Aerospace Science and Technology 116 (2021) 106822

Contents lists available at ScienceDirect

Aerospace Science and Technology

www.elsevier.com/locate/aescte

A novel hybrid method for flight departure delay prediction using Random Forest Regression and Maximal Information Coefficient

Zhen Guo [a], Bin Yu [a,b], Mengyan Hao [a], Wensi Wang [a], Yu Jiang [c,*], Fang Zong [d,*]

[a] School of Transportation Science and Engineering, Beihang University, Beijing 100191, PR China
[b] BDBC, Beihang University, Beijing 100191, PR China
[c] College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, PR China
[d] College of Transportation, Jilin University, Changchun 130022, PR China

ARTICLE INFO

Article history:
Received 24 December 2020
Received in revised form 6 May 2021
Accepted 11 May 2021
Available online 28 May 2021
Communicated by Jérôme Morio

Keywords:
Flight departure delay
Prediction
Random Forest Regression
Maximal Information Coefficient
Transportation systems

ABSTRACT

Flight departure delay prediction is one of the most critical components of intelligent aviation systems. The accurate prediction of flight departure delays can provide passengers with reliable travel schedules and enhance the service performance of airports and airlines. This article proposes a hybrid method of Random Forest Regression and Maximal Information Coefficient (RFR-MIC) for flight departure delay prediction. Random Forest Regression and Maximal Information Coefficient are inherently fused in terms of Information Consistency. Furthermore, this article focuses on utilizing flight information on multiple air routes for flight departure delay prediction. To validate the proposed flight departure delay prediction model, a numerical study is conducted using flight data collected from Beijing Capital International Airport (PEK). The proposed RFR-MIC model exhibits good performance compared with linear regression (LR), k-nearest neighbors (k-NN), artificial neural network (ANN), and standard Random Forest Regression (RFR). The results also show that flight information on multiple air routes can certainly improve the accuracy of flight departure delay prediction.

1. Introduction

Nowadays China has been the second-largest aviation market across the world in terms of the volumes of passengers [1]. In 2018, the whole Chinese aviation industry handled 612 million passengers, a 10.9% increase from the previous year [2]. However, flight delays have been rapidly increasing in the Chinese aviation industry during recent years [3]. Flight delay has been one of the major sources of passenger complaints in China, and the civil aviation authority has determined to improve the situation [4]. Moreover, flight delay usually spreads to the lower downstream flights through the air transportation network, causing unnecessary anxiety to passengers and poor performance of airports and airlines.

Due to the tight schedules that govern the complex air transportation network architecture, the development of accurate prediction models for flight delays has become a critical issue in intelligent aviation systems [5]. The prediction of flight departure delays with high accuracy is necessary. Passengers can efficiently arrange their schedules and alleviate their anxiety if they can obtain accurate flight departure delay information. The prediction results can also provide information about the future status of a flight system in the short term for airports and airlines. The Airports and airlines can adjust their schedules in advance to effectively respond to passengers' anxiety and enhance service performance. Therefore, flight departure delay prediction is required to provide accurate delay information for passengers, airports, and airlines.

Accurate prediction of flight departure delay is very difficult due to many stochastic variables involved (e.g., weather condition, air traffic control). The deployment of the flight departure delay prediction model is a challenging task. Flight delay prediction has been the topic of several previous efforts [6–9]. From a methodological perspective, modeling approaches for flight delay prediction can be classified into classic statistical and machine learning techniques. In the past decade, various classic statistical techniques have been applied to flight delay prediction. Tu et al. [10] identified major factors that impact flight departure delay, and developed a flight departure delay prediction model taking into account seasonal trend and daily propagation pattern, to estimate delay distributions. Srivastava [11] presented a linear regression model to predict taxi-out delay for John F. Kennedy International Airport. The explanatory variables used in the model included runway queue position, runway configuration, airline, and weather, etc. Based on a log-linear regression analysis, Lordan et al. [12]

* Corresponding authors.
E-mail addresses: jiangyu07@nuaa.edu.cn (Y. Jiang), zongfang@jlu.edu.cn (F. Zong).

https://doi.org/10.1016/j.ast.2021.106822
1270-9638/

developed a prediction model for taxi time for a specific airport, using explanatory variables that can be computed before prediction. Pérez-Rodríguez et al. [13] built a logit probability model to estimate the probabilities of aircraft delay. The proposed model was evaluated by comparing its result with those of frequentist logit and symmetric Bayesian logit models. Furthermore, the authors concluded that several explanatory variables, such as the previous delay, the size of the airline, and the day of the flight, are statistically significant factors to explain the probability of delay. Rodríguez-Sanz et al. [14] presented a flight delay prediction model based on a Bayesian Network approach, providing insights into the interdependencies between explanatory variables.

Compared with classic statistical techniques, machine learning techniques have become extremely popular in recent years. A great many machine learning models have been used in fields of aerospace research and transportation system, such as k-nearest neighbors (k-NN) [15], support vector regression (SVR) [16–18], and artificial neural network (ANN) models [19–23]. Some studies, seeking to improve the prediction performance, have introduced variant neural network models [24–27]. Recently, a variety of prediction models for flight delays have been developed. Balakrishna et al. [28] built a nonparametric reinforcement learning based method to predict taxi-out delays for individual flights. Kim et al. [29] developed a two-stage approach to predict delay status for different airports. In their study, the recurrent neural network (RNN), an artificial neural network modeling the behaviors of dynamic systems, was used for daily delay status and individual flight delay prediction. At the time, the best prediction accuracy of 87.42% was achieved for the delay status classification with only two classes. To support the crew pairing optimization, Chung et al. [30] proposed a prediction approach for flight delay applying a cascade neural network (CNN). The CNN model performed better than benchmark methods such as ARIMA and neural network, and obtained regression results showing root mean squared error (*RMSE*) of more than 20 minutes. It is imperative to develop a regression model to overcome the difficulty of the prediction task and achieve a superior prediction performance. Herrema et al. [31] assessed the performance of neural networks, reinforcement learning, regression tree, and multilayer perceptron methods on predicting taxi-out time. The authors concluded that the regression tree turned out to be the most efficient among the four methods. None of these prior studies have taken full account of flight information on multiple air routes in flight delay prediction. By contrast, in this study, we introduce several novel explanatory variables, which utilize flight information on multiple air routes and consider the impact of terminal performance to improve the prediction performance.

With the rapid development of ensemble learning techniques, Random Forest Regression (RFR) has been used in flight delay prediction. Rebollo and Balakrishnam [32] focused on departure delays at a particular airport or a particular link. The authors employed RFR to predict departure delay in the future, considering temporal and spatial delay states as explanatory variables. The regression tree achieved a mean absolute error (*MAE*) of about 20 minutes with a two hours prediction horizon. The use of RFR to predict flight delays is not yet widespread; however, it can be expected to increase due to the superior performance of RFR. Compared with other machine learning methods, RFR eliminates the overfitting to training datasets because of the use of multiple decision trees [33]. RFR has been proven to be suitable for real-time applications in various fields such as transportation systems [34,35] and aerodynamics[36]. Moreover, a variety of machine learning models (linear regression, logistic regression, bagging, boosting, neural network, and RFR) were evaluated, and RFR was chosen for prediction owing to its superior performance [37,38].

Several studies have shown that weather condition is one of the principal causes of flight delay [39,40]. However, understanding the impact of weather conditions along the airway is extremely challenging owing to data collection difficulties [9,41]. Similarity has been used as the basis for prediction and applied successively in several aspects of aviation. In the literature, a concept of similar days has been developed as a basis for measuring the impact of weather on air traffic [42–45]. For example, Grabbe et al. [42] identified similar days with similar weather conditions and airport traffic for air traffic flow prediction. Gorripaty et al. [45] investigated similar days for air traffic management based on a similarity metric between days. Moreover, Brooker [46] introduced a concept of “similar incident” to make risk prediction for aviation risk assessment. The similarity degree was set based on critical features of the event. Li et al. [47] proposed a similarity-based link prediction using several similarity indexes for an aviation network. In light of the idea of similarity, we utilize delays of multiple air routes with similar airways to estimate the weather condition of predicted flight. In practice, in countries or regions with narrow airways like China, it is very common that many flights share the same or similar airways, particularly during peak periods.

In the prediction of flight departure delays, potential collinearity among explanatory variables may induce feature redundancy and negatively impact the performance of the prediction model. And explanatory variables that are not strongly relevant to flight delay may also negatively impact the performance. Due to the complexity of the delay pattern, there may exist complicated nonlinear or even nonfunctional relationships between the variables. Recently, a novel Maximal Information Coefficient approach has been confirmed to be a powerful measure for discovering functional and nonfunctional correlations between variables [48], and successfully applied in various fields [49]. To improve the accuracy of flight delay prediction in this study, we propose a novel hybrid method of Random Forest Regression and Maximal Information Coefficient. Random Forest Regression and Maximal Information Coefficient are inherently fused in terms of Information Consistency. Information Consistency is defined and proved in Section 2. The predicted flight departure delay can be provided for passengers to alleviate anxiety and can also be used for airports and airlines to improve service performance.

This article aims to make two contributions. First, a novel method for the prediction of flight departure delay is proposed, that is, a hybrid method of Random Forest Regression and Maximal Information Coefficient (RFR-MIC). RFR-MIC involves Random Forest Regression and the concept of Maximal Information Coefficient, in which the roulette method and prohibitive list for feature selection are posed to enhance the performance of RFR. And the performance of RFR-MIC is evaluated by the comparison with several prediction methods (namely, RFR, ANN, k-NN, and linear regression). The results of RFR-MIC show higher accuracies in the correlation coefficient (R), mean absolute error (MAE), and root mean squared error (RMSE). RFR-MIC makes it possible for passengers, airports, and airlines to seek more accurate predictions of flight departure delays. So far according to our limited knowledge, none of the prior studies has embedded MIC into RFR to overcome the drawback of RFR and improve the prediction accuracy. Second, this article focuses on utilizing flight information on multiple air routes for flight departure delay prediction. Compared with the flight information on a single air route, multi-dimension flight information on multiple air routes can provide more benefit (e.g., timeliness and reliability of the information), which will certainly improve prediction accuracy.

The remainder of this article is organized as follows: Section 2 presents the methodology of the Maximal Information Coefficient approach and the proposed RFR-MIC model. Section 3 introduces the explanatory variables of flight delay prediction in detail. In

Section 4, a numerical study together with results and analysis including state-of-the-art comparison are presented. Finally, Section 5 concludes the article and looks forward to future work.

2. Methodology

In this section, a novel hybrid method using Random Forest Regression (RFR) and Maximal Information Coefficient (MIC) is introduced.

2.1. Maximal Information Coefficient

Maximal Information Coefficient (MIC) [48] is a powerful approach used to measure the correlation between two variables. Most of the correlation analysis approaches face the challenge of nonlinear or nonfunctional relationships between the variables [50], but MIC can not only discover the linear and nonlinear correlations in large data sets but also explore potential nonfunctional correlations [48]. The fundamental idea of MIC is that if a certain relationship exists between two variables, a grid can be drawn on the scatterplot of the variables to partition the data and encapsulate the relationship. The MIC approach is summarized below.

MIC is calculated based on mutual information and the grid partition method. Given two discrete variables with n elements, $x = \{x_i | i = 1, \ldots, n\}$ and $y = \{y_i | i = 1, \ldots, n\}$, a finite set $D = \{(x_i, y_i) | i = 1, \ldots, n\}$ of ordered pairs can be obtained. Given a grid G, we can partition the x_i values of D into x bins and the y_i values of D into y bins. MIC is obtained according to the following equations [48,49]:

$$MI^*(D, x, y) = \max MI(D|G) \tag{1}$$

where $MI^*(D, x, y)$ denotes the maximum mutual information of D over grids G. $D|G$ represents the distribution induced by the points in D on the cells of grid G. The characteristic matrix of D is defined by the following equation.

$$M(D)_{x,y} = \frac{MI^*(D, x, y)}{\log \min\{x, y\}} \tag{2}$$

The MIC of D with grid size less than $B(n)$ is defined as:

$$MIC(D)_{x,y} = \max_{xy < B(n)} \left\{ M(D)_{x,y} \right\} \tag{3}$$

where $\omega(1) < B(n) < O\left(n^{1-\varepsilon}\right)$ and $0 < \varepsilon < 1$. In general, MIC works well in practice when $B(n) = n^{0.6}$ [48]. Thus, we also use this value in this study.

MIC is normalized into a range [0, 1]. A higher MIC value indicates a stronger correlation between the variables. The calculation procedure of MIC is illustrated in Algorithm 1.

Algorithm 1 MIC calculation procedure

Inputs: Discrete variables
Step 1. Maximum mutual information:
for each grid G **do**
  Partition scatterplot of the variables with each grid
  Compute mutual information, $0 \leq MI(D, x, y) \leq logmin\{x, y\}$
end for
Step 2. Characteristic matrix:
Generate the characteristic matrix
Step 3. MIC:
Compute MIC with limited grid size, $B(n) = n^{0.6}$
Outputs: MIC

2.2. Proposed RFR–MIC hybrid method

For flight delay prediction, understanding the definite causes of flight delay is extremely challenging. Hence it is inevitable to introduce several explanatory variables that might contribute to flight delays. Incorporating high-dimensional explanatory variables into the flight delay prediction model may negatively impact the performance of the model owing to potential collinearity among explanatory variables. And explanatory variables that are not strongly relevant to flight delay may also negatively impact the performance of the prediction model. In the standard RFR model, the best split at each node in the decision tree is selected from a random selection of features. Incidentally, the term "feature" is commonly used in RFR, namely the explanatory variable in the prediction model. For instance, the explanatory variables introduced in this study may affect the prediction of flight departure delay, which are also referred to as features in RFR. To improve the accuracy of flight delay prediction in this study, we propose a novel hybrid method of Random Forest Regression and Maximal Information Coefficient (RFR-MIC). In theory, information divergence in RFR maintains Information Consistency with mutual information in MIC. Let x and y be two variables. $p(x, y)$ is the joint probability density of variables x and y. $p(x)$ and $p(y)$ are the marginal probability densities of variables, respectively. $p(x|y)$ and $p(y|x)$ are conditional probability density of variables, respectively. Information Consistency is defined as follows.

Definition 1 *(Information Consistency).* Information divergence in Random Forest Regression is equivalent to mutual information in Maximal Information Coefficient, and the equation $IG(x, y) = MI(x, y) = \sum_x \sum_y p(x, y) log \frac{p(x,y)}{p(x)p(y)}$ holds.

Proof. According to the theory of entropy, it follows:

$$H(x|y) = -\sum_x \sum_y p(x, y) \log p(x|y) \tag{4}$$

$$H(y|x) = -\sum_x \sum_y p(x, y) \log p(y|x) \tag{5}$$

$$H(x, y) = -\sum_x \sum_y p(x, y) \log p(x, y) \tag{6}$$

$$MI(x, y) = H(x; y) = H(x, y) - H(x|y) - H(y|x) \tag{7}$$

$$H(y) = H(x; y) + H(y|x) \tag{8}$$

Substituting Equations (4)-(6) into Equation (7), it follows:

$$\begin{aligned} MI(x, y) = & -\sum_x \sum_y p(x, y) \log p(x, y) \\ & + \sum_x \sum_y p(x, y) \log p(x|y) \\ & + \sum_x \sum_y p(x, y) \log p(y|x) \\ = & \sum_x \sum_y p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)} \end{aligned} \tag{9}$$

According to the definition of information divergence, it follows:

$$IG(x, y) = H(y) - H(y|x) \tag{10}$$

On the basis of Equations (8) and (10), Equation (11) can be obtained as follows:

$$IG(x, y) = H(x; y) + H(y|x) - H(y|x) = H(x; y) \tag{11}$$

On the basis of Equations (9) and (11), Information Consistency has been proved completely. □

RFR-MIC contains two main procedures, where the first process involves utilizing the roulette method to select features for each split at each node in the decision tree, and the second process involves imposing a prohibitive list in feature selection. Algorithm 2 shows the detailed procedure of the training of the proposed RFR-MIC method.

Algorithm 2 Proposed RFR–MIC training procedure

Inputs: Training dataset, $B(n) = n^{0.6}$
Step 1. Initialization:
Data normalization
Set prohibitive list $PL := \emptyset$
Step 2. MIC analysis:
Fitness set F
Prohibitive set P_i for each explanatory variable i
Step 3. Sample selection:
Bootstrap selection for samples
Step 4. Parameter calibration:
Calibration of main parameters (n_{tree}, $m_{feature}$) in RFR–MIC model
Step 5. Decision tree generation:
while $n \leq n_{tree}$ **do**
  for each node **do**
    Step 5.1. Feature selection:
    while $m \leq m_{feature}$ **do**
      Select feature i from explanatory variables based on the roulette method
      Set $PL := PL \cup P_i$
      for each feature $j \in PL$ **do**
        Set fitness $F_j := 0$
      end for
      Set $m := m + 1$
    end while
    Step 5.2. Best split selection:
    Select the best split from the selected features based on information gain
  end for
  Set $n := n + 1$
end while
Outputs: Trained RFR–MIC model

First, distinct from the standard RFR model, we attempt to identify high fitness features for each decision tree, which may improve the performance of the flight delay prediction model. The "fitness" is numerically measured by the F-FD Correlation between the features and flight delays. Let $(X_1, X_2, \ldots, X_m)$ be a feature vector. F-FD Correlation is defined as follows.

Definition 2 *(F-FD Correlation).* A feature X_i is correlated to flight delays Y if and only if there exist any feature subset $Sub_i \subseteq X - \{X_i\}$ and the property $P(Y|Sub_i, X_i) \neq P(Y|X_i)$ holds.

F-FD Correlation is approximated by $MIC(X_i, Y)$ between features and flight delay. That is, MIC between the features and flight delay is considered to be the fitness of features, where a higher value indicates higher fitness. And features for each decision tree are selected using the roulette method, which is based on the selection probability (i.e., the fitness) of each feature. Second, we combine feature selection with F-F Correlation to eliminate collinearity among explanatory variables. F-F Correlation can also be regarded as collinearity and is defined as follows.

Definition 3 *(F-F Correlation).* A feature X_i is correlated to feature X_j if and only if there exist any feature subset $Sub_i \subseteq X - \{X_i\}$ and the property $P(X_j|Sub_i, X_i) \neq P(X_j|Sub_i)$ holds.

For any features X_i and X_j, the F-F Correlation is approximated by $MIC(X_i, X_j)$. The closer the $MIC(X_i, X_j)$ value is to 1, the stronger the F-F Correlation between features is. Prohibitive set P_i for each explanatory variable i is extracted in terms of high $MIC(X_i, X_j)$ value with other explanatory variables. Whenever a high fitness feature i (i.e., explanatory variable i) is selected by roulette method in feature selection, the prohibitive list will be updated by uniting with prohibitive set P_i. And the fitness of features in the prohibitive list will be set as zero. Fig. 1 illustrates a simple feature selection procedure in RFR–MIC where two features are selected from all features.

In addition, the RFR-MIC model requires two main parameters: the number of decision trees in the forest (n_{tree}) and the number of features considered for each split at each node in the decision tree ($m_{feature}$). We use grid search in the numerical study to figure out the best value of parameters.

3. Explanatory variables

The objective of this study is to predict the departure delay of individual flights at a particular airport, for different prediction horizons. To provide an accurate flight departure delay prediction, it is indispensable to determine the appropriate explanatory variables. Rebollo and Balakrishnan [32] used a combination of categorical and continuous explanatory variables in their proposed delay prediction models. Yu et al. [9] summarized the variables used in the literature and introduced a set of micro variables. Furthermore, they also identified the explanatory variables that are of high importance for prediction accuracy. The existing explanatory variables considered in this study are summarized in Table 1.

A summary of brief explanations of existing explanatory variables is listed in Table 2. The used value of time interval is set as 60 minutes for both the crowdedness degree of the airport in the number of flights and the crowdedness degree of the airport in the number of passengers, referring to Ref. [9].

However, in these studies, they mainly focused on flight information on a single air route and neglected the impact of terminal performance on the flight delay. We introduce several novel explanatory variables, which utilize flight information on multiple air routes and consider the impact of terminal performance.

3.1. On-time performance of the terminal

The terminal is a critical component of an airport, connecting ground transportation with the airport apron. Within the terminal, passengers check in, go through security and wait for boarding. The on-time performance of the terminal tends to have a potential impact on the flight delay. For instance, owing to the sharing of temporal and spatial resources, flights from the same terminal sometimes suffer extensive delays or cancellations simultaneously. Flights from terminals with worse on-time performance have a higher possibility of being delayed or canceled. Moreover, the on-time performance of the terminal changes with time. Therefore, we utilize the average delay of flights per time interval on multiple air routes to represent the on-time performance of a terminal.

Flight cancellation is considered a flight delay in this study. Xiong and Hansen [51] suggested that cancellation should be taken into consideration as a flight delay metric for assessing the performance of aviation systems. The flight delay in this study is defined as shown in Equation (12).

$$d_f = \begin{cases} t_f^a - t_f^s & \text{if flight } f \text{ is not canceled} \\ h & \text{if flight } f \text{ is canceled} \end{cases} \tag{12}$$

where d_f is the delay of flight f; t_f^a and t_f^s represent the actual and scheduled departure time of flight f. h is the equivalent delay time of cancellation. According to the regulations of the Federal Aviation Administration (FAA), the European Aviation Safety Agency (EASA) and the Civil Aviation Administration of

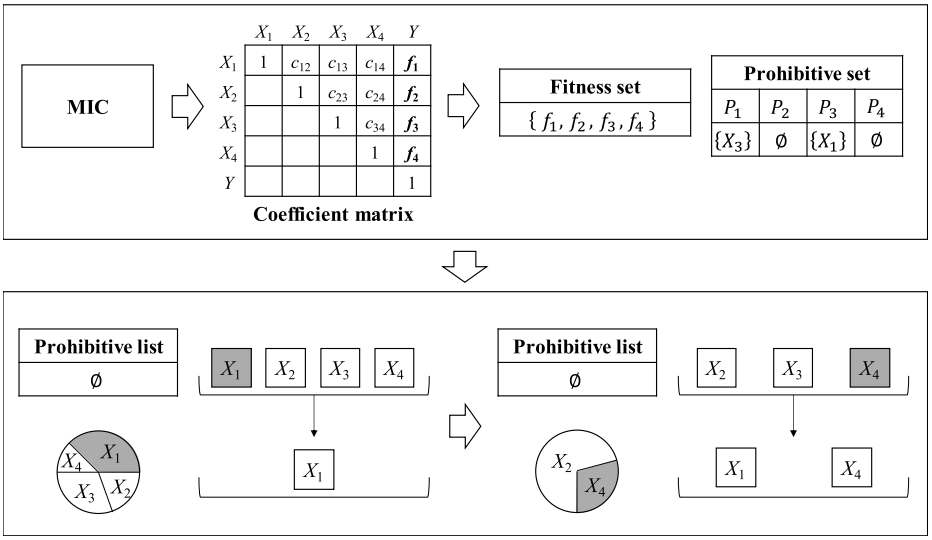

Fig. 1. A simple illustration of the feature selection procedure in RFR–MIC.

Table 1
Summary of existing explanatory variables.

Explanatory variable	Type of variable		Flight information	
	Categorical	Continuous	Single air route	Multiple air routes
Airline size	✓		✓	
Actual aircraft capacity		✓	✓	
Scheduled turnaround time		✓	✓	
Minimum turnaround time		✓	✓	
Origin or pass-by flight	✓		✓	
Delay of the previous flight		✓	✓	
Crowdedness degree of the airport in number of flights		✓		✓
Crowdedness degree of the airport in number of passengers		✓		✓

Table 2
List of explanations of existing explanatory variables.

Explanatory variable	Brief explanation
Airline size	A variable representing whether the flight belongs to a major airline
Actual aircraft capacity	The actual passenger capacity of the aircraft
Scheduled turnaround time	A planned time of turnaround operations for the flight
Minimum turnaround time	A minimum time of turnaround operations formulated according to the rated aircraft capacity
Origin or pass-by flight	A variable indicating whether the flight departs from the airport or passes by the airport
Delay of the previous flight	Departure delay of the previous flight
Crowdedness degree of the airport in number of flights	An index reflecting the number of flights per time interval
Crowdedness degree of the airport in number of passengers	An index reflecting the number of the passenger flows per time interval

China (CAAC), the equivalent delay time h is equal to 180 minutes.

For the predicted flight f, the on-time performance of terminal (P_f^t) is calculated by Equation (13).

$$P_f^t = \frac{\sum_{k \in F_k} d_k}{|F_k|} \tag{13}$$

where F_k represents the set of departure flights from the same terminal as flight f during $\left(t_f^s - \tau - \Delta t, t_f^s - \tau\right)$; t_f^s, τ, Δt and d_k represent scheduled departure time, prediction horizon, time interval, and delay of flight k, respectively. The time interval Δt is equal to 60 minutes, as suggested by Ref. [52] on the on-time performance of aviation.

3.2. Average delay of multiple air routes with similar airways

The aviation system does not immediately recover from high delay situations, e.g., a day with severe weather. Weather conditions along the airway tend to have a long-term impact on the flight delay. The airway is defined airspace with a specified width, connecting one location to another, along which aircraft may be flown. Thus, we also introduce an average daily delay to estimate weather conditions. D_f is the average daily delay of all preceding flights of multiple air routes with similar airways to flight f. For

the predicted flight f, D_f can be obtained by the following equations.

$$D_f = \sum_{g \in F_g} \frac{c_g^s}{\Gamma} \times d_g \tag{14}$$

$$\Gamma = \sum_{g \in F_g} c_g^s \tag{15}$$

where F_g represents the set of departure flights of multiple routes with similar airways to flight f during $\left(0, t_f^s - \tau\right)$; c_g^s and d_g represents similarity coefficient and delay of flight g. The similarity coefficient (c_g^s) is set by the similarity degree of airways between specific flights [9]. Γ denotes the sum of the similarity coefficient of each preceding flight.

In addition, air traffic control (*ATC*) is also identified as one of the key influential factors of flight delay. Due to the difficulty in measuring air traffic control (*ATC*) directly, we utilize a real-time expected waiting time (*EWT*) for each flight issued by the air traffic control tower of airports in China as an alternative measure of *ATC*. The expected waiting time is issued primarily according to flow rate control, weather condition, traffic congestion, and major events. In fact, passengers can obtain an up-to-date expected waiting time for each flight in mobile applications authorized by airports and airlines. In general, the closer expected waiting time contributes to a more accurate value. Owing to the advanced intelligent aviation system, we can access the collected expected waiting time issued by the control tower and utilize the time-dependent expected waiting time as the measure of *ATC* for different prediction horizons.

4. Numerical study

4.1. Data description and processing

Our numerical study uses departure flight data collected from Beijing Capital International Airport (PEK). PEK is a hub airport in China including three terminals, with approximately 1300 domestic and 300 international flights on a typical day. The dataset contains departure flight records covering the entire day (00:00–24:00) from January 1, 2017 to March 15, 2018. The dataset has a total of 528,471 flight records from 49 airlines. In this study, we select flights from PEK to HGH (Hangzhou International Airport), FOC (Fuzhou Changle International Airport), and YIW (Yiwu Airport) to test the model for flight departure delay prediction. The flight departure delay is predicted for a 2-h (i.e., two hours) prediction horizon, in other words, we provide a prediction of flight departure delay two hours before the scheduled departure time.

Standard data preprocessing includes data cleaning, data handling, data normalization, and feature extraction. Data cleaning and handling are necessary for any prediction model. The dataset is cleaned by removing errors, and the attributes of the dataset are handled into input features of the prediction model. For instance, the input feature "average delay of multiple air routes with similar airways" is calculated by Equation (14). Table 3 lists the top five routes with the highest similarity coefficient (SC). To avoid a magnitude difference between input features, we need to normalize the features using standard normalization. The normalized features can be obtained by Equation (16):

$$x^* = \frac{x - \mu}{\sigma} \tag{16}$$

where x^* is a normalized feature value; x, μ, and σ are the feature value, the mean value, and the standard deviation, respectively. Feature extraction is applied to the standard RFR model. To be specific, we use two iterations to improve the standard RFR model. In iteration 1, we use all features to train the RFR model and extract the top features which contribute 95% of importance to the model. The extracted features are used to train the RFR model again in iteration 2.

Table 3
Similarity coefficient of top five air routes.

PEK-HGH		PEK-FOC		PEK-YIW	
Air route	*SC*	Air route	*SC*	Air route	*SC*
PEK-HGH	1.00	PEK-FOC	1.00	PEK-YIW	1.00
PEK-YIW	0.93	PEK-JJN	0.91	PEK-HGH	0.93
PEK-NKG	0.82	PEK-TXN	0.84	PEK-CZX	0.78
PEK-CZX	0.81	PEK-NKG	0.75	PEK-NKG	0.77
PEK-HYN	0.75	PEK-YIW	0.74	PEK-FOC	0.74

Eventually, the resultant clean dataset is divided into training sets and test sets. We utilize the data of the first 21 days per month as the training set and the remaining days as the test set. Regarding parameter tuning, this study adopts grid search and 10-fold cross-validation fold in the training.

4.2. Performance indicators

In this study, the performance of the flight delay prediction model is evaluated in three indicators: the correlation coefficient (R), mean absolute error (MAE) and root mean square error (RMSE). The three indicators are calculated as Equations (17)–(19):

$$R = \frac{\sum_{i=1}^{n} \left(y_i - \overline{y}\right) \times \left(f_i - \overline{f}\right)}{\sqrt{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 \times \sum_{i=1}^{n} \left(f_i - \overline{f}\right)^2}} \tag{17}$$

$$MAE = \frac{1}{n} \times \sum_{i=1}^{n} |y_i - f_i| \tag{18}$$

$$RMSE = \sqrt{\frac{1}{n} \times \sum_{i=1}^{n} (y_i - f_i)^2} \tag{19}$$

where n is the number of samples. f_i is the predicted value, y_i is the actual value. $\overline{f}$ and $\overline{y}$ are the mean of predicted and actual value. The correlation coefficient R is used to measure the degree of linear correlation between the predicted value and the actual value. In addition, MAE and RMSE are the additional indicators to examine the performance of prediction models. MAE is the evaluation indicator for prediction models. RMSE is usually used to compare the stability of prediction models. Generally, the higher value of R and the lower value of MAE and RMSE indicate the higher accuracy of prediction models.

4.3. Results and analysis

For the sake of simplicity, the notations of variables are shown in Table 4. To train the proposed RFR-MIC model, the fitness set and prohibitive sets should be identified first. The fitness set F is obtained according to the MIC between the explanatory variables and flight delay. And prohibitive set P_i for each explanatory variable is extracted in terms of high collinearity with other explanatory variables. In general, a MIC higher than 0.8 indicates high collinearity between explanatory variables. Fig. 2 depicts the coefficient matrix. Thus, fitness set F is {0.021, 0.031, 0.092, 0.055, 0.025, 0.28, 0.082, 0.063, 0.37, 0.76, 0.35}. And prohibitive sets can be attained and summarized in Table 5.

As previously mentioned, there are two main parameters while using the RFR-MIC model: the number of decision trees in the forest (n_{tree}) and the number of features considered for each split at

Table 4
Notations of variables.

Notation	Variable
X_1	Airline size
X_2	Actual aircraft capacity
X_3	Scheduled turnaround time
X_4	Minimum turnaround time
X_5	Origin or pass-by flight
X_6	Delay of the previous flight
X_7	Crowdedness degree of the airport in number of flights
X_8	Crowdedness degree of the airport in number of passengers
X_9	On-time performance of the terminal
X_{10}	Expected waiting time
X_{11}	Average delay of multiple air routes with similar airways
Y	Flight departure delay

Table 5
Prohibitive sets.

P_1	Ø	P_7	$\{X_8\}$
P_2	Ø	P_8	$\{X_7\}$
P_3	Ø	P_9	Ø
P_4	Ø	P_{10}	Ø
P_5	Ø	P_{11}	Ø
P_6	Ø		

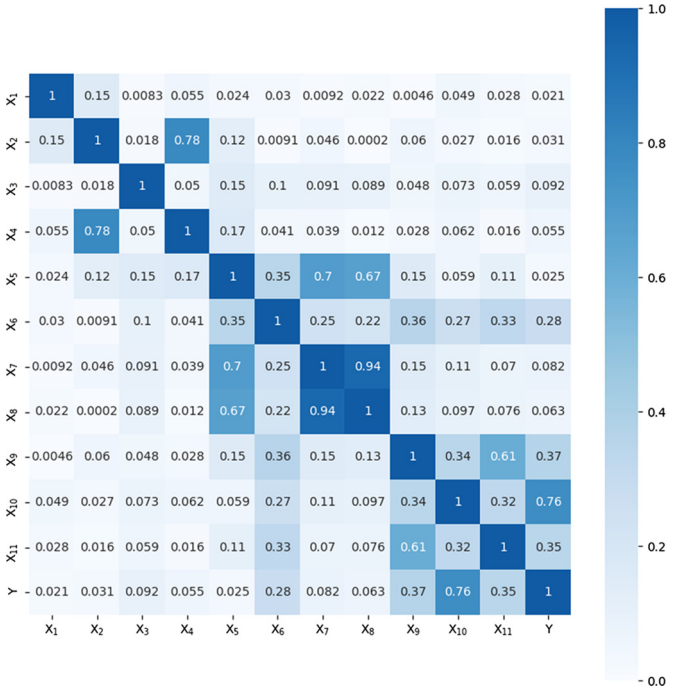

Fig. 2. Coefficient matrix. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

each node in the decision tree ($m_{feature}$). The default value of n_{tree} is 500. However, a higher value of n_{tree} can contribute to a more stable result [53]. The default value of $m_{feature}$ is one-third of the number of total features [33]. However, as the performance of RFR can be sensitive to $m_{feature}$ [54], we use grid search to figure out the best value of parameters for the flight departure delay prediction. For the RFR-MIC model, two parameters (n_{tree}, $m_{feature}$) are set as (600, 4) using grid search. The value of $m_{feature}$ is consistent with the inventors recommend D/3 (D is the number of total features) for regression problems.

This study is conducted on a computer with an Intel Core i7 CPU (3.2 GHz) and 16 GB memory. To compare the performance of the proposed RFR-MIC method, flight departure delay is also predicted by four benchmark methods, namely standard Random

Table 6
Comparison of the proposed and benchmark methods.

	RFR-MIC	RFR	ANN	k-NN	LR
R	**0.91**	0.87	0.83	0.85	0.74
MAE (min)	**8.73**	10.41	11.58	12.25	13.85
RMSE (min)	**17.08**	21.78	23.76	21.97	28.05

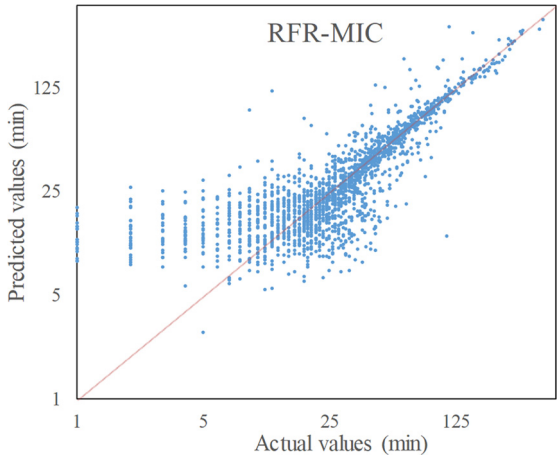

Fig. 3. Predictability of proposed RFR-MIC method on flight departure delay.

Forest Regression (RFR), artificial neural network (ANN), k-nearest neighbors (k-NN), and linear regression (LR) methods. We extract the first 6 features which contribute 95% of importance to the RFR model. The extracted features ($X_6 - X_{11}$) are used to train the standard RFR model. The best parameters (n_{tree}, $m_{feature}$) for the standard RFR model are (1000, 6). The value of R, MAE, and RMSE of the five methods are summarized in Table 6. It can be seen that the RFR-MIC method has the best prediction performance compared with the four other methods.

Fig. 3 shows that using the RFR-MIC method, 97.1% of the predicted values are within 30 minutes' deviation from the actual values, implying acceptable prediction performance. Please note that the axes are in logarithmic scales. Comparing Fig. 3 and Fig. 4 (4a, 4b, 4c, and 4d), it can be observed that, compared with the other four benchmarks, the prediction results of the RFR-MIC method are much closer to the actual values. In summary, the performance of the RFR-MIC method for flight departure delay prediction is better than that of the four other methods.

The prediction for different terminals is also observed to be stable, owing to the small gap between the three terminals' MAE values (MAE_2-$MAE_1 = 0.57$ min is the largest gap). As Fig. 5 shows, no strong correlation between MAE and the number of flights is detected, which implies that RFR-MIC is robust for different terminals regardless of their data sample size. The numerical study for a 3-h (i.e., three hours) prediction horizon is also evaluated. Fig. 6 shows the MAE comparison for 2-h and 3-h prediction horizons. When the 3-h prediction horizon is used for flight departure delay prediction, the performance of the proposed RFR-MIC is still better than those of the comparative methods, even if the error increases. According to the prediction results, the proposed RFR-MIC is effective for the prediction of departure delay.

Four scenarios with different input explanatory variables are set to validate the novel explanatory variables introduced in this study. The input explanatory variables and results of the four scenarios are listed in Table 7. S0 is the basic scenario, and only existing explanatory variables are applied to the proposed RFR-MIC model. The results show that these novel explanatory variables, which utilize flight information on multiple air routes and consider the impact of terminal performance, are indeed of high importance for flight departure delay prediction accuracy.

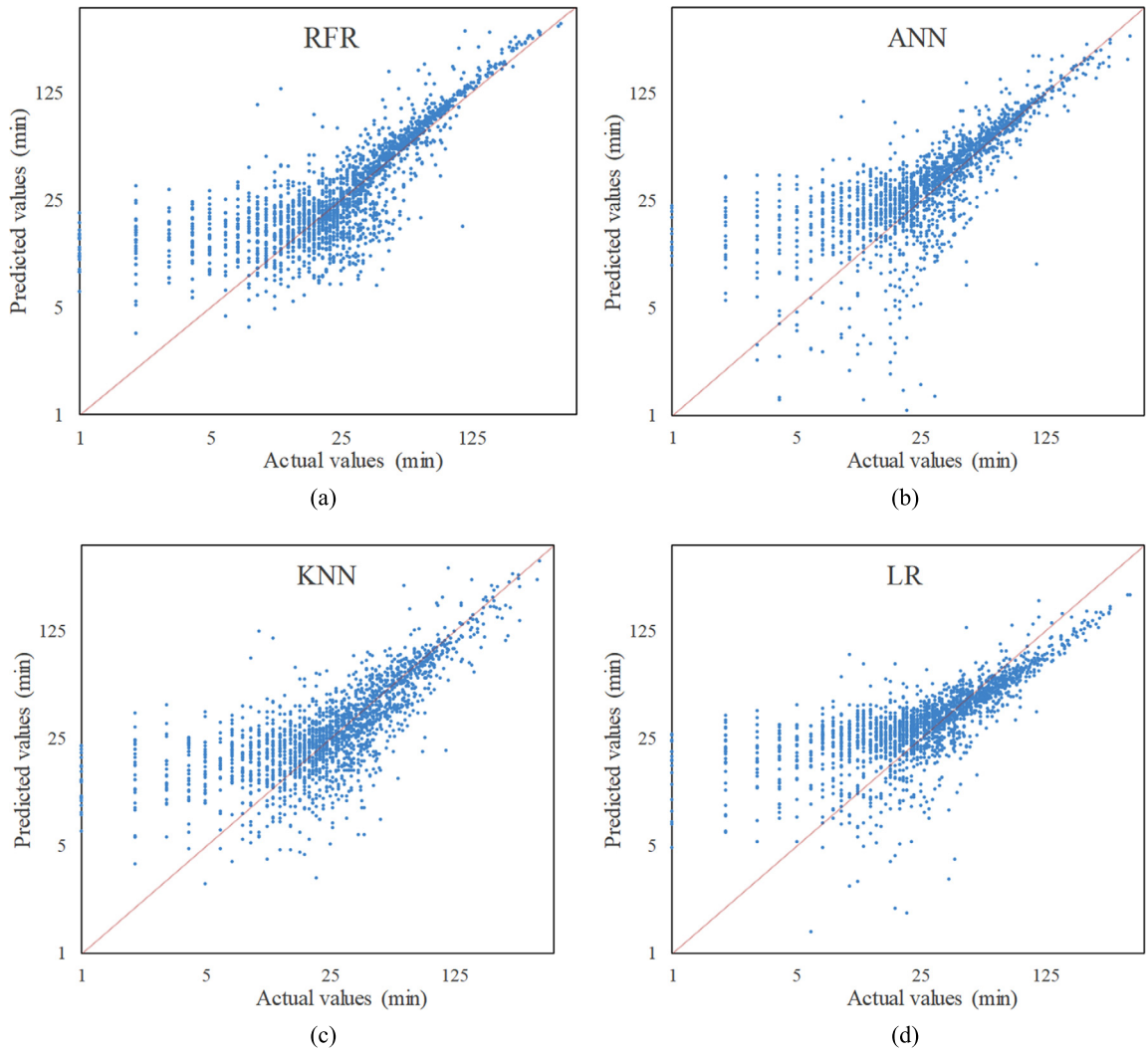

Fig. 4. Predictability of four benchmark methods on flight departure delay. (a) RFR method; (b) ANN method; (c) k-NN method; (d) LR method.

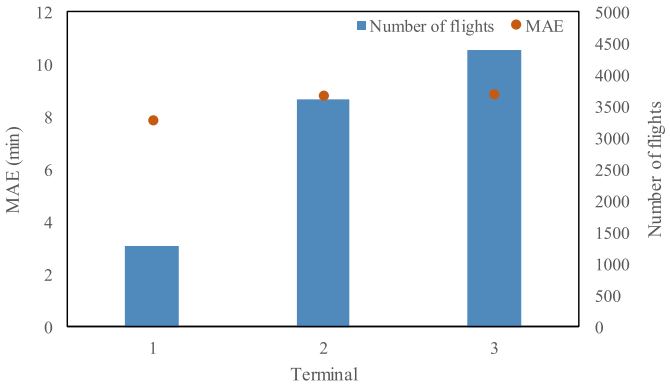

Fig. 5. MAE and number of flights per terminal.

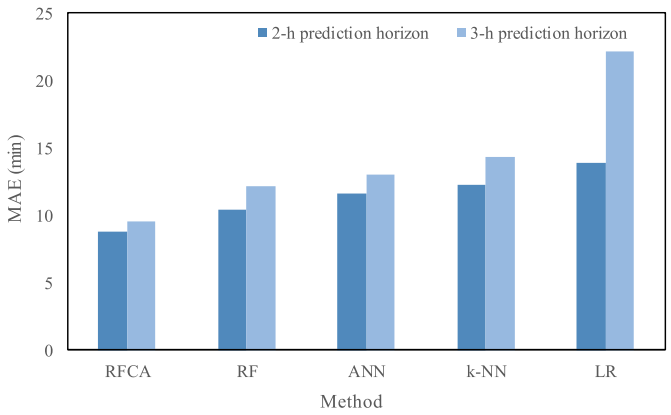

Fig. 6. MAE comparison for 2-h and 3-h prediction horizons.

Table 7
Performance comparison of different scenarios.

Scenario	Input explanatory variable			R	MAE (min)	RMSE (min)
	Existing variables	P_f^t	D_f			
S0	✓			0.84	10.25	22.73
S1	✓	✓		0.86	9.89	21.72
S2	✓		✓	0.87	9.74	20.97
S3	✓	✓	✓	**0.91**	**8.73**	**17.08**

A roulette method is utilized in the feature selection for each decision tree in this study. The roulette method is a random selection method where the probability for the selection of each feature is proportional to its fitness. To analyze the roulette method and the fitness set, we compare the results from the roulette selection method based on MIC fitness (RS-MF) with the results from the roulette selection method based on uniform fitness (RS-UF). As

Table 8
Comparison of the RS-MF and the RS-UF.

	R	MAE (min)	RMSE (min)
RS-MF	**0.91**	**8.73**	**17.08**
RS-UF	0.83	11.91	22.97

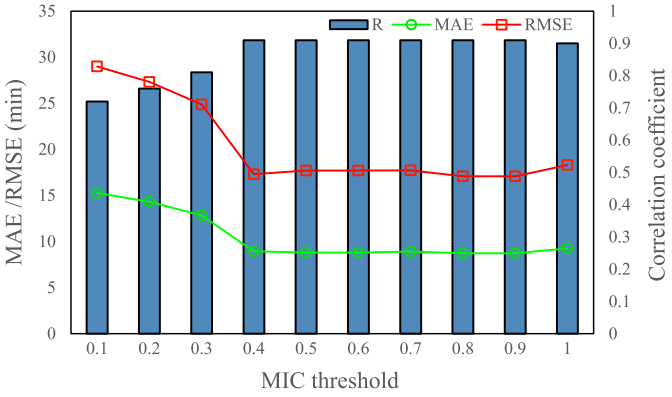

Fig. 7. Performance indicators comparison for different MIC thresholds.

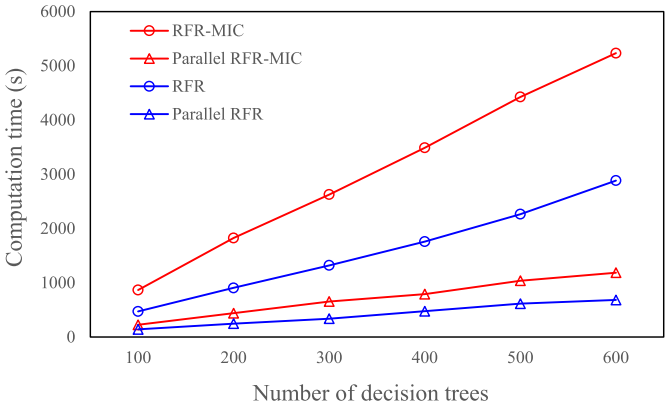

Fig. 8. Computing time comparison for the different number of decision trees.

shown in Table 8, the RS-MF significantly outperforms the RS-UF in terms of prediction accuracy.

There is no empirical MIC threshold that works for all numerical studies. Thus, we conduct a parameter analysis of the threshold for MIC. Fig. 7 illustrates the performance indicators variation of the RFR-MIC model when the threshold for MIC changes in the range of 0.1 to 1. We can find that the RFR-MIC model performs best under the MIC threshold of 0.8 and 0.9 with the same prohibitive sets. In addition, the RFR-MIC model shows good performance when the MIC threshold falls within the range of 0.4 to 0.9.

Existing third-party libraries cannot implement the hybrid of RFR and MIC. Therefore, we code the proposed RFR-MIC and the standard RFR in Python. A parallel technique is applied to accelerate the computing time. Due to the independence of decision trees, more processing can be used to parallelize the training of decision trees. Fig. 8 shows a computing time comparison of the RFR-MIC and the standard RFR. We can find that the hybrid of RFR and MIC results in more computing time for the prediction model compared to the standard RFR model. The introduction of the roulette method and prohibitive list increases the computing time. With a parallelization of decision trees based on six processing, the computing time of the RFR-MIC model decreases significantly and is within 20 minutes. In practice, the computing time is sufficient for the model training. Once be trained, the RFR-MIC model can be used to provide real-time predictions in seconds.

5. Conclusions

In an intelligent aviation system, it is vital to predict the flight departure delay in an accurate and timely manner. Flight departure delay prediction can provide passengers with reliable travel schedules to alleviate anxiety and provide more proactive operation strategies for airports and airlines to enhance service performance. This article focuses on utilizing flight information on multiple air routes for flight departure delay prediction. We introduce several novel explanatory variables, which utilize flight information on multiple air routes and consider the impact of terminal performance. In this article, we propose the RFR-MIC method to overcome the drawback of RFR and enhance the performance. To be noted, it is the first attempt to combine RFR and MIC according to our limited knowledge. The model eliminates potential collinearity among explanatory variables by imposing a prohibitive list in feature selection. The roulette method is used to select features for each decision tree, which can eliminate the negative impacts of irrelevant and noisy explanatory variables. In the numerical study using real-world data, the departure flight data collected from Beijing Capital International Airport were used to evaluate the performance of the proposed RFR-MIC model. Flights from PEK to HGH, FOC, and YIW were taken to test the model, and the prediction horizon was set to two hours. Additionally, compared with the other benchmark methods, the RFR-MIC model performs better performance than the RFR, ANN, k-NN, and LR models. Overall, despite the sophisticated air transportation conditions in the empirical study, the proposed RFR-MIC model achieves a good performance in flight departure delay prediction with a high correlation coefficient, low MAE, and RMSE. The results also show that the novel explanatory variables introduced in this study, are indeed of high importance for flight departure delay prediction accuracy.

While our study contributes to a better prediction of flight departure delay, there are some limitations in this study, which present fruitful areas for further research. For departure flights, arrival airport information tends to have a potential impact on the flight delays. Owing to a lack of reliable relevant data, arrival airport information is not used to predict flight departure delays. To fill the gap, further research can be conducted to enhance the performance of the proposed model by using both the arrival and departure airports information. In addition, further research can be supported by parallel computing and cloud computing to reduce computation time and provide a real-time prediction.

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.

Acknowledgements

This research was supported by National Natural Science Foundation of China U1811463. This work was partially funded by Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University.

References

[1] H. Jiang, Y. Zhang, An investigation of service quality, customer satisfaction and loyalty in China's airline market, J. Air Transp. Manag. 57 (2016) 80–88, https://doi.org/10.1016/j.jairtraman.2016.07.008.
[2] Civil Aviation Administration of China, The Communique of China's Civil Aviation Development and Statistics in 2018, 2019, Beijing.
[3] M.G. Tsionas, Z. Chen, P. Wanke, A structural vector autoregressive model of technical efficiency and delays with an application to Chinese airlines, Transp. Res., Part A, Policy Pract. 101 (2017) 1–10, https://doi.org/10.1016/j.tra.2017.05.003.

[4] I. Vlachos, Z. Lin, Drivers of airline loyalty: evidence from the business travelers in China, Transp. Res., Part E, Logist. Transp. Rev. 71 (2014) 1–17, https://doi.org/10.1016/j.tre.2014.07.011.
[5] P. Goedeking, Networks in Aviation: Strategies and Structures, Springer Science & Business Media, 2010.
[6] E. Mueller, G. Chatterji, Analysis of aircraft arrival and departure delay characteristics, in: AIAA's Aircr. Technol. Integr. Oper. 2002 Tech. Forum, 2002, p. 5866.
[7] R. Wesonga, F. Nabugoomu, P. Jehopio, Parameterized framework for the analysis of probabilities of aircraft delay at an airport, J. Air Transp. Manag. 23 (2012) 1–4, https://doi.org/10.1016/j.jairtraman.2012.02.001.
[8] T. Diana, Can machines learn how to forecast taxi-out time? A comparison of predictive models applied to the case of Seattle/Tacoma International Airport, Transp. Res., Part E, Logist. Transp. Rev. 119 (2018) 149–164, https://doi.org/10.1016/j.tre.2018.10.003.
[9] B. Yu, Z. Guo, S. Asian, H. Wang, G. Chen, Flight delay prediction for commercial air transport: a deep learning approach, Transp. Res., Part E, Logist. Transp. Rev. 125 (2019) 203–221, https://doi.org/10.1016/j.tre.2019.03.013.
[10] Y. Tu, M.O. Ball, W.S. Jank, Estimating flight departure delay distributions - a statistical approach with long-term trend and short-term pattern, J. Air Transp. Manag. 103 (2008) 112–125, https://doi.org/10.1198/016214507000000257.
[11] A. Srivastava, Improving departure taxi time predictions using ASDE-X surveillance data, in: 2011 IEEE/AIAA 30th Digit. Avion. Syst. Conf., IEEE, New York, 2011.
[12] O. Lordan, J.M. Sallan, M. Valenzuela-Arroyo, Forecasting of taxi times: the case of Barcelona-El Prat airport, J. Air Transp. Manag. 56 (2016) 118–122, https://doi.org/10.1016/j.jairtraman.2016.04.015.
[13] J.V. Perez-Rodriguez, J.M. Perez-Sanchez, E. Gomez-Deniz, Modelling the asymmetric probabilistic delay of aircraft arrival, J. Air Transp. Manag. 62 (2017) 90–98, https://doi.org/10.1016/j.jairtraman.2017.03.001.
[14] A. Rodriguez-Sanza, F. Gomez Comendador, R. Arnaldo Valdes, J. Perez-Castan, R. Barragan Montes, S. Camara Serrano, Assessment of airport arrival congestion and delay: prediction and reliability, Transp. Res., Part C, Emerg. Technol. 98 (2019) 255–283, https://doi.org/10.1016/j.trc.2018.11.015.
[15] Z.-X. Guo, P.-L. Shui, Anomaly based sea-surface small target detection using k-nearest neighbor classification, IEEE Trans. Aerosp. Electron. Syst. 56 (2020) 4947–4964, https://doi.org/10.1109/TAES.2020.3011868.
[16] C. Yan, Z. Yin, X. Shen, D. Mi, F. Guo, D. Long, Surrogate-based optimization with improved support vector regression for non-circular vent hole on aero-engine turbine disk, Aerosp. Sci. Technol. 96 (2020) 105332, https://doi.org/10.1016/j.ast.2019.105332.
[17] P.-P. Xi, Y.-P. Zhao, P.-X. Wang, Z.-Q. Li, Y.-T. Pan, F.-Q. Song, Least squares support vector machine for class imbalance learning and their applications to fault detection of aircraft engine, Aerosp. Sci. Technol. 84 (2019) 56–74, https://doi.org/10.1016/j.ast.2018.08.042.
[18] L.-H. Ren, Z.-F. Ye, Y.-P. Zhao, A modeling method for aero-engine by combining stochastic gradient descent with support vector regression, Aerosp. Sci. Technol. 99 (2020) 105775, https://doi.org/10.1016/j.ast.2020.105775.
[19] A.N. Haq, G. Kannan, Effect of forecasting on the multi-echelon distribution inventory supply chain cost using neural network, genetic algorithm and particle swarm optimisation, Int. J. Serv. Oper. Inform. 1 (2006) 1–22.
[20] M. Zhang, X. Zhou, Y. Zhang, L. Sun, M. Dun, W. Du, X. Cao, Propagation index on airport delays, Transp. Res. Rec. 2673 (2019) 536–543, https://doi.org/10.1177/0361198119844240.
[21] M.G. De Giorgi, M. Quarta, Hybrid MultiGene Genetic Programming - artificial neural networks approach for dynamic performance prediction of an aero-engine, Aerosp. Sci. Technol. 103 (2020) 105902, https://doi.org/10.1016/j.ast.2020.105902.
[22] A. Boutemedjet, M. Samardžić, L. Rebhi, Z. Rajić, T. Mouada, UAV aerodynamic design involving genetic algorithm and artificial neural network for wing preliminary computation, Aerosp. Sci. Technol. 84 (2019) 464–483, https://doi.org/10.1016/j.ast.2018.09.043.
[23] M.V. Mousavi, H. Khoramishad, The effect of hybridization on high-velocity impact response of carbon fiber-reinforced polymer composites using finite element modeling, Taguchi method and artificial neural network, Aerosp. Sci. Technol. 94 (2019) 105393, https://doi.org/10.1016/j.ast.2019.105393.
[24] S. Gao, J. Liu, Adaptive neural network vibration control of a flexible aircraft wing system with input signal quantization, Aerosp. Sci. Technol. 96 (2020) 105593, https://doi.org/10.1016/j.ast.2019.105593.
[25] H. Sheng, Q. Chen, J. Li, Z. Li, Z. Wang, T. Zhang, Robust adaptive backstepping active control of compressor surge based on wavelet neural network, Aerosp. Sci. Technol. 106 (2020) 106139, https://doi.org/10.1016/j.ast.2020.106139.
[26] M. Zhou, X. Qu, X. Li, A recurrent neural network based microscopic car following model to predict traffic oscillation, Transp. Res., Part C, Emerg. Technol. 84 (2017) 245–264, https://doi.org/10.1016/j.trc.2017.08.027.
[27] Z. Xu, T. Wei, S. Easa, X. Zhao, X. Qu, Modeling relationship between truck fuel consumption and driving behavior using data from Internet of vehicles, Comput.-Aided Civ. Infrastruct. Eng. 33 (2018) 209–219, https://doi.org/10.1111/mice.12344.
[28] P. Balakrishna, R. Ganesan, L. Sherry, Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: a case-study of Tampa Bay departures, Transp. Res., Part C, Emerg. Technol. 18 (2010) 950–962, https://doi.org/10.1016/j.trc.2010.03.003.
[29] Y.J. Kim, S. Choi, S. Briceno, D. Mavris, A deep learning approach to flight delay prediction, in: 2016 IEEE/AIAA 35th Digit. Avion. Syst. Conf., IEEE, New York, 2016.
[30] S.H. Chung, H.L. Ma, H.K. Chan, Cascading delay risk of airline workforce deployments with crew pairing and schedule optimization, Risk Anal. 37 (2017) 1443–1458, https://doi.org/10.1111/risa.12746.
[31] F. Herrema, R. Curran, H. Visser, D. Huet, R. Lacote, Taxi-out time prediction model at Charles de Gaulle airport, J. Aerosp. Inform. Syst. 15 (2018) 120–130, https://doi.org/10.2514/1.I010502.
[32] J.J. Rebollo, H. Balakrishnan, Characterization and prediction of air traffic delays, Transp. Res., Part C, Emerg. Technol. 44 (2014) 231–241, https://doi.org/10.1016/j.trc.2014.04.007.
[33] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.
[34] T. Feng, H.J.P. Timmermans, Detecting activity type from GPS traces using spatial and temporal information, Eur. J. Transp. Infrastruct. Res. 15 (2015) 662–674.
[35] B. Yu, H. Wang, W. Shan, B. Yao, Prediction of bus travel time using random forests based on near neighbors, Comput.-Aided Civ. Infrastruct. Eng. 33 (2018) 333–350, https://doi.org/10.1111/mice.12315.
[36] A. Afzal, A. Aabid, A. Khan, S. Afghan Khan, U. Rajak, T. Nath Verma, R. Kumar, Response surface analysis, clustering, and random forest regression of pressure in suddenly expanded high-speed aerodynamic flows, Aerosp. Sci. Technol. 107 (2020) 106318, https://doi.org/10.1016/j.ast.2020.106318.
[37] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Prediction, Inference and Data Mining, Springer-Verlag, New York, 2009.
[38] J.J. Rebollo, H. Balakrishnan, Characterization and Prediction of Air Traffic Delays, Massachusetts Institute of Technology, 2012.
[39] V. Deshpande, M. Arikan, The impact of airline flight schedules on flight delays, Manuf. Serv. Oper. Manag. 14 (2012) 423–440, https://doi.org/10.1287/msom.1120.0379.
[40] L. Belcastro, F. Marozzo, D. Talia, P. Trunfio, Using scalable data mining for predicting flight delays, ACM Trans. Intell. Syst. Technol. 8 (2016), https://doi.org/10.1145/2888402.
[41] X. Tang, P. Chen, Y. Zhang, 4D trajectory estimation based on nominal flight profile extraction and airway meteorological forecast revision, Aerosp. Sci. Technol. 45 (2015) 387–397, https://doi.org/10.1016/j.ast.2015.06.001.
[42] S. Grabbe, B. Sridhar, A. Mukherjee, Clustering days and hours with similar airport traffic and weather conditions, J. Aerosp. Inform. Syst. 11 (2014) 751–763, https://doi.org/10.2514/1.I010212.
[43] K.D. Kuhn, A methodology for identifying similar days in air traffic flow management initiative planning, Transp. Res., Part C, Emerg. Technol. 69 (2016) 1–15, https://doi.org/10.1016/j.trc.2016.05.014.
[44] M. Bloem, N. Bambos, Ground delay program analytics with behavioral cloning and inverse reinforcement learning, J. Aerosp. Inform. Syst. 12 (2015) 299–313, https://doi.org/10.2514/1.I010304.
[45] S. Gorripaty, Y. Liu, M. Hansen, A. Pozdnukhov, Identifying similar days for air traffic management, J. Air Transp. Manag. 65 (2017) 144–155, https://doi.org/10.1016/j.jairtraman.2017.06.005.
[46] P. Brooker, Experts, Bayesian Belief Networks, rare events and aviation risk estimates, Saf. Sci. 49 (2011) 1142–1155, https://doi.org/10.1016/j.ssci.2011.03.006.
[47] K. Li, L. Tu, L. Chai, Ensemble-model-based link prediction of complex networks, Comput. Netw. 166 (2020) 106978, https://doi.org/10.1016/j.comnet.2019.106978.
[48] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, P.C. Sabeti, Detecting novel associations in large data sets, Science 80 (334) (2011) 1518–1524, https://doi.org/10.1126/science.1205438.
[49] G. Sun, J. Li, J. Dai, Z. Song, F. Lang, Feature selection for IoT based on maximal information coefficient, Future Gener. Comput. Syst. 89 (2018) 606–616.
[50] Y. Liang, W. He, W. Zhong, F. Qian, Objective reduction particle swarm optimizer based on maximal information coefficient for many-objective problems, Neurocomputing 281 (2018) 1–11.
[51] J. Xiong, M. Hansen, Value of flight cancellation and cancellation decision modeling ground delay program postoperation study, Transp. Res. Rec. (2009) 83–89, https://doi.org/10.3141/2106-10.
[52] W.-B. Du, M.-Y. Zhang, Y. Zhang, X.-B. Cao, J. Zhang, Delay causality network in air transport systems, Transp. Res., Part E, Logist. Transp. Rev. 118 (2018) 466–476, https://doi.org/10.1016/j.tre.2018.08.014.
[53] R. Díaz-Uriarte, S. Alvarez de Andrés, Gene selection and classification of microarray data using random forest, BMC Bioinform. 7 (2006) 3, https://doi.org/10.1186/1471-2105-7-3.
[54] R. Grimm, T. Behrens, M. Märker, H. Elsenbeer, Soil organic carbon concentrations and stocks on Barro Colorado Island — digital soil mapping using Random Forests analysis, Geoderma 146 (2008) 102–113, https://doi.org/10.1016/j.geoderma.2008.05.008.