Transportation Research Part C 121 (2020) 102832

Contents lists available at ScienceDirect

Transportation Research Part C

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

Data-driven choice set generation and estimation of route choice models

Rui Yao, Shlomo Bekhor *

Department of Civil and Environmental Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel

ARTICLE INFO

Keywords:
Route choice
Data-driven
Discrete Choice
Feature selection
Random Forest

ABSTRACT

This paper proposes a novel combination of machine learning techniques and discrete choice models for route choice modeling. The data-driven choice set generation method identifies routes characteristics by clustering, and implicitly generates the choice set by sampling route characteristic attributes from the clusters. Important features are selected by random forests for route choice model development. With the selected features, the methodological-iterative approach is applied to specify the utility functions and to find significant explanatory variables automatically.

Results show that the proposed data-driven method produces a discrete route choice model not only with strong explanatory power, but also with high prediction accuracy compared to models estimated with conventional choice set generation methods.

1. Introduction

Route choice deals with individual's preference on alternative routes, and it is highly important for transportation network planning, traffic prediction, and travel behavior analysis. Route choice modeling is a nontrivial task, since it involves the determination of a relevant choice set, the specification of utility functions and the estimation (and application) of route choice models.

In the context of route choice, determining the choice set is particularly challenging, as many possible alternatives exist in the network. Moreover, choice set size and composition influence model estimation and prediction (Bekhor et al., 2008; Bovy, 2009). On one hand, a small choice set may not be able to reproduce the chosen route, and consequently may not capture the individual's behavior and preference. On the other hand, a very large choice set might cover the chosen route, but possibly lead to misinterpretation of the estimated route choice coefficients, as relatively few alternatives are actually perceived by individuals.

Route choice set is generated either by explicit (selective) methods which generates a subset of routes using pre-defined rules, or by implicit methods without specifying the alternative routes a priori (Bekhor et al., 2006). Both methods have advantages and shortcomings. The explicit methods are mainly based on variants of shortest path algorithms, which artificially construct a set of routes by changing link impedances. Explicit methods are generally computational efficient, but could lead to biased estimation results. Conversely, implicit methods do not construct a path-based choice set, but consider successive choices at each node at a link level. The implicit methods are theoretically appealing, because of its ability to reach unbiased results. However, implicit link-based methods require intensive computation, especially in large dense networks.

In recent years, new data collection methods provide researchers with detailed spatiotemporal movement data and rich individual information. Together with high resolution network data, it creates new opportunity to obtain better insights in individuals' route

* Corresponding author.
E-mail addresses: andyyao@campus.technion.ac.il (R. Yao), sbekhor@technion.ac.il (S. Bekhor).

https://doi.org/10.1016/j.trc.2020.102832
Received 18 June 2020; Received in revised form 6 October 2020; Accepted 7 October 2020
Available online 1 November 2020
0968-090X/

choice behaviors. However, specifying the utility function of the route choice models is a nontrivial task. Many models today are still specified using trial-and-error based on researcher's interpretations and experiences.

The evolution and rapid advance of machine learning methods, which increased the capabilities to unveil data patterns, and guidelines to the modeling process, are becoming increasingly popular. There are already several implementations of machine learning methods in the field of transportation, such as the identification of different travel modes and prediction of travel times in the network. However, in the context of route choice, only a handful of studies applied data-driven methods to identify route choice set for each origin-destination pair (Rieser-Schüssler et al., 2013; Ton et al., 2018), and to select route choice attributes (Tribby et al., 2017).

In this paper, we develop a systematic way for route choice modeling by combining data-driven techniques with discrete route choice models. The machine learning methods are applied to both route choice set generation and choice modeling. Machine learning techniques, such as clustering, decision trees and random forests (RF), have proven their robustness and ability to provide precise predictions. In parallel, discrete choice models have been widely researched, and their capabilities in explaining individual's route choice behaviors are well known. This paper applies a combination of both machine learning methods and traditional discrete choice models, which not only provides rich behavioral interpretation, but also high prediction accuracy. This novel combination will advance the state of art modeling approaches, and potentially open new directions in route choice modeling. In the following subsections, we briefly review the choice set generation methods and machine learning methods, to provide a more accurate insight on the contributions of the present paper.

1.1. Choice set generation

Many choice set generation methods were developed in the literature for route choice modeling, and they can be broadly divided into explicit selective methods and implicit methods (Bekhor et al., 2006).

Explicit choice set generation methods can be categorized into deterministic methods, stochastic and probabilistic methods. Both deterministic and stochastic explicit methods were developed based on variations of shortest path searches.

Typically, shortest paths are iteratively generated by changing different input variables, such as link impedances, search criteria and constraints. For example, Ben-Akiva et al. (1984) proposed the labelling approach, in which researchers pre-define a set of labels, formulate different generalized link costs based on their interpretation of the labels, and find minimum generalized cost routes. Unlike different deterministic k-shortest path methods, the labeling approach preserves the behavioral interpretation of the routes generated, but requires knowledge and experiences on defining appropriate labels a priori.

Stochastic explicit (simulation) methods (Ramming, 2001; Bekhor et al., 2006) randomly draw link impedances from different probability distributions. A comparison of different choice set generation methods is recently presented in Yao and Bekhor (2021). The results indicate that conventional methods are unable to fully reproduce the observed routes, and only combining different methods we can obtain high coverage with average choice set size of 50 routes.

Probabilistic methods are also termed as sampling methods, in which routes generated are considered sampled from a universal choice set. This method applies random walks to generate routes, starting from the origin node, and randomly select an outgoing link at each node (Frejinger et al., 2009) in acyclic network, or create detours for some randomly selected segment along the shortest path in a general network (Flötteröd and Bierlaire, 2013). However, Axhausen and Schüssler (2009) argued that the sampling method is sensitive to the parameter settings for sampling distributions, and inappropriate setting may generate unreasonable routes.

The explicit methods are in general computational attractive, but one shortcoming is that there is no guarantee to reproduce the observed route in the choice set, and may cause misleading interpretation of the estimation results. In contrast to explicit methods, implicit methods proposed by Fosgerau et al. (2013) do not require to generate routes as a first step to model estimation. Instead, a link-based modeling approach is adapted, by assuming individuals make successive decisions at each node. However, apart from heavy computations, the Bellman equation of the link-based implicit methods may not be solvable when the network contains cycles and the link cost is small enough (Oyama and Hato, 2017).

1.2. Machine learning methods

One of the major data sources for route choice modeling is GPS data. Many studies extract individuals' selected route from trajectory information, and generate alternative routes for discrete route choice modeling (Bierlaire et al., 2010; Dhakar and Srinivasan, 2014; Zimmermann et al., 2017). Recently, GPS data are further used to identify the route choice sets by combining with data-driven approaches. For example, clustering method is applied on paths' origins and destinations to identify groups of habitual paths based on GPS observations (Ciscal-Terry et al., 2016; Ton et al., 2017). The data-driven approach ensures that the chosen route is included in the choice set, and all the routes in the choice set are attractive alternatives observed from the data. However, this deterministic data-driven approach has one shortcoming, the alternative routes that are not chosen are not included in the choice set (Ton et al., 2017).

Discrete route choice modeling applies random utility maximization theorem to explicitly interpret individual's route choice behavior. However, with high resolution network, high dimensional and large volume of route choice data, specifying an appropriate utility function is nontrivial, and often requires researchers' experience and manual tuning. To tackle this problem, Shiftan and Bekhor (2021) recently developed a methodological-iterative (MI) approach to find significant explanatory variables among all available features, and serve as a starting point for further model development. The method was applied manually, and few iterations were performed, which did not provide satisfying prediction results.

In contrast to discrete choice models, some studies applied machine learning techniques to understand the route choice behavior, and enjoy the robust prediction ability provided by machine learning methods. Decision Trees are used in the literature to

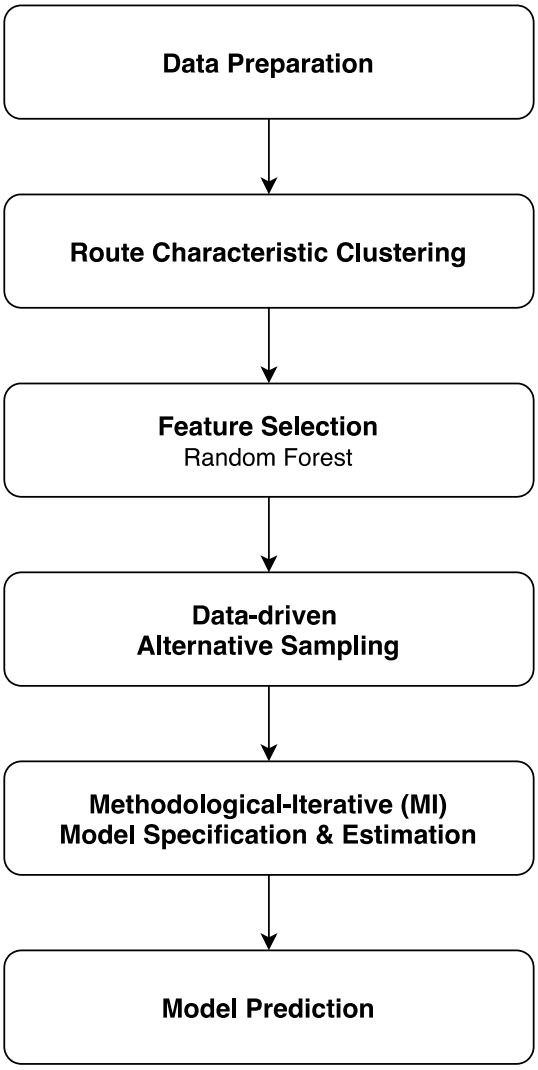

Fig. 1. 6-step modeling approach.

accommodate route choice preferences (Yamamoto et al., 2002; Park and Bell, 2011). Sun and Park (2017) utilized neural network (NN) and support vector machine (SVM) on a small dataset obtained from driving simulator, their study explores the possibilities of using non-parametric NN and SVM techniques for route choice modeling and results in good prediction power. Lai et al. (2018) evaluate and compare three different machine learning methods, NN, SVM and Random Forest, with econometric approach for route choice modeling. One of their results is that, machine learning methods generally outperform discrete route choice models in terms of models' goodness-of-fit to the data and prediction. Moreover, random forest performs the best among machine learning and econometric models in both estimation and prediction, and is also able to provide the importance of different variables. However, there are concerns of using machine learning approach for choice modeling. Non-parametric machine learning methods, for example hierarchical clustering method, may be too descriptive, and lead to inconsistent estimation results (Bhat and Dubey, 2014).

To the best of our knowledge, only few papers discussed combining machine learning techniques with traditional discrete choice models. Wong et al. (2018) applied generative method in machine learning to discover underlaying latent behavior in choice decision process without explicitly using attitudinal attributes. Tribby et al. (2017) utilized the feature importance information obtained by random forest, to select variables for route choice modeling. A similar approach is also adopted for mode choice modeling by Cheng et al. (2019) and Jahangiri and Rakha (2015). Their results indicate that the data-driven machine learning method improves goodness of fit of the models estimated, compared to commonly used utility functions.

In summary, although several route choice models were developed in the literature, there are still outstanding points. First, route choice set generation is a challenging task. Both explicit and implicit methods have their advantages and shortcomings. On one hand, explicit methods may not be able to reproduce observed routes. On the other hand, implicit methods can provide unbiased results, but require intensive computational power. Data driven approach to identify route choice set is appealing, because these methods can

guarantee to include all the observed routes in the choice set, however, these deterministic methods may not include alternatives which are considered but not chosen.

Secondly, traditional discrete route choice models often require manual tuning for model specification to achieve good estimation results. Machine learning models have shown their strength in both model estimation and prediction for route choice modeling. However, machine learning models are merely able to provide statistical relations of the samples (Lai et al., 2018).

In this paper, we aim to fill some of these gaps, by developing a systematic data-driven approach for choice set generation, route choice model specification, estimation and prediction. The developed approach is a novel combination of machine learning methods and traditional discrete choice models, and is expected to bring the following contributions:

(1) The developed data-driven choice set generation method implicitly generates routes with behavioral interpretations without the need to artificially define the choice set generation rules a prior.
(2) Random forest, one of the advanced machine learning techniques, is applied to extract important route attributes, and served as a starting point for model development with the potential to avoid over specification.
(3) The methodological-iterative approach is modified to construct utility functions in an automatic manner, and to find significant explanatory variables as well.
(4) The estimated model not only improves the interpretation of individual's behavior, but also improves the model prediction accuracy.

2. Methodology

The proposed methodology is a six-step procedure (Fig. 1) for choice set generation and route choice modeling, which incorporates machine learning methods and a methodological-iterative process for systematically specifying the utility functions and estimating the model.

The developed approach requires the extraction of route attributes from observations as inputs to the route characteristics clustering. Unlike previous studies, in which clustering is applied to identify the route choice set for each origin-destination pair, our clustering step identifies observed routes with common characteristics.

The next step uses the cluster labels together with all available attributes to train the random forest classifier. The random forest provides feature importance as criteria to select features with significant explanatory powers.

Based on the selected features provided by random forest, we apply the data-driven alternative sampling method to generate choice set. The sociodemographic and common features remain the same for different alternatives, while the alternative specific attributes are sampled and re-constructed from the fitted distributions.

After obtaining the choice set, the methodological-iterative step constructs the utility functions and estimate route choice model automatically. This step iteratively checks for each selected feature, by combining highly correlated coefficients and keeping only the significant coefficients.

In the last step, the prediction power of the estimated model is validated by performing cross-validation and comparing with other models. In the following sections, we will discuss the six steps in detail.

2.1. Data preparation

In data preparation step, sociodemographic attributes, trip specific attributes and route specific attributes are extracted from the dataset. Specifically, route specific attributes require additional processing for the route characteristic clustering step, while both sociodemographic attributes and trip specific attributes can enter directly into the later modeling steps.

For route specific attributes, we introduce normalized route characteristic attributes to account for observations with different trip length, origins and destinations that are widely spread in the network. The normalized route characteristic attributes are expected to capture the overall characteristics of the routes regardless the trip length, and to avoid endogeneity in the 6-step approach. We describe in detail the 11 route characteristic attributes in the following:

2.1.1. Route directness

The route directness is a ratio of the route distance L_{od}^{i} to the straight-line distance between origin $\boldsymbol{o}$ and destination $\boldsymbol{d}$, $\|\boldsymbol{o}-\boldsymbol{d}\|$, of observation i. The lowest value is 1.0, in which the route is the same distance as the Euclidean distance. Number closer to 1.0 indicates a more directed route, it is calculated using Eq. (1):

$$RouteDirectness = \frac{L_{od}^{i}}{\|\boldsymbol{o}-\boldsymbol{d}\|} \tag{1}$$

2.1.2. Route percentage delay

Route percentage delay measures the congestion on the route, which calculates the percentage difference between route time-dependent travel time $TT_{od}^{i,t}$ with respect to the departure period t (i.e. AM peak, PM peak and off peak) and the free flow route travel time TT_{od}:

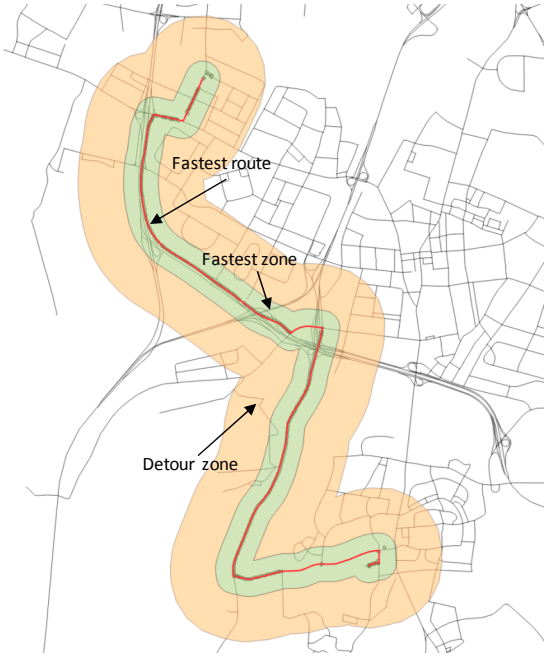

Fig. 2. Fastest zone and detour zone.

$$RoutePercentDelay = \frac{TT_{od}^{i,t} - TT_{od}}{TT_{od}} \tag{2}$$

The minimum value of route percentage delay is 0.0, which indicates the travel time on the route coincides with free flow travel time. With a smaller route percentage delay value, the route is less congested.

2.1.3. Route average speed

The route average speed is obtained by dividing the route length to route time-dependent travel time.

2.1.4. Route average number of links

The route average number of links is calculated by dividing the total number of links to route length. In general, high average number of links indicates there are more intersections in the route, and the route is likely to be more complicated.

2.1.5. Route left turn percentage

The route left turn percentage is calculated by dividing the number of left turns $LT_{od}^{i}$to the number of intersections of the route IT_{od}^{i}, using Eq. (3):

$$RouteLeftTurnPercent = \frac{LT_{od}^{i}}{IT_{od}^{i}} \tag{3}$$

The left turn percentage ranges from 0.0 to 1.0, a smaller value means the route is with less left turns, and may suggest less waiting time at the intersection (Prato and Bekhor, 2006).

2.1.6. Route city node percentage

Similar to left turn percentage, the route city node percentage measures how much the route passes through intersections in the city center, city node typically implies higher congestion, longer waiting time at the intersection. But it may also depend on trip purpose, origin and destination of the trip.

2.1.7. Route highway/expressway percentage

This attribute is calculated by dividing the length of highway and expressway links to the total route length. Route highway/expressway percentage ranges between 0.0 and 1.0, with a value close to 1.0 implying the route has higher capacity, less intersections and higher speed.

2.1.8. Route average operating cost

The route average operating cost calculates per kilometer fuel cost and toll cost if using toll ways. The fuel cost is a function of

speed, in which higher speed results in less fuel cost. While using toll ways requires extra payments but typically with higher speed and lower fuel cost.

2.1.9. Route average intersection time

This attribute calculates the average time spent at the intersections.

2.1.10. Route length detour

The route length detour attribute calculates the detour in percentage from the shortest path length SL_{od}:

$$RouteLengthDetour = \frac{L^i_{od}}{SL_{od}} \tag{4}$$

2.1.11. Route time detour

Similar to route length detour, route time detour finds the excess percentage route travel time from time-dependent fastest path travel time ST^t_{od}

$$RouteTimeDetour = \frac{TT^{i,t}_{od}}{ST^t_{od}} \tag{5}$$

Apart from the route characteristic attributes mentioned above and other travel attributes, network related features are also calculated for each observation. In the recent study of Yao and Bekhor (2021), it is observed that more than 60% of the GPS trajectories coincides with the fastest route for the same dataset. To account for routes that are similar to the fastest route, we introduce the fastest zone and detour zone around the fastest route (Fig. 2). Note that, the fastest route is explicitly found regardless if the actual trajectories are the fastest. The fastest zone is then defined with a buffer distance around the fastest route, by assuming a maximum distance of 300 m or 2% of the fastest route length:

$$BufferDistance_{FastestZone} = \max(300\ \text{m},\ 2\%\ \ Fastest\ \ Route\ \ Length) \tag{6}$$

Similarly, the buffer distance for the detour zone is defined as Eq. (7), note that the detour zone excludes the fastest zone:

$$BufferDistance_{DetourZone} = \max(1000\ \text{m},\ 10\%\ \ Fastest\ \ Route\ \ Length) \tag{7}$$

In summary, the data preparation step extracts sociodemographic attributes, trip specific attributes (e.g. trip purpose, trip departure time), fastest route and shortest route information and observed route characteristic attributes. The normalized route characteristic attributes are used in the next step for clustering, which identifies observed routes with common characteristics, and interprets the clusters with behavioral meanings.

2.2. Route characteristic clustering

The second step identifies routes with similar patterns in terms of their normalized route characteristic attributes, and serves as a base for choice set generation. We run K-means clustering method (Arthur and Vassilvitskii, 2006) on all the map matched routes. The K-means clustering method aims to separate the n observations $\boldsymbol{X} = \{x_1, \cdots, x_n\}$ into k disjoint clusters $= \{C_1, \cdots, C_k\}$, where each cluster is described by the mean μ_i of the samples in the cluster, such that the inertia (i.e. variance) criterion (Eq. (8)) is minimized:

$$\arg_C \min \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{8}$$

The K-means algorithm first chooses the initial centroids, which are randomly separated from each other. Then the algorithm repeatedly assigns data points to their nearest centroid, and calculates the mean value of all of the data points assigned to each previous centroid. This process is run till the centroid locations do not vary significantly.

One of the parameters required by the K-means algorithm is the number of clusters k. There are several methods proposed in the literature to determine the number of clusters. In this paper, we find the number of clusters k that maximize the silhouette score. The silhouette score measures how close the samples are to their cluster centroids, and how far away the samples are to their neighboring clusters. The score falls within the range of $[-1, 1]$, in which 1 indicates that the data point is far away from its neighborhood cluster, and -1 indicates that the data point has been assign to the "wrong" cluster:

$$SilhouetteScore = \frac{1}{n} \sum_{j=1}^{n} \frac{b(x_j) - a(x_j)}{\max\{a(x_j), b(x_j)\}} \tag{9}$$

where $a(x_j)$ measures the mean distance between data point x_j to all other data points in the same cluster C_i:

$$a(x_j) = \frac{1}{|C_i| - 1} \sum_{x_k \in C_i, j \neq k} d(x_j, x_k) \tag{10}$$

and $b(x_j)$ measures the minimum distance between data point x_j to all data points in other cluster:

$$b(x_j) = \min_{h \neq i} \frac{1}{|C_h|} \sum_{x_k \in C_h} d(x_j, x_k), \quad \forall x_j \in C_i \tag{11}$$

Apart from labeling each observation by their route characteristics, the clustering step also provides behavioral interpretation of the clusters. The clusters can help understanding the characteristics of the routes being chosen by the respondents, as illustrated and discussed later in the paper.

2.3. Feature selection

Feature selection techniques are widely used as a preprocessing step to select relevant features for model development, and it is expected to reduce overfitting and simplify the model for easier interpretation (Saeys et al., 2007).

A common feature selection method is random forest, which ensembles a collection of decision trees. A decision tree classifier requires each observation to be labeled; for this reason, the previous step of clustering route characteristics is performed, to provide such information for classification.

Decision tree successively performs a set of "if-then" operations to partition the space of features V into a set of mutually exclusive regions, and predict the class of the observation (Breiman, 2001). The Gini impurity measures the likelihood of misclassification of data. Given a set of data D reaching at node k in the decision tree, which contains data from n classes, and denoting p_j to be the relative frequency of class j in D, the Gini impurity is defined as Eq. (12):

$$Gini(D) = \sum_{j=1}^{n} p_j \cdot (1 - p_j) \tag{12}$$

At each node k in the decision tree, any potential binary split of D using different features $v \in V$ is evaluated with the weighted sum of the impurity of the resulting partitions $D_{1,v} \cup D_{2,v} = D$. Only if there is a reduction in Gini impurity by the split using feature v, the decision tree will continue the partition operation, and the feature v results maximum $\Delta Gini_k$ (i.e. maximum reduction in Gini impurity) is selected:

$$\Delta Gini_k = Gini(D) - \left(\frac{|D_{1,v}|}{|D|} Gini(D_{1,v}) + \frac{|D_{2,v}|}{|D|} Gini(D_{2,v}) \right) \tag{13}$$

The node importance ni_k (weighted Gini impurity reduction by the number of samples arriving at node k), measures how much the variable v contributes to improve the classification in the decision tree:

$$ni_k = |D| Gini(D) - \left(|D_{1,v}| Gini(D_{1,v}) + |D_{2,v}| Gini(D_{2,v}) \right) \tag{14}$$

With the same notion, the importance of each feature fi_v in the decision tree is averaged using Eq. (15):

$$fi_v = \frac{\sum_{k:nodes\ \ split\ \ using\ \ feature\ \ v} ni_k}{\sum_{l:all\ \ nodes} ni_l} \tag{15}$$

The normalized feature importance is obtained by dividing by the sum of all feature importance in the decision tree:

$$normfi_v = \frac{fi_v}{\sum_{u \in V} fi_u} \tag{16}$$

Random forest creates a large number of decision trees, by repeatedly selecting a random sample (with replacement) of the training set and a random subset of the features, and fits decision trees to the samples. The random forest classifier applies a "majority voting" mechanism over all the decision trees to obtain the final classification result, and it is expected to outperform a single decision tree. Feature importance using random forest is obtained by averaging the normalized feature importance over all the decision trees (Louppe, 2014):

$$RFfi_v = \frac{1}{T} \sum_{t \in alltrees} \text{norm}fi_{v,t} \tag{17}$$

where T is the total number of trees in the random forest, $\text{norm}fi_{v,t}$ is the normalized feature importance of feature v in decision tree t.

There are several parameters to be specified for random forests: 1) the number of decision trees; 2) maximum number of features to consider when looking for the best split; 3) minimum number of samples required to split a tree node; 4) minimum number of samples required to be at a leaf node; 5) maximum depth of the tree. These parameters are selected by running a randomized hyperparameter search with cross-validation.

The feature importance is obtained after running the random forest, features with an importance value smaller than 0.02 is considered insignificant and removed from the feature list. The rest of the features are sorted by importance order and given to the following steps.

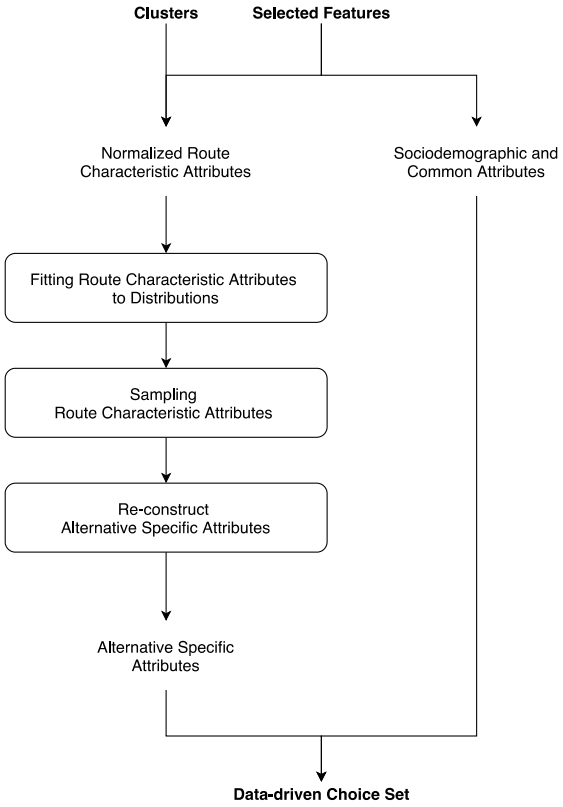

Fig. 3. Data-driven Choice Set Generation.

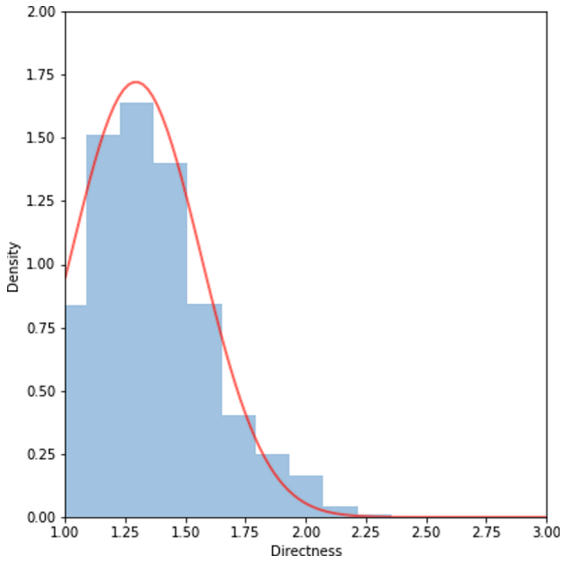

Fig. 4. Directness distribution fitting.

2.4. Data-driven alternative sampling

As discussed in the introduction section, conventional explicit choice set generation methods do not guarantee to reproduce the observed routes, and implicit methods are computational expensive. The data-driven alternative sampling method tries to tackle some of these issues.

The proposed data-driven alternative sampling approach is inspired by the labeling methods (Ben-Akiva et al., 1984). Instead of

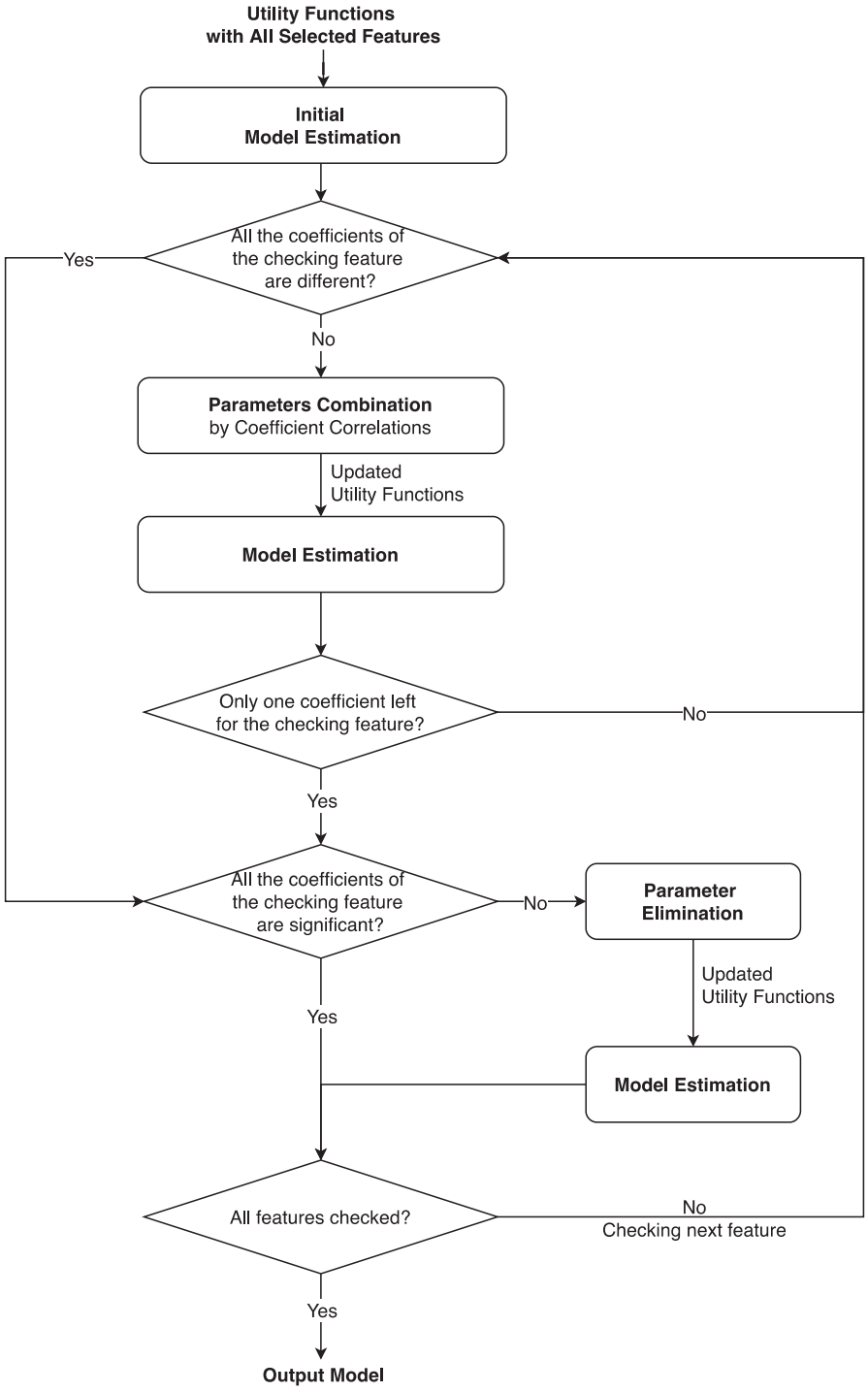

Fig. 5. Methodological-Iterative Model Specification and Estimation.

artificially defining the labels and explicitly searching for alternative routes in the network with these labels, we sample the normalized route characteristic attributes to describe the alternative routes. Our approach tries to capture individuals' interpretation on different routes by differentiating routes with their characteristics, rather than their explicit attribute values. The data-driven alternative sampling method is presented in Fig. 3.

The clusters are expected to cover the characteristics of the attractive routes considered by the individuals. Therefore, the number of clusters is the same as the number of alternatives in the choice set, and each alternative corresponds to one of the route

characteristics labels.

Data-driven alternative sampling method generate alternatives based on selected important features, which can be divided into two groups: Group 1 – normalized route characteristic attributes as mentioned in the data preparation section; Group 2 – sociodemographic and other common attributes shared by the same observation.

Group 2 features enter directly into the choice set, while Group 1 – normalized route characteristic attributes and cluster information are used to generate alternative specific attributes.

Each normalized route characteristic attribute (Group 1 features) is fitted to a cluster-specific distribution. Considering the physical meanings of these attributes, they are at least bounded from one side. For example, the minimum value of route directness is equal to 1, which means the length of the chosen route is equal to the Euclidean distance, and it is the most direct route. We consider fitting these attributes to truncated normal distributions, which are described by the mean μ, variance σ^2, lower bound a and upper bound b. Note that, other distributions might be considered as well, we consider truncated normal distributions in this paper.

Fig. 4 illustrates fitting directness attributes of one cluster to a truncated normal distribution. The lower bound a of directness attribute is 1.00, which indicates the route length is equal to the Euclidean distance, and it is unbounded from above (i.e. upper bound $b = \infty$). A truncated normal distribution defined at $[1.00, \infty)$ with mean μ of 1.29, and variance σ^2 of 0.07 calculated from the data.

We then sample the normalized route characteristic attributes from these distributions for each alternative and each observation. And re-construct the normalized attributes to absolute alternative specific attributes using the Eqs. (1)–(5) and their relations.

For example, the *RouteDirectness* attribute is sampled from the distribution shown in Fig. 4. Using relation of Eq. (1), the route length attribute of alternative i from origin o to destination d, L_{od}^{i} is then re-constructed by Eq. (18):

$$L_{od}^{i} = RouteDirectness \cdot \|\boldsymbol{o} - \boldsymbol{d}\| \tag{18}$$

Note that we purposely use normalized characteristics in the sampling process, and re-constructed to absolute attribute values in the choice set. This is needed because the observations are drawn from different origin-destination pairs, the normalized characteristics allow proper handling of the spatial attributes in the clustering step. For example, two routes passing mainly on highway/expressway links, with Euclidean distance of 10 km and 40 km, and route length of 15 km and 60 km respectively, will not be grouped together in the same cluster due to large differences in route length. But if normalized route characteristics are used, both of them have route directness of 1.5, and together with other normalized route characteristics, they can be clustered together.

The normalized route characteristics and the re-construction process also help avoiding endogeneity issue in model estimation and prediction step. In k-means clustering, distance measurements between data points are used to cluster data points. The normalized route characteristics will generate different distance measurements than using route attributes. Essentially, different datasets are used in the clustering step (normalized route characteristic dataset) and model classification/prediction step (route attribute dataset) to avoid endogenous issue.

Furthermore, the normalized route characteristics are able to provide cluster-specific distributions for all the OD pairs. These distributions are expected to aggregate information on diverse routes that share the same characteristics. And then we are able to generate attractive alternatives that are not observed for a specific OD pair, using information gathered from routes of different OD pairs in the same cluster.

By combining the alternative specific attributes, sociodemographic and common attributes, we generated a data-driven choice set with attractive alternatives and each alternative in the choice set is labeled with their characteristics. This label information with each alternative provides new possibilities in systematically specifying the utility functions and estimating the models.

2.5. Methodological-iterative (MI) model specification and estimation

The clustering step not only helps to generate an attractive choice set, but also provides labels to each alternative. These labels allow us to specify utility functions with alternative specific coefficients, and improve behavioral interpretation of the models. The feature selection step reduces the coefficient estimation search space, and consequentially reduce the computational power required. Moreover, these selected important features can serve as a starting point for the MI method, and potentially ease the overfitting issue.

The MI method aims to find statistically significant explanatory variables, and obtain a model with high goodness-of-fit and prediction power. The overall MI procedure is shown in Fig. 5. The MI method can be seen as a two-level backward elimination procedure, in which the upper level checks all the features (e.g. Highway/expressway length, route delay, etc.), and the lower level checks all the coefficients for a checking feature (i.e. combine/eliminate insignificantly different alternative specific coefficients for a checking feature). We describe the procedure in detailed in the following paragraphs.

Given J alternatives, the MI method starts by initializing utility function U_i with all the selected features for $J-1$ alternatives as defined in Eq. (19), and the Jth alternative is set as the reference:

$$U_i = ASC_i + \sum_{x \in V} \beta_{x,i} \cdot x \tag{19}$$

where

ASC_i: alternative specific constant of alternative i,
V: the set of selected features,
x: the selected feature,

Table 1
Example utility functions.

Alternative	ASC	x_{time}	x_{cost}
1	ASC_1	$\beta_{\text{time\_1}}$	$\beta_{\text{cost\_1}}$
2	ASC_2	$\beta_{\text{time\_2}}$	$\beta_{\text{cost\_2}}$
3	$ASC_3(0)$	$\beta_{\text{time\_3}}(0)$	$\beta_{\text{cost\_3}}(0)$

Table 2
MI example iteration estimation results.

Iteration	$\beta_{\text{time\_1}}$	$\beta_{\text{time\_2}}$	$\beta_{\text{cost\_1}}$	$\beta_{\text{cost\_2}}$	$\beta_{\text{cost\_1\_2}}$	$\lvert\beta_{\text{time\_1}} - \beta_{\text{time\_2}}\rvert$	$\lvert\beta_{\text{cost\_1}} - \beta_{\text{cost\_2}}\rvert$
0	**−1.50**	**−2.50**	**−0.03**	**−0.04**	–	**1.00**	**0.01**
	(0.00)	(0.00)	(0.03)	(0.03)	–	(0.00)	(0.80)
1	**−1.60**	**−2.70**	–	–	**−0.04**	**1.10**	–
	(0.00)	(0.00)	–	–	(0.01)	(0.00)	–
1′	**−1.60**	**−2.70**	–	–	**−0.04**	**1.10**	–
	(0.00)	(0.00)	–	–	(0.03)	(0.00)	–
2′	**−1.60**	**−2.75**	–	–	–	**1.15**	–
	(0.00)	(0.00)	–	–	–	(0.00)	–

$\beta_{x,i}$: alternative specific coefficient for feature $x \in V$ alternative i

Then a discrete choice model with the utility functions defined above is estimated. Different model forms can be used, but for simplicity, this paper considers the multinomial logit (MNL) model form.

The MI method now runs in a two-level iterative manner. The upper-level checks selected features from less important to most important one-by-one, and eliminate insignificant features. The lower-level iterative process combines highly correlated coefficients and eliminates insignificant coefficients of a checking feature, to further avoid overfitting and improve model performance.

Starting from the least important feature x, the MI method checks the coefficient correlations of all $\beta_{x,i}(\forall i, i \neq J-1)$. If the coefficients corresponding to the current feature x in different utility functions are not significantly different, the two highly correlated coefficients $\beta_{x,j}$ and $\beta_{x,k}$ are combined to one coefficient β_{x,j_k}. Then another discrete choice model is estimated with the updated utility functions, and the coefficient correlations are checked again.

This lower iterative process for combining coefficients of feature x is stopped when either all the coefficients of feature x are significantly different from each other, or there is only one coefficient left. The next step of the lower-level process eliminates all the coefficients that are not significantly different from zero, and estimate a new discrete choice model with the updated utility functions.

The MI method then runs in the upper level, and moves to the second worse feature y. The lower iterative process is then executed for feature y. The MI procedure terminates when all the features in the selected feature set V are checked. The final outputs of the MI process are the model specification and an estimated discrete choice model.

We provide an illustrative example of the MI procedure in the following. Consider 3 alternatives with 2 selected features – *time* and *cost*, in which *time* has higher feature importance than *cost*. Using Eq. (19), the utility functions of the 3 alternatives are shown in Table 1, with the 3rd alternative setting as reference. Note that, only β_{time} and β_{cost} of alternative 1 and 2 are considered in this MI procedure.

The estimation results of each MI iteration is shown in Table 2, in which the value of coefficients are shown in bold and the p-values are shown in the bracket.

We first explain the lower iterative process in detail. During initialization (iteration 0), a discrete choice model is estimated with all the coefficients defined in Table 1. We start checking from the least important feature: *cost*. As shown in $|\beta_{cost\_1} - \beta_{cost\_2}|$, the absolute difference between coefficients β_{cost_1}, β_{cost_2} is relatively small and the p-value is larger than the pre-defined 0.02 threshold, which suggests coefficient β_{cost_1} and β_{cost_2} are not significantly different.

We then try to combine these two alternative specific coefficients to be $\beta_{cost\_1\_2}$, and estimate a new discrete choice model with updated utility functions. Note that, at iteration 0, the p-values of both β_{cost_1} and β_{cost_2} are larger than the pre-defined threshold. At this iteration, we do not eliminate these two coefficients, instead we combine them. This is because the combination of these two coefficients might be significant, and the significance of the combined coefficient will be verified in the later iterations.

At iteration 1, we obtain estimation results of the new discrete choice model with the combined coefficient $\beta_{cost\_1\_2}$. In this example, we have only one coefficient $\beta_{cost\_1\_2}$ left for the checking feature: *cost*, the lower iterative combination process is finished. We then need to check if the combined coefficient $\beta_{cost\_1\_2}$ is significantly different from 0, and eliminate it otherwise.

Suppose all the coefficients of the checking feature are significant, the lower-level process is finished. As in iteration 1, the combined coefficient $\beta_{cost\_1\_2}$ is significantly different from 0, we can finish checking feature *cost* and move to the next feature *time*.

However, if any coefficient of the checking feature is not significantly different from 0, all of these insignificant coefficients will be removed, and a new discrete choice model with the updated utility functions should be estimated before moving to the next checking feature. We show this case in iteration 1′, in which the p-value of the combined coefficient $\beta_{cost\_1\_2}$ is larger than the threshold. A new discrete choice model without the coefficient $\beta_{cost\_1\_2}$ is estimated, and the results are shown in iteration 2′. The table shows that the

checking feature *cost* is actually removed from the utility functions.

We finish the lower-level process for the feature *cost*, and now move to the next feature *time*. In this example (both iteration 1 and iteration 2′), β_{time_1} and β_{time_2} are significantly different from each other, and no coefficient combination is needed. We now check if all the coefficients after the lower-level coefficient combination process are significantly different from 0. In this case, they are all significant, and then the MI process is finished.

The feature checking order agrees with feature importance, as more important features are expected to provide stronger explanatory power to the model. And less important features may only provide marginal improvement to the model, and this process will remove these insignificant coefficients to avoid overfitting.

2.6. Model prediction

In order to validate the model prediction power, the dataset of 5,002 samples is randomly split into a training set with 4,000 samples and a testing set with 1,002 samples for independent prediction.

The estimated discrete model is used to predict the label of each observation in the test set, the label corresponds to the highest probability is assigned to the observation. The model prediction accuracy is evaluated using the F1 score:

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{20}$$

where

t_p: true positive returned by the model
f_p: false positive returned by the model
f_n: false negative returned by the model

$$\text{precision} = \frac{t_p}{t_p + f_p}$$

$$\text{recall} = \frac{t_p}{t_p + f_n}$$

For the random forest forecasting, the predicted chosen alternative is defined as the one with highest mean probability estimate across all the trees. As for all the discrete choice models, the predicted chosen alternative is simulated given the probabilities of all the alternatives.

We compare the proposed model with conventional route choice models. For this comparison, we generate explicit choice sets using two different methods: link penalty and labeling. A maximum of 20 routes per OD pair are generated using link penalty method, and a maximum of 14 routes are generated with labeling method. The utility function specification for the labeling method is exactly as in the MI models, while the utility specification in link penalty method lacks the alternative specific constants (because the alternatives cannot be labeled).

3. Results

In this section, we present the results of the data-driven choice set generation and route choice modeling approach. The results are obtained using the Tel Aviv household travel survey data and map matched GPS trajectories with a detailed Tel Aviv metropolitan planning network.

The following subsections are organized in accordance with the 6-step proposed methodology. The first subsection presents the candidate features obtained from the data preparation step. The route characteristic clustering results based on normalized route characteristic attributes are shown in the second subsection. The third step applies random forest to select important features, and consequently the feature importance and selected features are shown in the third subsection. The fourth subsection presents results of the data-driven alternative sampling method with the selected features. The estimation results of the MI route choice models with the generated choice set are shown in the fifth subsection. Lastly, the proposed model prediction power is examined by comparing it with two MNL route choice models estimated with conventional choice set generation methods.

3.1. Data preparation

The GPS observation data is obtained from Tel Aviv household travel survey, which was conducted using a designated mobile phone application. The survey collected information from 28,530 individuals and 265,815 trips over a 2-day period. A detailed description of the respondent recruitment and data collection process can be found in Nahmias-Biran et al. (2018).

The Tel Aviv travel survey includes household information, individual sociodemographic data and their activities and travel diary with GPS records: the exact time and locations of their movements for 48 consecutive hours with 2 s time step on average, the purpose of the tours and the mode. Apart from providing the GPS data, the rich information in this survey allows us to better understand individuals' route choice behavior.

In this paper we focus on a subset of the survey: car trips with origin–destination Euclidean distance at least 2 km for meaningful

Table 3
Data statistics.

	Car Trips(Main activitiesTrip Length >= 2 km)	SelectedMap Matched Trips		Car Trips(Main activitiesTrip Length >= 2 km)	SelectedMap Matched Trips
Unique Individual	11,376	2488	**Household Size (persons)**		
Sample Size	28,984	5002	1	10%	11%
			2	28%	24%
Trip Purpose			3	18%	18%
Commute Trips	95%	96%	4	21%	23%
Maintenance Trips	4%	3%	5	15%	16%
Personal Trips	1%	1%	5+	8%	8%
Trip Length [km]			**Gender**		
<5	38%	36%	Male	57%	51%
5–10	29%	32%	Female	43%	49%
10–15	14%	15%	**Age**		
15+	19%	17%	15–30	13%	12%
Household Vehicles			30–40	21%	23%
1	40%	37%	40–50	25%	28%
2	49%	52%	50–60	19%	20%
2+	11%	11%	60+	22%	17%

Table 4
Route characteristic attribute general statistics.

Route characteristics	min	max	average	standard deviation
Route directness	1.00	5.57	1.46	0.33
Route percentage delay	1.01	16.45	1.85	0.61
Route average speed	2.33	105.65	36.89	13.35
Route average number of links	0.44	26.69	4.42	2.21
Route left turn percentage	0.00	1.00	0.11	0.09
Route city node percentage	0.00	1.00	0.10	0.19
Route highway/expressway percentage	0.00	1.00	0.71	0.29
Route average operating cost	0.56	9.52	1.06	0.39
Route average intersection time	0.00	0.71	0.18	0.07
Route length detour	1.00	4.28	1.11	0.21
Route time detour	1.00	3.16	1.08	0.17

route choice behavior interpretation. Moreover, we focus on the individual's trip for his/her main activity, classified as commute, maintenance (e.g. shopping) or personal trips (e.g. visiting friends). We then select the trips based on their GPS record availability, which results in a subset of 5,002 trips for map matching, the GPS trajectory filtering process is illustrated in the Appendix A.

In Table 3, we provide overall statistics of the car trip dataset and the subset of map matched trips. From these data, we can confirm that the map matched trips are representative of the overall car trip sample.

After cleaning and filtering the GPS data, the 5,002 car trips are map matched to a detailed planning network of Tel Aviv metropolitan area, which contains 8,583 nodes and 21,151 directed links. We apply Hidden Markov Model (Newson and Krumm, 2009) to perform map matching, by assuming the GPS data noise follows Gaussian distribution with mean $\mu = 0$ and standard deviation $\sigma = 20$, tolerance for non-direct route $\beta = 2$ and the maximal search distance for candidate road segment is 50 m. In addition, manual inspection on the samples of the matched routes was performed to ensure the matching quality.

The normalized route characteristic attributes are then extracted from the 5,002 map matched trajectories using Eqs. (1)–(5) and their relations. The general statistics of the 11 route characteristic attributes are shown in Table 4.

Note that these characteristics are network related attributes, which can be calculated for any dataset that includes network topology. The ranges of these attribute values are related to the resolution of the network. The attribute values may be different if using navigation networks, comparing to planning networks (as in this paper).

The data preparation step extracts candidate sociodemographic features, trip specific features (e.g. trip purpose, trip departure time), fastest route and shortest route features and observed route characteristic features, as presented in Table 5. The rich dataset provides 50 distinct features that are potential explanatory variables for the discrete choice model.

Among all these features, route information features are alternative specific. Therefore, these features are converted into 11 normalized route characteristic attributes using Eqs. (1)–(5) and the additional relations described in the methodology section. These normalized attributes provide unified indicators of the route characteristics, regardless the distance between the origins and destinations. In the next section, the normalized route characteristic attributes will be used for clustering routes with similar characteristics.

Table 5
Candidate features.

Sociodemographic features		Trip specific features		Route specific features			
Individual information	Age	OD information	Distance between OD	Shortest and fastest route information	Fastest route time	Route information	Route Length
	Gender		Network density around origin		Shortest path distance		Highway/ Expressway length
	Level of education		Network density around destination		Shortest path travel time		Route time
	Full time employer		Origin is in CBD dummy		Shortest path directness		Route intersection time
	Part time employer		Destination is in CBD dummy		Average delay in fastest zone		Route delay
	Unemployed	Trip departure time	Evening peak trip		Average delay in detour zone		Route cost
	Retired		Morning peak trip		Average speed in fastest zone		Route number of links
	Housewife		Off peak trip		Average speed in fastest zone		Route number of left turns
	Solider	Trip purpose	Commute trip		Highway/Expressway length percentage in fastest zone		Number of city nodes
	Student		Maintenance trip		Highway/Expressway length percentage in detour zone		Toll road
Household information	Household number of vehicles		Personal trip		Average intersection time in fastest zone		
	Household area		Number of on-board passengers		Average intersection time in detour zone		
	Household size				City node percentage in fastest zone		
	Number of children under 8 in the household				City node percentage in detour zone		

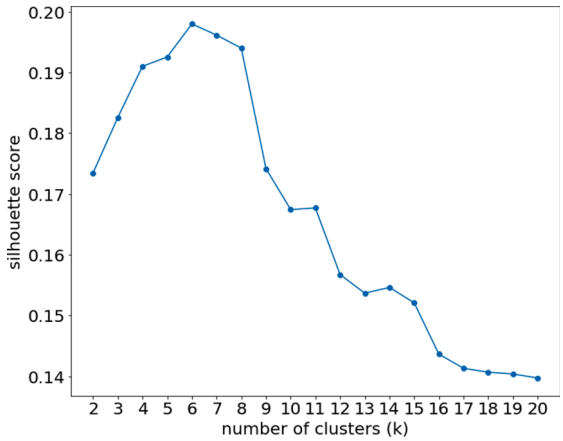

Fig. 6. Silhouette score to number of clusters.

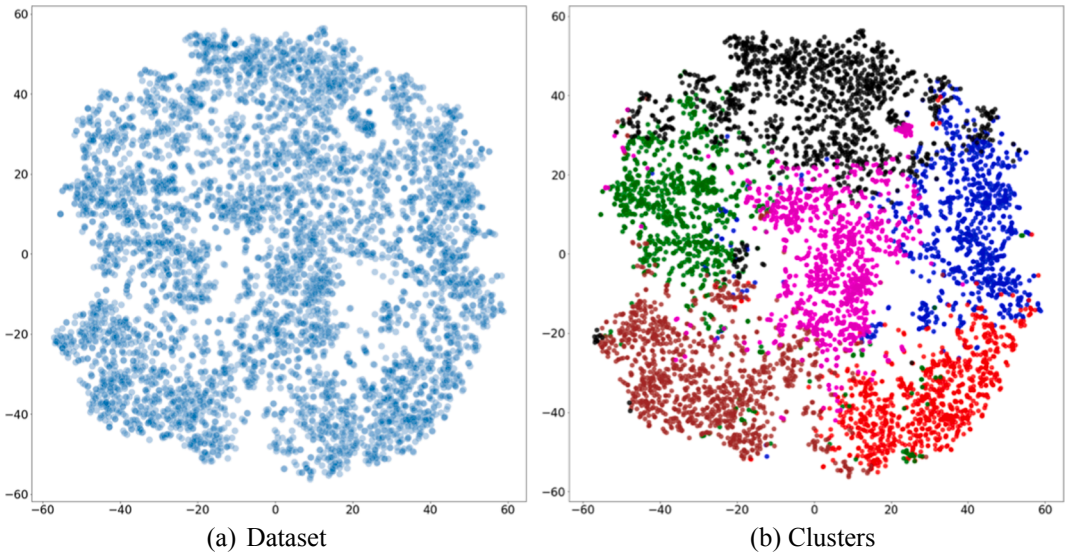

Fig. 7. t-SNE cluster visualization.

3.2. Route characteristic clustering

Using the normalized route characteristic attributes, this step groups routes with similar characteristics into clusters. As indicated in the methodology section, we select the number of clusters k that maximize the Silhouette score, by running k-means clustering with number of clusters ranging from 2 to 20. Fig. 6 shows the changes in Silhouette score with respect to different cluster numbers, in which 6 clusters appears to have the maximum Silhouette score. Together with considerations of the balanced clusters (i.e. all resulting clusters with similar number of elements), we choose to group the routes into 6 clusters. T-distributed Stochastic Neighbor Embedding (t-SNE) is applied to provide visualization of the dataset and clustering results in a two-dimensional space (Maaten and Hinton, 2008), as shown in Fig. 7.

The clustering step not only provides each route with a cluster label, but also provides us behavioral interpretations to route characteristic labels.

The route characteristic labels can be inferred by comparing the mean values of the attributes of each specific cluster relative to all observations. We present the percentage differences of each cluster to the overall means in Table 6, in which the worst characteristic values are marked in red, and best characteristic values are marked in green.

The behavioral interpretations of each cluster can be established by finding the most significant attribute comparing to other clusters (percentages comparing to overall averages are shown in the bracket). For example, cluster 1 can be labeled as maximizing highway/expressway routes because of its highest percentage of highway/expressway, highest average speed (29.4%), lowest number of percentage delay (−14.3%), number of links (−34.0%) and intersection time (−30.9%) among all clusters.

As opposed to cluster 1, cluster 2 has the lowest percentage of highway/expressway (−62.4%) with relatively lower speed

Table 6
Route characteristics attributes of clusters.

Route Characteristics	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Overall Average
Route directness	1.42 -3.6%	1.42 -3.8%	1.43 -3.0%	**1.32** **-10.3%**	**1.89** **28.6%**	1.37 -7.3%	1.47
Route percentage delay	**1.61** **-14.3%**	1.62 -13.9%	**2.60** **38.5%**	2.17 15.4%	1.71 -8.8%	1.77 -5.6%	1.88
Route average speed	**52.34** **45.8%**	27.60 -23.1%	30.55 -14.9%	**24.37** **-32.1%**	36.25 1.0%	37.80 5.3%	35.89
Route average number of links	**2.99** **-34.0%**	5.34 17.9%	4.03 -11.1%	6.04 33.2%	**6.15** **35.8%**	3.19 -29.7%	4.53
Route left turn percentage	**6.79%** **-39.9%**	15.21% 34.7%	7.59% -32.8%	9.13% -19.1%	14.02% 24.1%	**16.04%** **42.0%**	11.30%
Route highway/expressway percentage	**91.11%** **29.4%**	**26.51%** **-62.3%**	86.01% 22.2%	70.99% 0.8%	67.33% -4.4%	81.73% 16.1%	70.41%
Route average operating cost	0.95 -15.7%	1.06 -6.0%	1.03 -8.6%	1.04 -8.0%	**1.66** **46.7%**	**0.93** **-17.7%**	1.13
Route length detour	1.07 -4.0%	1.06 -5.1%	1.07 -3.7%	1.07 -3.7%	**1.41** **26.1%**	**1.03** **-7.8%**	1.11
Route time detour	1.04 -5.4%	1.07 -2.7%	1.11 0.6%	1.09 -0.7%	**1.30** **17.9%**	**1.02** **-7.3%**	1.10
Route city node percentage	2.44% -76.5%	3.91% -62.3%	7.78% -24.9%	**48.06%** **364.1%**	5.81% -43.9%	**2.01%** **-80.6%**	10.36%
Route average intersection time	**0.12** **-30.9%**	0.18 4.3%	0.15 -15.2%	0.20 16.1%	0.17 -4.2%	**0.26** **45.5%**	0.18
Number of observations	1103	932	753	699	779	736	5002
Interpretation	Highway routes	Urban routes	Highly congested routes	Routes at city center	Detour routes	Shortest and Fastest routes	

(−23.1%), and can be interpreted as routes within urban areas. Cluster 3 has the largest percentage delay (38.5%), and therefore characterizes very congested routes.

Cluster 4 corresponds to routes that are with more direct routes (−10.3%), mainly pass through city center nodes (364.1%) and low travel speed (−32.1%). These characteristics are well suited to describe routes at the city center.

Among all clusters, cluster 5 represents detour routes which have the least direct route (28.6%), meaning the route is more detour and longer. In general, a more detour (26.1%)/ longer (18.0%) route consists of more links (35.8%), and therefore the average operating cost (46.7%) is higher.

And lastly, cluster 6 represents shortest and fastest routes, in which the route length detour (−7.8%) and time detour (−7.3%) is the smallest, and consequently the average operational cost is low (−17.7%). The shortest and fastest routes avoid passing through city centers (−80.6%), but may require more left turns (42.0%) than other routes, thus higher average intersection time (45.5%).

3.3. Feature selection

As shown in Table 5, there are 50 candidate features to be considered for model specifications. The feature selection step eliminates

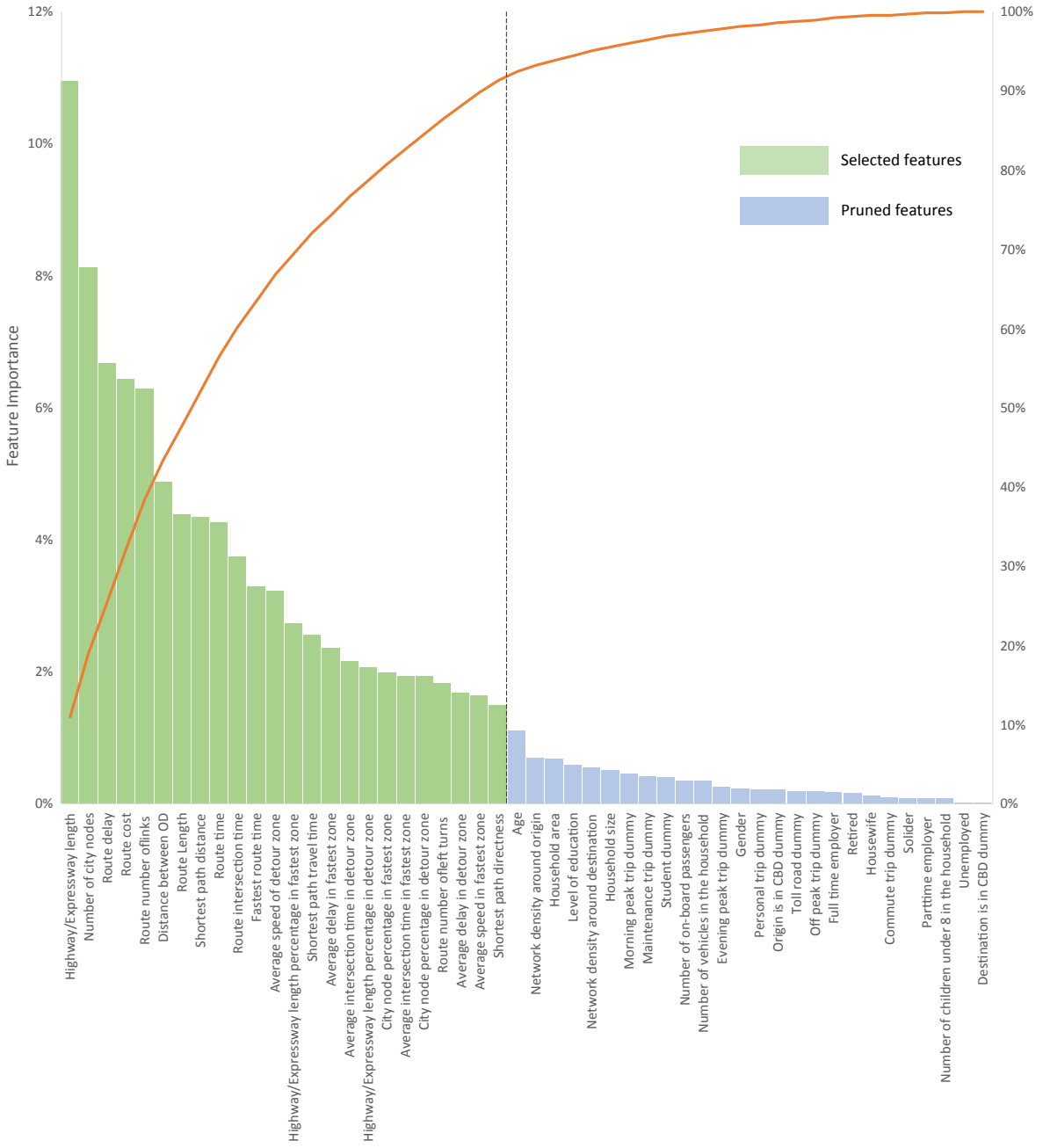

Fig. 8. Feature importance.

features that do not significantly contribute to reduce the impurity of decision trees in the random forest.

First, the 5,002 observations are randomly split into a training subset of size 4,000 and another test subset with the rest of the observations. In order to have a fair comparison between different models, the same training subset and test subset will be used in different models.

A randomized hyperparameter search with cross-validation is performed to determine the parameters used in the random forest. The following parameters are used in this study: number of decision trees = 500; maximum number of features to consider when looking for the best split = 8; minimum number of samples required to split a tree node = 10; minimum number of samples required to be at a leaf node = 4; maximum depth of the tree = 80. With these parameters, the random forest reaches an accuracy of 99.9% on the training set and 83.3% accuracy on the test set (in terms of weighted average F1 score).

Fig. 8 shows the feature importance rank on a bar chart, where the features with importance more than 2% are marker in green, and other features in blue. The cumulative feature importance is also plotted in Fig. 8. The 2% threshold for selecting features corresponds

Table 7
Selected features (with importance ranking).

1	**Highway/Expressway length**	13	Highway/Expressway length percentage in fastest zone
2	**Number of city nodes**	14	Shortest path travel time
3	**Route delay**	15	Average delay in fastest zone
4	**Route cost**	16	Average intersection time in detour zone
5	**Route number of links**	17	Highway/Expressway length percentage in detour zone
6	Distance between OD	18	City node percentage in fastest zone
7	**Route Length**	19	Average intersection time in fastest zone
8	Shortest path distance	20	City node percentage in detour zone
9	**Route time**	**21**	**Route number of left turns**
10	**Route intersection time**	22	Average delay in detour zone
11	Fastest route time	23	Average speed in fastest zone
12	Average speed of detour zone	24	Shortest path directness

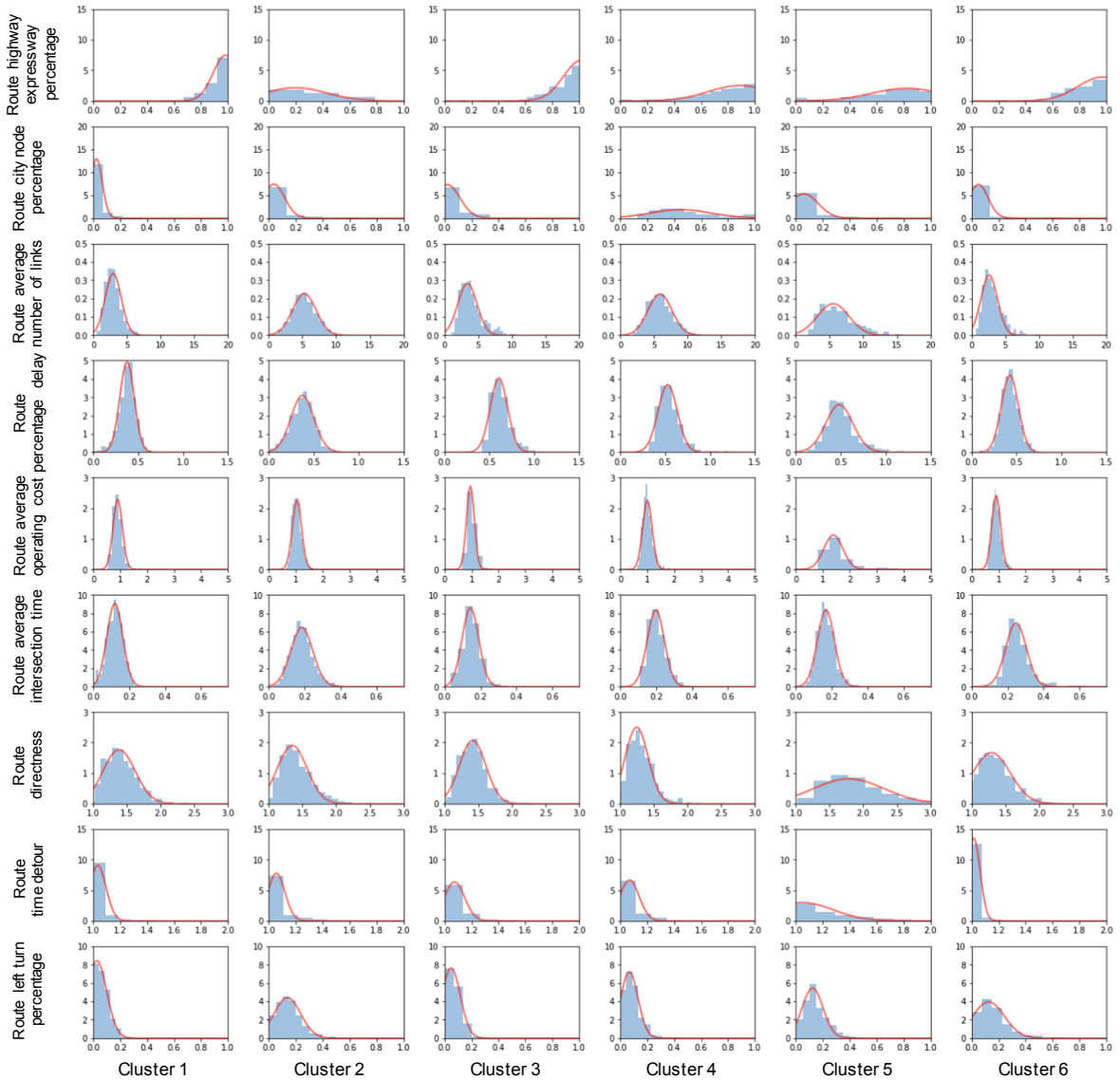

Fig. 9. Route characteristic distributions of clusters.

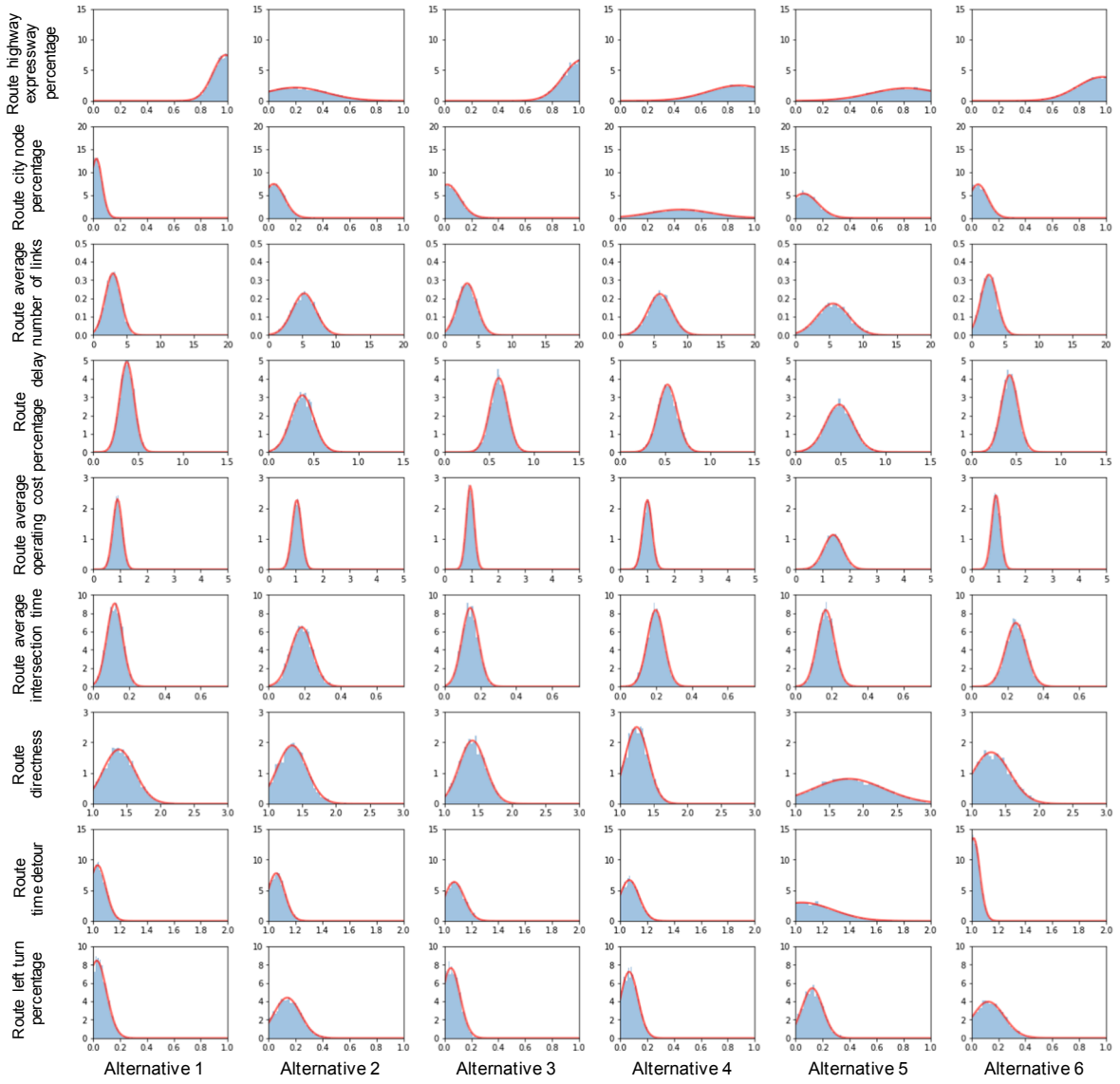

Fig. 10. Distributions of sampled alternatives.

to over 90% of the feature importance, with only 24 features out of 50 candidate features.

The selected features are shown in Table 7 with their importance ranking (descending rank). As expected, the majority of the selected features are network related, this may suggest that route choice behaviors are more related to network characteristics than sociodemographic characteristics.

Most selected features are in line with the literature. For example, individuals are more sensitive to the percentages of highway/expressway length. The number of city nodes also plays an important role, which suggests that a route passing in urban environment is an important decision factor affecting individuals' route choice. Table 7 also indicates that many fastest route related features are presented in the selected feature set, which implies that many individuals are aware of the fastest route, perhaps related to navigation apps (such as Waze and Google Map).

The 24 selected features provide a starting point for route choice model specification. Out of these features, 9 selected route information features (marked in bold in Table 4) will be used in the next step to generate alternatives. The other 15 features are the same for different alternatives of the same observation, and enter directly into the later modeling steps.

3.4. Data-driven alternative sampling

The data-driven approach generates alternative routes by sampling route characteristic attributes from the clusters, instead of explicitly searching for routes in the network. The proposed data-driven choice set generation method tries to capture individuals'

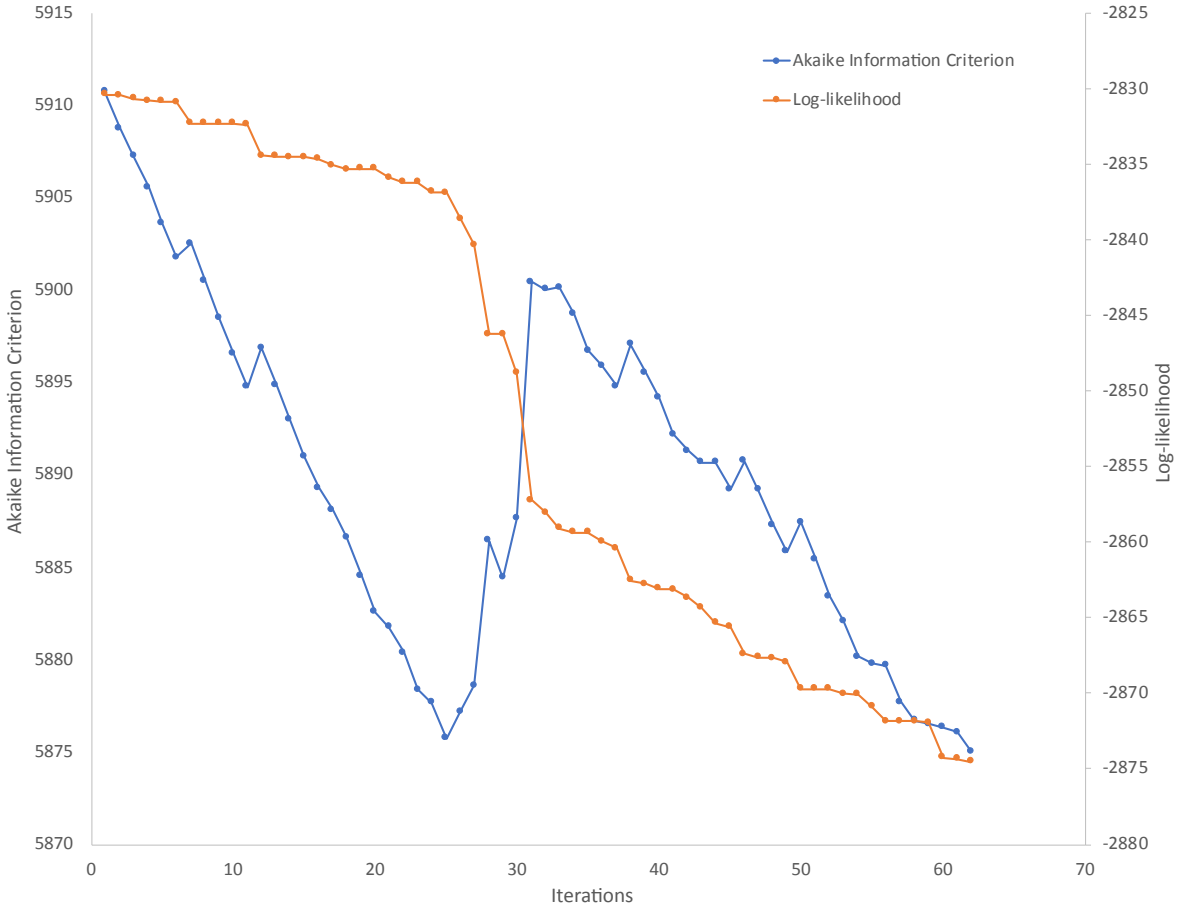

Fig. 11. MI route choice model AIC and log-likelihood values.

interpretation on different routes, in which individuals differentiate routes with the route characteristics rather than the attributes. The following 9 normalized route characteristic attributes which correspond to the 9 selected features are sampled: route highway expressway percentage, route city node percentage, route average number of links, route percentage delay, route average operating cost, route average intersection time, route directness, route time detour, route left turn percentage.

Each of these normalized route characteristic attributes is fitted to a cluster-specific truncated normal distribution as shown in Fig. 9. The red curve plotted in Fig. 9 represents the probability density function of the fitted truncated normal distribution. The different distributions of the route characteristic attribute among clusters also justify that the previous clustering step can generate distinguished clusters.

For each observation, the chosen alternative matches the observed normalized route characteristic attributes. Other alternatives are obtained by sampling the normalized route characteristic attributes from the corresponding fitted truncated normal distributions. In this paper, there are 6 alternatives, which correspond to the 6 clusters obtained in the route characteristic clustering step.

Fig. 10 presents the normalized route characteristic attribute distributions of the 6 alternatives (including the chosen one) for all 5,002 observations. The figure indicates that the sampled alternative attributes resemble the cluster distributions. For example, alternative 1 corresponds to the first cluster label (freeway routes), which has highest freeway/expressway percentage and smallest city node percentage (i.e. most highways do not pass through city centers). The other alternatives respectively correspond to their cluster label interpretations.

These normalized route characteristic attributes are then re-constructed to selected route information features using Eqs. (1)–(5) and their relations. Other common features, such as sociodemographic features and trip specific features, are shared between different alternatives and enter directly in the utility functions of each alternative. The resulting choice set includes all the selected features with alternative specific attributes obtained from data-driven alternative sampling. The next step utilizes this choice set for route choice modeling.

3.5. Methodological-iterative (MI) model specification and estimation

The MI model starts with an initial model with all 24 selected features. The utility functions of each alternative are specified using

Table 8
Final methodological-iterative route choice model (with alternative 6 as reference).

Coefficient		Value	*t*-test
ASC 1		2.06	1.93
ASC 2		−0.4	−0.41
ASC 3		−9.23	−5.76
ASC 4		8.44	4.91
ASC 5		−7.32	−6.72
Alt 1,2,3,4,5	Highway/expressway length	0.33	12.4
Alt 1,2,3,4,5	Route distance	−0.19	−6.49
Alt 1,2,3,4,5	Average speed in detour zone	−0.31	−8.5
Alt 1,2,3,4,5	Shortest route travel time	0.55	14.9
Alt 1,3	number of left turns	−0.65	−15.4
Alt 2,4,5		−0.45	−20.6
Alt 1,2,4	route cost	0.19	8.37
Alt 3,5		−0.02	−2.66
Alt 1	intersection time	0.36	9.12
Alt 4		−0.17	−4.85
Alt 2,3,5		−0.06	−3.46
Alt 1,4,5	OD distance	0.61	9.08
Alt 3		1.19	11.5
Alt 1,2	route travel time	−0.97	−24.1
Alt 3,5		−0.14	−7.03
Alt 4		−0.3	−10.6
Alt 1,2,4,5	route number of links	0.39	17
Alt 3		0.74	11.4
Alt 1,3	number of city nodes	−0.16	−8.37
Alt 2		−0.43	−9.61
Alt 4		−0.05	−4.92
Alt 5		−0.11	−6.82
Alt 1,3	route delay	−0.16	−12.5
Alt 4,5		−0.06	−4.37
Alt 2	fastest route travel time	0.86	19.7
Alt 3,4		0.29	10.3
Alt 1	intersection time in detour zone	19.9	6.23
Alt 2,4		13.4	4.57
Alt 3		21.1	4.83
Alt 5		6.26	1.88
Alt 1	shortest route distance	0.44	6.07
Alt 2,5		−0.55	−8.68
Alt 3,4		−1.25	−14.8
Alt 3,5	shortest route directness	2.77	10.5
Alt 4		−0.2	−0.45
Alt 1,2,5	number of city nodes in detour zone	1.37	2.81
Alt 3		3.71	3.76
Alt 4		−0.44	−0.56
Alt 1,2	average delay in detour zone	−1.31	−1.26
Alt 3,5		2.21	1.99
Alt 4		−2.74	−1.79
Alt 1	highway/expressway percentage in detour zone	2.34	3.26
Alt 2,5		1.59	2.51
Alt 3		4.31	3.29
Alt 4		5.04	4.72
Alt 1	intersection time in fastest zone	−20.9	−8.4
Alt 2		−11.9	−5.46
Alt 3		−31.3	−7.29
Alt 4		−9.79	−3.32
Alt 5		−8.36	−3.44
Alt 1,3	number of city nodes in fastest zone	−3.37	−6.62
Alt 2,4,5		−0.89	−2.1
Alt 3	average delay in fastest zone	1.98	3.16
Alt 4		−4.18	−3.6
Alt 2	highway/expressway percentage in fastest zone	−6.74	−13.7
Alt 5		−5.12	−9.7
Alt 1,3,4	average speed in fastest zone	−0.1	−6.08
Alt 2,5		0.07	4.15
Number of observations		4000	
Number of coefficients		63	
***L*(0)**		−7167.04	
***L*(β)**		−2874.51	
ρ(0)		0.6	

Eq. (19), which results in 125 coefficients (one alternative-specific coefficient for each selected feature and an alternative specific constant for each alternative). Since the alternatives are labeled, it is possible to set an alternative as reference. The results for the initial MI MNL route choice model are presented in the Appendix A Table 12. Starting from the initial MI model, the MI method will iteratively combine or eliminate the non-significant coefficients.

The MI method runs in an iterative manner. At each iteration, it checks if the coefficients for the same feature are significantly different from each other (and also from 0). For the non-significant coefficients, they are either combined with other alternative specific coefficients for the same feature, or eliminated. The changes of model Akaike Information Criterion (AIC) and log-likelihood values to the iterations are shown in Fig. 11.

A total of 62 iterations are performed to reach the final MI model, which has only 63 coefficients out of the initial 125 coefficients. The final MI model has the smallest AIC value among all 62 models, which suggests the final model reduces the risk of overfitting. Note that at iteration 25, there is a significant increase in AIC value: this is because the procedure of coefficient elimination (i.e. removing features with insignificant coefficients from some of the utility functions) results in a large decrease in log-likelihood value.

The estimation results of the final MI model are presented in Table 8. Remember that, the coefficients shown in the final models are obtained by setting alternative 6 as the reference, which is shortest and fastest route. The log-likelihood ratio between the final and initial model is:

$$LR = 2\left[L\left(\widehat{\beta}_{Final,63}\right) - L_{Initial,125}\left(\widehat{\beta}_{Initial,125}\right)\right] = -88.286 \tag{21}$$

The initial model is better than the final model in terms of log-likelihood ratio (LR). However, as we will show in the following section, the final model performs slightly better in terms of prediction accuracy, which may because most of the coefficients in the final model are significant and the MI method potentially avoid overfitting with less coefficients.

The behavioral interpretations, provided by the clustering step, are consistent with the alternative routes. This allows to include alternative specific constants in the utility functions. Results show that most alternative specific constants are significant, while there may exist some similarities between alternative 2 and 6 (set as reference) based on their alternative specific constant values.

The MI method iteratively eliminates insignificant coefficients for a checking feature; in case that all the coefficients of a feature are insignificant, this feature is also removed. Note that the final MI model includes all 24 selected features, which confirms that the random forest is very effective in selecting significant features.

The estimated model provides meaningful insights to the route choice behaviors. For example, individuals in general prefer routes with longer highway/expressway, shorter distances, less travel time and delay, less left turns, and less city nodes. Individuals who are sensitive to route travel times are willing to spend more to save time.

Many attributes which are related to the fastest route have strong influences on individuals' route preferences as well. With higher route travel time on the shortest path, which coincides with alternative 6, individuals are more likely to choose the other 5 alternatives.

Comparing to the ones choosing alternative 6, individuals choosing highway routes (alternative 1), urban routes (alternative 2) and routes at city center (alternative 4) are less sensitive to the route cost.

Individuals choosing highway routes are more sensitive to the average intersection time, this may be related to the lower number of intersections in highway routes.

The negative value of average speed in the fastest zone for alternatives 1, 3, 4 means that individuals are more likely to choose alternative 6 (shortest and fastest route) instead of highway routes, highly congested routes and routes at city centers, if the speed in alternative 6 is higher. In contrast, individuals choosing urban routes (alternative 2) and detour routes (alternative 5) are less sensitive to average speed on the fastest zone than the ones choosing shortest and fastest route.

We also examine two additional options with checking coefficient significancy and eliminating insignificant coefficients before parameter combination procedure:

(1) Before checking each feature, remove all insignificant coefficients, only then perform parameter combinations for the checking feature
(2) Before each parameter combination procedure, remove all insignificant coefficients

However, using our dataset, both models underperform the final MI model in terms of final log likelihood and prediction accuracy. The corresponding flowcharts and final model estimation results are shown in the Appendix A. Note that with a different dataset, these two options may perform better, for this reason we also provide flowcharts to illustrate the methods in the Appendix A.

3.6. Model prediction

The last step in the methodology is to verify the model prediction. The prediction accuracies of the estimated models are compared using the same test subset split in the previous subsection. Apart from the prediction accuracy of the random forest model, initial and final MI models, initial MI model without random forest, another two models with conventional choice set generation methods are also compared, the labeling model and link penalty model.

The labeling model generates a route choice set, by searching routes with the following 14 subjective criteria: least time, least distance, least free-flow time, least delay, minimize operating cost, minimize left turn, minimize V/C-weighted time, turn-penalty hierarchy path, maximize freeways path, maximize expressways path, maximize arterial path, maximize number of intersections in

Table 9
Model prediction accuracies.

	Random forest	Final MI model	Initial MI model	Initial MI model without RF	Labeling model	Link penalty model
Test sample size	1002	1002	1002	1002	1002	1002
Number of features	50	24	24	50	24	24
Number of alternatives	6	6	6	6	14 (max)	20 (max)
Number of coefficients	–	63	125	255	325	294
Prediction accuracy (F1 score)	83%	73%	72%	72%	55%	47%

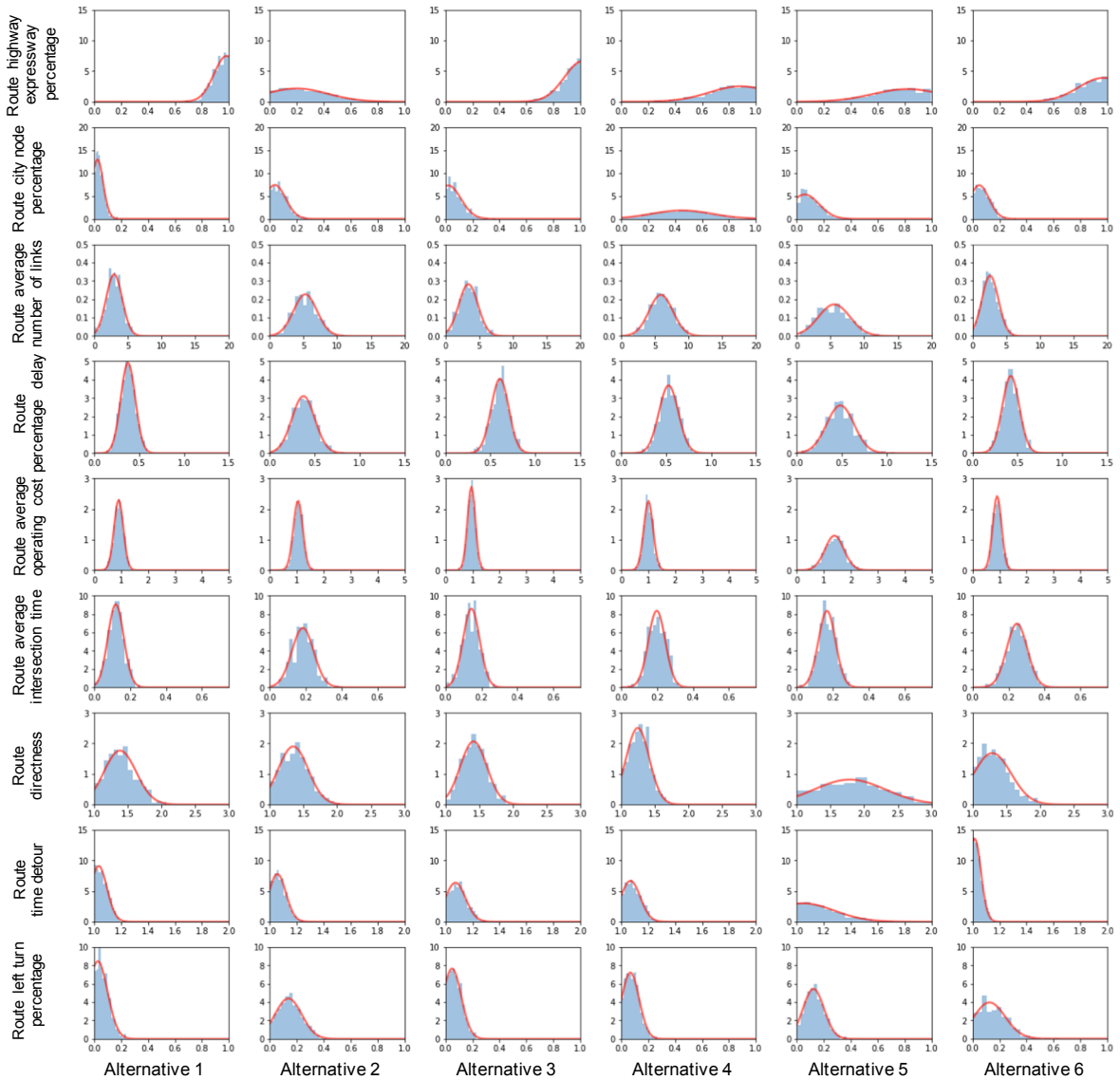

Fig. 12. Simulated observation attribute distribution (500 observations).

Table 10
Simulation study results.

Name		Sample										Avg.	Std.	Preset value	Abs. diff.
		1	2	3	4	5	6	7	8	9	10				
ASC 1		2.08	2.06	2.07	2.07	2.10	2.09	2.07	2.08	2.10	2.11	2.08	0.02	2.06	0.02
ASC 2		−0.40	−0.40	−0.40	−0.40	−0.40	−0.40	−0.40	−0.40	−0.40	−0.40	−0.40	0.00	−0.40	0.00
ASC 3		−9.25	−9.23	−9.24	−9.24	−9.23	−9.25	−9.24	−9.24	−9.25	−9.23	−9.24	0.01	−9.23	0.01
ASC 4		8.44	8.44	8.44	8.58	8.44	8.44	8.44	8.41	8.44	8.44	8.45	0.05	8.44	0.01
ASC 5		−7.32	−7.31	−7.32	−7.32	−7.32	−7.32	−7.32	−7.32	−7.32	−7.33	−7.32	0.00	−7.32	0.00
Alt 1,2,3,4,5	Highway/expressway length	0.31	0.34	0.33	0.33	0.29	0.28	0.31	0.29	0.32	0.28	0.31	0.02	0.33	0.03
Alt 1,2,3,4,5	Route distance	−0.22	−0.26	−0.19	−0.20	−0.24	−0.26	−0.23	−0.25	−0.21	−0.26	−0.23	0.03	−0.19	0.04
Alt 1,2,3,4,5	Average speed in detour zone	−0.31	−0.31	−0.31	−0.31	−0.31	−0.31	−0.31	−0.31	−0.31	−0.31	−0.31	0.00	−0.31	0.00
Alt 1,2,3,4,5	Shortest route travel time	0.52	1.04	0.55	0.55	0.47	0.45	0.51	0.49	0.53	0.43	0.55	0.18	0.55	0.00
Alt 1,3	number of left turns	−0.69	−0.65	−0.68	−0.68	−0.66	−0.69	−0.69	−0.68	−0.70	−0.66	−0.68	0.02	−0.65	0.03
Alt 2,4,5		−0.45	−0.48	−0.45	−0.45	−0.46	−0.45	−0.45	−0.45	−0.45	−0.47	−0.46	0.01	−0.45	0.01
Alt 1,2,4	route cost	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.00	0.19	0.00
Alt 3,5		−0.07	−0.51	−0.03	−0.03	−0.09	−0.11	−0.07	−0.11	−0.06	−0.14	−0.12	0.14	−0.02	0.10
Alt 1	intersection time	0.44	0.38	0.41	0.42	0.47	0.47	0.41	0.44	0.48	0.48	0.44	0.03	0.36	0.08
Alt 4		−0.17	−0.17	−0.17	−0.17	−0.17	−0.17	−0.17	−0.17	−0.17	−0.17	−0.17	0.00	−0.17	0.00
Alt 2,3,5		−0.09	−1.10	−0.06	−0.06	−0.10	−0.11	−0.09	−0.09	−0.07	−0.12	−0.19	0.32	−0.06	0.13
Alt 1,4,5	OD distance	0.59	0.64	0.61	0.61	0.58	0.57	0.59	0.58	0.60	0.56	0.59	0.02	0.61	0.02
Alt 3		1.00	1.17	1.02	1.00	1.16	1.02	1.02	1.02	0.96	1.16	1.05	0.08	1.19	0.14
Alt 1,2	route travel time	−0.97	−0.97	−0.97	−0.97	−0.97	−0.97	−0.97	−0.97	−0.97	−0.97	−0.97	0.00	−0.97	0.00
Alt 3,5		−0.18	0.04	−0.14	−0.14	−0.24	−0.26	−0.18	−0.22	−0.16	−0.29	−0.18	0.09	−0.14	0.04
Alt 4		−0.30	−0.30	−0.30	−0.30	−0.30	−0.30	−0.30	−0.30	−0.30	−0.31	−0.30	0.00	−0.30	0.00
Alt 1,2,4,5	route number of links	0.38	0.47	0.39	0.39	0.36	0.36	0.38	0.37	0.38	0.34	0.38	0.03	0.39	0.00
Alt 3		0.65	0.73	0.68	0.69	0.72	0.66	0.68	0.67	0.64	0.71	0.68	0.03	0.74	0.06
Alt 1,3	number of city nodes	−0.25	−0.16	−0.22	−0.22	−0.17	−0.24	−0.21	−0.25	−0.24	−0.18	−0.21	0.03	−0.16	0.06
Alt 2		−0.43	−0.43	−0.43	−0.43	−0.43	−0.43	−0.43	−0.43	−0.43	−0.43	−0.43	0.00	−0.43	0.00
Alt 4		−0.05	−0.05	−0.05	−0.05	−0.05	−0.05	−0.05	−0.05	−0.05	−0.05	−0.05	0.00	−0.05	0.00
Alt 5		−0.12	−0.75	−0.11	−0.11	−0.13	−0.13	−0.12	−0.13	−0.12	−0.15	−0.19	0.20	−0.11	0.08
Alt 1,3	route delay	−0.33	−0.21	−0.24	−0.30	−0.20	−0.32	−0.27	−0.30	−0.40	−0.21	−0.28	0.06	−0.16	0.11
Alt 4,5		−0.07	−0.23	−0.06	−0.06	−0.08	−0.09	−0.07	−0.08	−0.07	−0.10	−0.09	0.05	−0.06	0.03
Alt 2	fastest route travel time	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.00	0.86	0.00
Alt 3,4		0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.00	0.29	0.00
Alt 1	intersection time in detour zone	19.90	19.89	19.90	18.79	19.90	19.90	19.90	17.90	19.90	20.20	19.62	0.71	19.90	0.28
Alt 2,4		13.40	13.40	13.40	13.40	13.40	13.40	13.40	13.40	13.40	13.40	13.40	0.00	13.40	0.00
Alt 3		21.10	21.17	21.10	20.80	21.10	21.12	21.10	21.10	21.10	21.10	21.08	0.10	21.10	0.02
Alt 5		6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	6.26	0.00	6.26	0.00

(*continued on next page*)

Table 10 (*continued*)

Name		Sample										Avg.	Std.	Preset value	Abs. diff.
		1	2	3	4	5	6	7	8	9	10				
Alt 1	shortest route distance	0.77	0.47	0.69	0.68	0.96	0.88	0.71	0.76	0.97	1.05	0.79	0.17	0.44	0.35
Alt 2,5		−0.57	−0.50	−0.55	−0.55	−0.59	−0.60	−0.57	−0.59	−0.56	−0.63	−0.57	0.04	−0.55	0.02
Alt 3,4		−1.25	−1.25	−1.25	−1.25	−1.25	−1.25	−1.25	−1.25	−1.25	−1.25	−1.25	0.00	−1.25	0.00
Alt 3,5	shortest route directness	2.76	2.77	2.77	2.77	2.76	2.76	2.76	2.76	2.76	2.75	2.76	0.01	2.77	0.01
Alt 4		−0.20	−0.20	−0.20	−0.20	−0.20	−0.20	−0.20	−0.20	−0.20	−0.20	−0.20	0.00	−0.20	0.00
Alt 1,2,5	number of city nodes in detour zone	1.37	1.37	1.37	1.37	1.37	1.37	1.37	1.37	1.37	1.37	1.37	0.00	1.37	0.00
Alt 3		3.70	3.71	3.71	3.71	3.71	3.71	3.71	3.70	3.70	3.71	3.71	0.00	3.71	0.00
Alt 4		−0.44	−0.44	−0.44	−0.44	−0.44	−0.44	−0.44	−0.44	−0.44	−0.44	−0.44	0.00	−0.44	0.00
Alt 1,2	average delay in detour zone	−1.31	−1.31	−1.31	−1.31	−1.31	−1.31	−1.31	−1.31	−1.31	−1.31	−1.31	0.00	−1.31	0.00
Alt 3,5		2.21	2.21	2.21	2.21	2.21	2.21	2.21	2.21	2.21	2.21	2.21	0.00	2.21	0.00
Alt 4		−2.74	−2.74	−2.74	−2.74	−2.74	−2.74	−2.74	−2.74	−2.74	−2.74	−2.74	0.00	−2.74	0.00
Alt 1	highway/expressway percentage in	2.35	2.34	2.34	2.34	2.35	2.35	2.35	2.35	2.35	2.36	2.35	0.01	2.34	0.01
Alt 2,5	detour zone	1.59	1.59	1.59	1.59	1.59	1.59	1.59	1.59	1.59	1.59	1.59	0.00	1.59	0.00
Alt 3		4.31	4.31	4.31	4.31	4.31	4.30	4.30	4.31	4.30	4.31	4.31	0.00	4.31	0.00
Alt 4		5.04	5.04	5.04	5.04	5.04	5.04	5.04	5.04	5.04	5.04	5.04	0.00	5.04	0.00
Alt 1	intersection time in fastest zone	−20.90	−20.90	−20.90	−20.90	−20.90	−20.90	−20.90	−20.90	−20.90	−20.90	−20.90	0.00	−20.90	0.00
Alt 2		−11.90	−11.90	−11.90	−11.90	−11.90	−11.90	−11.90	−11.90	−11.90	−11.90	−11.90	0.00	−11.90	0.00
Alt 3		−31.30	−31.30	−31.30	−31.30	−31.30	−31.30	−31.30	−31.30	−31.30	−31.30	−31.30	0.00	−31.30	0.00
Alt 4		−9.79	−9.79	−9.79	−9.79	−9.79	−9.79	−9.79	−9.79	−9.79	−9.79	−9.79	0.00	−9.79	0.00
Alt 5		−8.36	−8.36	−8.36	−8.36	−8.36	−8.36	−8.36	−8.36	−8.36	−8.36	−8.36	0.00	−8.36	0.00
Alt 1,3	number of city nodes in fastest zone	−3.39	−3.38	−3.38	−3.38	−3.38	−3.38	−3.38	−3.39	−3.39	−3.38	−3.38	0.00	−3.37	0.01
Alt 2,4,5		−0.89	−0.89	−0.89	−0.89	−0.90	−0.89	−0.89	−0.89	−0.89	−0.89	−0.89	0.00	−0.89	0.00
Alt 3	average delay in fastest zone	1.97	1.97	1.97	1.97	1.97	1.97	1.97	1.97	1.97	1.97	1.97	0.00	1.98	0.01
Alt 4		−4.18	−4.18	−4.18	−4.18	−4.18	−4.18	−4.18	−4.18	−4.18	−4.18	−4.18	0.00	−4.18	0.00
Alt 2	highway/expressway percentage in	−6.74	−6.74	−6.74	−6.74	−6.74	−6.74	−6.74	−6.74	−6.74	−6.74	−6.74	0.00	−6.74	0.00
Alt 5	fastest zone	−5.12	−5.12	−5.12	−5.12	−5.12	−5.12	−5.12	−5.12	−5.12	−5.12	−5.12	0.00	−5.12	0.00
Alt 1,3,4	average speed in fastest zone	−0.10	−0.10	−0.10	−0.10	−0.10	−0.10	−0.10	−0.10	−0.10	−0.10	−0.10	0.00	−0.10	0.00
Alt 2,5		0.02	0.23	0.07	0.06	−0.03	−0.05	0.02	−0.01	0.03	−0.19	0.02	0.11	0.07	0.05

Table 11
Runtime performance comparisons.

	Random forest	MI Model without Random Forest	MI Model with Random Forest	Labeling model	Link penalty model
Number of features	50	50	24	24	24
Number of coefficients	–	255 (initial) 63 (final)	125(initial) 63 (final)	325	294
Runtime	34 min	364 h 48 min	32 h 27 min	44 h 09 min	9 h 17 min

the city center, minimize number of intersections in the city center and minimize stop lights. The detailed definitions of these labels are referred to Bekhor et al. (2006). The same 24 selected features and utility function Eq. (19) is applied in the labeling model, which results in 325 coefficients. Note that, for some observations, different labels may generate identical routes, and result in less unique routes and less alternatives.

Another model generates route choice set with link penalty method and sorted by the penalty factor, the detailed descriptions of the generated choice set are provided in Yao and Bekhor (2021). The link penalty model is also specified with 24 selected features for fair comparison. However, the utility functions are modified as following:

$$U_i = \sum_{x \in V_1} \beta_x \cdot x + \sum_{x \in V_2} \beta_{x,i} \cdot x \tag{22}$$

where

$V = \{x\}$: the set of selected features x,
V_1 and $V_1 \subseteq V$: the set of selected alternative specific features,
$V_2 = \{x | x \in V \text{and} x \notin V_1\}$: the set of selected common features,
β_x: generic coefficient for feature $x \in V_1$
$\beta_{x,i}$: alternative specific coefficient for feature $x \in V_2$ alternative i

For details of the route generation methods, see Yao and Bekhor (2021). The model estimation results of the labeling model and link penalty model are shown in the Appendix A. All discrete choice models have MNL form.

The prediction accuracy results are presented in Table 9.

Among all the models, the random forest model with 50 features outperforms all other models in terms of prediction accuracy. Remind that the random forest model provides feature importance based on impurity reduction, but it does not directly provide coefficient estimates and elasticities as part of the model outputs.

The final MI model provides the best prediction accuracy, 73%, among all four discrete choice models. Although the log-likelihood ratio of the final MI model is slightly higher than the initial MI model, the final model provides slightly better prediction results. This indicates that the initial MI model is overfitted, and the MI method improves the prediction accuracy.

We also provide an initial MI model with 50 features (i.e. without random forest to select important features) to examine the effectiveness of random forest in selecting important features. Results suggest that, initial MI model with 50 features has the same prediction performance as initial MI model with 24 features.

The conventional models exhibit much lower prediction accuracies in comparison to MI. Note that the utility functions are similarly defined, but the number of alternatives of the two conventional models is larger.

4. Discussion

The essence of the proposed data-driven choice set generation method is the clustering. The clustering step identifies routes with similar characteristics based on the normalized attributes. Preliminary experiments suggested that without normalization procedure, the spatial distribution of the OD pairs and the different attribute units can result in clusters based only on trip length. The normalized route characteristic attributes are essential to generate behavioral interpretable clusters.

We are also interested in examining the consistency of our proposed data-driven MI model. For this task, we performed a simulated observation experiment similar to Fosgerau et al. (2013). We apply the same model specification as in the final MI model, and use it to simulate 500 observations for each of 10 randomly selected OD pairs in the dataset.

For each randomly selected OD pair, the selected common features remain the same as the dataset, while alternative specific features for all 6 alternatives are sampled from the fitted truncated normal distribution. The final MI model is applied to simulate the probability of choosing each alternative, and the one with highest probability is chose.

Before showing the estimation results, Fig. 12 shows the route characteristic attribute distributions. With 500 simulated observations for an OD pair, the sampled attributes are well-fitted to the truncated distribution.

Table 10 reports the estimated coefficients for each of the samples. Results show that the standard deviation of the estimated coefficients are relatively small among different sample runs, and the averages of the estimated coefficients are close to the chosen coefficient values. These results suggest that the data-driven MI model can produce consistent and robust estimations with only 6 implicitly generated alternatives.

A possible explanation for the robustness of the proposed model is that, with multiple samplings as in this experiment, the sampled

attributes resemble the cluster distributions and result in consistent alternatives corresponding to the clusters.

Furthermore, we are also interested in the runtime performances of different models (Table 11), and examine the effectiveness of random forest in the MI process. All the runtime performances of different models are obtained using a 6-core PC with 16 GB RAM.

Results show that, random forest has the best runtime performance among all route choice models. Among the discrete choice models, the link penalty model has the shortest runtime. The relative longer runtimes of MI models are due to the iterative process, in which many discrete choice models are estimated to select significant coefficients, and to reach a final MI model.

Results confirms that the random forest is able to remove unimportant features and improve the runtime performance of MI models. Furthermore, we are able to obtain in the end of the MI process (without random forest) an identical final MI model (with random forest). This suggests our MI process is consistent with or without the help of random forest.

5. Conclusion

This paper develops a novel choice set generation and route choice model specification approach by combining machine learning techniques and discrete choice models. The data-driven choice set generation method does not explicitly search for routes in the network, but identifies the main route characteristics using clustering. The choice set is implicitly generated by sampling route characteristic attributes from the clusters. The proposed data-driven choice set generation method tries to capture individuals' interpretation on different routes, in which individuals differentiate routes with the route characteristics rather than the attributes.

The success of the proposed method is highly dependent on the clustering step. Without properly defining the clustering attributes, it is not possible to generate high quality clusters. The normalized route characteristics applied in the paper provide a unified measurement for identifying the clusters. Since these characteristics are normalized, they can be applied using other datasets.

This paper applies random forest to select important features, which serve as a starting point for route choice model development. Methodological-iterative (MI) method is adapted to systematically specify the utility functions for the route choice model. The MI method combines highly correlated coefficients, and eliminates insignificant coefficients to avoid overfitting.

Results show that the random forest classifier performs best in terms of prediction power. This result is consistent with findings of Lai et al. (2018). Conventional discrete choice models provide insights to route choice behavior, but not necessary provide satisfying prediction results. Our proposed data-driven method produces a discrete route choice model with higher prediction accuracy by combining machine learning techniques with discrete choice models. The estimated MI model outperforms two conventional discrete choice models with link penalty and labeling for choice set generation methods. The consistency of the proposed approach is also verified with a simulated observation experiment.

In this paper, we applied the method for the simple MNL choice model. The methodology proposed in this paper can accommodate more complex model forms: for example, Shiftan and Bekhor (2021) implemented the MI method with the cross-nested logit model. In future works, more sophisticated route choice model that considers route overlapping will be incorporated to the proposed method. In terms of utility function specifications, combination of features and non-linear specification will be considered.

CRediT authorship contribution statement

Rui Yao: Methodology, Software, Validation, Formal analysis, Investigation, Writing - original draft. **Shlomo Bekhor:** Conceptualization, Methodology, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Israeli Ministry of Science and Technology (grant number 3-15606).

Appendix A

Fig. 13 shows the GPS trajectory filtering process, from initial of 268,825 individual trips, and processed with several selection criteria and logical validation, to the final 5002 selected trips for map matching. And the estimation results of different models (see Figs. 14–16 and Tables 13–17).

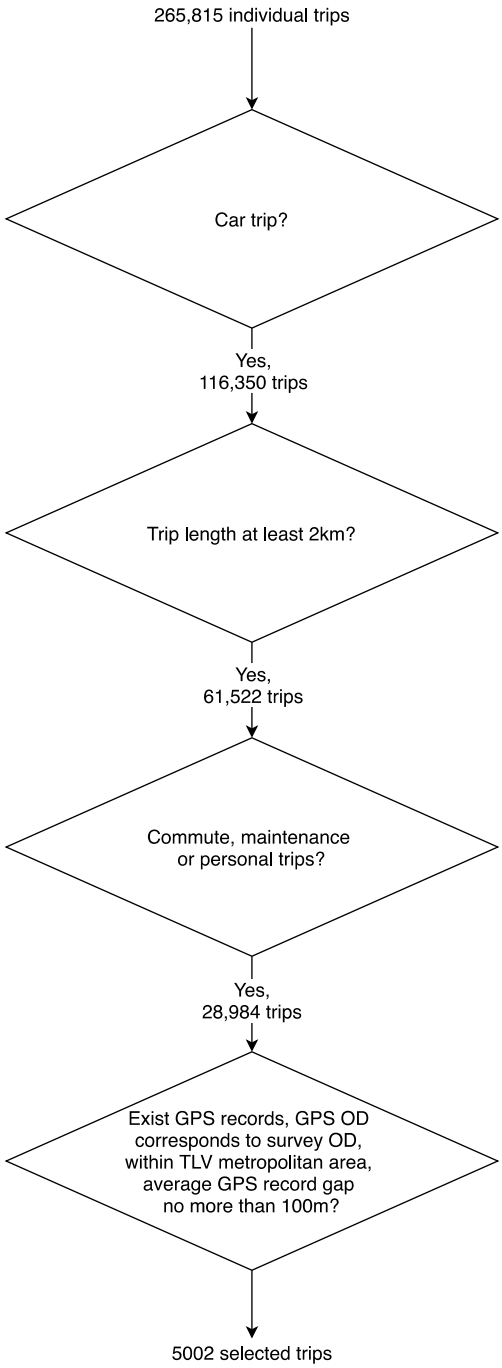

Fig. 13. Trip filtering process.

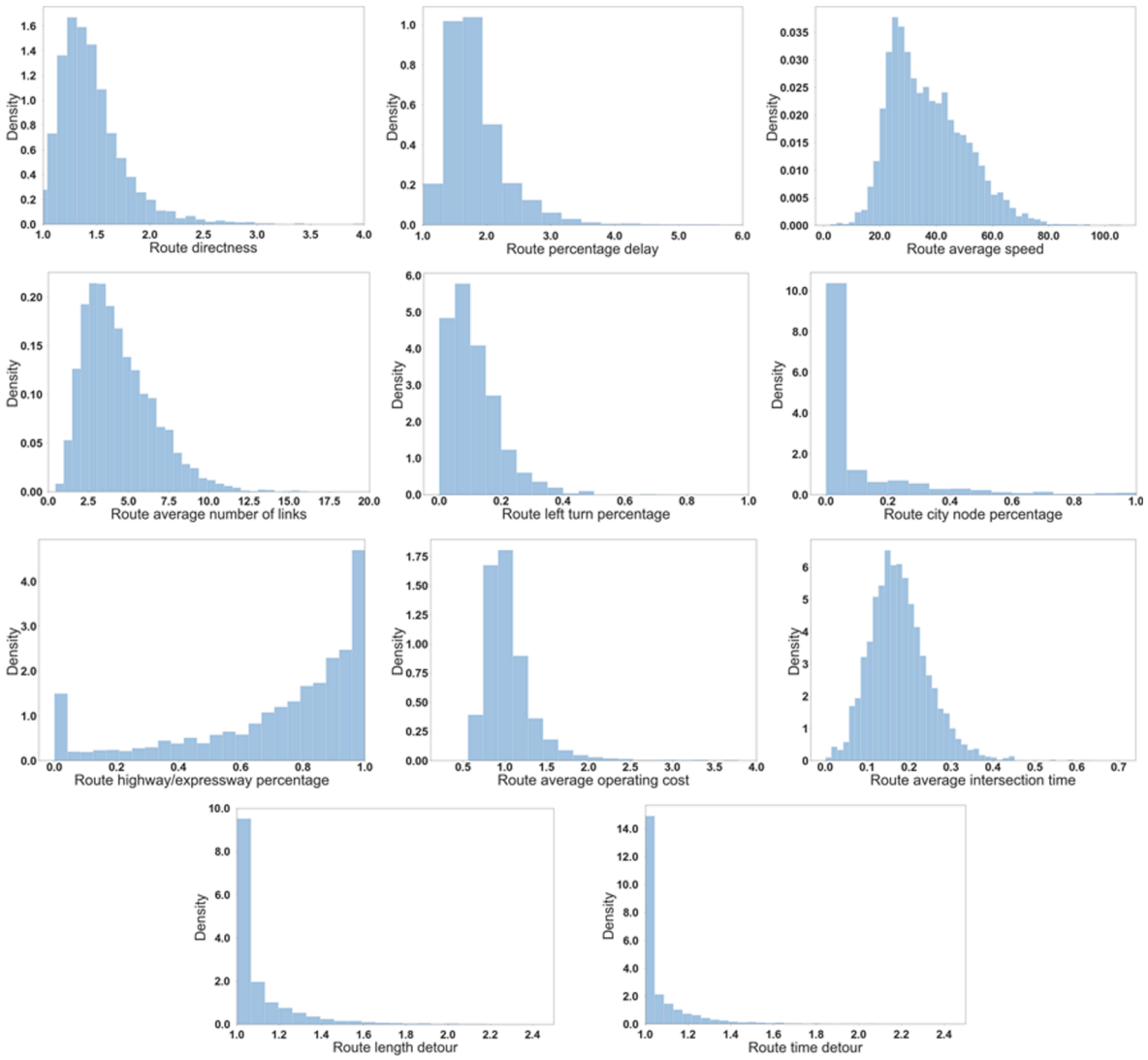

Fig. 14. Distributions of normalized route characteristic attributes.

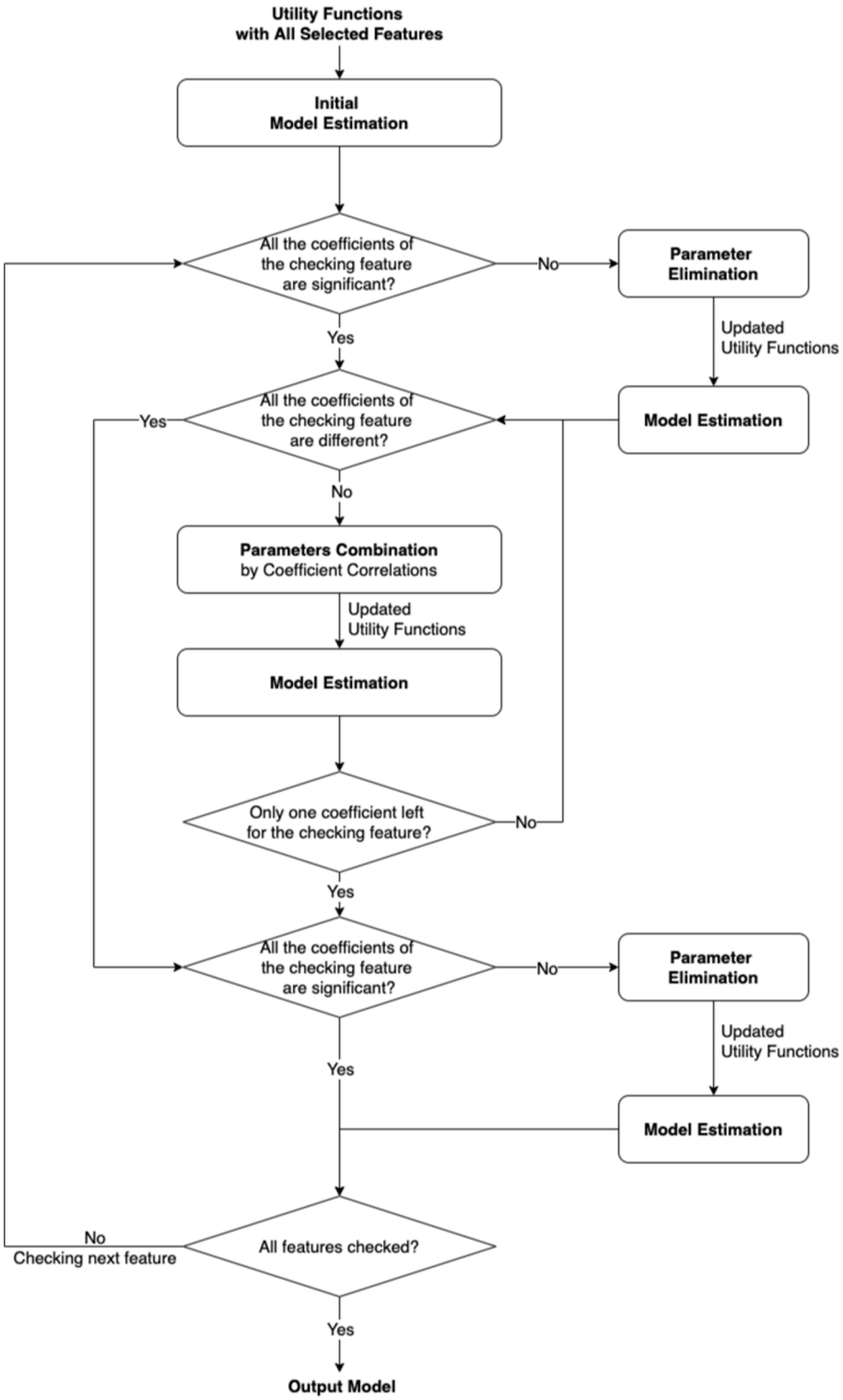

Fig. 15. Eliminate insignificant coefficients before checking each feature.

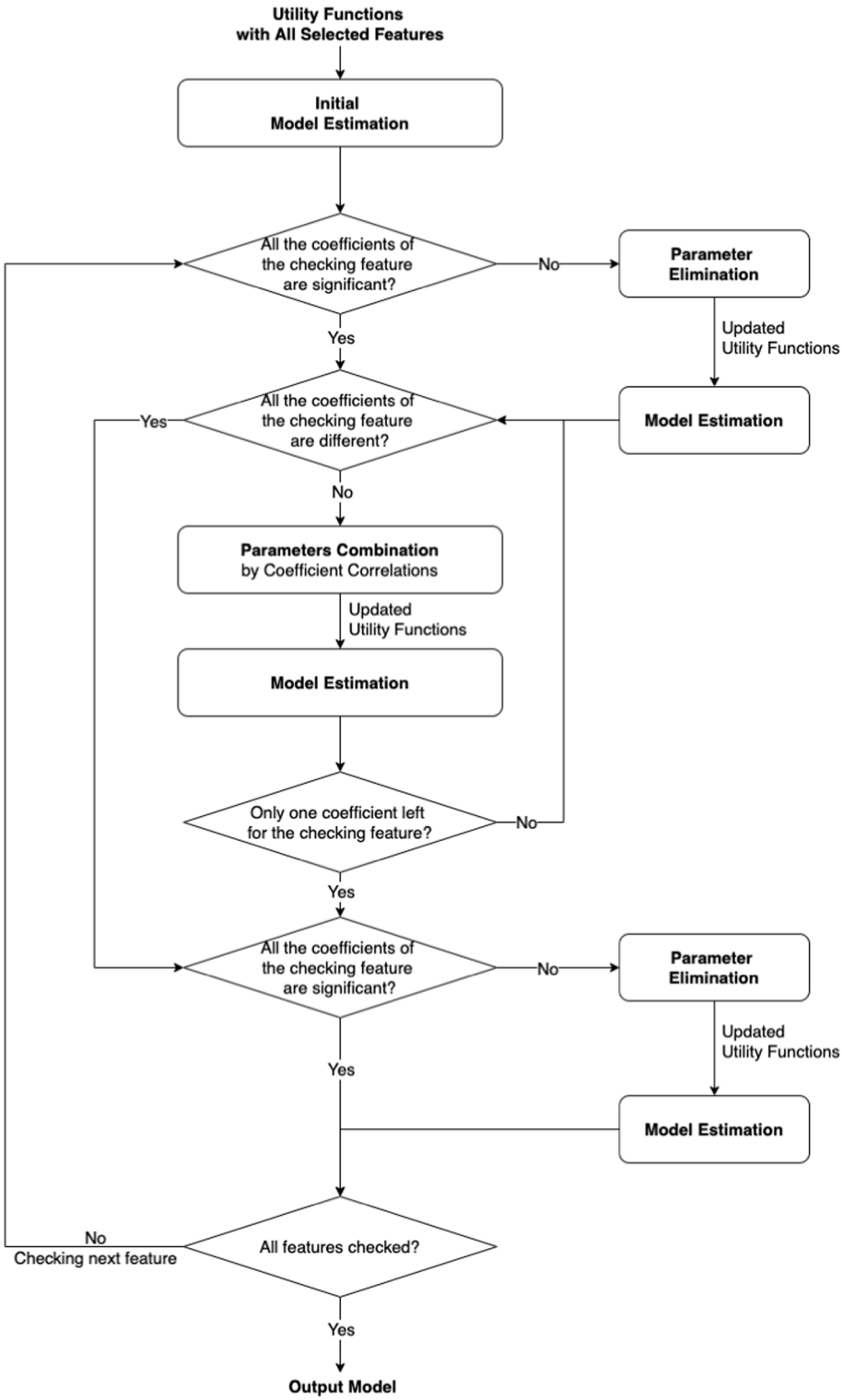

Fig. 16. Eliminate insignificant coefficients before each parameter combination.

Table 12
Initial methodological-iterative route choice model (with alternative 6 as reference).

Coefficient	Value	*t*-test	Coefficient	Value	*t*-test
ASC 1	2.23	12.40	Alt 1 shortest route travel time	0.50	4.53
ASC 2	−1.44	−8.26	Alt 2 shortest route travel time	0.44	3.26
ASC 3	−4.54	−35.40	Alt 3 shortest route travel time	0.50	4.53
ASC 4	9.65	58.40	Alt 4 shortest route travel time	0.40	3.15
ASC 5	−4.92	−36.80	Alt 5 shortest route travel time	0.50	4.53
Alt 1 highway/expressway length	−0.01	−0.11	Alt 1 shortest route directness	0.64	1.43
Alt 2 highway/expressway length	0.06	0.80	Alt 2 shortest route directness	−0.02	−0.08
Alt 3 highway/expressway length	0.50	5.30	Alt 3 shortest route directness	2.73	7.94
Alt 4 highway/expressway length	0.33	5.04	Alt 4 shortest route directness	−1.31	−4.96
Alt 5 highway/expressway length	0.30	7.23	Alt 5 shortest route directness	2.52	6.95
Alt 1 number of left turns	−0.74	−12.00	Alt 1 intersection time in detour zone	16.40	288.00
Alt 2 number of left turns	−0.44	−10.50	Alt 2 intersection time in detour zone	14.60	181.00
Alt 3 number of left turns	−0.57	−9.21	Alt 3 intersection time in detour zone	10.30	202.00
Alt 4 number of left turns	−0.47	−9.00	Alt 4 intersection time in detour zone	15.20	222.00
Alt 5 number of left turns	−0.45	−14.90	Alt 5 intersection time in detour zone	4.95	67.80
Alt 1 route cost	0.21	5.75	Alt 1 number of city nodes in detour zone	0.86	2.64
Alt 2 route cost	0.07	0.82	Alt 2 number of city nodes in detour zone	1.27	4.94
Alt 3 route cost	−0.01	−0.34	Alt 3 number of city nodes in detour zone	2.76	11.00
Alt 4 route cost	0.19	2.62	Alt 4 number of city nodes in detour zone	−0.54	−1.63
Alt 5 route cost	−0.02	−2.25	Alt 5 number of city nodes in detour zone	1.11	3.73
Alt 1 intersection time	0.37	7.56	Alt 1 average delay in detour zone	−0.80	−5.55
Alt 2 intersection time	−0.06	−1.31	Alt 2 average delay in detour zone	−1.10	−5.68
Alt 3 intersection time	−0.03	−0.67	Alt 3 average delay in detour zone	1.68	8.65
Alt 4 intersection time	−0.19	−4.30	Alt 4 average delay in detour zone	−2.97	−14.70
Alt 5 intersection time	−0.06	−2.63	Alt 5 average delay in detour zone	1.47	6.07
Alt 1 OD distance	0.84	6.13	Alt 1 highway/expressway percentage in detour zone	3.33	17.30
Alt 2 OD distance	−0.19	−0.88	Alt 2 highway/expressway percentage in detour zone	1.34	5.13
Alt 3 OD distance	1.03	7.34	Alt 3 highway/expressway percentage in detour zone	6.17	52.90
Alt 4 OD distance	0.04	0.25	Alt 4 highway/expressway percentage in detour zone	5.05	22.50
Alt 5 OD distance	0.53	3.41	Alt 5 highway/expressway percentage in detour zone	1.78	7.75
Alt 1 route travel time	−1.14	−14.20	Alt 1 average speed in detour zone	−0.06	−3.06
Alt 2 route travel time	−0.80	−8.95	Alt 2 average speed in detour zone	0.03	1.77
Alt 3 route travel time	−0.16	−2.75	Alt 3 average speed in detour zone	−0.15	−4.79
Alt 4 route travel time	−0.37	−5.59	Alt 4 average speed in detour zone	−0.30	−11.60
Alt 5 route travel time	−0.13	−5.56	Alt 5 average speed in detour zone	−0.04	−2.12
Alt 1 route distance	−0.20	−1.82	Alt 1 intersection time in fastest zone	−18.40	−402.00
Alt 2 route distance	−0.04	−0.24	Alt 2 intersection time in fastest zone	−10.80	−147.00
Alt 3 route distance	−0.32	−2.74	Alt 3 intersection time in fastest zone	−30.40	−630.00
Alt 4 route distance	0.21	1.49	Alt 4 intersection time in fastest zone	−10.20	−171.00
Alt 5 route distance	−0.19	−3.82	Alt 5 intersection time in fastest zone	−9.19	−119.00
Alt 1 route number of links	0.49	6.92	Alt 1 number of city nodes in fastest zone	−3.45	−9.28
Alt 2 route number of links	0.32	5.98	Alt 2 number of city nodes in fastest zone	−0.69	−1.85
Alt 3 route number of links	0.72	8.53	Alt 3 number of city nodes in fastest zone	−3.14	−10.70
Alt 4 route number of links	0.43	7.39	Alt 4 number of city nodes in fastest zone	−1.00	−2.54
Alt 5 route number of links	0.40	11.70	Alt 5 number of city nodes in fastest zone	−1.05	−3.13
Alt 1 number of city nodes	−0.17	−4.75	Alt 1 average delay in fastest zone	−0.05	−0.17
Alt 2 number of city nodes	−0.39	−7.92	Alt 2 average delay in fastest zone	−0.67	−1.39
Alt 3 number of city nodes	−0.16	−6.89	Alt 3 average delay in fastest zone	1.38	3.42
Alt 4 number of city nodes	−0.05	−4.88	Alt 4 average delay in fastest zone	−4.50	–23.50
Alt 5 number of city nodes	−0.11	−6.68	Alt 5 average delay in fastest zone	−0.05	−0.14
Alt 1 route delay	−0.19	−7.09	Alt 1 highway/expressway percentage in fastest zone	−1.76	−5.28
Alt 2 route delay	0.06	1.65	Alt 2 highway/expressway percentage in fastest zone	−6.47	−15.30
Alt 3 route delay	−0.15	−9.94	Alt 3 highway/expressway percentage in fastest zone	2.63	22.70
Alt 4 route delay	−0.06	−3.02	Alt 4 highway/expressway percentage in fastest zone	0.70	1.49
Alt 5 route delay	−0.06	−2.96	Alt 5 highway/expressway percentage in fastest zone	−4.66	−10.20
Alt 1 fastest route travel time	0.19	1.33	Alt 1 average speed in fastest zone	−0.06	−3.25
Alt 2 fastest route travel time	0.83	5.03	Alt 2 average speed in fastest zone	0.07	3.80
Alt 3 fastest route travel time	0.34	2.66	Alt 3 average speed in fastest zone	−0.11	−4.24
Alt 4 fastest route travel time	0.55	3.85	Alt 4 average speed in fastest zone	−0.11	−4.36
Alt 5 fastest route travel time	0.02	0.21	Alt 5 average speed in fastest zone	0.06	3.11
Alt 1 shortest route distance	0.35	3.31	**Number of observations**	4000	
Alt 2 shortest route distance	−0.53	−3.77	**Number of coefficients**	125	
Alt 3 shortest route distance	−1.13	−10.40	$L(0)$	−7167.04	
Alt 4 shortest route distance	−1.01	−7.01	$L(\beta)$	−2830.37	
Alt 5 shortest route distance	−0.43	−3.79	$\rho(0)$	0.61	

Table 13
Initial MI route choice model without random forest (with alternative 6 as reference).

Coefficient	Value	*t*-test	Coefficient	Value	*t*-test
ASC 1	−0.45	0.00	Alt 1 network density around origin	0.30	0.32
ASC 2	−2.09	−8.84	Alt 2 network density around origin	0.24	0.26
ASC 3	−3.23	−22.50	Alt 3 network density around origin	0.40	0.33
ASC 4	2.92	19.30	Alt 4 network density around origin	0.14	0.14
ASC 5	−5.20	−41.70	Alt 5 network density around origin	0.27	0.28
Alt 1 highway/expressway length	−0.10	−1.15	Alt 1 household area	−0.03	−1.51
Alt 2 highway/expressway length	0.08	1.16	Alt 2 household area	0.03	1.65
Alt 3 highway/expressway length	0.41	4.58	Alt 3 household area	−0.04	−1.45
Alt 4 highway/expressway length	0.27	4.05	Alt 4 household area	0.00	−0.08
Alt 5 highway/expressway length	0.28	6.63	Alt 5 household area	−0.01	−0.61
Alt 1 number of left turns	−0.62	−9.80	Alt 1 level of education	−0.10	−1.76
Alt 2 number of left turns	−0.45	−10.10	Alt 2 level of education	−0.04	−0.77
Alt 3 number of left turns	−0.62	−9.76	Alt 3 level of education	−0.07	−0.82
Alt 4 number of left turns	−0.45	−8.44	Alt 4 level of education	0.02	0.29
Alt 5 number of left turns	−0.49	−15.20	Alt 5 level of education	−0.04	−0.69
Alt 1 route cost	0.14	3.67	Alt 1 network density around destination	0.39	4.22
Alt 2 route cost	0.13	1.56	Alt 2 network density around destination	0.10	1.06
Alt 3 route cost	−0.13	−2.94	Alt 3 network density around destination	0.38	3.08
Alt 4 route cost	−0.06	−0.82	Alt 4 network density around destination	0.10	0.95
Alt 5 route cost	−0.05	−4.66	Alt 5 network density around destination	0.14	1.41
Alt 1 intersection time	0.23	4.82	Alt 1 household size	0.07	1.04
Alt 2 intersection time	−0.07	−1.57	Alt 2 household size	0.03	0.50
Alt 3 intersection time	−0.01	−0.15	Alt 3 household size	0.07	0.72
Alt 4 intersection time	−0.26	−6.15	Alt 4 household size	0.09	1.21
Alt 5 intersection time	−0.08	−3.50	Alt 5 household size	0.11	1.52
Alt 1 OD distance	0.72	5.02	Alt 1 morning peak trip dummy	−0.28	−0.21
Alt 2 OD distance	−0.44	−2.25	Alt 2 morning peak trip dummy	−0.95	−0.70
Alt 3 OD distance	0.86	5.89	Alt 3 morning peak trip dummy	−0.91	−0.51
Alt 4 OD distance	0.07	0.38	Alt 4 morning peak trip dummy	0.61	0.39
Alt 5 OD distance	0.51	3.02	Alt 5 morning peak trip dummy	−1.67	−1.23
Alt 1 route travel time	−1.16	−11.70	Alt 1 maintenance trip dummy	−0.25	−0.24
Alt 2 route travel time	−0.77	−8.38	Alt 2 maintenance trip dummy	−0.77	−0.70
Alt 3 route travel time	−0.10	−1.45	Alt 3 maintenance trip dummy	−0.95	−0.56
Alt 4 route travel time	−0.35	−5.05	Alt 4 maintenance trip dummy	1.00	0.77
Alt 5 route travel time	−0.16	−6.41	Alt 5 maintenance trip dummy	−1.52	−1.24
Alt 1 route distance	0.12	1.08	Alt 1 student dummy	−0.49	−1.53
Alt 2 route distance	−0.08	−0.57	Alt 2 student dummy	0.13	0.46
Alt 3 route distance	−0.10	−0.85	Alt 3 student dummy	−0.40	−0.86
Alt 4 route distance	0.57	3.79	Alt 4 student dummy	−0.16	−0.48
Alt 5 route distance	−0.14	−2.42	Alt 5 student dummy	0.11	0.37
Alt 1 route number of links	0.57	7.20	Alt 1 number of on-board passengers	−0.19	−1.25
Alt 2 route number of links	0.34	6.62	Alt 2 number of on-board passengers	−0.27	−1.90
Alt 3 route number of links	0.66	8.38	Alt 3 number of on-board passengers	0.23	0.90
Alt 4 route number of links	0.54	9.19	Alt 4 number of on-board passengers	−0.02	−0.10
Alt 5 route number of links	0.49	13.10	Alt 5 number of on-board passengers	−0.08	−0.48
Alt 1 number of city nodes	−0.21	−5.98	Alt 1 number of vehicles in the household	0.22	1.76
Alt 2 number of city nodes	−0.42	−8.71	Alt 2 number of vehicles in the household	0.27	2.29
Alt 3 number of city nodes	−0.15	−6.11	Alt 3 number of vehicles in the household	0.16	0.92
Alt 4 number of city nodes	−0.06	−5.17	Alt 4 number of vehicles in the household	0.27	2.02
Alt 5 number of city nodes	−0.10	−5.86	Alt 5 number of vehicles in the household	0.24	1.97
Alt 1 route delay	−0.20	−7.30	Alt 1 evening peak trip dummy	−0.40	−0.35
Alt 2 route delay	0.05	1.27	Alt 2 evening peak trip dummy	−0.67	−0.57
Alt 3 route delay	−0.14	−8.53	Alt 3 evening peak trip dummy	−0.82	−0.51
Alt 4 route delay	−0.06	−3.03	Alt 4 evening peak trip dummy	1.10	0.83
Alt 5 route delay	−0.07	−3.49	Alt 5 evening peak trip dummy	−1.71	−1.45
Alt 1 fastest route travel time	0.31	2.33	Alt 1 gender	0.03	0.16
Alt 2 fastest route travel time	1.02	6.58	Alt 2 gender	0.02	0.16
Alt 3 fastest route travel time	0.41	3.34	Alt 3 gender	0.00	0.00
Alt 4 fastest route travel time	0.67	4.73	Alt 4 gender	0.08	0.45
Alt 5 fastest route travel time	0.23	2.17	Alt 5 gender	0.17	1.06
Alt 1 shortest route distance	0.38	3.27	Alt 1 personal trip dummy	−0.06	−0.01
Alt 2 shortest route distance	−0.40	−2.84	Alt 2 personal trip dummy	−0.60	−0.06
Alt 3 shortest route distance	−1.00	−8.62	Alt 3 personal trip dummy	−1.22	−0.12
Alt 4 shortest route distance	−1.07	−7.62	Alt 4 personal trip dummy	1.04	0.10
Alt 5 shortest route distance	−0.41	−3.35	Alt 5 personal trip dummy	−1.98	−0.20
Alt 1 shortest route travel time	0.36	3.65	Alt 1 origin is in CBD dummy	0.30	0.32
Alt 2 shortest route travel time	0.24	2.02	Alt 2 origin is in CBD dummy	0.24	0.26
Alt 3 shortest route travel time	0.36	3.65	Alt 3 origin is in CBD dummy	0.40	0.33

(*continued on next page*)

Table 13 (*continued*)

Coefficient	Value	*t*-test	Coefficient	Value	*t*-test
Alt 4 shortest route travel time	0.29	2.45	Alt 4 origin is in CBD dummy	0.14	0.14
Alt 5 shortest route travel time	0.36	3.65	Alt 5 origin is in CBD dummy	0.27	0.28
Alt 1 shortest route directness	1.16	2.56	Alt 1 toll road dummy	0.01	0.04
Alt 2 shortest route directness	−0.25	−0.89	Alt 2 toll road dummy	0.16	0.24
Alt 3 shortest route directness	2.18	7.34	Alt 3 toll road dummy	−0.18	−0.27
Alt 4 shortest route directness	−0.83	−3.35	Alt 4 toll road dummy	0.68	0.87
Alt 5 shortest route directness	2.82	7.23	Alt 5 toll road dummy	−0.01	−0.04
Alt 1 intersection time in detour zone	12.50	201.00	Alt 1 off peak trip dummy	−0.04	−0.03
Alt 2 intersection time in detour zone	13.50	183.00	Alt 2 off peak trip dummy	−0.52	−0.45
Alt 3 intersection time in detour zone	4.02	81.60	Alt 3 off peak trip dummy	−1.51	−0.78
Alt 4 intersection time in detour zone	11.70	169.00	Alt 4 off peak trip dummy	1.27	0.94
Alt 5 intersection time in detour zone	8.25	120.00	Alt 5 off peak trip dummy	−1.71	−1.37
Alt 1 number of city nodes in detour zone	0.94	2.56	Alt 1 full time employer	−0.26	−0.13
Alt 2 number of city nodes in detour zone	1.58	5.30	Alt 2 full time employer	−0.55	−0.32
Alt 3 number of city nodes in detour zone	2.95	12.50	Alt 3 full time employer	0.25	0.11
Alt 4 number of city nodes in detour zone	−1.10	−3.32	Alt 4 full time employer	0.50	0.25
Alt 5 number of city nodes in detour zone	0.48	1.43	Alt 5 full time employer	0.60	0.35
Alt 1 average delay in detour zone	1.27	7.62	Alt 1 retired	−0.09	−0.31
Alt 2 average delay in detour zone	1.08	5.20	Alt 2 retired	−0.26	−1.01
Alt 3 average delay in detour zone	2.62	15.10	Alt 3 retired	−0.54	−1.25
Alt 4 average delay in detour zone	−1.77	−8.03	Alt 4 retired	0.10	0.32
Alt 5 average delay in detour zone	0.73	3.38	Alt 5 retired	0.60	2.20
Alt 1 highway/expressway percentage in detour zone	1.41	6.29	Alt 1 housewife	0.04	0.07
Alt 2 highway/expressway percentage in detour zone	−0.15	−0.62	Alt 2 housewife	−0.47	−1.07
Alt 3 highway/expressway percentage in detour zone	4.98	55.60	Alt 3 housewife	−0.51	−1.41
Alt 4 highway/expressway percentage in detour zone	3.50	15.90	Alt 4 housewife	0.50	1.40
Alt 5 highway/expressway percentage in detour zone	0.64	2.85	Alt 5 housewife	−0.77	−2.21
Alt 1 average speed in detour zone	0.00	−0.03	Alt 1 commute trip dummy	−0.27	−0.20
Alt 2 average speed in detour zone	0.07	3.62	Alt 2 commute trip dummy	−0.81	−0.62
Alt 3 average speed in detour zone	−0.16	−4.75	Alt 3 commute trip dummy	−1.09	−0.60
Alt 4 average speed in detour zone	−0.28	−9.95	Alt 4 commute trip dummy	0.89	0.60
Alt 5 average speed in detour zone	−0.03	−1.53	Alt 5 commute trip dummy	−1.68	−1.25
Alt 1 intersection time in fastest zone	−18.40	−313.00	Alt 1 solider	−0.09	−0.19
Alt 2 intersection time in fastest zone	−11.80	−187.00	Alt 2 solider	−0.16	−0.38
Alt 3 intersection time in fastest zone	−25.10	−568.00	Alt 3 solider	−0.23	−2.83
Alt 4 intersection time in fastest zone	−10.60	−178.00	Alt 4 solider	0.93	2.22
Alt 5 intersection time in fastest zone	−8.36	−112.00	Alt 5 solider	1.44	3.68
Alt 1 number of city nodes in fastest zone	−3.04	−7.30	Alt 1 part time employer	−0.15	−0.54
Alt 2 number of city nodes in fastest zone	−0.50	−1.32	Alt 2 part time employer	−0.53	−2.26
Alt 3 number of city nodes in fastest zone	−3.77	−14.00	Alt 3 part time employer	0.36	1.11
Alt 4 number of city nodes in fastest zone	−0.32	−0.91	Alt 4 part time employer	0.35	1.29
Alt 5 number of city nodes in fastest zone	−1.03	−2.76	Alt 5 part time employer	0.63	2.65
Alt 1 average delay in fastest zone	0.28	0.62	Alt 1 number of children under 8 in the household	0.00	0.03
Alt 2 average delay in fastest zone	−0.31	−0.66	Alt 2 number of children under 8 in the household	0.05	0.46
Alt 3 average delay in fastest zone	0.85	2.88	Alt 3 number of children under 8 in the household	0.01	0.06
Alt 4 average delay in fastest zone	−2.95	−15.10	Alt 4 number of children under 8 in the household	−0.10	−0.73
Alt 5 average delay in fastest zone	−0.09	−0.23	Alt 5 number of children under 8 in the household	0.00	0.03
Alt 1 highway/expressway percentage in fastest zone	−1.87	−4.77	Alt 1 unemployed	−0.18	−0.55
Alt 2 highway/expressway percentage in fastest zone	−6.79	−15.70	Alt 2 unemployed	−0.15	−0.54
Alt 3 highway/expressway percentage in fastest zone	0.93	8.08	Alt 3 unemployed	0.87	2.01
Alt 4 highway/expressway percentage in fastest zone	0.37	0.90	Alt 4 unemployed	0.54	1.61
Alt 5 highway/expressway percentage in fastest zone	−4.91	−10.60	Alt 5 unemployed	0.25	0.84
Alt 1 average speed in fastest zone	−0.04	−2.09	Alt 1 destination is in CBD dummy	0.25	0.38
Alt 2 average speed in fastest zone	0.09	4.98	Alt 2 destination is in CBD dummy	0.21	0.39
Alt 3 average speed in fastest zone	−0.05	−1.73	Alt 3 destination is in CBD dummy	−0.23	0.11
Alt 4 average speed in fastest zone	−0.07	−2.54	Alt 4 destination is in CBD dummy	0.48	0.52
Alt 5 average speed in fastest zone	0.09	4.47	Alt 5 destination is in CBD dummy	0.18	0.36
Alt 1 age	−0.04	−0.49	**Number of observations**	4000	
Alt 2 age	−0.02	−0.31	**Number of coefficients**	255	
Alt 3 age	0.09	0.81	$L(0)$	−7167.04	
Alt 4 age	−0.09	−1.05	$L(\beta)$	−2667.39	
Alt 5 age	0.12	1.52	$\rho(0)$	0.63	

Table 14
Conventional link penalty route choice model.

Coefficient		Value	*t*-test	Coefficient		Value	*t*-test
Route travel time		−0.42	−15.10	Alt 6	average delay in detour zone	2.52	9.92
Route distance		−0.86	−17.20	Alt 7	average delay in detour zone	1.41	7.95
Route number of links		0.09	3.24	Alt 8	average delay in detour zone	1.15	5.99
Route number of city nodes		−0.06	−6.37	Alt 9	average delay in detour zone	0.11	0.76
Route delay		−0.35	−7.36	Alt 10	average delay in detour zone	1.38	9.43
Route intersection time		0.24	9.68	Alt 11	average delay in detour zone	3.25	16.50
Route highway/expressway length		0.44	16.90	Alt 12	average delay in detour zone	−0.48	−2.44
Route number of left turns		−0.28	−20.40	Alt 13	average delay in detour zone	−0.59	−2.87
Route cost		−0.10	−9.50	Alt 14	average delay in detour zone	−2.36	−48.50
Alt 1	fastest route travel time	0.12	2.38	Alt 15	average delay in detour zone	1.21	27.90
Alt 2	fastest route travel time	0.14	7.84	Alt 16	average delay in detour zone	−1.17	−18.60
Alt 3	fastest route travel time	0.13	2.61	Alt 17	average delay in detour zone	−2.45	−34.50
Alt 4	fastest route travel time	0.10	1.96	Alt 18	average delay in detour zone	5.14	472.00
Alt 5	fastest route travel time	0.20	2.85	Alt 19	average delay in detour zone	−1.26	−69.20
Alt 6	fastest route travel time	0.10	1.84	Alt 1	highway/expressway percentage in detour zone	0.23	0.86
Alt 7	fastest route travel time	0.29	3.52	Alt 2	highway/expressway percentage in detour zone	−0.05	−0.22
Alt 8	fastest route travel time	0.13	2.22	Alt 3	highway/expressway percentage in detour zone	1.09	7.07
Alt 9	fastest route travel time	0.26	2.28	Alt 4	highway/expressway percentage in detour zone	−0.65	−4.16
Alt 10	fastest route travel time	0.15	3.05	Alt 5	highway/expressway percentage in detour zone	1.91	8.51
Alt 11	fastest route travel time	0.26	1.92	Alt 6	highway/expressway percentage in detour zone	−2.08	−10.70
Alt 12	fastest route travel time	0.14	2.48	Alt 7	highway/expressway percentage in detour zone	0.89	8.01
Alt 13	fastest route travel time	0.19	1.45	Alt 8	highway/expressway percentage in detour zone	0.86	7.70
Alt 14	fastest route travel time	0.43	2.10	Alt 9	highway/expressway percentage in detour zone	−1.93	−18.20
Alt 15	fastest route travel time	0.19	1.77	Alt 10	highway/expressway percentage in detour zone	−0.89	−8.25
Alt 16	fastest route travel time	0.08	0.59	Alt 11	highway/expressway percentage in detour zone	−1.42	−11.20
Alt 17	fastest route travel time	0.04	0.28	Alt 12	highway/expressway percentage in detour zone	3.10	23.10
Alt 18	fastest route travel time	−0.16	−1.18	Alt 13	highway/expressway percentage in detour zone	−1.10	−7.99
Alt 19	fastest route travel time	0.31	1.15	Alt 14	highway/expressway percentage in detour zone	2.94	94.20
Alt 1	OD distance	−0.58	−4.10	Alt 15	highway/expressway percentage in detour zone	−1.57	−42.10
Alt 2	OD distance	−0.61	−4.18	Alt 16	highway/expressway percentage in detour zone	−1.10	−39.00
Alt 3	OD distance	−0.29	−1.94	Alt 17	highway/expressway percentage in detour zone	4.29	148.00
Alt 4	OD distance	−0.44	−2.90	Alt 18	highway/expressway percentage in detour zone	−0.91	−391.00
Alt 5	OD distance	−0.34	−2.20	Alt 19	highway/expressway percentage in detour zone	−1.80	–223.00
Alt 6	OD distance	−0.43	−2.87	Alt 1	average speed in detour zone	−0.05	−1.30
Alt 7	OD distance	−0.19	−1.21	Alt 2	average speed in detour zone	−0.04	−1.13
Alt 8	OD distance	−0.27	−1.75	Alt 3	average speed in detour zone	−0.07	−2.00
Alt 9	OD distance	−0.44	−3.17	Alt 4	average speed in detour zone	−0.05	−1.31
Alt 10	OD distance	−0.09	−0.67	Alt 5	average speed in detour zone	−0.12	−3.23
Alt 11	OD distance	−0.09	−0.62	Alt 6	average speed in detour zone	−0.03	−0.94
Alt 12	OD distance	−0.21	−1.49	Alt 7	average speed in detour zone	−0.05	−1.53
Alt 13	OD distance	−0.42	−3.06	Alt 8	average speed in detour zone	−0.06	−1.62
Alt 14	OD distance	−0.52	−3.37	Alt 9	average speed in detour zone	−0.05	−1.27
Alt 15	OD distance	0.10	0.65	Alt 10	average speed in detour zone	−0.07	−1.84
Alt 16	OD distance	−0.21	−1.28	Alt 11	average speed in detour zone	−0.05	−1.24
Alt 17	OD distance	−0.12	−0.73	Alt 12	average speed in detour zone	−0.05	−1.38
Alt 18	OD distance	−0.29	−1.44	Alt 13	average speed in detour zone	−0.04	−1.03
Alt 19	OD distance	0.44	1.72	Alt 14	average speed in detour zone	−0.17	−3.97
Alt 1	shortest route distance	0.49	4.37	Alt 15	average speed in detour zone	0.02	0.36
Alt 2	shortest route distance	0.47	4.16	Alt 16	average speed in detour zone	−0.02	−0.35
Alt 3	shortest route distance	0.26	2.19	Alt 17	average speed in detour zone	−0.14	−2.79
Alt 4	shortest route distance	0.39	3.31	Alt 18	average speed in detour zone	0.07	1.27
Alt 5	shortest route distance	0.26	2.17	Alt 19	average speed in detour zone	−0.15	−2.35
Alt 6	shortest route distance	0.40	3.32	Alt 1	intersection time in fastest zone	0.43	5.75
Alt 7	shortest route distance	0.20	1.61	Alt 2	intersection time in fastest zone	0.61	8.02
Alt 8	shortest route distance	0.23	1.93	Alt 3	intersection time in fastest zone	0.14	2.16
Alt 9	shortest route distance	0.38	3.52	Alt 4	intersection time in fastest zone	1.92	32.10
Alt 10	shortest route distance	0.09	0.80	Alt 5	intersection time in fastest zone	1.88	21.60
Alt 11	shortest route distance	0.11	0.93	Alt 6	intersection time in fastest zone	−1.77	−23.10
Alt 12	shortest route distance	0.22	1.98	Alt 7	intersection time in fastest zone	−0.71	−8.53
Alt 13	shortest route distance	0.32	2.93	Alt 8	intersection time in fastest zone	−1.41	−18.50
Alt 14	shortest route distance	0.33	2.74	Alt 9	intersection time in fastest zone	2.83	53.90
Alt 15	shortest route distance	−0.10	−0.82	Alt 10	intersection time in fastest zone	1.72	38.00
Alt 16	shortest route distance	0.16	1.21	Alt 11	intersection time in fastest zone	−1.36	−28.90
Alt 17	shortest route distance	0.02	0.15	Alt 12	intersection time in fastest zone	−2.47	−40.00
Alt 18	shortest route distance	0.35	2.24	Alt 13	intersection time in fastest zone	1.10	16.80
Alt 19	shortest route distance	−0.44	−2.13	Alt 14	intersection time in fastest zone	4.80	167.00
Alt 1	shortest route travel time	−0.28	−5.17	Alt 15	intersection time in fastest zone	−2.22	−74.50
Alt 2	shortest route travel time	−0.28	−6.36	Alt 16	intersection time in fastest zone	−0.23	−7.05

(*continued on next page*)

Table 14 (*continued*)

Coefficient		Value	*t*-test	Coefficient		Value	*t*-test
Alt 3	shortest route travel time	−0.47	−3.58	Alt 17	intersection time in fastest zone	−1.90	−46.10
Alt 4	shortest route travel time	−0.29	−5.17	Alt 18	intersection time in fastest zone	−2.71	−466.00
Alt 5	shortest route travel time	−0.30	−2.37	Alt 19	intersection time in fastest zone	1.71	174.00
Alt 6	shortest route travel time	−0.54	−2.75	Alt 1	number of city nodes in fastest zone	1.05	3.25
Alt 7	shortest route travel time	−0.28	−2.96	Alt 2	number of city nodes in fastest zone	0.45	1.45
Alt 8	shortest route travel time	−0.18	−1.47	Alt 3	number of city nodes in fastest zone	1.45	6.77
Alt 9	shortest route travel time	−0.09	−0.77	Alt 4	number of city nodes in fastest zone	1.09	4.42
Alt 10	shortest route travel time	−0.05	−0.40	Alt 5	number of city nodes in fastest zone	1.13	3.64
Alt 11	shortest route travel time	−0.40	−1.56	Alt 6	number of city nodes in fastest zone	1.49	5.64
Alt 12	shortest route travel time	−0.29	−18.90	Alt 7	number of city nodes in fastest zone	0.55	2.49
Alt 13	shortest route travel time	−0.28	−6.36	Alt 8	number of city nodes in fastest zone	1.09	4.43
Alt 14	shortest route travel time	−0.28	−6.36	Alt 9	number of city nodes in fastest zone	1.61	6.72
Alt 15	shortest route travel time	−0.35	−5.18	Alt 10	number of city nodes in fastest zone	1.07	4.65
Alt 16	shortest route travel time	−0.28	−6.36	Alt 11	number of city nodes in fastest zone	0.20	0.85
Alt 17	shortest route travel time	−0.47	−5.79	Alt 12	number of city nodes in fastest zone	0.26	1.01
Alt 18	shortest route travel time	−0.28	−4.53	Alt 13	number of city nodes in fastest zone	0.94	3.47
Alt 19	shortest route travel time	−0.42	−3.83	Alt 14	number of city nodes in fastest zone	1.32	6.28
Alt 1	shortest route directness	−0.95	−2.72	Alt 15	number of city nodes in fastest zone	−0.43	−1.94
Alt 2	shortest route directness	−0.87	−2.48	Alt 16	number of city nodes in fastest zone	0.28	1.12
Alt 3	shortest route directness	−0.52	−1.56	Alt 17	number of city nodes in fastest zone	3.69	13.10
Alt 4	shortest route directness	−0.54	−1.48	Alt 18	number of city nodes in fastest zone	1.22	37.90
Alt 5	shortest route directness	−0.37	−1.07	Alt 19	number of city nodes in fastest zone	0.62	14.30
Alt 6	shortest route directness	−0.59	−2.10	Alt 1	average delay in fastest zone	0.35	1.03
Alt 7	shortest route directness	−0.14	−0.39	Alt 2	average delay in fastest zone	0.62	2.16
Alt 8	shortest route directness	−0.36	−1.05	Alt 3	average delay in fastest zone	0.55	1.63
Alt 9	shortest route directness	−0.87	−2.80	Alt 4	average delay in fastest zone	1.47	3.80
Alt 10	shortest route directness	0.01	0.03	Alt 5	average delay in fastest zone	−0.93	−3.78
Alt 11	shortest route directness	0.29	0.89	Alt 6	average delay in fastest zone	1.44	4.62
Alt 12	shortest route directness	−0.16	−0.52	Alt 7	average delay in fastest zone	0.57	2.09
Alt 13	shortest route directness	−0.81	−2.53	Alt 8	average delay in fastest zone	1.22	3.70
Alt 14	shortest route directness	−0.87	−4.25	Alt 9	average delay in fastest zone	0.95	4.74
Alt 15	shortest route directness	−0.13	−0.68	Alt 10	average delay in fastest zone	−0.07	−0.45
Alt 16	shortest route directness	−1.42	−6.35	Alt 11	average delay in fastest zone	−0.06	−0.34
Alt 17	shortest route directness	−1.39	−5.42	Alt 12	average delay in fastest zone	0.94	4.00
Alt 18	shortest route directness	−1.73	−39.20	Alt 13	average delay in fastest zone	1.14	4.50
Alt 19	shortest route directness	1.02	19.60	Alt 14	average delay in fastest zone	1.35	36.10
Alt 1	intersection time in detour zone	5.79	68.90	Alt 15	average delay in fastest zone	−0.34	−8.06
Alt 2	intersection time in detour zone	1.26	14.50	Alt 16	average delay in fastest zone	1.24	45.90
Alt 3	intersection time in detour zone	−3.03	−60.60	Alt 17	average delay in fastest zone	−1.62	−46.10
Alt 4	intersection time in detour zone	−1.62	−29.20	Alt 18	average delay in fastest zone	−3.35	−278.00
Alt 5	intersection time in detour zone	−0.97	−10.70	Alt 19	average delay in fastest zone	0.55	44.50
Alt 6	intersection time in detour zone	0.85	11.20	Alt 1	highway/expressway percentage in fastest zone	−1.07	−2.69
Alt 7	intersection time in detour zone	−1.36	−21.60	Alt 2	highway/expressway percentage in fastest zone	0.00	0.00
Alt 8	intersection time in detour zone	−0.70	−11.00	Alt 3	highway/expressway percentage in fastest zone	−0.29	−0.73
Alt 9	intersection time in detour zone	1.36	27.10	Alt 4	highway/expressway percentage in fastest zone	−1.37	−3.73
Alt 10	intersection time in detour zone	0.28	6.45	Alt 5	highway/expressway percentage in fastest zone	0.39	1.07
Alt 11	intersection time in detour zone	0.77	14.90	Alt 6	highway/expressway percentage in fastest zone	−1.00	−2.78
Alt 12	intersection time in detour zone	−2.59	−42.10	Alt 7	highway/expressway percentage in fastest zone	−1.13	−3.88
Alt 13	intersection time in detour zone	2.08	29.40	Alt 8	highway/expressway percentage in fastest zone	−1.85	−6.81
Alt 14	intersection time in detour zone	−0.60	−19.80	Alt 9	highway/expressway percentage in fastest zone	−1.05	−6.66
Alt 15	intersection time in detour zone	1.59	53.00	Alt 10	highway/expressway percentage in fastest zone	−1.59	−9.88
Alt 16	intersection time in detour zone	−1.39	−36.30	Alt 11	highway/expressway percentage in fastest zone	−0.90	−5.54
Alt 17	intersection time in detour zone	−1.53	−34.00	Alt 12	highway/expressway percentage in fastest zone	−0.13	−0.66
Alt 18	intersection time in detour zone	0.32	54.60	Alt 13	highway/expressway percentage in fastest zone	−2.87	−17.00
Alt 19	intersection time in detour zone	−0.59	−57.70	Alt 14	highway/expressway percentage in fastest zone	−1.43	−48.10
Alt 1	number of city nodes in detour zone	−0.73	−2.25	Alt 15	highway/expressway percentage in fastest zone	−1.77	−41.90
Alt 2	number of city nodes in detour zone	0.47	1.50	Alt 16	highway/expressway percentage in fastest zone	−1.45	−40.60
Alt 3	number of city nodes in detour zone	0.71	2.41	Alt 17	highway/expressway percentage in fastest zone	−1.25	−67.20
Alt 4	number of city nodes in detour zone	−0.34	−1.13	Alt 18	highway/expressway percentage in fastest zone	−0.30	−572.00
Alt 5	number of city nodes in detour zone	0.48	1.48	Alt 19	highway/expressway percentage in fastest zone	−2.43	−239.00
Alt 6	number of city nodes in detour zone	−1.01	−3.06	Alt 1	average speed in fastest zone	0.04	1.26
Alt 7	number of city nodes in detour zone	0.09	0.38	Alt 2	average speed in fastest zone	0.01	0.43
Alt 8	number of city nodes in detour zone	0.25	1.30	Alt 3	average speed in fastest zone	0.02	0.59
Alt 9	number of city nodes in detour zone	−1.35	−7.85	Alt 4	average speed in fastest zone	0.04	1.29
Alt 10	number of city nodes in detour zone	−1.19	−7.12	Alt 5	average speed in fastest zone	0.04	1.01
Alt 11	number of city nodes in detour zone	−0.03	−0.16	Alt 6	average speed in fastest zone	0.04	1.23
Alt 12	number of city nodes in detour zone	2.22	11.60	Alt 7	average speed in fastest zone	0.03	0.89
Alt 13	number of city nodes in detour zone	−0.37	−1.89	Alt 8	average speed in fastest zone	0.03	0.89
Alt 14	number of city nodes in detour zone	0.19	1.09	Alt 9	average speed in fastest zone	0.06	1.89

(*continued on next page*)

Table 14 (*continued*)

Coefficient		Value	*t*-test	Coefficient		Value	*t*-test
Alt 15	number of city nodes in detour zone	0.79	4.21	Alt 10	average speed in fastest zone	0.06	1.60
Alt 16	number of city nodes in detour zone	2.02	9.84	Alt 11	average speed in fastest zone	0.04	0.99
Alt 17	number of city nodes in detour zone	−0.43	−1.84	Alt 12	average speed in fastest zone	−0.04	−1.02
Alt 18	number of city nodes in detour zone	2.14	81.10	Alt 13	average speed in fastest zone	0.07	1.83
Alt 19	number of city nodes in detour zone	−0.85	−21.10	Alt 14	average speed in fastest zone	0.12	3.14
Alt 1	average delay in detour zone	−1.01	−4.23	Alt 15	average speed in fastest zone	−0.01	−0.26
Alt 2	average delay in detour zone	0.33	1.36	Alt 16	average speed in fastest zone	0.04	0.80
Alt 3	average delay in detour zone	0.58	3.71	Alt 17	average speed in fastest zone	0.11	2.55
Alt 4	average delay in detour zone	0.75	4.98	Alt 18	average speed in fastest zone	−0.08	−1.46
Alt 5	average delay in detour zone	1.93	6.77	Alt 19	average speed in fastest zone	0.15	2.44
Number of observations				4000			
Number of coefficients				294			
***L*(0)**				−10010.79			
***L*(*β*)**				−5423.77			
***ρ*(0)**				0.46			

Table 15
Conventional labeling route choice model.

Coefficient		Value	*t*-test	Coefficient		Value	*t*-test
ASC 1		2.56	0.00	Alt 1	number of city nodes in detour zone	0.04	0.14
ASC 2		3.58	24.80	Alt 2	number of city nodes in detour zone	1.68	5.44
ASC 3		1.12	9.22	Alt 3	number of city nodes in detour zone	2.35	7.46
ASC 4		0.08	0.71	Alt 4	number of city nodes in detour zone	−0.55	−2.71
ASC 5		−2.10	−16.60	Alt 5	number of city nodes in detour zone	−2.97	−16.30
ASC 6		0.99	8.45	Alt 6	number of city nodes in detour zone	0.19	0.70
ASC 7		5.27	41.10	Alt 7	number of city nodes in detour zone	1.46	7.07
ASC 8		2.05	15.10	Alt 8	number of city nodes in detour zone	0.51	2.20
ASC 9		3.99	62.20	Alt 9	number of city nodes in detour zone	−1.26	−13.10
ASC 10		−4.00	0.00	Alt 10	number of city nodes in detour zone	7.14	188.00
ASC 11		4.33	34.70	Alt 11	number of city nodes in detour zone	−0.96	−3.98
ASC 12		−2.20	−9.26	Alt 12	number of city nodes in detour zone	3.31	35.60
ASC 13		−3.44	−18.00	Alt 13	number of city nodes in detour zone	−1.51	−6.60
Alt 1	route highway/expressway length	0.20	6.08	Alt 1	average delay in detour zone	0.50	2.03
Alt 2	route highway/expressway length	0.28	7.11	Alt 2	average delay in detour zone	0.84	3.26
Alt 3	route highway/expressway length	0.20	4.99	Alt 3	average delay in detour zone	2.28	8.77
Alt 4	route highway/expressway length	0.06	0.80	Alt 4	average delay in detour zone	2.96	22.40
Alt 5	route highway/expressway length	0.22	2.96	Alt 5	average delay in detour zone	−0.24	−4.79
Alt 6	route highway/expressway length	0.24	4.72	Alt 6	average delay in detour zone	0.87	4.22
Alt 7	route highway/expressway length	0.08	1.42	Alt 7	average delay in detour zone	2.75	15.70
Alt 8	route highway/expressway length	0.02	0.43	Alt 8	average delay in detour zone	3.70	21.10
Alt 9	route highway/expressway length	0.13	1.52	Alt 9	average delay in detour zone	0.13	6.32
Alt 10	route highway/expressway length	−0.11	−1.11	Alt 10	average delay in detour zone	4.98	604.00
Alt 11	route highway/expressway length	0.21	2.13	Alt 11	average delay in detour zone	0.82	30.70
Alt 12	route highway/expressway length	0.24	1.99	Alt 12	average delay in detour zone	−1.33	−51.10
Alt 13	route highway/expressway length	0.06	0.52	Alt 13	average delay in detour zone	8.26	165.00
Alt 1	fastest route travel time	0.22	2.67	Alt 1	highway/expressway percentage in detour zone	−3.72	−8.61
Alt 2	fastest route travel time	0.20	2.38	Alt 2	highway/expressway percentage in detour zone	−2.95	−10.30
Alt 3	fastest route travel time	0.12	1.50	Alt 3	highway/expressway percentage in detour zone	−3.24	−14.80
Alt 4	fastest route travel time	0.67	3.91	Alt 4	highway/expressway percentage in detour zone	−3.61	−38.70
Alt 5	fastest route travel time	−0.01	−0.06	Alt 5	highway/expressway percentage in detour zone	−1.64	−36.30
Alt 6	fastest route travel time	0.27	2.47	Alt 6	highway/expressway percentage in detour zone	−1.96	−12.00
Alt 7	fastest route travel time	0.12	1.39	Alt 7	highway/expressway percentage in detour zone	−2.47	−23.60
Alt 8	fastest route travel time	0.14	1.50	Alt 8	highway/expressway percentage in detour zone	−2.97	−26.70
Alt 9	fastest route travel time	0.82	3.40	Alt 9	highway/expressway percentage in detour zone	0.79	32.80
Alt 10	fastest route travel time	0.19	1.73	Alt 10	highway/expressway percentage in detour zone	0.21	15.20
Alt 11	fastest route travel time	0.26	2.23	Alt 11	highway/expressway percentage in detour zone	−4.33	−55.50
Alt 12	fastest route travel time	0.23	1.91	Alt 12	highway/expressway percentage in detour zone	−1.20	−49.30
Alt 13	fastest route travel time	0.14	0.88	Alt 13	highway/expressway percentage in detour zone	−6.93	−158.00
Alt 1	route intersection time	0.09	1.68	Alt 1	average speed in detour zone	0.07	1.99
Alt 2	route intersection time	0.00	0.02	Alt 2	average speed in detour zone	0.07	1.66
Alt 3	route intersection time	−0.06	−0.95	Alt 3	average speed in detour zone	0.09	2.27
Alt 4	route intersection time	0.02	0.18	Alt 4	average speed in detour zone	0.07	1.67
Alt 5	route intersection time	0.16	1.44	Alt 5	average speed in detour zone	0.08	1.70

(*continued on next page*)

Table 15 (*continued*)

Coefficient		Value	*t*-test	Coefficient		Value	*t*-test
Alt 6	route intersection time	0.16	1.92	Alt 6	average speed in detour zone	0.09	2.10
Alt 7	route intersection time	0.13	1.35	Alt 7	average speed in detour zone	0.01	0.25
Alt 8	route intersection time	0.20	1.88	Alt 8	average speed in detour zone	0.06	1.37
Alt 9	route intersection time	0.43	3.04	Alt 9	average speed in detour zone	−0.07	−1.19
Alt 10	route intersection time	0.29	2.05	Alt 10	average speed in detour zone	0.18	2.64
Alt 11	route intersection time	0.26	1.14	Alt 11	average speed in detour zone	0.09	1.54
Alt 12	route intersection time	0.00	−0.01	Alt 12	average speed in detour zone	0.08	1.35
Alt 13	route intersection time	−0.04	−0.21	Alt 13	average speed in detour zone	0.19	3.07
Alt 1	OD distance	0.23	3.32	Alt 1	number of left turns	−0.36	−11.70
Alt 2	OD distance	0.16	1.74	Alt 2	number of left turns	−0.41	−10.20
Alt 3	OD distance	0.18	1.96	Alt 3	number of left turns	−0.25	−6.22
Alt 4	OD distance	0.43	3.11	Alt 4	number of left turns	−0.28	−3.99
Alt 5	OD distance	0.16	1.32	Alt 5	number of left turns	−0.41	−5.03
Alt 6	OD distance	0.28	2.60	Alt 6	number of left turns	−0.30	−5.97
Alt 7	OD distance	0.24	2.22	Alt 7	number of left turns	−0.30	−5.18
Alt 8	OD distance	0.17	1.35	Alt 8	number of left turns	−0.35	−5.43
Alt 9	OD distance	−0.31	−2.43	Alt 9	number of left turns	−0.26	−2.87
Alt 10	OD distance	0.35	2.13	Alt 10	number of left turns	−0.33	−2.82
Alt 11	OD distance	1.04	6.18	Alt 11	number of left turns	−0.64	−5.80
Alt 12	OD distance	−0.08	−0.61	Alt 12	number of left turns	−0.29	−2.97
Alt 13	OD distance	0.11	0.72	Alt 13	number of left turns	−0.24	−2.48
Alt 1	shortest route distance	0.49	6.55	Alt 1	route cost	−0.01	−0.87
Alt 2	shortest route distance	−0.16	−2.82	Alt 2	route cost	0.00	0.11
Alt 3	shortest route distance	0.39	3.88	Alt 3	route cost	−0.01	−0.78
Alt 4	shortest route distance	0.26	1.70	Alt 4	route cost	0.05	2.77
Alt 5	shortest route distance	−0.24	−1.67	Alt 5	route cost	−0.09	−0.58
Alt 6	shortest route distance	0.20	2.01	Alt 6	route cost	0.01	0.62
Alt 7	shortest route distance	0.40	3.93	Alt 7	route cost	−0.04	−1.60
Alt 8	shortest route distance	0.50	3.60	Alt 8	route cost	−0.02	−0.87
Alt 9	shortest route distance	0.88	6.19	Alt 9	route cost	0.00	−0.20
Alt 10	shortest route distance	0.39	2.23	Alt 10	route cost	0.01	0.12
Alt 11	shortest route distance	−0.38	−1.56	Alt 11	route cost	−0.38	−1.45
Alt 12	shortest route distance	0.86	3.88	Alt 12	route cost	−0.04	−0.54
Alt 13	shortest route distance	0.24	1.62	Alt 13	route cost	−0.35	−1.69
Alt 1	shortest route travel time	0.03	0.36	Alt 1	intersection time in fastest zone	13.40	251.00
Alt 2	shortest route travel time	0.03	0.36	Alt 2	intersection time in fastest zone	13.50	247.00
Alt 3	shortest route travel time	0.03	0.36	Alt 3	intersection time in fastest zone	13.90	230.00
Alt 4	shortest route travel time	−0.09	−0.69	Alt 4	intersection time in fastest zone	14.20	322.00
Alt 5	shortest route travel time	0.18	2.03	Alt 5	intersection time in fastest zone	16.50	544.00
Alt 6	shortest route travel time	−0.03	−0.28	Alt 6	intersection time in fastest zone	13.30	273.00
Alt 7	shortest route travel time	0.03	0.36	Alt 7	intersection time in fastest zone	5.82	110.00
Alt 8	shortest route travel time	0.03	0.36	Alt 8	intersection time in fastest zone	15.80	306.00
Alt 9	shortest route travel time	−0.41	−2.09	Alt 9	intersection time in fastest zone	2.12	151.00
Alt 10	shortest route travel time	0.03	0.36	Alt 10	intersection time in fastest zone	−4.85	−839.00
Alt 11	shortest route travel time	0.03	0.36	Alt 11	intersection time in fastest zone	5.97	232.00
Alt 12	shortest route travel time	0.03	0.36	Alt 12	intersection time in fastest zone	17.30	1010.00
Alt 13	shortest route travel time	0.07	0.57	Alt 13	intersection time in fastest zone	24.40	685.00
Alt 1	route travel time	−0.08	−2.32	Alt 1	number of city nodes in fastest zone	0.03	0.09
Alt 2	route travel time	−0.04	−1.62	Alt 2	number of city nodes in fastest zone	−1.45	−4.95
Alt 3	route travel time	−0.03	−1.81	Alt 3	number of city nodes in fastest zone	−0.81	−2.70
Alt 4	route travel time	−0.41	−4.17	Alt 4	number of city nodes in fastest zone	0.66	2.30
Alt 5	route travel time	−0.03	−0.76	Alt 5	number of city nodes in fastest zone	3.27	14.00
Alt 6	route travel time	−0.09	−2.20	Alt 6	number of city nodes in fastest zone	0.34	1.17
Alt 7	route travel time	−0.02	−1.02	Alt 7	number of city nodes in fastest zone	−0.80	−2.80
Alt 8	route travel time	−0.04	−0.75	Alt 8	number of city nodes in fastest zone	−1.79	−5.65
Alt 9	route travel time	−0.25	−2.03	Alt 9	number of city nodes in fastest zone	2.42	22.80
Alt 10	route travel time	−0.16	−1.90	Alt 10	number of city nodes in fastest zone	0.15	4.24
Alt 11	route travel time	−0.07	−0.74	Alt 11	number of city nodes in fastest zone	−2.84	−9.75
Alt 12	route travel time	−0.02	−0.24	Alt 12	number of city nodes in fastest zone	−0.43	−3.98
Alt 13	route travel time	−0.06	−0.78	Alt 13	number of city nodes in fastest zone	2.30	7.38
Alt 1	route distance	−0.76	−11.30	Alt 1	average delay in fastest zone	−1.38	−5.08
Alt 2	route distance	−0.15	−2.69	Alt 2	average delay in fastest zone	−0.71	−2.18
Alt 3	route distance	−0.61	−6.92	Alt 3	average delay in fastest zone	−1.88	−6.42
Alt 4	route distance	−0.72	−5.44	Alt 4	average delay in fastest zone	−2.32	−15.90
Alt 5	route distance	0.17	1.23	Alt 5	average delay in fastest zone	0.68	18.80
Alt 6	route distance	−0.58	−6.68	Alt 6	average delay in fastest zone	−1.06	−3.56
Alt 7	route distance	−0.41	−4.14	Alt 7	average delay in fastest zone	−1.52	−7.05
Alt 8	route distance	−0.38	−2.66	Alt 8	average delay in fastest zone	−0.45	−1.90
Alt 9	route distance	−0.61	−4.29	Alt 9	average delay in fastest zone	−2.02	−73.90

(*continued on next page*)

Table 15 (*continued*)

Coefficient		Value	*t*-test	Coefficient		Value	*t*-test
Alt 10	route distance	−0.42	−1.94	Alt 10	average delay in fastest zone	−3.68	−461.00
Alt 11	route distance	0.21	0.47	Alt 11	average delay in fastest zone	0.09	2.40
Alt 12	route distance	−0.74	−2.77	Alt 12	average delay in fastest zone	−3.52	−109.00
Alt 13	route distance	0.30	0.73	Alt 13	average delay in fastest zone	−4.58	−85.20
Alt 1	route number of links	−0.29	−5.04	Alt 1	highway/expressway percentage in fastest zone	0.31	0.99
Alt 2	route number of links	−0.35	−4.92	Alt 2	highway/expressway percentage in fastest zone	0.70	1.75
Alt 3	route number of links	−0.40	−4.79	Alt 3	highway/expressway percentage in fastest zone	1.06	3.17
Alt 4	route number of links	−0.53	−3.84	Alt 4	highway/expressway percentage in fastest zone	0.53	2.42
Alt 5	route number of links	−0.29	−1.74	Alt 5	highway/expressway percentage in fastest zone	3.46	65.30
Alt 6	route number of links	−0.60	−4.52	Alt 6	highway/expressway percentage in fastest zone	1.61	6.39
Alt 7	route number of links	−0.29	−2.58	Alt 7	highway/expressway percentage in fastest zone	0.32	1.77
Alt 8	route number of links	−0.07	−0.53	Alt 8	highway/expressway percentage in fastest zone	1.05	9.48
Alt 9	route number of links	−0.59	−2.80	Alt 9	highway/expressway percentage in fastest zone	3.67	172.00
Alt 10	route number of links	−0.21	−0.72	Alt 10	highway/expressway percentage in fastest zone	2.35	227.00
Alt 11	route number of links	−0.03	−0.14	Alt 11	highway/expressway percentage in fastest zone	−1.37	−25.00
Alt 12	route number of links	−0.37	−2.15	Alt 12	highway/expressway percentage in fastest zone	0.95	35.60
Alt 13	route number of links	−0.02	−0.10	Alt 13	highway/expressway percentage in fastest zone	2.65	59.20
Alt 1	route number of city nodes	−0.06	−4.22	Alt 1	average speed in fastest zone	−0.04	−1.03
Alt 2	route number of city nodes	−0.04	−2.38	Alt 2	average speed in fastest zone	−0.06	−1.65
Alt 3	route number of city nodes	−0.02	−1.38	Alt 3	average speed in fastest zone	−0.06	−1.61
Alt 4	route number of city nodes	−0.05	−1.51	Alt 4	average speed in fastest zone	0.00	−0.10
Alt 5	route number of city nodes	−0.13	−2.87	Alt 5	average speed in fastest zone	−0.05	−1.27
Alt 6	route number of city nodes	−0.04	−1.44	Alt 6	average speed in fastest zone	−0.09	−2.33
Alt 7	route number of city nodes	0.02	0.87	Alt 7	average speed in fastest zone	−0.06	−1.39
Alt 8	route number of city nodes	−0.03	−1.08	Alt 8	average speed in fastest zone	−0.09	−2.22
Alt 9	route number of city nodes	−0.13	−1.87	Alt 9	average speed in fastest zone	−0.03	−0.55
Alt 10	route number of city nodes	−0.16	−2.21	Alt 10	average speed in fastest zone	−0.17	−2.81
Alt 11	route number of city nodes	0.01	0.27	Alt 11	average speed in fastest zone	−0.11	−2.16
Alt 12	route number of city nodes	−0.02	−0.54	Alt 12	average speed in fastest zone	−0.11	−2.10
Alt 13	route number of city nodes	−0.15	−1.82	Alt 13	average speed in fastest zone	−0.04	−0.82
Alt 1	route delay	−0.26	−5.48	Alt 1	shortest route directness	−0.18	−0.90
Alt 2	route delay	−0.23	−5.14	Alt 2	shortest route directness	−0.28	−0.93
Alt 3	route delay	−0.15	−3.36	Alt 3	shortest route directness	−0.07	−0.19
Alt 4	route delay	−0.16	−1.88	Alt 4	shortest route directness	0.36	1.01
Alt 5	route delay	−0.29	−3.87	Alt 5	shortest route directness	0.54	2.32
Alt 6	route delay	−0.28	−4.09	Alt 6	shortest route directness	0.19	0.58
Alt 7	route delay	−0.31	−4.12	Alt 7	shortest route directness	−0.83	−2.58
Alt 8	route delay	−0.54	−6.08	Alt 8	shortest route directness	−0.56	−1.70
Alt 9	route delay	−0.43	−3.28	Alt 9	shortest route directness	−2.20	−20.80
Alt 10	route delay	−0.28	−1.88	Alt 10	shortest route directness	−0.08	−1.35
Alt 11	route delay	−0.66	−3.16	Alt 11	shortest route directness	2.92	11.10
Alt 12	route delay	−0.56	−3.76	Alt 12	shortest route directness	−0.86	−8.25
Alt 13	route delay	−0.13	−1.00	Alt 13	shortest route directness	−0.16	−0.64
Alt 1	intersection time in detour zone	3.39	59.80				
Alt 2	intersection time in detour zone	0.69	9.39				
Alt 3	intersection time in detour zone	3.09	43.30				
Alt 4	intersection time in detour zone	1.44	40.80				
Alt 5	intersection time in detour zone	6.07	235.00	**Number of observations**		4000	
Alt 6	intersection time in detour zone	10.90	201.00	**Number of coefficients**		325	
Alt 7	intersection time in detour zone	−6.13	−131.00	$L(0)$		−135560.30	
Alt 8	intersection time in detour zone	−0.61	−12.80	$L(\beta)$		−4408.18	
Alt 9	intersection time in detour zone	19.00	1760.00	$\rho(0)$		0.97	
Alt 10	intersection time in detour zone	13.50	2400.00				
Alt 11	intersection time in detour zone	−5.26	−275.00				
Alt 12	intersection time in detour zone	29.30	2480.00				
Alt 13	intersection time in detour zone	−17.30	−540.00				

Table 16
Eliminate insignificant coefficients before checking each feature final estimated model result.

Coefficient	Value	p-value
ASC 1	−0.86	0.00
ASC 2	1.51	0.13
ASC 3	−1.78	0.00
ASC 4	−0.01	0.00
ASC 5	0.15	0.03
Alt 1,2,3,4,5 fastest route travel time	−0.11	0.00
Alt 1 number of city nodes	−0.14	0.00
Alt 2 number of city nodes	−0.63	0.00
Alt 3 number of city nodes	−0.10	0.00
Alt 4 number of city nodes	−0.04	0.00
Alt 5 number of city nodes	−0.17	0.00
Alt 1,2 route delay	−0.15	0.00
Alt 3,4,5 route delay	−0.03	0.00
Alt 1,2,5 highway/expressway percentage in detour zone	−1.32	0.00
Alt 1 highway/expressway length	0.40	0.00
Alt 2 highway/expressway length	−0.40	0.00
Alt 3 highway/expressway length	0.32	0.00
Alt 4 highway/expressway length	0.13	0.00
Alt 5 highway/expressway length	0.17	0.00
Alt 4 average speed in fastest zone	−0.17	0.00
Number of observations	4000	
Number of coefficients	20	
***L*(0)**	−7167.04	
***L*(β)**	−5280.92	
ρ(0)	0.26	
Test sample size	1002	
Prediction Accuracy (F1 score)	32%	

Table 17
Eliminate insignificant coefficients before each parameter combination final estimated model result.

Coefficient	Value	p-value
ASC 1	−0.86	0.00
ASC 2	1.51	0.13
ASC 3	−1.78	0.00
ASC 4	−0.01	0.00
ASC 5	0.15	0.03
Alt 1 number of city nodes	−0.20	0.00
Alt 2 number of city nodes	−0.62	0.00
Alt 3 number of city nodes	−0.10	0.00
Alt 4 number of city nodes	−0.04	0.00
Alt 5 number of city nodes	−0.17	0.00
Alt 1,2 route delay	−0.17	0.00
Alt 3,4,5 route delay	−0.02	0.00
Alt 1 highway/expressway length	0.31	0.00
Alt 2 highway/expressway length	−0.36	0.00
Alt 3 highway/expressway length	0.29	0.00
Alt 4 highway/expressway length	0.08	0.00
Alt 5 highway/expressway length	0.11	0.00
Number of observations	4000	
Number of coefficients	17	
***L*(0)**	−7167.04	
***L*(β)**	−5672.68	
ρ(0)	0.21	
Test sample size	1002	
Prediction Accuracy (F1 score)	27%	

References

Arthur, D., Vassilvitskii, S., 2006. k-means++: The advantages of careful seeding. Stanford.
Axhausen, K.W., Schüssler, N., 2009. Accounting for route overlap in urban and suburban route choice decisions derived from GPS observations. Arbeitsberichte Verkehrs-und Raumplanung 590.
Ben-Akiva, M.E., Bergman, M.J., Daly, A.J., Ramaswamy, R., 1984. Modelling inter-urban route choice behaviour. In: Proceedings of the Ninth International Symposium on Transportation and Traffic Theory, VNU Science Press, Utrecht, The Netherlands, pp. 299–330.

Bekhor, S., Ben-Akiva, M.E., Ramming, M.S., 2006. Evaluation of choice set generation algorithms for route choice models. Ann. Oper. Res. 144 (1), 235–247.
Bekhor, S., Toledo, T., Prashker, J.N., 2008. Effects of choice set size and route choice models on path-based traffic assignment. Transportmetrica 4 (2), 117–133.
Bhat, C.R., Dubey, S.K., 2014. A new estimation approach to integrate latent psychological constructs in choice modeling. Transport. Res. Part B: Methodol. 67, 68–85.
Bierlaire, M., Chen, J., Newman, J., 2010. Modeling route choice behavior from smartphone GPS data (No. REP_WORK).
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Bovy, P.H.L., 2009. On modelling route choice sets in transportation networks: a synthesis. Trans. Rev. 29 (1), 43–68.
Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F., 2019. Applying a random forest method approach to model travel mode choice behavior. Travel Behav. Soc. 14, 1–10.
Ciscal-Terry, W., Dell'Amico, M., Hadjidimitriou, N.S., Iori, M., 2016. An analysis of drivers route choice behaviour using GPS data and optimal alternatives. J. Transp. Geogr. 51, 119–129.
Dhakar, N.S., Srinivasan, S., 2014. Route choice modeling using GPS-based travel surveys. Transp. Res. Rec. 2413 (1), 65–73.
Flötteröd, G., Bierlaire, M., 2013. Metropolis–Hastings sampling of paths. Transport. Res. Part B: Methodol. 48, 53–66.
Fosgerau, M., Frejinger, E., Karlstrom, A., 2013. A link based network route choice model with unrestricted choice set. Transport. Res. Part B: Methodol. 56, 70–80.
Frejinger, E., Bierlaire, M., Ben-Akiva, M., 2009. Sampling of alternatives for route choice modeling. Transport. Res. Part B: Methodol. 43 (10), 984–994.
Jahangiri, A., Rakha, H.A., 2015. Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. IEEE Trans. Intell. Transport. Syst. 16 (5), 2406–2417.
Louppe, G., 2014. Understanding random forests: From theory to practice. arXiv preprint arXiv:1407.7502.
Lai, X., Fu, H., Li, J., Sha, Z., 2018. Understanding drivers' route choice behaviours in the urban network with machine learning models. IET Intell. Trans. Syst. 13 (3), 427–434.
Maaten, L.V.D., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (Nov), 2579–2605.
Nahmias-Biran, B.-H., Han, Y., Bekhor, S., Zhao, F., Zegras, C., Ben-Akiva, M., 2018. Enriching activity-based models using smartphone-based travel surveys. Transp. Res. Rec. 2672 (42), 280–291.
Newson, P., Krumm, J., 2009. Hidden Markov map matching through noise and sparseness. In: Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems, pp. 336–343.
Oyama, Y., Hato, E., 2017. A discounted recursive logit model for dynamic gridlock network analysis. Transport. Res. Part C: Emerg. Technol. 85, 509–527.
Park, K., Bell, M.G., 2011. Learning user preferences of route choice using fuzzy decision tree induction (No. 11-3646).
Prato, C.G., Bekhor, S., 2006. Applying branch-and-bound technique to route choice set generation. Transp. Res. Rec. 1985 (1), 19–28.
Ramming, M.S., 2001. Network knowledge and route choice. Unpublished Ph. D. Thesis, Massachusetts Institute of Technology.
Rieser-Schüssler, N., Balmer, M., Axhausen, K.W., 2013. Route choice sets for very high-resolution data. Transportmetrica A: Trans. Scie. 9 (9), 825–845.
Saeys, Y., Inza, I., Larranaga, P., 2007. A review of feature selection techniques in bioinformatics. Bioinformatics 23 (19), 2507–2517.
Shiftan, Y., Bekhor, S., 2021. Applying random forest in discrete choice modeling: a case study of household car allocation. In: Proceeding of hEART 2021: 9th Symposium of the European Association for Research in Transportation.
Sun, B., Park, B.B., 2017. Route choice modeling with Support Vector Machine. Transp. Res. Procedia 25, 1806–1814.
Ton, D., Cats, O., Duives, D., Hoogendoorn, S., 2017. How do people cycle in Amsterdam, Netherlands?: Estimating cyclists' route choice determinants with GPS data from an Urban Area. Transp. Res. Rec. 2662 (1), 75–82.
Ton, D., Duives, D., Cats, O., Hoogendoorn, S., 2018. Evaluating a data-driven approach for choice set identification using GPS bicycle route choice data from Amsterdam. Travel Behav. Soc. 13, 105–117.
Tribby, C.P., Miller, H.J., Brown, B.B., Werner, C.M., Smith, K.R., 2017. Analyzing walking route choice through built environments using random forests and discrete choice techniques. Environ. Plann. B: Urban Anal. City Sci. 44 (6), 1145–1167.
Wong, M., Farooq, B., Bilodeau, G.-A., 2018. Discriminative conditional restricted Boltzmann machine for discrete choice and latent variable modelling. J. Choice Modell. 29, 152–168.
Yamamoto, T., Kitamura, R., Fujii, J., 2002. Drivers' route choice behavior: analysis by data mining algorithms. Transp. Res. Rec. 1807 (1), 59–66.
Yao, R., Bekhor, S., 2021. Experiments on route choice set generation using a large GPS trajectory set. In: Proceeding of hEART 2021 : 9th Symposium of the European Association for Research in Transportation.
Zimmermann, M., Mai, T., Frejinger, E., 2017. Bike route choice modeling using GPS data without choice sets of paths. Transport. Res. Part C: Emerg. Technol. 75, 183–196.