Res. Record, J. Transp. Res. Board*, vol. 2672, no. 47, pp. 35–45, Dec. 2018.

[7] T. Hillel, M. Bierlaire, M. Z. E. B. Elshafie, and Y. Jin, "A systematic review of machine learning classification methodologies for modelling passenger mode choice," *J. Choice Model.*, vol. 38, Mar. 2021, Art. no. 100221.

[8] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. IEEE World Congr. Comput. Intell.*, Jun. 2008, pp. 1322–1328.

[9] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data Mining Knowl. Discovery*, vol. 28, no. 1, pp. 92–122, Jan. 2014.

[10] J. Hagenauer and M. Helbich, "A comparative study of machine learning classifiers for modeling travel mode choice," *Exp. Syst. Appl.*, vol. 78, pp. 273–282, Dec. 2017.

[11] M. Pirra and M. Diana, "A study of tour-based mode choice based on a support vector machine classifier," *Transp. Planning Technol.*, vol. 42, no. 1, pp. 23–36, Jan. 2019.

[12] E.-J. Kim, "Analysis of travel mode choice in Seoul using an interpretable machine learning approach," *J. Adv. Transp.*, vol. 2021, Mar. 2021, Art. no. 6685004.

[13] S. Wang, B. Mo, and J. Zhao, "Predicting travel mode choice with 86 machine learning classifiers: An empirical benchmark study," in *Proc. 99th Annu. Meeting Transp. Res. Board*, 2020, pp. 279–296.

[14] Y. Qian et al., "Classification of imbalanced travel mode choice to work data using adjustable SVM model," *Appl. Sci.*, vol. 11, no. 24, p. 11916, Dec. 2021.

[15] U. Gazder and N. T. Ratrout, "A new logit-artificial neural network ensemble for mode choice modeling: A case study for border transport," *J. Adv. Transp.*, vol. 49, no. 8, pp. 855–866, Dec. 2015.

[16] S. Rasouli and H. J. P. Timmermans, "Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates," *Eur. J. Transp. Infrastruct. Res.*, vol. 14, no. 4, Sep. 2014.

[17] M. G. Karlaftis, "Predicting mode choice through multivariate recursive partitioning," *J. Transp. Eng.*, vol. 130, no. 2, pp. 245–250, Mar. 2004.

[18] T. Hillel, M. Bierlaire, M. Elshafie, and Y. Jin, "A new framework for assessing classification algorithms for mode choice prediction," Univ. Cambridge, Cambridge, U.K., Tech. Rep., 2018, pp. 1–5. [Online]. Available: https://transp-or.epfl.ch/heart/2018/abstracts/5446.pdf

[19] S. Rezaei, A. Khojandi, A. M. Haque, C. Brakewood, M. Jin, and C. Cherry, "Performance evaluation of mode choice models under balanced and imbalanced data assumptions," *Transp. Lett.*, vol. 14, no. 8, pp. 920–932, Sep. 2022, doi: 10.1080/19427867.2021.1955567.

[20] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. ICML*, vol. 97, 1997, pp. 179–186.

[21] N. Japkowicz, "The class imbalance problem: Significance and strategies," *Proc. Int. Conf. Artif. Intell.*, 2000, pp. 111–117.

[22] D. D. Lewis, "A comparison of two learning algorithms for text categorization 1 introduction 2 text categorization: Nature and approaches," *Proc. 3rd Annu. Symp. Doc. Anal. Inf. Retr.*, 1994, pp. 1–14.

[23] C. X. Ling and C. Li, "Data mining for direct marketing: Problems and solutions," in *Proc. 4th Int. Conf. Knowl. Discovery Data Mining*, vol. 98, 1998, pp. 73–79.

[24] M. Pazzani, C. Merz, P. Murphy, K. Ali, T. Hume, and C. Brunk, "Reducing misclassification costs," in *Mach. Learn. Proc.*, Amsterdam, The Netherlands: Elsevier, 1994, pp. 217–225.

[25] P. Domingos, "MetaCost: A general method for making classifiers cost-sensitive," in *Proc. 5th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 1999, pp. 155–164.

[26] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. Int. Joint Conf. Artif. Intell.*, vol. 17, no. 1, 2001, pp. 973–978.

[27] K. M. Ting, "An instance-weighting method to induce cost-sensitive trees," *IEEE Trans. Knowl. Data Eng.*, vol. 14, no. 3, pp. 659–665, May 2002.

[28] Z.-H. Zhou and X.-Y. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 1, pp. 63–77, Jan. 2006.

[29] F. Provost, "Machine learning from imbalanced data sets 101," in *Proc. AAAI Workshop Imbalanced Data Sets*, vol. 68, 2000, pp. 1–3.

[30] S. Ertekin, J. Huang, L. Bottou, and C. Lee Giles, "Learning on the border: Active learning in imbalanced data classification," in *Proc. Int. Conf. Inf. Knowl. Manag.*, 2007, pp. 127–136.

[31] S. Ertekin, J. Huang, and C. L. Giles, "Active learning for class imbalance problem," in *Proc. 30th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. (SIGIR)*, vol. 7, 2007, pp. 823–824.

[32] N. Abe, "Invited talk: Sampling approaches to learning from imbalanced datasets: Active learning, cost sensitive learning and beyond," *Proc. ICML Workshop, Learning Imbalanced Data Sets*, vol. 22, 2003.

[33] G. Wu and E. Y. Chang, "KBA: Kernel boundary alignment considering imbalanced data distribution," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 786–795, Jun. 2005.

[34] X. Hong, S. Chen, and C. J. Harris, "A kernel-based two-class classifier for imbalanced data sets," *IEEE Trans. Neural Netw.*, vol. 18, no. 1, pp. 28–41, Jan. 2007.

[35] M. E. Ben-Akiva, S. R. Lerman, and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand*, vol. 9. Cambridge, MA, USA: MIT Press, 1985.

[36] X. Zhao, X. Yan, A. Yu, and P. Van Hentenryck, "Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models," *Travel Behav. Soc.*, vol. 20, pp. 22–35, Jul. 2020.

[37] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[38] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, "An up-to-date comparison of state-of-the-art classification algorithms," *Expert Syst. Appl.*, vol. 82, pp. 128–150, Oct. 2017.

[39] S. Wang, Q. Wang, and J. Zhao, "Deep neural networks for choice analysis: Extracting complete economic information for interpretation," *Transp. Res. Part C, Emerg. Technol.*, vol. 118, Sep. 2020, Art. no. 102701.

[40] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning From Imbalanced Data Sets*. Berlin, Germany: Springer, 2018.

[41] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Exp. Syst. Appl.*, vol. 73, pp. 220–239, May 2017.

[42] J. Stefanowski, *Overlapping, Rare Examples and Class Decomposition in Learning Classifiers From Imbalanced Data BT—Emerging Paradigms in Machine Learning*, S. Ramanna, L. C. Jain, and R. J. Howlett, Eds. Berlin, Germany: Springer, 2013, pp. 277–306.

[43] I. Tomek, "Two modifications of CNN," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-6, no. 11, pp. 769–772, Nov. 1976.

[44] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in *Artificial Intelligence in Medicine*, vol. 1. Berlin, Germany: Springer, pp. 63–66.

[45] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Trans. Syst., Man, Cybern.*, vols. SMC-2, no. 3, pp. 408–421, Jul. 1972.

[46] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[47] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1819–1837, Jun. 2014.

[48] *Cycling and Walking Investment Strategy*, Dept. Transp., London, U.K., 2017.

[49] *Declaration on Cycling as a Climate Friendly Transport Mode*, Informal Meeting EU Ministers Transp. Luxembourg, Europe, 2015.

[50] Greater London Authority. (2018). *Mayor's Transport Strategy*. Accessed: Dec. 30, 2020. [Online]. Available: https://www.london.gov.uk/sites/default/files/mayors-transport-strategy-2018.pdf

[51] *Share the Road: Investment in Walking and Cycling Road Infrastructure*, United Nations Environment Programme Transport Unit, Nairobi, Kenya, 2010.

[52] K. Martens, *Transport Justice: Designing Fair Transportation Systems*. New York, NY, USA: Routledge, 2017.

[53] J. Chen, Z. Wu, and J. Zhang, "Driving safety risk prediction using cost-sensitive with nonnegativity-constrained autoencoders based on imbalanced naturalistic driving data," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 12, pp. 4450–4465, Dec. 2019.

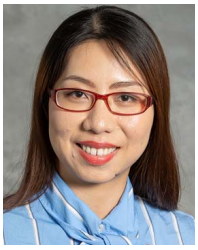
[54] C. S. Silveira et al., "Importance of subject-dependent classification and imbalanced distributions in driver sleepiness detection in realistic conditions," *IET Intell. Transp. Syst.*, vol. 13, no. 2, pp. 398–405, 2019.

[55] L. Oliveira et al., "Driver drowsiness detection: A comparison between intrusive and non-intrusive signal acquisition methods," in *Proc. 7th Eur. Workshop Vis. Inf. Process. (EUVIP)*, 2018, pp. 1–6.

[56] S. Wager, S. Athey, S. Wager, and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *J. Amer. Stat. Assoc.*, vol. 113, no. 523, pp. 1228–1242, Jul. 2018, doi: 10.1080/01621459.2017.1319839.

[57] D. Lee, S. Derrible, and F. C. Pereira, "Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling," *Transp. Res. Record, J. Transp. Res. Board*, vol. 2672, no. 49, pp. 101–112, Dec. 2018.
[58] L. Cheng, X. Chen, J. De Vos, X. Lai, and F. Witlox, "Applying a random forest method approach to model travel mode choice behavior," *Travel Behaviour Soc.*, vol. 14, pp. 1–10, Jan. 2019.
[59] Y. Zhang and Y. Xie, "Travel mode choice modeling with support vector machines," *Transp. Res. Record, J. Transp. Res. Board*, vol. 2076, no. 1, pp. 141–150, Jan. 2008.
[60] H. Omrani, O. Charif, P. Gerber, A. Awasthi, and P. Trigano, "Prediction of individual travel mode with evidential neural network model," *Transp. Res. Record, J. Transp. Res. Board*, vol. 2399, no. 1, pp. 1–8, Jan. 2013.
[61] C. R. Sekhar, Minal, and E. Madhu, "Mode choice analysis using random Forrest decision trees," *Transp. Res. Procedia*, vol. 17, pp. 644–652, 2016.
[62] X. Chang, J. Wu, H. Liu, X. Yan, H. Sun, and Y. Qu, "Travel mode choice: A data fusion model using machine learning methods and evidence from travel diary survey data," *Transportmetrica A, Transp. Sci.*, vol. 15, no. 2, pp. 1587–1612, Nov. 2019.
[63] M. Bierlaire, "A short introduction to PandasBiogeme," Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, Tech. Rep. TRANSP-OR 200605, 2020. [Online]. Available: https://transp-or.epfl.ch/documents/technicalReports/Bier20.pdf
[64] M. Abadi et al., (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://www.tensorflow.org/
[65] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011, doi: 10.5555/1953048.2078195.
[66] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox, "Hyperopt: A Python library for model selection and hyperparameter optimization," *Comput. Sci. Discov.*, vol. 8, no. 1, 2015, Art. no. 014008.
[67] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017.

Huanfa Chen received the B.S. degree in chemistry and the M.S. degree in cartography and geographical information systems from Peking University, China, in 2011 and 2014, and the Ph.D. degree in geographical information science from University College London, U.K., in 2019. He is currently a Lecturer in spatial data science with the Centre for Advanced Spatial Analysis (CASA), University College London. His research interests span spatial optimization, GeoAI, and agent-based simulation, with applications in transport, crime, and public health.



Yan Cheng received the B.Sc. degree in transportation engineering and the Ph.D. degree in road and railway engineering from Tongji University, China, in 2012 and 2018, respectively.

She worked as an Associate Lecturer (Teaching) at the Centre for Transport Studies, University College London, U.K., from 2018 to 2021. She is currently a Distinguished Research Fellow at Tongji University. Her research interests include railway planning and design, travel demand prediction, and travel behavior analysis.

Dr. Cheng is a member of the Special Interest Group A3—Rail Transport and F1—Transport and Spatial Development of World Conference on Transport Research Society and a member of Transport and Geography Commission of International Geographical Union.