



Research Article

A Customized Deep Neural Network Approach to Investigate Travel Mode Choice with Interpretable Utility Information

Zhengchao Zhang ¹, Congyuan Ji,² Yineng Wang,² and Yanni Yang ³

¹Department of Civil Engineering, Tsinghua University, Beijing 100084, China

²Department of Industrial Engineering, Tsinghua University, Beijing 100084, China

³School of Management and Engineering, Capital University of Economics and Business, Beijing 100070, China

Correspondence should be addressed to Yanni Yang; yangyanni@cueb.edu.cn

Received 5 September 2019; Revised 6 August 2020; Accepted 26 August 2020; Published 16 September 2020

Academic Editor: Antonio Comi

Copyright © 2020 Zhengchao Zhang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Discrete choice modeling of travel modes is an essential part of traffic planning and management. Thus far, this field has been dominated by multinomial logit (MNL) models with a linear utility specification. However, deep neural networks (DNNs), owing to their powerful capacity of nonlinear fitting, are now rapidly replacing these models. This is because, by using DNNs, mode choice can be assimilated with the classification problems within the machine learning community. This article proposes a newly designed DNN framework for traffic mode choice in the style of two hidden layers. First, a local-connected layer automatically extracts an effective utility specification from the available data, and then, a fully connected layer augments the feature representation. Validated by a practical city-wide multimodal traffic dataset in Beijing, our model significantly outperforms the random utility models and simple fully connected neural network in terms of the prediction accuracy. Besides the comparison of the predictive power, we also present the interpretability of the proposed model.

1. Introduction

Discrete choice models (DCMs) have emerged as powerful theoretical frameworks to analyze individual travel behavior among a given set of discrete alternatives (e.g., taking a subway instead of selecting a private car or bus). Since the seminal paper by McFadden [1], for decades, the multinomial logit (MNL) model has been widely adopted for exploring individual decision-making. Despite its oversimplified assumption of a linear utility specification in regard to complex human choice behavior [1], this type of a model still realizes practical applications because it enables a high level of interpretability. However, the interpretability gained from a linear expression is frequently at the expense of the predictive power. Indeed, an assumed linear statistical structure cannot adequately capture the potential regulations in a dataset and will encounter the issue of dealing with categorical explanatory variables (e.g., income level and departure time range).

Typically, different from the MNL model, machine learning methods, particularly deep learning models using a data-oriented approach, are becoming increasingly prominent in numerous research fields, including transportation [2–4]. Deep neural networks (DNNs) are mathematical tools that are loosely inspired by the functional aspects of biological neural systems. These models have repeatedly demonstrated excellent performances in an extensive range of specific transportation tasks, such as short-term traffic flow prediction [5], license plate recognition [6], automobile driving risk detection [7], ownership demand estimation [8], and movement pattern inference [9]. Although some traditional algorithms with hand-crafted features are very effective for the given problems [10], well-recognized guidelines to choose the appropriate features are not available in general. Thus, the deep learning technology which leverages automatic feature learning is more scalable and robust. Despite the success in the abovementioned aspects, the achievements of DNNs in the subfield of travel

behavior study are still reasonably limited. For this topic, several previous works utilized a conventional fully connected neural network (FCN) [11–14]. However, the versatile architecture of DNNs particularly makes them well equipped to deal with large volumes of (even unstructured) data [15].

In this study, to compensate for the abovementioned deficiency, we aim to propose a distinctive DNN approach to understand transportation mode choice. Specifically, the contributions of this paper are threefold. (i) An ingenious four-layer DNN model is established, with the first layer taking the input data, the second layer learning the utility function for each traffic mode and exogenous information, the third layer mining the correlation rules in the former layer, and the last layer calculating the choice probability of each mode. (ii) Testing on a real-world large-scale dataset in Beijing (including the travels by subway, bus, private car, and taxi), the proposed model remarkably achieves a better predictive performance than the random utility and FCN models. (iii) Based on the empirical results, we reveal that the valuable insights about identifying the characteristic factors of traffic mode choice can be obtained from our model.

The remainder of this paper is organized as follows. Section 2 overviews the recent relevant studies. Section 3 describes the proposed model. Section 4 exhibits the dataset and experimental results. Section 5 demonstrates the evidence of model interpretability. Finally, Section 6 provides the summary.

2. Literature Review

In the area of travel demand analysis, since the 1980s, the MNL model has been predominantly used to examine various types of choices, such as travel mode, driving route, travel frequency, departure time, and usage of new transit lines [16–21]. The key components elicited from the MNL include choice prediction, choice probabilities, probability derivatives, and marginal rate of substitutes, all of which are critical for policy formulation. Overall, both the prediction accuracy and economic explanation are worth studying in depth.

The machine learning community has also generated tremendous interest in predicting various choices. A mainstream practice holds the traffic mode choice as a distinct case of the general classification problems. For instance, Pirra and Diana employed the support vector machine (SVM) to recognize tour-based mode choice patterns [22], whereas the random forests (RFs) and boosting method are relied to predict the availability of sharing bikes and airline itinerary choice [23, 24]. Subsequently, the data-driven approaches were presented for modeling the multimodal travel choice, which derived the explanatory variables from travel diary data [25–27]. Pekel and Soner Kara performed a comprehensive review of the findings related to a DNN relying on the public transport choice [28]. Not surprisingly, these works yield outstanding results owing to the big data fitting ability while neglecting tackling the issue of interpretability.

Some of the most recent efforts have attempted to go beyond simply targeting the accuracy and have conducted innovative behavioral studies based on data-driven methods. Wong et al. presented latent behavior attributes using a restricted Boltzmann machine (BM) [29], and Van Cranenburgh and Alwosheel initiated DNN-based approaches to investigate the decision-rule heterogeneity between travelers [14]. Two notable research studies bridge the gap between a DCM and DNN. Wang and Zhao illustrated that the MNL could be expressed as a shallow and sparse neural network by deriving the equivalent mathematical expressions and enabling the extraction of important economic information from a DNN [12, 13]. Sifringer et al. integrated a new nonlinear representation originating from a neural network with the MNL to enhance the accuracy of the prediction and parameter estimation [30].

These pioneering works prove that a DNN has the potential for exploring the choice behavior. However, they do not focus on the construction of a unique DNN structure that allows high predictability while maintaining interpretability. Inspired by the recent breakthroughs in the deep learning domain, we orient to put forward a flexible and general DNN framework for traffic mode choice that can increase the overall predictive performance and acquire shinning points to account for the choice behavior.

3. Methodology

3.1. Deep Neural Network. In a DNN, computations are performed in terms of interconnected groups of artificial neurons (also known as nodes), and processing information is obtained by the so-called connectionist approach [31]. Three types of layers are commonly distinguished: the input layer, hidden layer, and output layer. Explanatory variables are injected into the input layer, and the output layer contains the dependent variables. In the context of the selected models, the input nodes are concerned with the attributes of the alternatives, whereas the output nodes include the choice probabilities. The data stream propagates in a forward direction through links that connect the nodes with learnable weights and biases. At each node, the weights are multiplied with the input values from the previous layer and then summed, and finally, the results are propagated to the next layer, after passing through the activation function. By default, a DNN includes more than one type of heterogeneous hidden layers.

Although the fact that an extensive variety of DNNs has been invented to deal with numerous challenges, there seems to be no particularly suitable solution for traffic mode choice. The household standard 2D convolution neural networks (2D CNNs) and recurrent neural networks (RNNs) lead the ground-breaking progress in the fields of computer vision (CV) and natural language processing (NLP). However, they are unsuitable for our problem since the designs of their structures clearly deviate from our target. An FCN is frequently considered as a generic method to solve any classification problem. Nevertheless, without incorporating problemwise specific knowledge into the composition of a DNN, the performance will definitely degrade and is

typically unexplainable. In the absence of a suitable “off-the-shelf” tool, the next subsection proposes a new DNN topology that is particularly designed for multimodal traffic choice.

3.2. DNN for Traffic Mode Choice. Figure 1 shows the architecture of our conceived DNN, which is a multilayer perceptron with two distinct hidden layers. The inputs are divided into five classes: exogenous features (e.g., the descriptions of the origin and destination (OD) and personal information) and attributes of taking subway, taxi, bus, and car (e.g., the travel time, cost, and variability for each travel mode). Note that the discrete variables (e.g., the ID of an OD pair, departure time period, age level, and income level) cannot directly feed to neural networks. To deal with this problem, we exploit the embedding method [32] to transform each categorical attribute into a low-dimensional real vector. Specifically, each categorical value, $v \in |V|$, is mapped to a real space, $\mathbb{R}^{E \times 1}$ (known as the embedding space), by multiplying a parameter matrix, $\mathbf{W}_{\text{embed}} \in \mathbb{R}^{V \times E}$. Here, V represents the vocabulary size of the original categorical value, whereas E is the dimension of the embedding space (usually $E \ll V$). Thus, our model effectively reduces the input dimension (compared to that of a one-hot encoding) and is computationally more efficient when encountering categorical values [33]. Following the embedding preprocessing, the first hidden layer behaviorally imitates the MNL to form the categorized utility specification by an adaptive linear transformation, which can be expressed as

$$U_k = \langle \mathbf{X}_k^*, \mathbf{W}_k^{(1)} \rangle + b_k^{(1)}, \quad k \in \mathbf{S}, \quad (1)$$

where \mathbf{S} denotes the set of all feature categories, e.g., $\mathbf{S} = \{\text{exogenous, car, bus, taxi, subway}\}$. $\mathbf{X}_k^* \in \mathbb{R}^{d_k}$ is the input feature vector of k^{th} category after embedding, if discrete variables are involved, the original features are noted as \mathbf{X}_k with dimensions of d_k . $\mathbf{W}_k^{(1)} \in \mathbb{R}^{d_k}$, $b_k^{(1)}$ are the learnable weight and intercept owned by the first hidden layer to extract the utility specification of k^{th} feature category, i.e., U_k . $\langle \cdot, \cdot \rangle$ is the inner product of two vectors.

On receiving the utility terms, the second hidden layer refines the feature representation by a nonlinear fully connected mapping. Therefore, it considers the relevance of different types of utilities. Finally, the output layer calculates the choice probabilities by the softmax activation [31], which ensures that the sum of the outcomes is 1. This procedure is given as

$$\mathbf{H}^{(1)} = [U_{\text{exo}}, U_{\text{sub}}, U_{\text{bus}}, U_{\text{taxi}}, U_{\text{car}}], \quad (2)$$

$$\mathbf{H}^{(2)} = \sigma(\mathbf{H}^{(1)} \mathbf{W}^{(2)} + \mathbf{b}^{(2)}), \quad (3)$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \quad (4)$$

$$\mathbf{H}^{(3)} = \mathbf{H}^{(2)} \mathbf{W}^{(3)}, \quad (5)$$

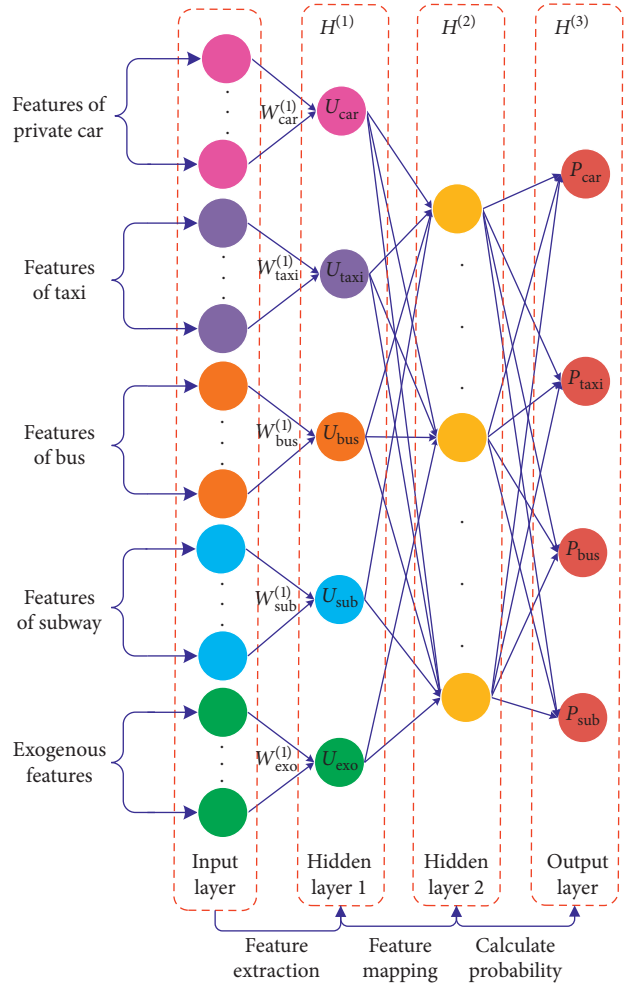


FIGURE 1: Designed DNN topology.

$$P_n = \frac{e^{H_n^{(3)}}}{\sum_{j \in \mathbf{C}} e^{H_j^{(3)}}}, \quad n \in \mathbf{C}, \quad (6)$$

where $[\cdot, \cdot]$ concatenates the scalars to compose the vector, $\mathbf{H}^{(1)} \in \mathbb{R}^5$, of the first hidden layer. $\mathbf{W}^{(2)} \in \mathbb{R}^{5 \times l}$, $\mathbf{b}^{(2)} \in \mathbb{R}^l$ are the weight and intercept parameters of the second hidden layer, respectively, which mirror the utility function into an l -dimensional hyper vector, $\mathbf{H}^{(2)}$, along with the sigmoid activation, $\sigma(\cdot)$. $\mathbf{W}^{(3)} \in \mathbb{R}^{l \times 4}$ adjusts the output dimension to match the number of options. $H_n^{(3)}$ is the n^{th} element of $\mathbf{H}^{(3)}$, and P_n is the choice probability of mode n . $\mathbf{C} = \{\text{car, bus, taxi, subway}\}$ is the choice set.

3.3. Model Training. The objective function of the above-defined network is to minimize the cross-entropy loss [34] between the true choice and estimated choice probability, by which the weighted and intercept parameters can be learnt. It is expressed in the following equation:

$$\min_{W, b} - \sum_i \sum_n y_n(i) \cdot \log P_n(i), \quad (7)$$

where $y_n(i)$ is the observed choice variable (or true label) and is equal to 1 if the individual, i , chooses the alternative, n ; otherwise it is 0. $\mathbf{P}_n(i)$ is the corresponding output probability of the model.

It is worth pointing out that a DNN always encounters a vanishing gradient during the training stage. To overcome this drawback, a well-known strategy called batch normalization (BN) [35] is applied to the two hidden layers. Therefore, equations (1) and (3) are accordingly changed to

$$\begin{aligned}\hat{U}_k &= \text{BN}_{\gamma_1\beta_1}(\langle \mathbf{X}_k^*, \mathbf{W}_k^{(1)} \rangle + b_k^{(1)}), \\ \hat{\mathbf{H}}^{(2)} &= \sigma(\text{BN}_{\gamma_2\beta_2}(\mathbf{H}^{(1)}\mathbf{W}^{(2)} + \mathbf{b}^{(2)})).\end{aligned}\quad (8)$$

Focusing on a batch of samples, $\mathbf{B} = \{x_1, x_2, \dots, x_m\}$, the BN is implemented as

$$\begin{aligned}\text{BN}_{\gamma,\beta}(x_r) &= \gamma \frac{x_r - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \varepsilon}} + \beta, \\ \mu_{\mathcal{B}} &= \frac{1}{m} \sum_{r=1}^m x_r, \\ \sigma_{\mathcal{B}}^2 &= \frac{1}{m} \sum_{r=1}^m (x_r - \mu_{\mathcal{B}})^2,\end{aligned}\quad (9)$$

where $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}^2$ are the mean and variance of the underlying batch, respectively. γ and β are the parameters to be learnt (commonly with initial values of 1 and 0), which focus on reducing the internal covariate shift when forward propagating through each layer [36].

All the training steps are provided in Algorithm 1.

4. Case Study

4.1. Dataset

4.1.1. Study Area. The study site is chosen as a closed region bounded by the 4th Ring Road in Beijing, which covers a 302 km² downtown area. We partition this area into 123 traffic zones which regard the subway stations as centroids, and the positions of all subway stations are shown in Figure 2. The ODs of the trips within each traffic zone are assumed as the centroids.

4.1.2. Data Collection and Preprocess. In this research, the data are derived from the public transportation system, taxi orders, and anonymous navigation users of AMAP [37] in Beijing, comprising the travels from a bus, subway, taxi, and private car. To reveal the regular and relatively stable patterns of travelers, we set the morning peak hours (06:00–11:00) of weekdays as the departure time interval. Because the content, format, structure, and information redundancy vary with the raw data source and different traffic modes, we process the massive heterogeneous data to ensure that the effective choice set for each traveler is composed of four modes (i.e., bus, subway, taxi, and private car). By indexing each piece of data, eliminating entries with abrupt changes, and implementing a

joint query among multiple datasets, we match the OD information of all the modes to the centroids. Thus, we acquire the dataset satisfying the requirement that all the mode choices are available for each traveler. In the operation of data cleaning, we eliminate some entries that are beyond the research scope, those with missing attributes, and those that are duplicates.

4.1.3. Data Fusion. In the last step, we fuse the data from the different sources, having different formats and with different characteristics, into one comprehensive multimode transportation choice dataset, with fields as listed in Table 1. The features {f1–f4, f17–f21} are exogenous information and {f5, f9, f13}, {f6, f10, f14}, {f7, f11, f15}, and {f8, f12, f16} are the attributes of the bus, taxi, private car, and subway, respectively.

To make up the different penetrations between each data source, namely, the records of buses, subways, taxis, and private cars accounting for 90%, 85%–90%, 50%, and 5%–10% of the total number, respectively, a sampling expansion method based on the urban-trip summation [38] is availed to balance them. Although our dataset is formed from a sampled data source, its quantity far exceeds the traditional questionnaires by a sample ratio of approximately 2–3%. From the perspectives of expense, timeliness, and convenience, this approach is superior to a traffic survey.

4.2. Experiment Settings

4.2.1. Benchmark Models. To test the effect of our model, it is compared with four prevalent baselines.

MNL: The standard MNL model constitutes a linear-in-parameter utility specification with properties f5–f16, as listed in Table 1. It is formulated as

$$\mathbf{V}_n(i) = \langle \mathbf{X}_n, \boldsymbol{\alpha}_n \rangle, \quad n \in \mathbf{C}, \quad (10)$$

$$P'_n(i) = \frac{e^{V_n(i)}}{\sum_{j \in \mathbf{C}} e^{V_j(i)}}, \quad (11)$$

where $\mathbf{V}_n(i)$ is the utility of individual i associated with alternative n . $\boldsymbol{\alpha}_n$ is the vector of the preference parameters, which will be estimated by the maximum likelihood.

NL: The nested logit model considers the correlation between alternative choices, which is the most widely known relaxation of the MNL model [39]. Because the travel time of private car, bus, and taxi highly depend on the traffic state of road network, we divide them into the same nest and subway into another. It is formulated as

$$P''_n(i) = \frac{e^{(V_n(i)/\lambda_l)}}{\sum_{k \in B_l} e^{(V_k(i)/\lambda_l)}} \times \frac{e^{\lambda_l \Gamma_l}}{\sum_{m=1}^M e^{\lambda_m \Gamma_m}}, \quad (12)$$

$$\Gamma_l = \ln \sum_{k \in B_l} e^{V_k(i)/\lambda_l},$$

Input The feature vectors, $\{X_{\text{exo}}, X_{\text{car}}, X_{\text{bus}}, X_{\text{taxi}}, X_{\text{sub}}\}$.
Observations of the individual choice, $Y = \{y(1), y(2), \dots, y(R)\}$.

Output The model with learnt parameters.

- (1) **Procedure** DNN model Train.
- (2) Initialize the parameter matrix: W_{embed} for embedding.
- (3) Embedding categorical values: $X_k^* \leftarrow X_k, k \in \{\text{exo}, \text{car}, \text{bus}, \text{taxi}, \text{sub}\}$.
- (4) Initialize a null set: $Q \leftarrow \emptyset$.
- (5) **for** all available individual sample $i (1 \leq i \leq M)$ **do**.
- (6) $X_{\text{input}}(i) \leftarrow [X_{\text{exo}}^*(i), X_{\text{car}}^*(i), X_{\text{bus}}^*(i), X_{\text{taxi}}^*(i), X_{\text{sub}}^*(i)]$.
- (7) $Y_{\text{label}}(i) \leftarrow y(i), y(i) \in \{\text{car}, \text{bus}, \text{taxi}, \text{subway}\}$.
- (8) A training sample $(X_{\text{input}}(i), Y_{\text{label}}(i))$ is placed in Q .
- (9) **end for**
- (10) Initialize all the weight and intercept parameters.
- (11) Initialize $\gamma_1, \gamma_2 = 1, \beta_1, \beta_2 = 0$ for BN.
- (12) **repeat**.
- (13) Randomly extract a batch of samples Q^b from Q .
- (14) Update the parameters by minimizing the equation (7) by the mini-batch gradient descent algorithm within Q^b .
- (15) **until** convergence criterion is met.
- (16) **end procedure**.

ALGORITHM 1: Training the model.



FIGURE 2: Centroids in the traffic zone.

where B_l denotes the nest l , i.e., $B_1 = \{\text{car}, \text{bus}, \text{taxi}\}, B_2 = \{\text{subway}\}$. M is the total number of nests, i.e., 2. λ_l measures the correlation of alternatives in the nest l . $V_n(i)$ keeps the same with definition in equation (10).

FCN: A three-layer fully connected neural network is set up, and the details are shown in Figure 3. The model structure is very similar to [12, 13], except for the number of hidden layers.

RF: Random forest is a well-known ensemble decision tree model, which is proven to achieve superior performance in a board range of data mining tasks. Since the RF model is able to deal with discrete features, we directly utilize all the features (i.e., f2–f21). The number

TABLE 1: Fields of the traffic mode choice dataset.

Index	Field
f1	Observation of mode choice
f2	Length level of trip
f3	Departure time interval
f4	House price of origin
f5	Bus travel time
f6	Taxi travel time
f7	Private car travel time
f8	Subway travel time
f9	Bus ticket
f10	Taxi price
f11	Private car fare
f12	Subway ticket
f13	Bus travel time fluctuation
f14	Taxi travel time fluctuation
f15	Private car travel time fluctuation
f16	Subway travel time fluctuation
f17	No. of origin rings
f18	No. of destination rings
f19	Distance between ODs
f20	Index of origin
f21	Index of destination

of decision trees is set as 100, and the maximum depth of the tree is set as 10, which achieves a balance between accuracy and efficiency on the validation set.

4.2.2. Hyperparameter. In our DNN framework, the number of hidden layers and its nodes, learning rate, and batch size are crucial hyperparameters for the model performance. We referred to the suggestions by standard practices [40, 41] and tuned the hyperparameters manually through a 5-fold stratified cross validation. Figure 4 shows the heatmap of model performances with different hidden layers and nodes in cross validation. Although the model performances

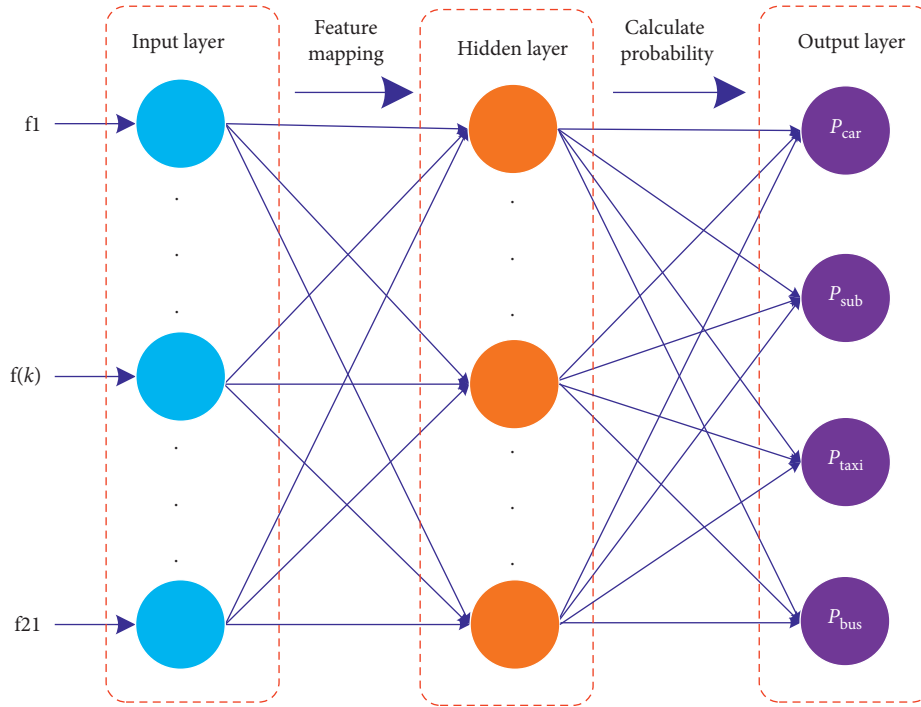


FIGURE 3: Schematic of the FCN.

slightly improve with more hidden layers and nodes, a complex structure will considerably increase the training time. As the result, our model is facilitated by one fully connected hidden layer with 32 nodes, the Xavier initialization [42], and Adam optimization [43] with a learning rate of 0.01 and weak to 0.9 after 100 training steps, early stopping as regularization, and mini-batch size = 2¹⁷.

4.2.3. Measurement. To evaluate the performance of each model, the dataset is divided into two independent subsets: the data from 80% OD pairs serve as the training set, whereas the remaining 20% is leveraged as the testing set, which are approximately 800,000 and 200,000 observations, respectively. The prediction accuracy is defined as the proportion of correctly predicted samples to that of the total number of samples in the testing set, which can be computed as

$$\text{accuracy} = \frac{\sum_i \tau_i}{|T|},$$

$$\tau_i = \begin{cases} 1, & \arg \max_n P_n(i) = y(i), \\ 0, & \text{else,} \end{cases} \quad (13)$$

where τ_i is a binary variable to identify whether a sample i is correctly predicted and $|T|$ is the cardinality of the testing set.

4.3. Results. The prediction accuracies of the five compared methods are listed in Table 2. Because the neural networks and random forest model are influenced by the randomness, we separately run these models five times and report the

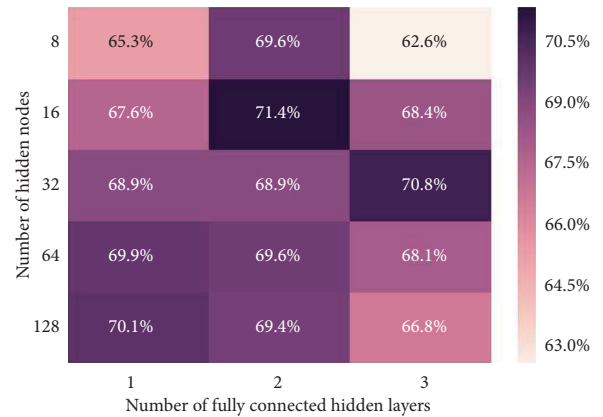


FIGURE 4: Heatmap of model performances with different hidden layers and nodes in cross validation.

average results and standard deviation. We further utilize the t -test to verify whether the performances between each model are significantly different. The results are shown in Table 3. Since the P values are all less than 0.05, the model performances are statistically variant. From the comparison, the following conclusions can be drawn: (1) Apparently, our model exhibits significant superiority than the MNL, NL, FCN, and RF in terms of the prediction accuracy. (2) The neural network-based method surpasses the random utility models in terms of the predictive performance, which is consistent with the conclusions from [12–14, 30]. Figure 5 displays the curves of the accuracy achieved on the training set with the training steps. It can be seen that the FCN rapidly reaches saturation, whereas our model continuously makes breakthroughs. Both the models are approximately

TABLE 2: Experimental results.

Model	Accuracy on testing set
MNL	53.5%
NL	55.8%
FCN	57.8% ($\pm 0.33\%$)
RF	64.2% ($\pm 0.13\%$)
Customized DNN	69.5% ($\pm 0.56\%$)

TABLE 3: P value of welch's t -test of performances between each model.

	MNL	NL	FCN	RF
FCN	1.4×10^{-5}	2.8×10^{-4}		
RF	9.0×10^{-9}	2.4×10^{-8}	4.3×10^{-10}	
Customized DNN	9.6×10^{-12}	3.3×10^{-11}	4.0×10^{-10}	7.7×10^{-8}

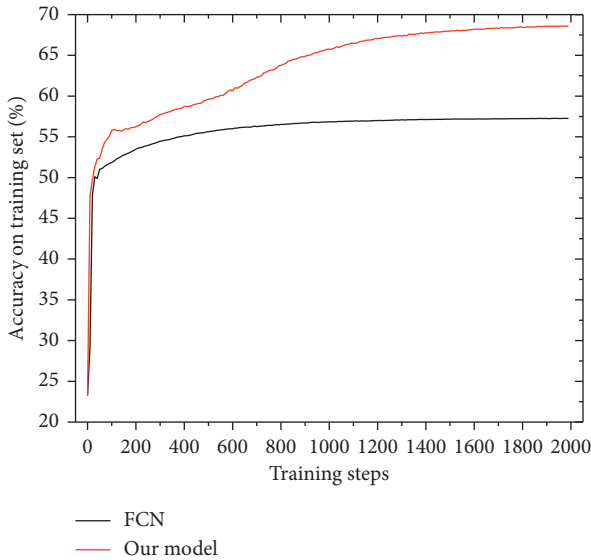


FIGURE 5: Curve of the training accuracy.

accurate on the training set and testing set, which testifies they are not overfitted.

5. Model Interpretation

In this section, based on the empirical results, we further confirm the interpretability of the developed model. This is discussed from the perspectives of the utility term and economic explanation.

5.1. Influential Input Factors. Focusing on equation (1), the explanatory variables with higher absolute values of weight $\widetilde{\mathbf{W}}_k^{(1)}$ will contribute more to the utility specification, U_k . Thus, the absolute values of a local-connected weight can measure the relative feature importance [5, 44, 45]. The relative importance of the q^{th} attribute belonging to the feature category, k , is delineated as

$$R_k^q = \frac{|\widetilde{\mathbf{W}}_k^{(1)}[q]|}{\|\widetilde{\mathbf{W}}_k^{(1)}\|_1}, \quad k \in \mathbf{S}, \quad (14)$$

where $\widetilde{\mathbf{W}}_k^{(1)}$ is the local-connected weight vector after the training stage and $\widetilde{\mathbf{W}}_k^{(1)}[q]$ is the corresponding q^{th} element. $\|\cdot\|_1$ and $|\cdot|$ return the L_1 norm of the vector and absolute value of the scalar, respectively.

Figure 6 shows the relative importance of the explanatory variables pertaining to each traffic mode (reliability is determined by the volatility of the travel time). The following interesting phenomena can be observed: (1) All the surface transports (i.e., bus, taxi, and private car) are sensitive to the travel time owing to the severe traffic jams during morning peak hours. (2) For the bus mode, the inexpensive tickets along with the travel time and its uncertainty are the distinguished characters. (3) In contrast, as the other option for public transit, the choice of a subway is only affected by the ticket because the travel time and variation are highly stable. (4) The expensive fare to hail a taxi is concerned about when the commute trip is planned. (5) Private car owners do not care about the fuel charges, which is much less than the fee of a taxi.

5.2. Economic Information. In this subsection, we discuss the use of numerical methods to probe how the behaviorally intuitive choice probabilities change with economic factors (e.g., travel time and travel cost). A short trip from origin No. 1 to destination No. 3 with a 3.8 km distance between 7:00 and 8:00 and a long trip from origin No. 60 to destination No. 65 with a 9.7 km distance between 6:00 and 7:00 are treated as two examples. The detailed attributes of each transportation mode in these two trips are presented in Tables 4 and 5.

First, we alter the bus travel time from 12.7 min and 16.5 min (the minimum travel time among the four modes) to 25.7 min and 36.7 min in the short and long trip instances, respectively. Figure 7 depicts the variation curve of the choice probability with the bus travel time. In general, the choice probabilities of a bus, taxi, and private car exhibit a similar descending trend, whereas that of a subway is opposite. From this, we can infer that the developed model takes the travel time of a bus as one index for the condition of the surface traffic. Therefore, a subway is a substitution for a congested surface transportation. It is worth emphasizing that the travel times of the surface transports identically increase or decrease in the training set. Based on this prior fact, the abovementioned results actually reflect that our model can capture the correlation rules in the dataset. By a simple analysis, it is easy to find that the designed DNN avoids the shortcoming of the independence of irrelevant alternatives (IIA). The aforementioned two properties are exhibited by the second fully connected hidden layer. The nonlinearity of the variation curve is also discriminative to the MNL, which is more substantial in the long trip.

Figure 8 visualizes the relationship between the probabilities of choosing a bus and its ticket. The colored dotted

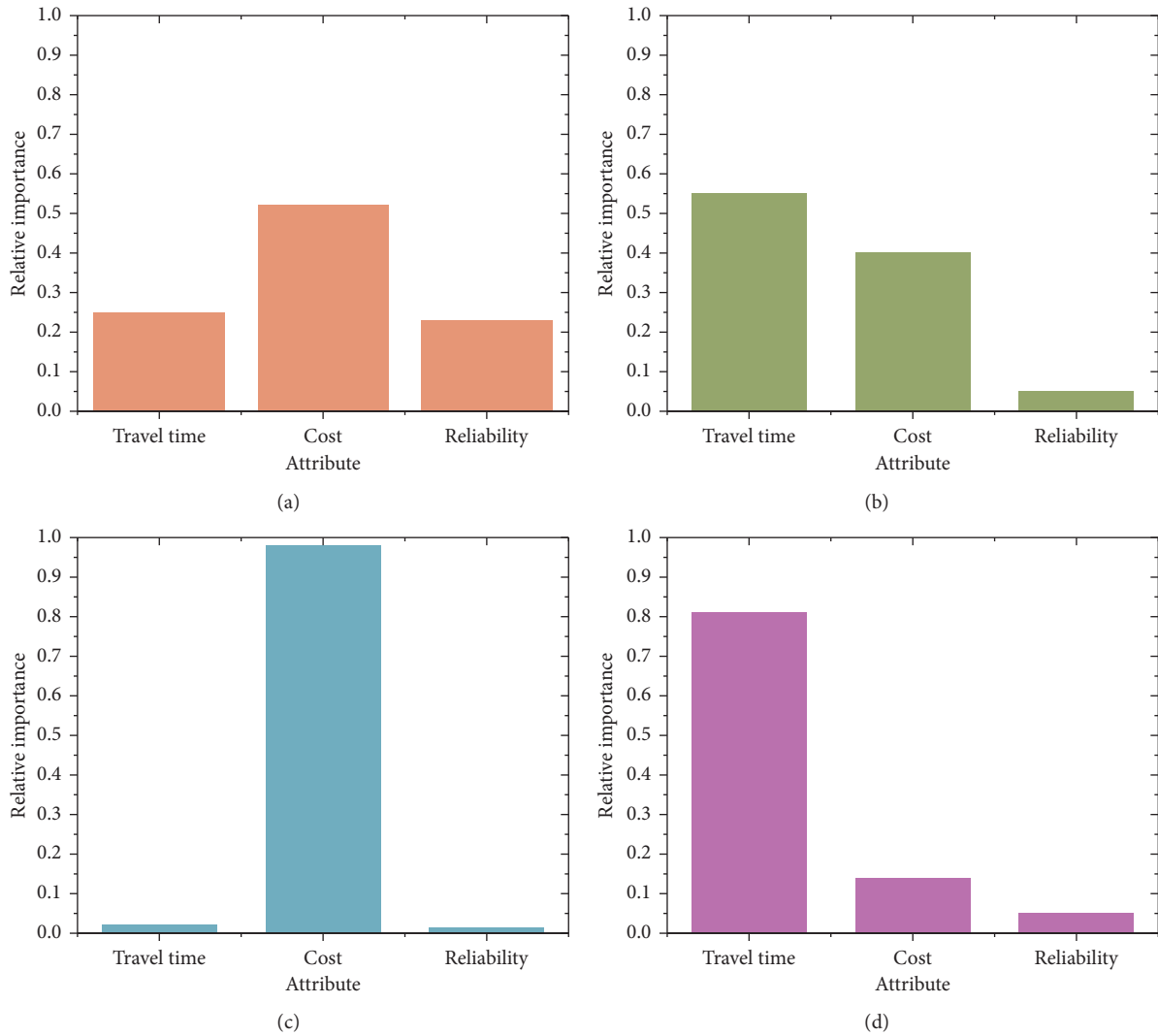


FIGURE 6: Relative feature importance of the four travel modes. (a) Bus. (b) Taxi. (c) Subway. (d) Private car.

TABLE 4: Travel information of a short trip.

Mode/attribute	Travel time	Fare	Relative standard deviation of travel time
Bus	25.7 min	¥1	68.7%
Taxi	12.7 min	¥15.9	13.5%
Private car	12.7 min	¥5.06	13.5%
Subway	10.4 min	¥3	0

TABLE 5: Travel information of a long trip.

Mode\attribute	Travel time	Fare	Relative standard deviation of travel time
Bus	36.4 min	¥1	60.6%
Taxi	16.5 min	¥25.5	19.5%
Private car	16.5 min	¥10.5	19.5%
Subway	18.6 min	¥4	0

curves are the results of five separate estimations, and the black one is from the ensembled model aggregated over them. All the curves are generated by varying the bus ticket

while holding the other variables at their original levels. The majority curves are intuitive and reasonable, resembling the standard S-shaped curve from the MNL model. However,

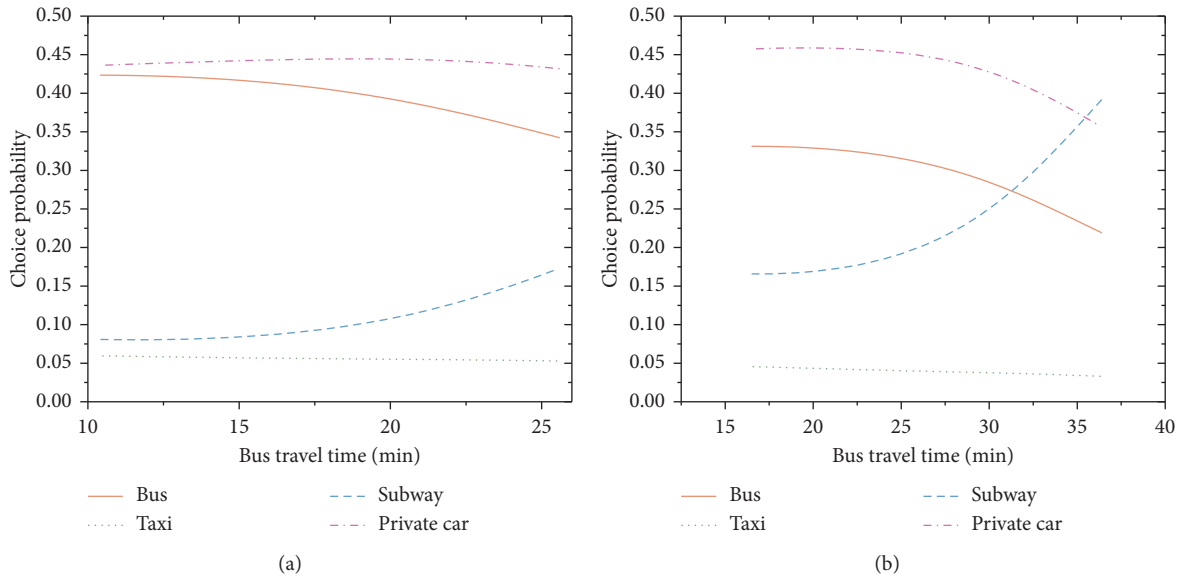


FIGURE 7: Variation curve of the choice probability with the bus travel time. (a) Short trip. (b) Long trip.

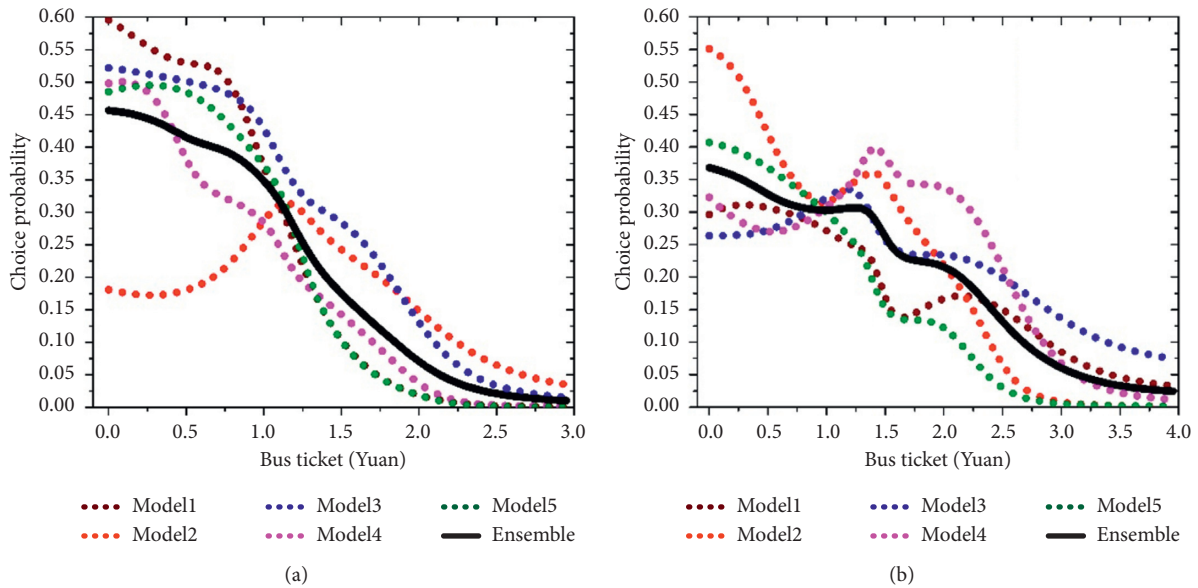


FIGURE 8: Variation curve of the bus choice probability with the bus ticket. (a) Short trip. (b) Long trip.

some particular individual estimations suffer the interpretability problem (e.g., red line) because the choice probability increases with the ticket increase, which is inconsistent with our general knowledge. Compared to each individual estimation, the ensemble model is more monotonic and smoother, showing a nonlinear decrease with the ticket increase. Furthermore, we can deduce that ¥1 to ¥2 and ¥1.5 to ¥3 are the sensitive ranges of the bus ticket in the short and long trip travels, respectively. Concurrently, the choice probability of a bus is less sensitive to the price in a long trip because its slope is relatively

less. Wang and Zhao [12, 13] also offered the evidence that, at the disaggregate level, the choice probability curves of DNN models could be nonmonotonically decreasing with the costs and be highly sensitive to the particular estimation owing to the irregularity of the probability fields and large estimation variances.

On basis of the abovementioned results, when facing large-scale datasets along with numerous categorical features, it is considerable to implement the ensemble DNN model for the travel choice behaviors analysis and unknown cases prediction.

6. Conclusions and Future Work

This paper develops a new general and flexible DNN framework that integrates two elaborate hidden layers for traffic mode choice. Using real-world multimodal transportation data, we demonstrate our model significantly improves the prediction performance compared to the random utility models, FCN and RF. It stresses the necessity of designing a particular DNN architecture according to the problemwise requirements, which is exactly our motivation.

Another substantive contribution is that we examine the model interpretability in depth. The important findings can be concluded into three sides: (1) the first local-connected hidden layer partially replaces the manual utility specification and allows for automatically discovering the influential explanatory variables for each traffic mode from the available data. (2) The second fully connected hidden layer enables the model to capture the correlated relationship in the dataset and eliminate the IIA problem. (3) Researchers and practitioners can obtain the stable economic information from complex human decision-making processes with the aid of the ensemble DNN model.

Subject to the dataset, the effects of individual properties on traffic mode choice are not involved in this study. Future research may be directed to apply our model to the stated preference (SP) survey data, in which the individual attributes will be treated as a new feature category when feeding to the input layer. The results can help us better understand the role of individual properties in the travel choice decision. In addition, it is worth exploring how to use the DNN for analyzing travelers' path choice behavior [46]. From the application aspect, predicting the passenger flow of the public transit via the proposed model may be helpful to urban safety management during the important events [47].

Data Availability

The data used to support the findings of this study have not been made available because the authors have signed the confidentiality agreement with the data providers.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research benefited substantially from the supervision of Dr. Fang He at Tsinghua University. This research was supported by grants from the National Key R&D Program of China (2018YFB1601600). This research was supported in part by the Tsinghua University-Toyota Research Center.

References

- [1] D. McFadden, *Conditional Logit Analysis of Qualitative Choice Behavior*, Academic Press, Cambridge, MA, USA, 1973.
- [2] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: a survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.
- [3] T. Tanprasert, C. Saiprasert, and S. Thajchayapong, "Combining unsupervised anomaly detection and neural networks for driver identification," *Journal of Advanced Transportation*, vol. 2017, Article ID 6057830, 13 pages, 2017.
- [4] Y. Wang, D. Zhang, Y. Liu, B. Dai, and L. H. Lee, "Enhancing transportation systems via deep learning: a survey," *Transportation Research Part C: Emerging Technologies*, vol. 99, pp. 144–163, 2019.
- [5] Z. Zhang, M. Li, X. Lin, Y. Wang, and F. He, "Multistep speed prediction on traffic networks: a deep learning approach considering spatio-temporal dependencies," *Transportation Research Part C: Emerging Technologies*, vol. 105, pp. 297–322, 2019.
- [6] H. Li, P. Wang, and C. Shen, "Toward end-to-end car license plate detection and recognition with deep neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 3, pp. 1126–1136, 2018.
- [7] W. Qi, Z. Wang, R. Tang, and L. Wang, "Driving risk detection model of deceleration zone in expressway based on generalized regression neural network," *Journal of Advanced Transportation*, vol. 2018, Article ID 8014385, 6 pages, 2018.
- [8] M. Paredes, E. Hemberg, U. M. O'Reilly, and C. Zegras, "Machine learning or discrete choice models for car ownership demand estimation and prediction?" in *Proceedings of the 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, June 2017.
- [9] K. Siła-Nowicka, J. Vandrol, T. Oshan, and J. A. Long, "Analysis of human mobility patterns from GPS trajectories and contextual information," *International Journal of Geographical Information Science*, vol. 30, no. 5, pp. 881–906, 2016.
- [10] J. Nalepa and M. Blocho, "Adaptive guided ejection search for pickup and delivery with time windows," *Journal of Intelligent & Fuzzy Systems*, vol. 32, no. 2, pp. 1547–1559, 2017.
- [11] G. E. Cantarella and S. De Luca, "Multilayer feedforward networks for transportation mode choice analysis: an analysis and a comparison with random utility models," *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 2, pp. 121–155, 2005.
- [12] S. Wang and J. Zhao, "Framing discrete choice model as deep neural network with utility interpretation," 2018.
- [13] S. Wang, Q. Wang, and J. Zhao, "Deep neural networks for choice analysis: extracting complete economic information for interpretation," *Transportation Research Part C Emerging Technologies*, vol. 118, Article ID 102701, 2020.
- [14] S. Van Cranenburgh and A. Alwosheel, "An artificial neural network based approach to investigate travellers' decision rules," *Transportation Research Part C: Emerging Technologies*, vol. 98, pp. 152–166, 2019.
- [15] A. J. Maren, C. T. Harston, and R. M. Pap, *Handbook of Neural Computing Applications*, Academic Press, New York City, NY, USA, 2014.
- [16] M. E. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand*, MIT Press, Cambridge, MA, USA, 1985.
- [17] J. De Dios Ortuzar and L. G. Willumsen, *Modelling Transport*, John Wiley & Sons, Hoboken, NJ, USA, 2011.
- [18] K. Small, *Urban Transportation Economics*, Taylor & Francis, Oxford, UK, 2013.
- [19] A. Soltani and A. Shams, "Analyzing the influence of neighborhood development pattern on modal choice,"

- Journal of Advanced Transportation*, vol. 2017, Article ID 4060348, 11 pages, 2017.
- [20] X. Ma, J. Yang, C. Ding, J. Liu, and Q. Zhu, "Joint analysis of the commuting departure time and travel mode choice: role of the built environment," *Journal of Advanced Transportation*, vol. 2018, Article ID 4540832, 2018.
- [21] S. A. Zargari and F. Safari, "Using clustering methods in multinomial logit model for departure time choice," *Journal of Advanced Transportation*, vol. 2020, pages, 2020.
- [22] M. Pirra and M. Diana, "A study of tour-based mode choice based on a support vector machine classifier," *Transportation Planning and Technology*, vol. 42, no. 1, pp. 23–36, 2019.
- [23] H. I. Ashqar, M. Elhenawy, M. H. Almanna, A. Ghanem, H. A. Rakha, and L. House, "Modeling bike availability in a bike-sharing system using machine learning," in *Proceedings of the 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, Naples, Italy, June 2017.
- [24] A. Lhéritier, M. Bocamazo, T. Delahaye, and R. Acuna-Agost, "Airline itinerary choice modeling using machine learning," *Journal of Choice Modelling*, vol. 31, pp. 198–209, 2019.
- [25] J. Hagenauer and M. Helbich, "A comparative study of machine learning classifiers for modeling travel mode choice," *Expert Systems with Applications*, vol. 78, pp. 273–282, 2018.
- [26] M. Ferrara, C. Liberto, M. Nigro, M. Trojani, and G. Valenti, "Multimodal choice model for e-mobility scenarios," *Transportation Research Procedia*, vol. 37, pp. 409–416, 2019.
- [27] 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: Transport Science*, vol. 15, no. 2, pp. 1687–1612, 2019.
- [28] E. Pekel and S. Soner Kara, "A comprehensive review for artificial neural network application to public transportation," *Journal of Engineering & Natural Sciences*, vol. 35, no. 1, 2017.
- [29] M. Wong, B. Farooq, and G.-A. Bilodeau, "Discriminative conditional restricted boltzmann machine for discrete choice and latent variable modelling," *Journal of Choice Modelling*, vol. 29, pp. 152–168, 2018.
- [30] B. Sifringer, V. Lurkin, and A. Alahi, "Let me not Lie: learning multi nomial logit," 2018, <https://arxiv.org/abs/1812.09747>.
- [31] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.
- [32] Y. Gal and G. Zoubin, *A Theoretically Grounded Application of Dropout in Recurrent Neural Networks*, University of Cambridge, Cambridge, UK, 2016.
- [33] D. Wang, J. Zhang, W. Cao, J. Li, and Y. Zheng, "When will you arrive? estimating travel time based on deep neural networks," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, Toronto, Canada, February 2018.
- [34] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [35] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," 2015, <https://arxiv.org/abs/1502.03167>.
- [36] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?" in *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montréal, Canada, 2018.
- [37] <https://www.amap.com/>.
- [38] <http://www.bjtrc.org.cn/List/index/cid/7.html>.
- [39] H. C. W. L. Williams, "On the formation of travel demand models and economic evaluation measures of user benefit," *Environment and Planning A: Economy and Space*, vol. 9, no. 3, pp. 285–344, 1977.
- [40] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, New York City, NY, USA, 2016.
- [41] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, O'Reilly Media, Newton, MA, USA, 2019.
- [42] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Sardinia, Italy, 2010.
- [43] D. P. Kingma and J. Ba, "A method for stochastic optimization," 2014, <https://arxiv.org/abs/1412.6980>.
- [44] Y. Wang, Y. Zhang, X. Piao, H. Liu, and K. Zhang, "Traffic data reconstruction via adaptive spatial-temporal correlations," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 4, pp. 1531–1543, 2018.
- [45] L. Sun and K. W. Axhausen, "Understanding urban mobility patterns with a probabilistic tensor factorization framework," *Transportation Research Part B: Methodological*, vol. 91, pp. 511–524, 2016.
- [46] A. Nuzzolo and A. Comi, "Individual utility-based path suggestions in transit trip planners," *Iet Intelligent Transport Systems*, vol. 10, no. 4, pp. 219–226, 2016.