International Journal of Computational Science and Engineering

ISSN online: 1742-7193 - ISSN print: 1742-7185
https://www.inderscience.com/ijcse

Statistical analysis for predicting residents' travel mode based on random forest

Lei Chen, Zhengyan Sun, Shunxiang Zhang, Guangli Zhu, Subo Wei

DOI: 10.1504/IJCSE.2024.10061857

Article History:

Received:	21 May 2022
Last revised:	07 June 2022
Accepted:	11 July 2022
Published online:	25 January 2024

Statistical analysis for predicting residents' travel mode based on random forest

Lei Chen

School of Computer Science,
Huainan Normal University,
Huainan, China
Email: leichen@hnnu.edu.cn
and
Artificial Intelligence Research Institute of
Hefei Comprehensive National Science Center,
Hefei, China

Zhengyan Sun, Shunxiang Zhang*,
Guangli Zhu and Subo Wei

School of Computer Science and Engineering,
Anhui University of Science and Technology,
Huainan, China
Email: 1948576279@qq.com
Email: sxzhang@aust.edu.cn
Email: glzhu@aust.edu.cn
Email: 1767773043@qq.com
*Corresponding author

Abstract: Random forest has achieved good results in the prediction task, but due to the complexity of travel mode and the uncertainty of random forest, the prediction accuracy of travel mode is low. To improve the accuracy of prediction, this paper proposes a residents' travel modes prediction method based on the random forest. To extract valuable feature information, the questionnaire survey data is collected, which is pre-processed by three kinds of appropriate methods. Then, each feature is analysed by the statistical learning method to obtain the important feature of transportation selection. Finally, a random forest is constructed to predict the travel mode of residents' selection of transportation. The parameters of random forest are modified and improved to achieve higher prediction accuracy of travel mode. The experimental results show that the method proposed in this paper effectively improves the prediction accuracy of the travel mode.

Keywords: residents' travel mode; statistical analysis; random forest.

Reference to this paper should be made as follows: Chen, L., Sun, Z., Zhang, S., Zhu, G. and Wei, S. (2024) 'Statistical analysis for predicting residents' travel mode based on random forest', *Int. J. Computational Science and Engineering*, Vol. 27, No. 1, pp.9–19.

Biographical notes: Lei Chen received his Master's degree from the School of Energy and Safety, Anhui University of Science and Technology, in 2008. He is a Professor at the Huainan Normal University. His research interests focus on knowledge graphs and recommendation systems.

Zhengyan Sun received her Bachelor's degree from the Qingdao Institute of Technology in 2020, China. Currently, she is an MS candidate in Computer Science and Engineering at the Anhui University of Science and Technology. Her current research interests are in natural language processing and data mining.

Shunxiang Zhang received his PhD from the School of Computing Engineering and Science, Shanghai University, Shanghai, in 2012. He is a Professor at the Anhui University of Science and Technology, China. His current research interests include web mining, sentimental analysis of social media, and complex network.

Guangli Zhu received her MS degree from the School of Computing Engineering and Science, Anhui University of Science and Technology, in 2005, China. She is an Associate Professor at the Anhui University of Science and Technology, China. Her current research interests include web mining, semantic search, and calculation theory.

Subo Wei received his Bachelor's degree from the Qingdao Institute of Technology in 2020, China. Currently, he is an MS candidate in Computer Science and Engineering at the Anhui University of Science and Technology. Her current research interests are in natural language processing, complex network and sentiment analysis.

1 Introduction

Many researchers have investigated that residents' travel mode is affected by three factors: personal living standards, local traffic conditions, and low-carbon travel in response to the national call. Due to a variety of factors induced by the selection of transportation diversity, it dramatically increases the difficulty of prediction. At the same time, the lack of true and effective data makes the prediction of travel mode even more difficult.

Because of the above problems, the following points are mainly considered:

1 Data processing to prepare for subsequent data analysis. Wu et al. (2022) used real data for processing and achieved good results.

2 Data analysis is analyse the importance of feature variables that affect travel modes.

3 Construction of a prediction model.

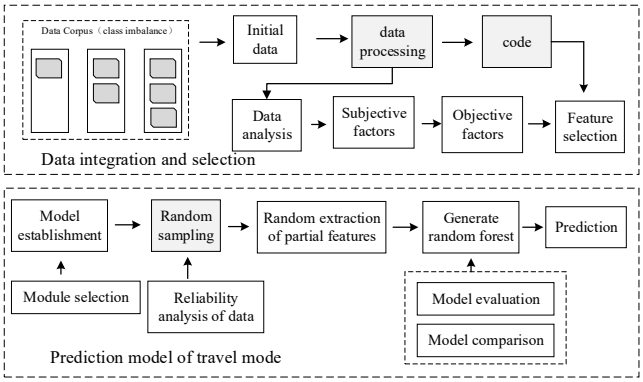
The obtained features construct a random forest to predict the travel mode of residents' selection of transportation. Different from the traditional random forest model, sample screen and class imbalance (Wang et al., 2019) are considered to improve the random forest. A higher accuracy model can be achieved.

Based on the above considerations, this work proposes a travel mode prediction model based on the random forest model. The framework for predicting residents' travel mode is shown in Figure 1. The motivation of the model is to learn the features of high discrimination for improving the accuracy of prediction. By modifying a random forest model, the accuracy of prediction is improved. This paper mainly makes the following two contributions.

- *Data integration and selection:* Based on the analysis of the data obtained from the questionnaire survey, multiple feature variables that can affect traffic are the selection. Analyse each feature variable element, find out the potential relationship of feature variables, and provide data preparation for variable selection.
- *Prediction of travel mode:* Variable selection is used in the construction of a decision tree to reduce the complexity of the decision tree. Each decision tree of the random forest adopts parallel computing. Different weights are given to each decision tree to improve the prediction ability of the model.

This paper is organised as follows. Section 2 introduces the related work and Section 3 gives the methods of data processing and analysis. Section 4 provides the construction of the transportation mode prediction model. Section 5 presents the experimental analysis of the model. Section 6 is the conclusions and looking forward to future work.

Figure 1 The framework of residents' travel mode prediction



2 Related works

2.1 Research on travel mode

At present, the existing research on travel modes is mainly divided into two types. One is based on travel behaviour theory, and the other is travel mode prediction.

According to travel behaviour theory, Shaaban and Maher (2020) used the theory of planned behaviour (TPB) to explore the impact of psychological factors on public transportation residents' selection behaviour. Based on the technology acceptance model (TAM) and TPB, Chen (2016) discussed the impact of different perception indicators and subjective norms on public bicycle loyalty. Basis the impact of travel behaviour, the research shows that factors such as travel time and cost (Bösehans and Walker, 2020), travel satisfaction and happiness (Arroyo et al., 2020), and comfort (Gao et al., 2017) have a significant impact on travel behaviour selection. In addition, special conditions such as the built environment (Yu et al., 2019) and suspension of public transportation (Zanni and Ryley, 2015) are also significantly correlated with residents' travel behaviour.

Nguyen-Phuoc et al. (2018) mainly consider the impact of mode conversion from public transportation to a private car. Chakrabarti and Shin (2017) have researched the relationship between auto-dependence and the inactivity epidemic to encourage people for using public transportation. From the perspective of subjective well-being, Li et al. (2020) get the relationship between the evolution of well-being and low-carbon travel that will help promote green travel.

For the prediction of traffic patterns, Hoque et al. (2021) improved prediction accuracy. Zhou et al. (2020) and Kim

et al. (2020) consider the activities of various factors, the optimal calculation of travel mode is carried out, and the optimal scheme is provided. Cheng et al. (2020) reduce the uncertainty of the model and the random feature is selected to capture the uncertainty of the model. Compared with the traditional static dependence, Xu et al. (2019) analyse the dynamic changes in residents' travel features in the future. To improve the prediction performance of short-term travel, Guo and Zhang (2019) redefined the prediction problem as a learning residual function and reduce the residual root function by 17.87% on datasets. Slik and Bhulai (2020) proposed a unique dataset for travel mode prediction, which significantly improves the accuracy of prediction results.

2.2 Research on random forest

In recent years, random forest because of its superiority, has seen more and more researchers use it as a prediction model. Wang and Chen (2020) introduce the concept of distance weight to optimise the parallel computing algorithm of all decision trees in random forests. By improving the sampling method to enrich random forest, Ghosh and Cabrera (2021) moved from the original random sampling to weighted random sampling to improve the accuracy of prediction. In the small dataset, Han et al. (2021) improve the prediction ability of small datasets by variable selection and inverse sampling probability weighting. Han et al. (2020), Mantero and Ishwaran (2021) and Jain and Phophalia (2022) improved the random forest model and proposed new models for different scenarios.

The field of random forest applications is also expanding. D'Amato et al. (2022) focused on the research of ESG investment in the financial field. Khandelwal et al. (2020) proposed the random forest model to predict the dynamic price of spot price to help people make bidding decisions. Behr et al. (2020) used the random forest of conditional inference trees to predict university dropouts and discover students with dropout risks in time.

Arya and Sastry (2022) predict the price trend of the stock index by combining deep learning and extra tree ensemble, thus the accuracy of the optimised prediction model was significantly improved. Capitaine et al. (2021) use the random forest model to analyse high-dimensional gene data, which allows the covariance structure to change over time to make the model more flexible.

Due to the diversification of urban transportation modes and the spatial imbalance of infrastructure construction, residents' selection willingness and behaviour show complex heterogeneity. Because of these factors, residents' impact on public transportation dependence degree is different. Therefore, it is of great significance to explore the relationship between multi-dimensional feature factors and residents' reliance on public transportation. Although the general random forest model has a good prediction effect, the complexity of residents' travel mode and the uncertainty of the model itself, this paper improves the random forest model.

3 Data processing and simple feature analysis

3.1 Credibility analysis of data

Because the data used in this paper is real data, including some subjective data, this need to evaluate these data and measure residents' travel mode objectively. It is necessary to set relevant indicators and indicator systems. On this basis, raw data are obtained through corresponding statistical surveys to provide basic support for subsequent evaluation and measurement. This paper mainly measures from both qualitative and quantitative aspects. It carries out several large-scale statistical surveys, which provide an important basis for reasonably evaluating the current situation of residents' travel mode selection. This data collection is mainly based on the questionnaire survey, which has accumulated a large number of people's travel-related information. The data include 27 feature genera such as trip purpose, age, mode of transportation, weather, travel time, travel times, etc.

Due to the different travel times of each age group, the transportation modes adopted are also different, and the value of the questionnaire is different. Therefore, the survey object is mainly based on the age group. The credibility of the content of the questionnaire is the primary guarantee. To test the credibility of the questionnaire, the questionnaire is analysed as follows.

Through the credibility analysis of the questionnaire, a questionnaire can be judged whether stable and credible. Firstly, the credibility of residents' travel mode is analysed to obtain Cronbach α. The coefficients are shown in Table 1. It can be seen from that Cronbach of the residents' travel mode questionnaire α. The coefficient is $0.776 > 0.7$, and the credibility is good. That is, the questionnaire has stability and credibility.

Table 1 Questionnaire credibility

Cronbach α coefficient	*Number of items*
0.776	50

3.2 Data sorting

The real data collected in the questionnaire survey is often affected by personal privacy protection, inconvenient filling in, false fill, and other factors, which will lead to a small amount of erroneous data. To make the data more effective at the same time, this paper needs to make the subsequent data analysis and model prediction scientific and reliable, so the original information needs to be processed.

1 *Judgment and processing of abnormal data:* Abnormal data mainly includes apparent error data and hidden error data. Obvious errors can be identified immediately. For example, single-selection questions become multiple-selection questions. Hidden errors need to be deleted after manual judgment.

2 *Processing of missing data:* There are various reasons for data missing. The original data shall not be modified or deleted to the greatest extent. For some data columns, the average value shall be used to fill the gaps as far as possible to ensure that there is enough data for subsequent data analysis.

3 *Simplicity of redundant data:* Due to human operation errors and other reasons, the collected survey volumes may have some redundancy problems. Redundant data may increase the weight of an influencing factor, resulting in the poor ability of the prediction model. Therefore, if redundant data are similar, the average value shall be taken. If all the data in the two questionnaires are the same, only one valid data shall be randomly selected.

3.3 Data statistical analysis

Many feature factors are affecting the selection of travel modes. This paper mainly analyses it from four aspects: family factors, personal factors, objective factors, and subjective factors.

1 *The influence of the family factors on the selection of travel mode:* Family factors will directly affect the selection of transportation modes. In cities, families with high income often choose private cars, while families with low income choose public transportation. Families with a large population often buy cars, so private cars are often used as usual means of transportation.

2 *Influence of personal factors on the selection of travel mode:* For the relationship between the selection of transportation mode and personal factors, different people have different travel motivations and choose different transportation tools. Individual age, education level, work and other factors are closely related. In-service staff usually uses public transport and private cars to work. Retirement posts and students choose to walk more leisurely.

3 *Influence of subjective factors on the selection of travel mode:* Subjective factors have a certain influence on the selection of transportation modes, such as personal space comfort and personal mood are the reasons why residents choose different transportation modes. But in most cases, objective factors affect subjective factors, which eventually lead to the selection of different transportation.

4 *Influence of objective factors on the selection of travel mode:* Generally, the traffic in the city centre is very heavy. To avoid traffic jams, residents usually choose walking or cycling. Weather conditions are also a factor in the selection of transportation modes. As far as the objective factors of transportation are concerned, there are mainly five kinds: traffic congestion, weather factors, road quality, the distance of the route, and certainty of travel time.

4 Construction of travel mode selection model

4.1 Selection of prediction model

In traditional research, prediction is based on a group of samples, which is mainly divided into supervised learning and unsupervised learning according to whether there are labels. Reinforcement learning is based on feature selection and optimisation. For a large number of real datasets, it needs reasonable methods to mine information in the shortest possible time. There are many models to predict residents' travel modes, such as the basic linear regression model, support vector machine, decision tree, and so on. However, most of these models are used for the binary classification task. They are single classification methods, which often do not work well for the diversity and complexity of data in real life. At the same time, there may be linear or nonlinear relationships between variables, which are easy to be ignored.

Given the above problems, this paper uses the method of constructing a random forest model to predict the travel mode of residents. Stochastic forest adopts the integrated approach, which can strategically combine multiple simple decision trees to optimise the prediction performance. The selection of transportation is affected by various factors, and the prediction accuracy will be reduced due to the imbalance of data categories, the diversity of variables themselves, and the independent parallel calculation for each decision tree. Therefore, the traditional random forest effect is not good. This paper proposes an improved random forest model, which can distinguish and explain relevant variables and interactions. It can also improve the accuracy of model estimation and prediction and improve the generalisation ability of the overall model.

4.2 Basic thinking of random forest algorithm

The basic idea of the algorithm: the random forest model trains n decision trees (Nagra et al., 2020) based on the decision tree algorithm.

- *Decision tree construction:* For *M* samples, two-stage stratified sampling is used to obtain a training set. Due to the differences in the categories of data in the training set, the class balance method is used to make all category data as consistent as possible, which is used to train n decision trees in the random forest.
- *Feature selection:* When the decision tree is split, some features (e.g., age, occupation, income, etc.) are randomly selected at the root node for classification (e.g., walking, cycling, bus, etc.), the feature variables (Liu et al., 2022) are screened. And the inverse sampling probability weighting method is adopted to further increase the difference between the decision trees. Due to the linear relationship between some variables. Therefore, a method combining superposition GLM learner and RF learner is proposed to improve the model's generalisation ability.

- GLM learners can better find the relationship between independent variables and dependent variables, and some variables will be deleted in the variable selection part of the random forest model. Therefore, to ensure that the variable selection section deletes unimportant variables, the GLM learner is used to prove them. Combined with the GLM learner, the importance of each variable feature can be more accurately determined. When new data is input, the model has a stronger ability to learn and select features.
- *Out of package estimation:* In each round of random sampling of random forest, about 36.8% of the training set data are not collected. These data are called 'out of the bag', which use the data in the package to build the decision tree and use the data outside the package to test.
- *Result prediction:* In each decision tree, the splitting continues until the tree reaches the maximum depth. After the memory estimation of the primary decision tree model, the majority voting strategy is used to predict the final result.

The specific algorithm process is in Algorithm 1.

Algorithm 1 Construction of random forest

Input: a training dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, sample subset T and k *CART*

Output: k models classify X_t

1 for ($i = 1$; $i <= T$; i_{++}){
2 extracting t sample form a training set D_t;
3 Train a *CART* decision tree with set D_t;
4 select m feature subsets $\{M_1, M_2, \ldots, M_m\}$ from M features;
5 select the optimal feature for splitting;}
6 for ($cart = 1$; $cart <= N$; $cart_{++}$){
7 selects D_t samples to form training samples;
8 Training $k = \{1, 2, 3, \ldots, K\}$ *CART* decision trees;
9 Computing *Similarity*(*cart*);
10 Cut CART tree with *Similarity*(*cart*);
11 Calculate weight $w(cart_k)$;
12 Predict x_t;}
13 for ($i = 1$; $i <= N$; i_{++}){
14 $X_t = \dfrac{\sum x_i w_i}{w_i}$;
15 calculation error $Err(X_t)$;
$Err(x_t) = Bias^2 + Variance + IrreducibleError$;
16 if $Err(X_t) < 0.05$:
17 detraining;
18 else:
19 subsequent training.

When the algorithm is input, the size of the sample subset is T, which is composed of k *CART* numbers. Steps 1–5 train a CART tree by randomly selecting the subset of t samples. In step 2, m samples are randomly selected to form a training subset. In step 3, the selected training subset is used to train a CART tree. In step 4, m features are randomly selected from M features. In step 5, m features are used to maximise the growth of each tree, and the best feature of segmentation is selected. Steps 6–12 trains k *CART* decision trees. Steps 7–8 train each CART database, steps 9–10 calculate the similarity of the decision tree, remove the tree with larger similarity, and reduce the correlation that the decision tree only detects. Step 11 trains the weight of each decision tree in parallel. Step 12 weighted the results of each decision tree to get the final prediction. Steps 13–19 are the error calculation of the whole model, and the error of the model is minimised by iterative training of the training set. Steps 14–15 calculate the model error. Steps 16–17 stop training if the model error is less than the threshold. Steps 18–19, on the contrary, continuing training.

4.3 Variable selection

The data selected in this paper have certain limitations, not standard datasets, but more authenticity. The purpose of this section is to determine a small part of the variables that can predict the results well. First, grow a random forest to determine the importance of each variable. The previous Section 3 has discussed the correlation of each variable, so the unimportant variables can be ignored, and delete the proportion of the least important variables. Then, starting again, a new random forest is growing, and other variables are growing except the deleted variables. In this way, it's to delete unimportant variables every time, until the error of other variables is not reduced.

The main problem is how large the proportion of variables should be deleted. By modifying the resolution of the number of variables selected, the proportion of deleted variables is getting smaller and smaller. At the same time, to maximise the prediction accuracy, it is necessary to select as many variables as possible. Although the proportion of some variables is deleted, most variables still retain or reduce their importance. In this paper, the performance of the model is measured by the error rate. At the same time, this work gets through analysis that 20 explanatory variables have the best effect and the lowest error rate. The same also needs to weaken the correlation between variables as much as possible. Therefore, the uncertainty of residents' travel mode prediction decreases with the increase of the set.

4.4 Construction process of random forest

Regarding residents' travel mode prediction, there is a high correlation and interaction between features, which is allowed in the construction of random forests. During the construction of the decision tree, k data points are randomly selected without replacement. There is a response variable $\{y_i, x_{i1}, x_{i2}, \ldots, x_{ij}\}$, $\{x_{i1}, x_{i2}, \ldots, x_{ij}\}$ which is a variable affecting the travel mode and forms a subset. Because random data extraction can randomly assign weight to each subset, the initial prediction result is shown in equation (1).

$$p(y_i) = w_i \sum_{j=1}^{K} T(x_{ij}) \tag{1}$$

where x_{ij} represents the variable, y_i represents the prediction result of the i^{th} subset, and w_i represents the weight of the subset. $T(\ )$ is a discriminant function. For the data x_{ij}, a certain threshold is satisfied to choose which branch to go to.

For the prediction of travel mode, the amount of observation data is small, with low dimension features and high latitude features. If all features are used in the prediction model, it may lead to overfitting. In the introduction of the previous section, it can be seen that some feature variables are highly correlated and others are high-order interactive. Therefore, this paper discusses the process of variables selection, which enhances the performance of the model by using weighted random sampling. Variables containing more information are added to the decision tree to minimise variables containing less information. Unimportant variables can be ignored to make them useless or unimportant, which will be deleted in the next growth. Therefore, except for the deleted variables, other variables continue to grow and continue to select ultimately insignificant variables to delete until the error is minimised.

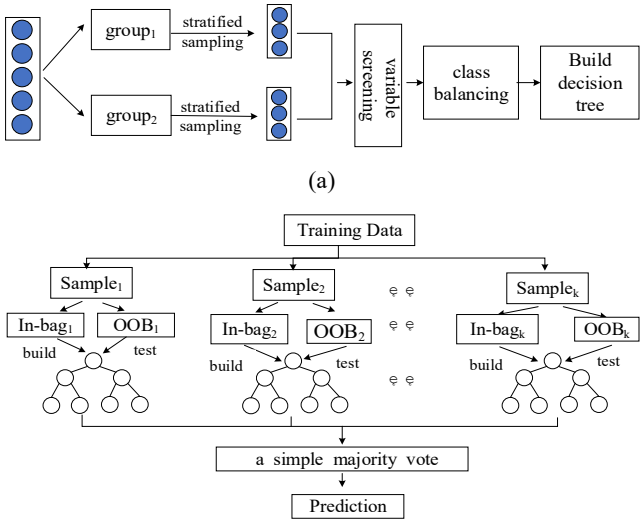
Each tree in the random forest is constructed independently using the samples of training data, which also shows that the bias of each decision tree is the same. If we want to improve the prediction ability, variation must be reduced. In this paper, a large number of irrelevant trees are constructed to reduce correlation. In addition to variable selection, the variance is reduced by reducing the correlation of trees, since the decision tree itself has noise and the bias is relatively low. It is assumed that the correlation coefficient between different trees is ρ. Then the variance of each tree is σ^2. The formula is shown in equation (2).

$$Avg(Var) = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2 \tag{2}$$

where B represents the number of trees. With the increase of B, the average value of the variance becomes smaller to reduce the variance.

The given observations are randomly divided into two samples: in-bag samples (68% of samples) and out-of-bag samples (32% of samples). Next, the classification and regression trees (CART) algorithm construct a tree on the samples in the bag. The motivation of CART is to predict the result variables of the samples outside the bag. A single decision tree is constructed by the grouping method, and multiple decision trees can form a random forest. Finally, according to the prediction results of K decision trees, the classification result with the largest number of votes is selected by voting. That is, the prediction of residents' travel mode is realised. The measurement results are shown in Figure 2.

Figure 2 Construction process from decision tree to random forest, (a) building decision tree (b) constructing random forest (see online version for colours)



Since each decision tree is independent, if the results of two decision trees are the same every time, it means that the two decision trees are the same, however, the model needs the diversity of decision trees. Therefore, it is needed to evaluate the consistency between the outputs of more than two decision trees, give a lower weight to the decision trees with high consistency, and use the kappa coefficient to express the diversity. The calculation formula is shown in equation (3).

$$Kappa = \frac{p_r(a) - p_r(e)}{1 - p_r(e)} \tag{3}$$

where $p_r(a)$ is the relative observed data consistency between the outputs of two decision trees, and $p_r(e)$ represents the hypothetical probability of accidental consistency. $p_r(a)$ is the accuracy of prediction, and $p_r(e)$ is obtained by the confusion matrix. For $p_r(a)$ formula and $p_r(e)$ formula to show in equations (4) and (5).

$$p_r(a) = \frac{\sum a_{ii}}{N} \tag{4}$$

$$p_r(e) = \frac{\sum_{j=1 \wedge j \neq i}^{m} \mathbf{a}_{ij} \cdot \mathbf{a}_{ji}}{N} \tag{5}$$

where a_{ij} represents the number of samples that are i predicted to j. Where a_{ji} represents the number of samples that are j predicted to i. N represents the total number of samples.

Each decision tree in the random forest is relatively independent, and parallel computing is implemented between them. For each input sample, each decision tree must be judged. Finally, the majority voting strategy is used to calculate the results, which can improve the convergence speed of the model by adjusting the super parameters of the

model. The main direction of this paper is to adjust the size of the terminal node, which is equivalent to adjusting the depth of the tree. In each decision tree, the path length of each terminal node is calculated, and the terminal nodes with considerable path lengths are pruned to achieve the optimal feature selection of the decision tree.

The classification tree algorithm uses the Gini coefficient instead of the information gain ratio in a random forest. Gini coefficient represents the impure of the model. The smaller the Gini coefficient, the lower the impure, and the better the feature selection. Therefore, this paper uses the Gini impurity index to determine the explanatory variables affecting travel mode selection. At node T, attribute t is used as the partition attribute to estimate the probability of belonging to different classes, expressed by $P(k|t)$, $k = 1, \ldots, Q$, Q is the number of sample types, and the definition formula of Gini index is shown in equation (6).

$$G(t) = 1 - \sum_{k=1}^{Q} p^2(k \mid t) \tag{6}$$

Gini coefficient is used to measure the uncertainty of random variables. The greater the $G(t)$ is, the higher the uncertainty of the data is. The smaller the $G(t)$ is, the lower the uncertainty of the data is. $G(t) = 0$, all samples in the dataset are the same class.

To optimise the model, out-of-pocket (OOB) samples are used to quantify the prediction performance without cross-validation. They used the prediction error rate Error OOB to measure the model performance under different parameters. The calculation method is shown in formula (7).

$$Error_{OOB} = \left\{1 - \frac{1}{N} \sum_{i \in OOB} \delta_i\right\} \times 100\% \tag{7}$$

where $Error_{OOB}$ is the prediction error rate of OOB samples and δ_i indicates the indicator variable. N is the number of observations. If $\delta_i = 1$ indicates correct prediction, otherwise $\delta_i = 0$ is incorrect. OOB data for each feature will get a rate of change, and finally can be sorted according to the rate of change to quantify the importance of features.

The random forest adopts the method of voting, and the minority obeys the majority. Each tree will vote for a category, and the category with the largest number of votes of all trees is taken as the output of the random forest. In this paper, the mean of all decision trees is taken as the output weighted average method of the random forest, where w_i is the weight of each decision tree. The calculation formula is shown in equation (8).

$$F(x) = \sum_{i=1}^{m} w_i R_i(x) \tag{8}$$

where w_i is the weight of each decision tree, which is the final weight obtained by iterative calculation. $R(x)$ represents the result of each decision tree for the dataset x_i, and finally sums it up.

5 Experiments

5.1 Dataset analysis

The advantages of the random forest prediction model proposed in this paper are illustrated by studying the travel mode selection behaviour of active groups (e.g., workers, farmers students, etc.) in Huainan City. The data collection is divided into two parts. In the first part, the network survey is conducted to obtain the travel information of residents in the family unit. In the second part, the flow of people in various regions is monitored through big data analysis to monitor the travel mode of residents. Through two parts of information, to find out the influence factors of travel mode selection.

The network survey includes two parts:

1 *Family and personal features:* Families as a whole, questionnaires data include personal information such as specific age features and all personal travel details. There is also information about family income, residential areas and so on.

2 *Travel information within 24 hours of a day:* Including departure time, destination, travel times and travel reasons.

Finally, after data clean-up, 4,391 trips were made using 1,562 people from 500 families. In addition, for the data of network population flow, the statistical analysis software is used to analyse this part of the data mainly for the collection of urban environmental data. The current land use conditions and traffic features are the main considerations.

The respondents had five main modes of transport: walking, bicycles, electric vehicle, bus and private cars. Due to the imbalance of data, the share of each category is different, 8.9%, 6.5%, 23.9%, 26.1% and 34.6% respectively. The data is giving some explanations, as shown in Table 2.

5.2 Experimental analysis

According to the analysis of big travel data, there is a correlation or even a strong correlation between variables. For example, there is a correlation between weather conditions and personal mood. For different feature variables, the importance of each feature variable is different. The higher the relative importance value, the greater the impact of feature variables in the selection of travel mode. The number of decision trees also affects the experimental effect. This paper experiments on the selection of the number k of decision trees, as shown in Figure 3.

The experiment shows that the number of decision trees in the random forest is 200, and the effect is the best. And the error rate gradually stabilises at a lower value in the number of decision trees 400. But the time also gradually increases. To realise the trade-off between prediction time and performance, 200 decision trees are selected in this paper.

Table 2 Explanatory variables for data

Variable	*Specific object*	*Description*
Family variable	Population	Mainly includes the number of family members [0, 6]
	Income	The average monthly income of households [0, 2,000], [2,000, ∞)
	Quantity	The total number of vehicles owned by families [0, 6]
	Vehicle type	Types of vehicles owned, (cars, bicycles, electric vehicles)
Personal variable	Age	Personal age [0, 15], [15, 25], [25, 60], [60,100]
	Work	Personal current working conditions (employed, retired)
	Education	Individual education level, (primary school, middle school, university)
	Accommodation	Main living areas for individuals (urban, suburban)
Subjective variable	Mood	The personal mood (bad, happy)
	Comfort	Safety and comfort of transport (bad, comfortable)
	Purpose	Travel purpose (business trip, shopping trip, entertainment trip)
Objective variable	Condition	Current traffic conditions, (smooth, congested)
	Building	Traffic construction in the urban area (lack, complete)
	Weather	Weather conditions at that time (sunny, rainy)
	Cost	Travel costs [0, 10], [10, 30], [30, ∞)
	Time	Travel time in hours [0–0.5], [0.5, 1], [1, ∞)
	Distance	Travel distance in kilometres [0, 3], [3, 10], [10, ∞)

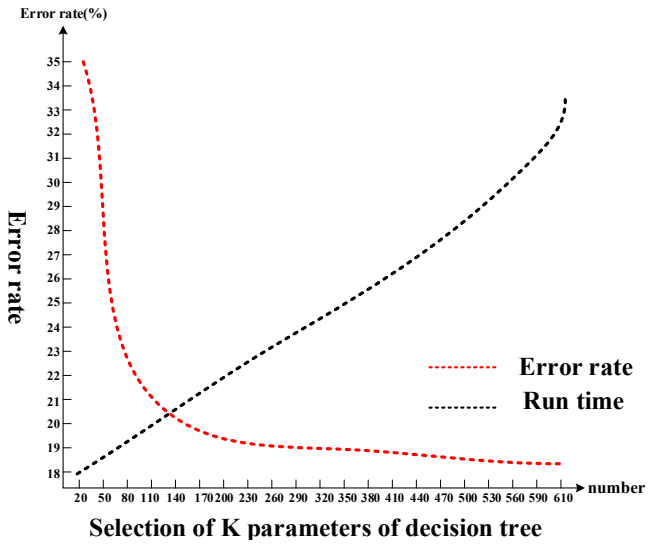
To test the effectiveness of the method, the precision, accuracy, and mean average precision (MAP) are usually selected to measure (Mercy et al., 2018). In the prediction task, accuracy is a commonly used method to measure the model. Since the class of sample data is very unbalanced, simple accuracy cannot accurately measure the prediction effect of a small number of categories. Therefore, this paper uses precision to measure the accuracy of samples and the precision of the formula as shown in equation (9).

$$Precision = \frac{TP}{TP + FP} \times 100\% \tag{9}$$

where TP denotes the samples that are predicted to be positive, FP predicts to be negative, and $TP + FP$ denotes all samples that are predicted to be positive.

The simple processing of data and variable selection are introduced in Section 3. In addition to the error of the model itself, the quality of variables and the selection process of variables also affect the accuracy of the final classification. Therefore, the results of variable selection in the dataset are compared, and the results are shown in Table 3.

Figure 3 Selection of K parameters of decision tree (see online version for colours)



The inverse probability weighting (IPW) method assigns the probability for each observation sample to be processed and weights each observation value according to its opposite probability. Due to the class imbalance of sample data, the effect of sampling actual data is usually affected. Therefore, this paper compares the IPW method with other sampling methods based on inverse sampling.

The experimental results are shown in Table 3. The inverse sampling IPW is still significantly higher than oversampling and under sampling without selection. In filtering, under-sampling leads to data loss in most classes and oversampling (Weng et al., 2020) leads to duplicate data in a few classes, resulting in a decline in inaccuracy. Experiments show that the effect of using IPW is better in variable selection.

Travel mode prediction modelling is affected by data and model uncertainty. Especially for class imbalance data, although the random forest model works better than other integrators, in general, there is a gap between the predicted results and the actual situation, as shown in Table 4.

To illustrate the effectiveness of the improved random forest model, this paper compares it with the existing classification prediction models, such as support vector machine, adaptive boosting (Shi et al., 2018), and decision tree and so on. The results are shown in Table 5. The prediction accuracy is selected for effect evaluation. It can be seen that the overall accuracy of random forest is relatively high, and the running time is short, which saves time and ensures accuracy.

Table 3 Influence of variable selection on experimental results

Travel mode	*Selection*			*No selection*		
	RF_ipw	*RF_under*	*RF_over*	*RF_ipw*	*RF_under*	*RF_over*
Walk	80.34%	52.35%	80.25%	76.45%	51.25%	81.76%
Bicycle	83.28%	63.13%	82.53%	67.43%	62.87%	65.84%
Electric vehicle	87.37%	52.89%	84.64%	75.66%	54.15%	56.78%
Bus	85.49%	70.68%	83.86%	78.46%	71.46%	74.62%
Private car	89.66%	64.73%	84.56%	68.56%	65.35%	65.16%

Table 4 Mode market share

Mode	*Actual*	*Predicted*
Walk	8.9%	6.2%
Bicycle	6.5%	4.3%
Electric vehicle	23.9%	24.7%
Bus	26.1%	27.4%
Private car	34.6%	37.4%

Table 5 Comparison of models

Travel mode	*RF*	*AdaBoost*	*SVM*	*MNL*	*Decision tree*	*Bayesian belief network*
Walk	80.34%	48.73%	79.25%	50.41%	65.49%	79.43%
Bicycle	83.28%	62.10%	80.63%	64.93%	71.83%	74.68%
Electric vehicle	87.37%	51.35%	83.27%	62.34%	70.15%	81.76%
Bus	85.49%	68.44%	82.76%	71.80%	74.53%	78.19%
Private car	89.66%	61.97%	83.68%	64.73%	77.46%	80.57%

The decision tree algorithm is easy to ignore the correlation of attributes in the dataset, and it is difficult to achieve the optimal selection of pruning. Compared with the decision tree, the accuracy of the random forest is higher, and the voting selection is more convincing. SVM and Bayesian belief network are more suitable for two classifications, and multi-classification needs to be further improved. For data imbalance, AdaBoost leads to a decline in classification accuracy and high time cost. In general, the effect of random forest is better for the prediction of travel mode.

5.3 Model specification and estimation

In general, there are errors in the prediction of observation data by a single decision tree, including the error of data and the error of the model itself, which affect the variance and deviation of the model. Therefore, considering the intrinsic correlation of travel variables, the integrated model of random forests is better. However, it is worth considering that for real data, different people are not invariable rules when making travel decisions, and each rule will be different. And because the data class used in this paper is unbalanced, it is easy to produce over-fitting, which has a certain impact on the experimental results. If the rule set with variable conditions is combined, the uncertainty of the model will be reduced, and the accuracy of prediction will increase.

In addition, using some sample-free models such as random polynomials to estimate the basic predictive value can also capture the uncertainty of the model. After estimating the predicted value, the final voting result obtained by using the random forest model is more ideal. To improve the prediction accuracy and weaken the correlation of feature variables, the correlation detection of each feature variable is considered in the pre-processing step. Although it can achieve a certain effect, it needs more time, and this part needs further improvement.

The complexity of the model constructed in this paper is close to the sum of the complexity of each decision tree in the method. Assuming that all decision trees have the same complexity, the overall complexity of the model is the number of decision trees CART multiplied by the complexity of each decision tree. If there are a total of n samples and m feature variables, the computational cost of the decision tree is $O(mn \log n)$. There are M trees in the total forest, so the complexity is $O(M(mn \log n))$. However, this is not the final complexity. Since the variable selection process is considered, the variables with small importance will be deleted, thus the actual variable feature is less than m. Similarly, each decision tree is conducted by parallel computing for reduces the correlation between decision trees. Considering the correlation of decision trees, different decision trees have different contributions, and the number of final decisive decision trees is less than M. Therefore, the model complexity proposed in this paper is less than $O(M(mn \log n))$.

6 Conclusions

This paper discusses the prediction of residents' travel mode based on random forest and combined with inverse sampling probability and variable selection. After predicting the whole dataset, it is evident that the overall effect of random forest is the best. Our contributions include the following two aspects.

1 Improve the prediction accuracy of the random forest. Improved random forest algorithm, make the model more suitable for our dataset. For the selection and construction of a single decision tree, this paper reduces the complexity by feature variable selection to delete the variables with small importance. At the same time,

for each decision tree using parallel computing, each decision tree has different weights, reducing the correlation between decision trees to improve the accuracy of the model.

2 Class imbalance and variable selection are considered for data adjustment. The data is pre-processed by statistical analysis to obtain high-quality data. At the same time, the relationship between variables is analysed to facilitate pruning and variable selection. In the establishment of a decision tree before and after the change of class imbalance data, and further reduce the impact of class imbalance on number prediction results.

The results show that residents' travel mode selection is indirectly dependent on objective variables. At the same time, it is directly dependent on subjective variables. The modification algorithm of the random forest improves the prediction accuracy of residents' travel mode and achieves good results.

In recent years, due to the continuous advancement of urbanisation, reasonable transportation planning has become increasingly important. Therefore, although the existing prediction methods have a certain degree of uncertainty and need to be improved to increase the accuracy of prediction, it also provides support for traffic planning.

On the one hand, the model described in this paper can help residents choose more convenient and efficient means of transportation and provide academic help. On the other hand, it can provide technical support for the traffic management department to provide scientific and reasonable traffic planning and decision-making.

Acknowledgements

This research work was supported in part by the University Synergy Innovation Program of Anhui Province (GXXT-2021-008).

References

Arroyo, R., Ruiz, T., Mars, L., Rasouli, S. and Timmermans, H. (2020) 'Influence of values, attitudes towards transport modes, and companions on travel behavior', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 71, pp.8–22 [online] https://doi.org/10.1016/j.trf.2020.04.002.

Arya, M. and Sastry, H.G. (2022) 'Stock indices price prediction in real time data stream using deep learning with extra-tree ensemble optimisation', *International Journal of Computational Science and Engineering*, Vol. 25, No. 2, pp.140–151.

Behr, A., Giese, M., Teguim, H.D.K. and Theune, K. (2020) 'Early prediction of university dropouts-a random forest approach', *Jahrbücher für Nationalökonomie und Statistik*, Vol. 240, No. 6, pp.743–789.

Bösehans, G. and Walker, I. (2020) 'Do supra-modal traveler types exist? A travel behaviour market segmentation using Goal framing theory', *Transportation*, Vol. 47, No. 1, pp.243–273.

Capitaine, L., Genuer, R. and Thiebaut, R. (2021) 'Random forests for high-dimensional longitudinal data', *Statistical Methods in Medical Research*, Vol. 30, No. 1, pp.166–184.

Chakrabarti, S. and Shin, E.J. (2017) 'Automobile dependence and physical inactivity: insights from the California household travel survey', *Journal of Transport and Health*, Vol. 6, pp.262–271 [online] https://doi.org/10.1016/j.jth.2017.05.002.

Chen, S.Y. (2016) 'Using the sustainable modified TAM and TPB to analyze the effects of perceived green value on loyalty to a public bike system', *Transportation Research Part A: Policy and Practice*, Vol. 88, No. 1, pp.58–72.

Cheng, L., Lai, X., Chen, X., Yang, S., De Vos, J. and Witlox, F. (2020) 'Applying an ensemble-based model to travel choice behavior in travel demand forecasting under uncertainties', *Transportation Letters The International Journal of Transportation Research*, Vol. 12, No. 6, pp.375–385.

D'Amato, V., D'Ecclesia, R. and Levantesi, S. (2022) 'ESG score prediction through random forest algorithm', *Computational Management Science*, Vol. 19, No. 2, pp.347–373.

Gao, Y., Rasouli, S., Timmermans, H. and Wang, Y. (2017) 'Understanding the relationship between travel satisfaction and subjective well-being considering the role of personality traits: a structural equation model', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 49, pp.110–123 [online] https://doi.org/10.1016/j.trf.2017.06.005.

Ghosh, D. and Cabrera, J. (2021) 'Enriched random forest for high dimensional genomic data', *IEEE-ACM Transactions on Computational Biology and Bioinformatics*, Vol. 17, No. 6, pp.1545–5963.

Guo, G. and Zhang, T. (2019) 'A residual spatio-temporal architecture for travel demand forecasting', *Transportation Research Part C-Emerging Technologies*, Vol. 115, p.102639 [online] https://doi.org/10.1016/j.trc.2020.102639.

Han, S., Kim, H. and Lee, Y.S. (2020) 'Double random forest', *Machine Learning*, Vol. 109, No. 8, pp.1569–1586.

Han, S., Williamson, B.D. and Fong, Y. (2021) 'Improving random forest predictions in small datasets from two-phase sampling designs', *BMC Medical Informatics and Decision Making*, Vol. 21, No. 1, pp.1–9.

Hoque, J.M., Erhardt, G.D., Schmitt, D., Chen, M., Chaudhary, A., Wachs, M. and Souleyrette, R.R. (2021) 'The changing accuracy of traffic forecasts', *Transportation*, Vol. 49, No. 2, pp.445–466.

Jain, V. and Phophalia, A. (2022) 'M-ary random forest – a new multidimensional partitioning approach to random forest', *Multimedia Tools and Applications*, Vol. 80, No. 28, pp.35217–35238.

Khandelwal, V., Chaturvedi, A.K. and Gupta, C.P. (2020) 'Amazon EC2 spot price prediction using regression random forests', *IEEE Transactions on Cloud Computing*, Vol. 8, No. 1, pp.59–72.

Kim, J., Bae, Y.K. and Chung, J.H. (2020) 'Modeling social distance and activity-travel decision similarity to identify influential agents in social networks and geographic space and its application to travel mode choice analysis', *Transportation Research Record*, Vol. 2674, No. 6, pp.466–479.

Li, Q.R., Li, Y.Q., Li, K., Chen, L., Zheng, Q. and Chen, K. (2020) 'Study on the influence of subjective well-being on travel mode selection', *Physics Letters A*, Vol. 384, No. 34, p.126867.

Liu, G., Ma, J., Hu, T. and Gao, X. (2022) 'A feature selection method with feature ranking using genetic programming', *Connection Science*, Vol. 34, No. 1, pp.1146–1168.

Mantero, A. and Ishwaran, H. (2021) 'Unsupervised random forests', *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Vol. 14, No. 2, pp.144–167.

Mercy, R.B.P., Manjula, D. and Sugumaran, V. (2018) 'Categorization of images using autoencoder hashing and training of intra bin classifiers for image classification and annotation', *Journal of Medical Systems*, Vol. 42, No. 7, pp.1–15.

Nagra, A.A., Han, F., Ling, Q.H., Abubaker, M., Ahmad, F., Mehta, S. and Apasiba, A.T. (2020) 'Hybrid self-inertia weight adaptive particle swarm optimisation with local search using C4.5 decision tree classifier for feature selection problems', *Connection Science*, Vol. 32, No. 1, pp.16–36.

Nguyen-Phuoc, D.Q., Currie, G., De Gruyter, C. and Young, W. (2018) 'How do public transport users adjust their travel behaviour if public transport ceases? A qualitative study', *Transportation Research Part F: Traffic Psychology and Behaviour*, Vol. 54, pp.1–14 [online] https://doi.org/10.1016/j.trf.2018.01.009.

Shaaban, K. and Maher, A. (2020) 'Using the theory of planned behavior to predict the use of upcoming public transportation service in Qatar', *Case Studies on Transport Policy*, Vol. 8, No. 2, pp.484–491.

Shi, L., Duan, Q., Dong, P., Xi, L. and Ma, X. (2018) 'Signal prediction based on boosting and decision stump', *International Journal of Computational Science and Engineering*, Vol. 16, No. 2, pp.117–122.

Slik, J. and Bhulai, S. (2020) 'Transaction-driven mobility analysis for travel mode choices', *Procedia Computer Science*, Vol. 170, pp.169–176 [online] https://doi.org/10.1016/j.procs.2020.03.022.

Wang, Q. and Chen, H. (2020) 'Optimization of parallel random forest algorithm based on distance weight', *Journal of Intelligent & Fuzzy Systems*, Vol. 39, No. 2, pp.1951–1963.

Wang, S., Minku, L.L., Chawla, N. and Yao, X. (2019) 'Learning from data streams and class imbalance', *Connection Science*, Vol. 31, No. 2, pp.103–104.

Weng, T.H., Chiu, C.C., Hsieh, M.Y., Lu, H. and Li, K.C. (2020) 'Parallelisation of practical shared sampling alpha matting with OpenMP', *International Journal of Computational Science and Engineering*, Vol. 21, No. 1, pp.105–115.

Wu, S., Liu, Y., Zou, Z. and Weng, T.H. (2022) 'S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis', *Connection Science*, Vol. 34, No. 1, pp.44–62.

Xu, H.J., Li, W.Y., Wang, T. and Yang, A.L. (2019) 'Research on dynamic prediction method for traffic demand based on trip generation analysis', *Advances in Mechanical Engineering*, Vol. 11, No. 6, pp.1–9.

Yu, L., Xie, B. and Chan, E.H. (2019) 'Exploring impacts of the built environment on transit travel: distance, time and mode choice, for urban villages in Shenzhen, China', *Transportation Research Part E: Logistics and Transportation Review*, Vol. 132, pp.57–71 [online] https://doi.org/10.1016/j.tre.2019.11.004.

Zanni, A.M. and Ryley, T.J. (2015) 'The impact of extreme weather conditions on long-distance travel behaviour', *Transportation Research Part A: Policy and Practice*, Vol. 77, pp.305–319 [online] https://doi.org/10.1016/j.tra.2015.04.025.

Zhou, F., Wu, J., Xu, Y. and Yi, C. (2020) 'Optimization scheme of tradable credits and bus departure quantity for travelers' travel mode choice guidance', *Journal of Advanced Transportation*, Vol. 2020, pp.1–8 [online] https://doi.org/10.1155/2020/6665161.