RAIRO-Oper. Res. 56 (2022) 1429–1451
https://doi.org/10.1051/ro/2022052

**RAIRO Operations Research**
www.rairo-ro.org

# A HYBRID MACHINE LEARNING-OPTIMIZATION APPROACH TO PRICING AND TRAIN FORMATION PROBLEM UNDER DEMAND UNCERTAINTY

Atiye Yousefi and Mir Saman Pishvaee*

**Abstract.** Due to the complexity of pricing in the service industry, it is important to provide an efficient pricing framework for real-life and large-sized applications. To this end, we combined an optimization approach with a regression-based machine learning method to provide a reliable and efficient framework for integrated pricing and train formation problem under hybrid uncertainty. To do so, firstly, a regression-based machine learning model is applied to forecast the ticket price of the passenger railway, and then, the obtained price in is used as the input of a train formation optimization model. Further, in order to deal with the hybrid uncertainty of demand parameters, a robust fuzzy stochastic programming model is proposed. Finally, a real transportation network from the Iran railway is applied to demonstrate the efficiency of the proposed model. The analysis of numerical results indicated that the proposed framework is able to state the optimal price with less complexity in comparison to traditional models.

**Mathematics Subject Classification.** 62A86, 90C17, 62J05.

Received March 25, 2021. Accepted April 6, 2022.

## 1. Introduction

The train formation problem (TFP) is one of the most important research areas in rail transportation planning. Among different transportation modes, the railway has important advantages over others in the aspects of safety, flexible capacity, and low emissions [49,54]. In order to provide the best services in the railway system and to use the maximum capacity, it is acceptable that the transportation plan adapts to the changing conditions. Price changes are one of the most important issues affecting optimal transportation planning, therefore, comprehensive and extensive researches exist in the literature (*e.g.*, [56]).

Pricing is the process whereby a business determines the price at which it will sell its products and services. Price is the key factor for revenue-generating, playing a fundamental role in railway operation planning. Therefore, the price should be determined at the right level, not so high that the potential buyers put off and not so low that the potential profits lose out [36]. Generally, ticket pricing in public transportation can be considered as a game between the company and consumers where each party tries to maximize its own profit [40]. Railway companies are trying to keep their overall revenue as high as possible and optimize their fundamental decisions, while customers are seeking to get the best price for their tickets [1]. Previous research on train ticket pricing indicated that factors such as the number of days passed from booking to departure, departure or arrival time,

*Keywords.* Regression-based machine learning, pricing, train formation problem, hybrid uncertainty, robust fuzzy stochastic programming.

School of Industrial Engineering, Iran University of Science and Technology, Tehran, Iran.
*Corresponding author: pishvaee@iust.ac.ir

etc. significantly influence the desired price of customers [4, 5]. So, one of the approaches that can adjust pricing considering customer satisfaction, is to pricing regarding factors that significantly influence pricing from a customer point of view.

Also, pricing influences customer demand, playing a significant role in attracting customers, and increasing sales [50]. Since the demand changes dynamically (especially during seasons of holidays and festivals), rail transportation corporations are facing complicated and different strategies and techniques to assign ticket prices [29]. Also, demand has a high degree of uncertainty [28], which influences fare prices and thereby affects railway planning; therefore, it is important to use an appropriate method to deal with the uncertainty of demand. Generally, some methods are developed to control the ambiguity of parameters including stochastic programming, robust programming, and fuzzy programming. In most of the previous researches related to pricing in railway transportation, the demand is considered as a scenario-based parameter that in each of them, the amount of demand is considered as a deterministic value. But in most real-life problems there is no or limited knowledge about the real amount of demand in each scenario and it is necessary to use experts' viewpoints. So, we faced the parameter with two sources of uncertainty. The first source is that this uncertain parameter is based on scenarios that are considered according to the probability of their occurrence. The second source is that the values of this parameter in each scenario are usually imprecise and can be specified by possibilistic distributions. Therefore, it is necessary to consider a framework to cope with the hybrid uncertainty simultaneously.

Due to this high complexity, it is important to use appropriate methods for predicting the ticket price. Considering the common applications of the pricing and the relevant complexity, one of the main contributions of this research is to provide a framework applicable to real-life and large-sized pricing problems. To this end, for the first time, we have combined an optimization approach with the regression-based machine learning method to provide a reliable and efficient framework for the integrated pricing-train formation problem. Considering a framework that integrated machine learning methods with optimization models can provide a reliable and efficient framework to deal with the complex market. Due to the ability of machine learning methods to examine large volumes of data and discover certain trends and patterns that are not achieved by using optimization models, developing the framework that used both of them concomitantly can improve the accuracy of the results.

Based on the above-mentioned descriptions, this study aims to provide a comprehensive model for pricing passengers' tickets in the railway system and to combine it with the train formation problem under demand uncertainty. The objective function of the proposed model is to maximize the profits of companies providing rail services and determining the optimal number of the train. In order to achieve a reliable and efficient framework to deal with the complex railway market, a framework that integrated machine learning methods with optimization models is developed. Also, to the best of our knowledge, this is the first attempt to ticket pricing considering both customers and railway sides by developing hybrid methods of machine learning and optimization approach.

The main contribution of our model is the effort to develop a framework that can deal with the weaknesses of each model of machine learning and optimization. Machine learning models cannot optimize and they are only predicting factors based on some available information. On the other hand, in optimization models, decisions are made without considering all the affecting factors. To be more precise, in the issue of train formation plan, price is one of the most important factors that affected the principal decisions of its. If the price is considered as a variable without considering the real factors such as travel time, distance, etc., the output of the model determines unrealistic prices that do not be efficient. Also, these unreliable prices negatively affect other important decisions, such as allocating trains over a period of time. Therefore, providing a comprehensive model for pricing passengers tickets in the railway system and combine it with the train formation problem provide more realistic results compared to previous approaches.

### 1.1. Contributions

The other contributions of this research are as follows:

(1) Provide a comprehensive model for pricing passengers' tickets in the railway system and to combine it with the train formation problem. The main decision in the train formation problem is to determine the optimal number of trains in each time period with the aim of maximizing the profits of railway companies. Previous articles on the train formation problem have neglected the combination of pricing with this issue (please see Sect. 2.1). However, mismatches between train allocated and passenger demand (which is directly affected by the price) usually lead to either the customer paying more or the railway company losing revenue which will lead to dissatisfaction on both sides. Therefore, our paper is the first attempt to combine the ticket pricing with the train formation problem.

(2) Using machine learning techniques to evaluate the effectiveness of the pricing-related factors in the railway system. Previous existing models in ticket pricing generally rely on a limited number of features which are not effective enough in predicting ticket price on both the customer and company sides. Also, given the huge amount of data available in the rail system, the use of data-driven methods can help decision-makers to extract useful knowledge from the relevant data [16].

(3) Using the robust fuzzy stochastic programming to handle the uncertainty of demand in TFP is the other contribution of the paper. Due to the uncertain nature of demand and its significant impact on pricing decisions, it is very important to address the uncertainty of input parameters.

The remainder of this paper is organized as follows. Section 2 describes the literature review. The model assumptions and problem descriptions are addressed in Section 3. Section 4 represents the machine learning technique to determine the price. The proposed mathematical model is elaborated in Section 5. Explanation of the developed programming approach used to cope with the uncertainty of demand is addressed in Section 6. Section 7 represents numerical examples and discusses the computational results. Finally, the conclusions and future works are presented in Section 8.

## 2. Literature review

Since the main purpose of this article is to design a data-driven approach to combine the pricing problem and train formation problem, in this section, we reviewed the previous papers related to pricing in railway transportation and the train formation plan problem. To classify the past articles related to railroad transportation pricing in an appropriate way, we considered the relative papers in two categories: (1) Railroad. (2) Combinational Transportation (refers to articles that consider railroad and another transportation mode, simultaneously).

### 2.1. Train formation problem

In this section, the works related to the train formation plan are reviewed. For example, Yaghini *et al.* [47] developed a mixed-integer programming model for train formation plan in Iranian railway. To solve the problem, a local branching algorithm is used. The results show the efficiency of the proposed approach in comparison to the exact approach. But'ko and Prokhorchenko [6] proposed an approach to analyzing the train formation problem based on interrelations between railway stations. To do this, the train formation plan was presented as a network structure with a directed graph and tested on Ukraine's railways for 2012–2013. Yaghini *et al.* [48] proposed a hybrid algorithm of the Simplex method and simulated annealing for the TFP problem in the Iranian railway. To adjust the best parameter values in the proposed algorithm, the design of experiments (DOE) method is used. Deng *et al.* [8] developed an optimization model to determine the train formation and service frequency in inter-city rail transport networks. The aim of this paper is to minimize the passengers' total travel costs and maximize the operator's benefit. The results show that the short train and high-frequency mode have greater ability to upgrade the service quality. Yaghini *et al.* [49] proposed a mathematical model with fuzzy costs for train formation planning in the Iranian railway. In this fuzzy model, the costs are considered in three scenarios, namely optimistic, normal and pessimistic. Also, a hybrid algorithm combining local branching and relaxation-induced neighborhood search methods is presented to solve the model. Xiao and Lin [43] present an optimization model of the train formulation plan which considers the one-block train and two-block train and delivers all of the commodities with the minimum car-hour consumption. In this paper, three sub-problems including shipment routing, shipment-to-block assignment, and block-to-train assignment are presented. Also, a heuristic optimization approach based on the ant colony system is proposed to solve the model. Lin [20] proposes a non-linear binary programming model to address the integrated railcar itinerary and train formation plan optimization problem. A simulated annealing-based heuristic solution approach is developed to solve the mathematical model. The results show the efficiency of the integrated model to determine the railcar path optimization and the train formation plan. Chen *et al.* [7] addressed a linear binary programming model to minimize total costs in the train formation plan problem in the China railway. In this paper, two subproblems are proposed to specify which blocks are built in each yard which railcars are allocated to which blocks. In this paper, the exact model is proposed for small- and medium-scale, and a novel two-stage tree-based decomposition approach is proposed to solve large-scale. Xiao *et al.* [44] considered the train formation problem in an actual 19-yard railway sub-network in China using both the single-block trains and the two-block trains. The main aim of this paper is to develop an optimization model of the train formulation plan to minimization of the total car-hour consumption at all yards. In order to solve the model, a hybrid algorithm of genetic algorithm and tabu search and a greedy algorithm are developed. Belošević *et al.* [3] proposed a new fuzzy multi-criteria approach to evaluate the train formation methods. The aim of this paper is to determine the rational sorting schedules based on different quantitative and qualitative indicators. Lin and Zhao [21] present a non-linear binary programming model for the train formation problem in order to minimize the total car-hour cost. In this paper, three train formation patterns including direct single-commodity trains, direct multi-commodity trains originating from loading stations, and direct trains from reclassification yards are discussed. Lin *et al.* [23] presented a linear 0–1 integer programming model to optimization of the total cost for the supplier and customer as well as the transportation costs of an entire train and non-direct train using data from the China rail system. In order to handle the uneven passenger demand in bi-directions, Zhao *et al.* [57] addressed a mixed-integer linear programming model to optimize the train formation plan and rolling stock scheduling. The proposed model is designed based on known passenger demand and timetable. Kozachenko *et al.* [17] presented a model for the multi-group train formation problem in order to minimize the time spent on shunting operations. In this paper, a discrete deterministic-controlled system is used to simulate the functioning of a flat yard. Lin *et al.* [24] proposes a non-linear binary programming model to address the simultaneous optimization of the railcar itinerary and train formation plan problems. The main purpose of this study is to minimize the total costs, travel distances and satisfy various practical requirements. A simulated annealing-based heuristic solution approach is developed to solve the mathematical model tested in a real-world railway network in China.

## 2.2. Pricing in railway transportation

### *2.2.1. Railroad transportation pricing*

This section presents an overview of related works on pricing in the passenger section of the railway system. For example, Mingbao *et al.* [25] used a discrete choice model to determine the passenger mode choice in the public transport system. Also, the rail and bus transit pricing game models are considered to determine the optimal price of urban rail transit. Yang and Chang [51] used the multinomial logit model to determine the cabin choice preferences. This paper also examined cabin attributes designed by the stated-preference method. The result shows that integrating the stated-preference method and revealed-preference data obtain more informative models. Xueyu and Jiaqi [46] considered the bi-level programming model to interests of both sides of the public transport companies and the passengers. The upper-level consider maximizing the overall revenue and the lower-level consider minimizing the generalized travel cost. Jin *et al.* [15] proposed a pricing model that is based on the value of travel time and Bi-level programming. The aim of this paper is to maximize the benefit of the railway agencies and the passengers' utility with different income levels. Hetrakul and Cirillo [14] incorporated the passenger choice models and the demand functions into a revenue management optimization system to maximize expected ticket revenue per each train trip. Qin *et al.* [30] developed the mixed-integer programming model to a combination of market segmentation, market competition, and dynamic pricing policy is a railway passenger fare policy portfolio. The aim of this paper is to maximize the expected revenue of China railway passenger transportation. Xiaoqiang *et al.* [45] considered a dynamic pricing model to provide the optimal price for passenger groups in the high-speed railway in China. The main aim of this study is to maximize the expected revenue of selling tickets by considering the best price for passenger groups. Vigren [39] investigated the effect of entry on the private railway companies in the state railways of Sweden. The main purpose of this study is to investigate the effect of entering the private railway company on the ticket price on the Stockholm–Gothenburg line. The results show significant price changes in the competitive markets. Gong *et al.* [11] established a discriminate Ramsey suboptimal pricing model to corporate the profit and the social welfare in Chinese high-speed rail. Gong *et al.* [12] adopted the theory of Ramsey suboptimal pricing, and build a sub-time pricing model of high-speed rail based on the operation cost and weights of peak and off-peak periods. Noordin and Mohd Ali Amran [26] used Monte Carlo Simulation to optimization the ticket pricing of electric train service in Malaya. In order to determine the minimum price for the ticket, the relationship between price and type of passenger, distance traveled and type of seat is considered. Wu *et al.* [42] proposed the model to combine pricing with seat allocation problems in the high-speed railway in China. The main goal of this paper is to maximize the ticket revenue, based on pricing constraints and revenue management strategies. To solve the proposed model, a two-stage algorithm is developed. The output of the first stage is the optimal price and the second is the optimal seat allocation. Beria *et al.* [4] investigated the pricing strategies of the two domestic rail passenger companies in Italy. The focus of this study is on determining the effect of some factors on price determination. The results indicated that in addition to distance, other factors such as demand, capacity, users' willingness to pay, are important. Qin *et al.* [31] used the differential pricing strategy to determine the best price for the high-speed railway in China. The results express the importance of using a dynamic pricing method in increasing the revenue of selling tickets, especially during the peak period. Lin *et al.* [22] considering the ticket pricing in investigated the problem of passenger assignment in a railway transportation system. A mathematical program and a heuristic approach were proposed as considering the price effect on passenger assignment problems Ali *et al.* [2] propose a hybrid method to resolve capacity conflicts between publicly controlled traffic and commercial traffic. The main goal of this paper is to calculate a reservation price for a commercial operator's path request by estimating the societal costs.

### *2.2.2. Combinational transportation pricing*

This section considered a review of related works on combinational transportation pricing. As mentioned previously, this section refers to articles that consider railroad and other transportation mode, simultaneously. In the field of articles considered pricing in the passenger transportation, Sato and Sawaki [33] presented a revenue management model of dynamic pricing for a competitive market of high-speed railways and airlines. In this paper, by using the multinomial logit model to describe the customer's discrete choice, we derive an optimal pricing policy determine to maximize the expected total revenue for the high-speed railway. Yang and Zhang [52] investigate the effects of competition between air transport and high-speed rail. The results indicated that for homogeneous passengers and given schedule frequencies, that the ticket prices of air transport and HSR decrease in the weight of welfare. Van den Berg and Verhoef [38] analyzed congestion pricing in a road and rail network with heterogeneous users, where the train and car are imperfect substitutes. The bottleneck congestion and the crowding congestion are used respectively to pricing in road and rail. Gremm [13] investigated the intermodal competition between a railway company and intercity bus companies in Germany. The results show that the railway prices are lower on routes with intermodal competition compared to monopolistic routes. Wang *et al.* [41] investigated the effect of travel time and safety of high-speed rail speed on airline traffic and price, taking into account the degree of substitutability between the two services. The results show the significant effect of reducing the travel time of high-speed railways airline traffic and price. Zhang *et al.* [56] analyzed the effect of the high-speed railway system on air ticket pricing and flight frequency in China. The results indicated on the routes with HSR services, airlines' prices and frequencies were reduced greatly. Su *et al.* [35] investigated the effect of the development of high-speed rail on airlines price. The results show a change in price strategy for leisure and business airline travelers.

Based on the above-mentioned overview, most of the previous work used stochastic programming to handle the demand uncertainty. Despite the fact that comprehensive and extensive researches are existed in the literature, using machine learning techniques along with optimization methods is neglected. Finally, the development of a mathematical model for ticket-pricing in the field of passenger train formation planning is the other issue that is not considered. Therefore, in order to enrich the literature, this paper presents a mathematical model for the integrated pricing-train formation problem that used a framework that integrated machine learning methods with optimization models. The objective function of this paper is to maximize the profits of companies providing rail services and determining the optimal number of the train. Moreover, the proposed model employs the Robust Fuzzy Stochastic Programming (RFSP) approach to handle the demand uncertainty.

## 3. Problem definition

According to the above-mentioned discussion, this study aims to integrate the pricing of the rail passenger transportation system and the train formation problem, which has not been addressed in related literature. So, the main contribution of this study is to provide a comprehensive model for pricing passengers' tickets in the railway system and combine it with the train formation problem that can deal with the weaknesses of each model of machine learning and optimization. The objective of this paper is to maximize the profits of companies providing rail services and determine the optimal number of the train. In order to achieve this goal, we proposed a framework illustrated in Section 4.

### 3.1. Assumptions

- Multiple periods are taken into account under 30 days' time horizon.
- Train capacity is pre-determined and limited. In other words, the goal is to allocate demand to the trains, and it is assumed that trains cannot change the wagon's arrangement along the origin-destination path.
- The travel time is definite and the demand is uncertain and dynamic.
- Unsatisfied demand in each time period is considered as lost demand.
- No new line will be added to the railway network during the investigation period. Also, no station will be removed from the line or added to it.
- The type and the quality of service provided by each train are known and do not change during the investigation period.

## 4. Research method and framework

The proposed optimization framework solves a joint pricing and train formation problem. This framework includes two stages (see Fig. 1). Our proposed framework has two phases including the machine learning-based stage and optimization stage.

The machine learning-based stage phase estimated the ticket price of the railway system by using the regression-based machine learning approaches. To ticket pricing, a machine learning workflow is considered which is known as the Cross-Industry Standard Process for Data Mining (CRISP-DM) and is described by Shearer [34]. The process of machine learning starts with data collection. In our case, we used historical data about ticket reservations in the railway system (see Sect. 4.1). Then, the data is analyzed to realize and investigate the data structure in Section 4.2. Finally, a regression-based machine learning algorithm is used to fit the model and to predict the price in Section 4.3.

In the optimization stage, first of all, we used previous stage estimated price as a parameter of the train formation problem to define the optimal number of trains (see Sect. 5). Then, the robust fuzzy stochastic programming approach is used to deal with the uncertainty of demand and to analyze the results (see Sect. 6).

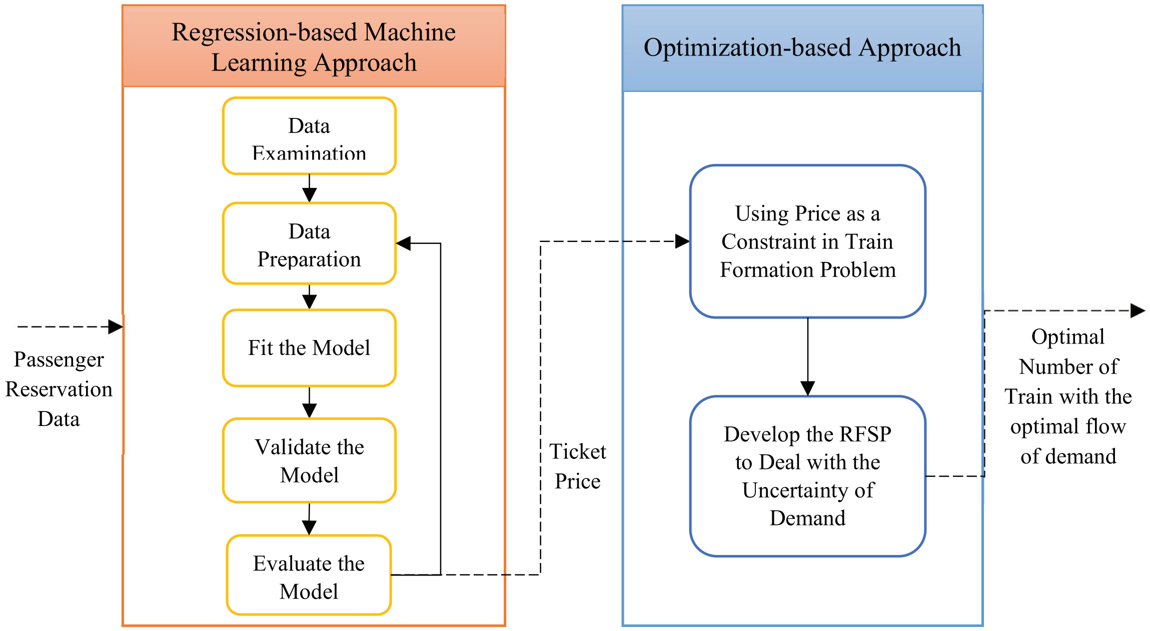

Regression-based Machine Learning Approach
Data Examination
Data Preparation
Fit the Model
Validate the Model
Evaluate the Model
Passenger Reservation Data
Ticket Price
Optimization-based Approach
Using Price as a Constraint in Train Formation Problem
Develop the RFSP to Deal with the Uncertainty of Demand
Optimal Number of Train with the optimal flow of demand

FIGURE 1. The research framework.

## 4.1. Data collection

In order to collect the passenger reservation data, the real data of a railway company in Iran, is collected. The trains that belong to this railway company are divided into 5 classes, each of which is slightly different in terms of service quality. This classification includes general class (type 1), economy class (type 2), business class (type 3), special class (type 4), and super special class (type 5). Type 1 is the cheapest and type 5 is the most expensive in comparison to other trains. This railway company provides most of its services on the Tehran–Mashhad route. Therefore, in this research, information related to the passenger ticket reservation on the Tehran–Mashhad route has been used. Exactly, our dataset includes data collected from April 2019 to July 2019, both on weekdays and weekends. Given the detailed booking and ticketing information available to this study, we are able to incorporate the following variables in the model:

(1) Price.
(2) Origin.
(3) Destination.
(4) Departure time.
(5) Arrival time.
(6) Travel time.
(7) Number of Days from Departure.
(8) Day of week.
(9) Month of Year.
(10) Type of train.

Table 1 provides the definitions and summary statistics for some variables used in estimation. There are several features of the data depicted in Table 1 that are worth mentioning. First, the highest number of trips is from Tehran to Mashhad or Mashhad to Tehran. In the considered route, the volume of the passenger is considerable in Tehran and Mashhad, which is also clear in Figure 6. Another note is the distance between the

TABLE 1. Some features of the data.

| Variable | Variable definition | Mean | Minimum | Maximum |
|---|---|---|---|---|
| Price | Ticket price | 362.9819 | 312.85 | 514.2 |
| Origin | Equals 1 if the origin is Tehran | 0.55 | 0 | 1 |
| | Equals 1 if the origin is Semnan | 0.12 | 0 | 1 |
| | Equals 1 if the origin is Shahrood | 0.14 | 0 | 1 |
| | Equals 1 if the origin is Neghab | 0.02 | 0 | 1 |
| | Equals 1 if the origin is Neyshabur | 0.15 | 0 | 1 |
| Destination | Equals 1 if the Destination is Mashhad | 0.44 | 0 | 1 |
| | Equals 1 if the Destination is Neyshabur | 0.20 | 0 | 1 |
| | Equals 1 if the Destination is Neghab | 0.02 | 0 | 1 |
| | Equals 1 if the Destination is Shahrood | 0.16 | 0 | 1 |
| | Equals 1 if the Destination is Semnan | 0.17 | 0 | 1 |
| Number of days from departure | Distance between ticket booking and departure day. | 17 | 1 | 28 |
| Type of train | Equals 1 if the train is type 1 | 215 564 | 0 | 1 |
| | Equals 1 if the train is type 2 | 194 567 | 0 | 1 |
| | Equals 1 if the train is type 3 | 153 457 | 0 | 1 |
| | Equals 1 if the train is type 4 | 143 727 | 0 | 1 |
| | Equals 1 if the train is type 5 | 134 662 | 0 | 1 |

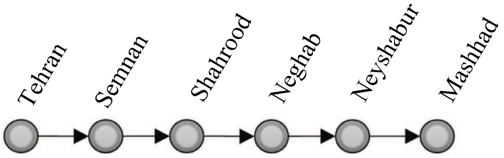


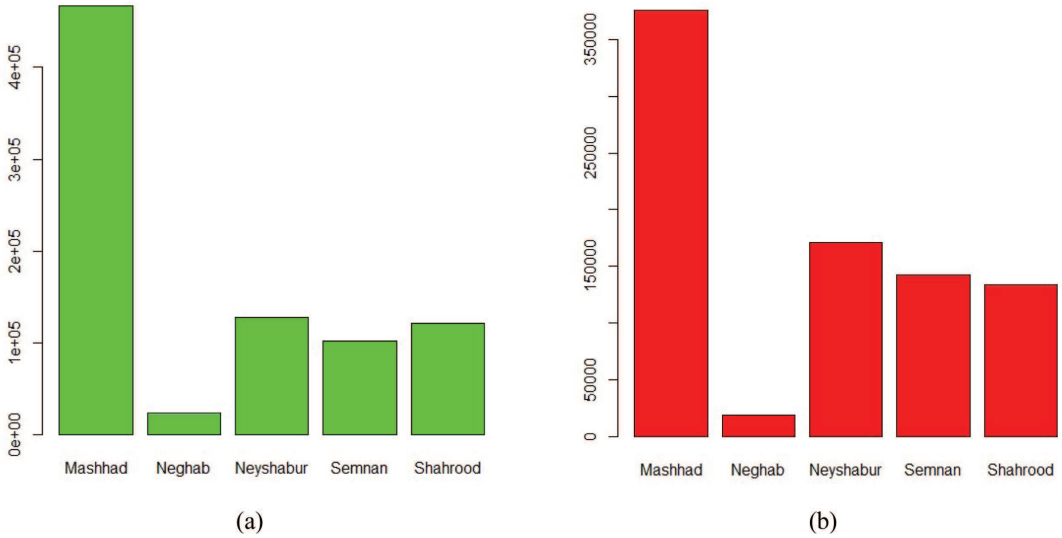
FIGURE 2. A considered network.

ticket booking and departure day. According to observations, most travelers book tickets in the days close to departure time. Indeed, tickets are purchased on average 17 days before departure day.

### *4.1.1. OD pairs analyzed*

Iran is a vast country with hundreds of kilometers between major cities. Therefore, trains are nice tools to bridge those long distances. One of the long-distance paths with a huge number of passengers is the route between Mashhad and Tehran. Mashhad is one of the biggest cities in Iran, which entices hosts thousands of travelers annually due to tourist and religious attractions. Considering the Tehran–Mashhad rail network, this line totally covers 926 km and connects 50 cities, including Garmsar, Semnan, Damghan, Shahrood, and other cities. Depending on the train type, the duration of trips varies between 8 and 12 h. There are a total of 15 stations on this route, that we exclude 9 stations because passenger demand in these markets is low and insufficient for estimation. The remaining 6 stations are shown in Figure 2.

Through this data, concerning the origin and destination, we can calculate the number of passengers per month for the considered route (see Fig. 3). Accordingly, the volume of passengers' commutes per month from Tehran and Mashhad are considerable in comparison to other cities, which representations the strategic position of these cities in the proposed route.



(a)

(b)

FIGURE 3. The number of passengers commutes per month for each city. (a) Number of passengers arrival each city. (b) Number of passengers departure each city.

## 4.2. Data preparation

All of the collected data need a lot of work before using. The collected data should be cleaned and prepared according to the model requirements. All the unnecessary data is removed and missing value is investigated.

### *4.2.1. Missing data*

In our dataset, approximately 20% percent of the data is missing. The data relating to Tehran, Mashhad, Neyshabur, and Shahrood are more complete. The pattern of missing data is related to the less commuting city such as Semnan and Neghab. There are many methods that can be used to account for missing data. In this paper, we used the Predictive Mean Matching (PMM) for imputation missing values. PMM is widely used for imputation missing values. It aims to reduce the bias introduced in a dataset through imputation, by drawing real values sampled from the data. This is achieved by building a small subset of observations where the outcome variable matches the outcome of the observations with missing values [32].

## 4.3. Fit the models

In order to develop a model for price forecasting, linear regression has been used, which is presented below. Linear regression is a subdivision of supervised learning, investigating the relationship between dependent and independent variable/variables. The goal of regression analysis is to create a mathematical model that can be used to predict the values of a dependent variable based upon the values of an independent variable.

Let $\{x_1, x_2, \cdots, x_n\}$ be a set of $n$ observations. The linear regression outputs a prediction for the target value using the function $y$:

$$y(\beta \cdot x) = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon \tag{4.1}$$

where the $\beta_0$ is the intercept, the $\beta_i (i = 1, \cdots, n)$ is the slope and $\varepsilon$ is the regression residual. The slope indicates the relationship between the dependent variable and a series of other independent ones. The $\varepsilon$ is the difference between the predicted value $(\hat{y})$ and the observed value. Based on the multiple linear regression, our prediction model is defined as follows:

$$p_r^t = \beta_0 + \beta_1 O_r + \beta_2 D_r + \beta_3 \mathrm{DT}_r + \beta_4 \mathrm{AT}_r + \beta_5 \mathrm{NDD}_r + \beta_6 \mathrm{DW}_r + \beta_7 \mathrm{MY}_r + \beta_8 \mathrm{ToT}_r + \varepsilon_r^t \tag{4.2}$$

where:

- $p_r^t$ is the point-to-point O&D fare for each ticket that is purchased for train $r$ in time $t$;
- $O_r$ is a vector of variables accounting for determining the effect of origin in the tickets that are purchased;
- $D_r$ is a vector of variables accounting for determining the effect of destination in the tickets that are purchased;
- $\mathrm{DT}_r$ is a vector of variables accounting for determining the effect of departure time in the tickets that are purchased;
- $\mathrm{AT}_r$ is a vector of variables accounting for determining the effect of arrival time in the tickets that are purchased;
- $\mathrm{NDD}_r$ is a vector of variables considered for determining the effect of the number of days from departure in the tickets that are purchased;
- $\mathrm{DW}_r$ is a vector of variables considered for determining the effect of day of the week in the tickets that are purchased;
- $\mathrm{MY}_r$ is a vector of variables considered for determining the effect of month of the year in the tickets that are purchased;
- $\mathrm{ToT}_r$ is a vector of variables considered for determining the effect of type of train in the tickets that are purchased;
- $\beta_0$ is the intercept term;
- $\beta_i$ is the regression coefficient for regressor $i$;
- $\varepsilon_r^t$ is the error term.

### 4.4. Evaluation and validation

Before discussing the statistical description of the regression model, some points should be made. As can be seen, only two indices related to the train type and time are considered in the developed regression model. Our justification for considering the $r$ index for all variables is the fact that all variables expressed in the regression model depend on the train type. In other words, all information, including origin, destination, departure time, etc., are determined according to train number. It should also be noted that, since the price predicted by the regression model is used as an input in the optimization model, increasing the indices results in more computational complexity of the mathematical model.

Table 2 indicates the statistics description of the model. There are 841 977 observations considered in this sample from the specific route control variables that are unreported for the sake of brevity. According to Table 2,

Table 2. The statistics description of the model.

| Variable | P_value* |
|---|---|
| $O_r$ | ✓ |
| $D_r$ | ✓ |
| $\mathrm{DT}_r$ | ✓ |
| $\mathrm{AT}_r$ | ✓ |
| $\mathrm{NDD}_r$ | ✓ |
| $\mathrm{DW}_r$ | ✓ |
| $\mathrm{MY}_r$ | ✓ |
| $\mathrm{ToT}_r$ | ✓ |
| Multiple R-squared | 0.8165 |
| Adjusted R-squared | 0.8383 |
| F-statistic | 5.876e+04 |
| $p$-value | <2.2e−16 |

**Notes.** *Significance level of 0.05.

Table 3. The predicted price (Toman).

| | Train | | | | |
|---|---|---|---|---|---|
| Period | 1 | 2 | 3 | 4 | 5 |
| 1 | 248 504 | 275 344 | 328 927 | 409 382 | 459 383 |
| 2 | 243 645 | 274 645 | 325 378 | 412823 | 460 386 |
| 3 | 252 596 | 274 437 | 337 363 | 428 231 | 465 848 |
| 4 | 243 747 | 275 839 | 345 477 | 427 373 | 475 659 |
| 5 | 251 029 | 279 470 | 363 629 | 432 722 | 479 933 |
| 6 | 250 385 | 280 473 | 354 544 | 438 864 | 480 377 |
| 7 | 249 277 | 282 730 | 359 479 | 437 466 | 482 643 |
| 8 | 249 623 | 28 1523 | 353 630 | 443 438 | 489 647 |
| 9 | 250 377 | 280 365 | 358 460 | 441 582 | 490 382 |
| 10 | 248 232 | 282 737 | 359 461 | 450 384 | 504 841 |

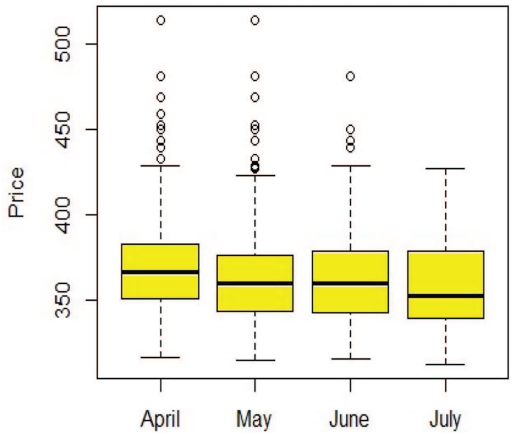


Figure 4. Price *vs.* Booking Month.

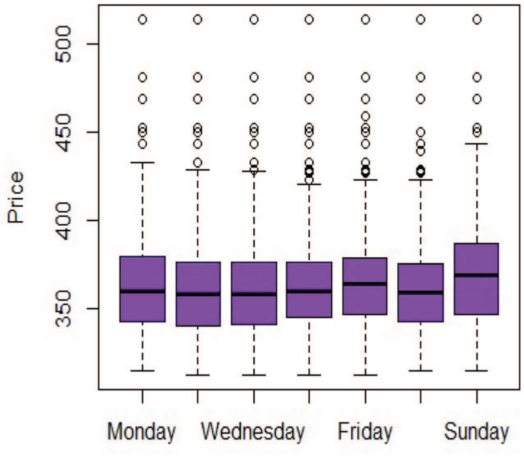


Figure 5. Price *vs.* Booking Day.

the values reported for R-squared and F-statistic indicate the performance of the fitted model. Based on the fitted model, the prediction price for 10 periods obtained as demonstrated in Table 3.

According to Table 3, the prices predicted for train Type 5 are higher than other trains. Also, the price of this train fluctuates more than the other ones. Based on initial information received from the railway company, due to variable demand and the high price of Type 5 trains, price fluctuations for this type of train are due to the demand coverage and revenue management goals.

Figures 4 and 5 demonstrate the price fluctuation due to change in the day and the month of the booking ticket. Since July is the beginning of the summer holidays in Iran, price fluctuation is considerable than in other months. But the ticket booking day seems to have little effect.

## 5. Model formulation

In this section, a mixed-integer programming (MIP) is developed to maximize the total revenue of selling tickets (second phase of our research framework). It should be noted that we were inspired by the model introduced by Yaghini *et al.* [47] and developed it based on our problem characteristics.

**Indices**

| | | |
|---|---|---|
| $r \cdot r'$ | Index of trains | $r \cdot r' = 1, \ldots, R$ |
| $t$ | Index of time periods | $t = 1, \ldots, T$ |
| $y$ | Index of yards | $y = 1, \ldots, Y$ |
| $d$ | Index of demand from origin node $o(d)$ Destination node $d(d)$ | $d = 1, \ldots, D$ |
| $s^{+}(y)$ | Index of trains outward from yard $\boldsymbol{y}$ | $y = 1, \ldots, Y$ |
| $s^{-}(y)$ | Index of trains inward to yard $\boldsymbol{y}$ | $y = 1, \ldots, Y$ |

**Parameters**

$u_r$ Capacity of train $\boldsymbol{r}$

$\tau_y$ Maximum number of cars that can be assembled in yard $\boldsymbol{y}$

$o(d^t)$ Origin node for demand $\boldsymbol{d}$ in period $\boldsymbol{t}$

$d(d^t)$ Destination node for demand $\boldsymbol{d}$ in period $\boldsymbol{t}$

$c_{dr}^t$ Shipping cost one unit of demand $\boldsymbol{d}$ on train $\boldsymbol{r}$ in period $\boldsymbol{t}$

$f_r^t$ Fixed cost to be paid if train $\boldsymbol{r}$ runs in period $\boldsymbol{t}$

$Ad_r^t$ Actual demand for on train $\boldsymbol{t}$ in period $\boldsymbol{t}$

$d_r^t$ Demand for train $\boldsymbol{r}$ in period $\boldsymbol{t}$

$p_r^t$ Ticket predicted prices on train $\boldsymbol{r}$ in period $\boldsymbol{t}$

$\gamma$ Price sensitivity

$\delta_{rr'}$ Number of customers switching from the train $\boldsymbol{r}$ to the train $\boldsymbol{r'}$ per unit increase in the price difference between $\boldsymbol{p_r}$ and $\boldsymbol{p_{r'}}$

**Variables**

$Y_r^t$ $\begin{cases} 1 \text{ If train } \boldsymbol{r} \text{ is allocated in period } \boldsymbol{t}; \\ 0 \text{ Otherwise;} \end{cases}$

$X_{dr}^t$ Amount of demand $\boldsymbol{d}$ that is satisfied by train $\boldsymbol{r}$ in period $\boldsymbol{t}$

The Objective function (5.1) considered the total cost and profit, calculated as the sum of the fixed costs of the trains and shipping cost of demand decrease from the profit of selling the ticket. The variable and fixed costs are dependent on many factors such as time, season, weather conditions [49], and economical changes such as inflation fluctuation. The fixed cost mainly consists of the cost of a unit of the locomotive such as train crew that is independent of the number of cars assigned to the train. While, shipping or variable cost is related to the number of locomotives assigned to the train [47].

$$\text{Max } Z = \sum_{r \in R} \sum_{t \in T} d_r^t p_r^t - \sum_{r \in R} \sum_{t \in T} f_r^t Y_r^t - \sum_{d \in D} \sum_{r \in R} \sum_{t \in T} c_{dr}^t X_{dr}^t. \tag{5.1}$$

In the above-mentioned equation, the first term calculates the revenue for selling tickets. The second term shows the fixed cost for the formation train and, the last term indicated the shipping cost for transferring passengers.

$$\sum_{d\in D}\sum_{r\in S^{+}(y)}\sum_{t\in T}X_{dr}^{t}-\sum_{d\in D}\sum_{r\in S^{-}(y)}\sum_{t\in T}X_{dr}^{t}=\begin{cases}\sum_{r\in R}\sum_{t\in T}d_{r}^{t} & \text{if } \ y=o(d^{t}) \quad \forall y\in y\\ -\sum_{r\in R}\sum_{t\in T}d_{r}^{t} & \text{if } y=d(d^{t}) \quad \forall y\in y\\ 0 & \text{otherwise} \quad \forall y\in y.\end{cases} \tag{5.2}$$

Constraints (5.2) ensure that the train containing the passenger must leave the origin and enter the destination. Also, demand satisfaction at each origin-destination is controlled by this equation. As mentioned previously, $o(d^t)$ refer to origin nodes for demand $d$ in period $t$, while $d(d^t)$ mention to destination nodes for demand $d$ in period $t$. Also, each demand $d \in D$ is characterized by a given amount $d_r^t$ of flow to be shipped from an origin node $o(d^t)$ to a destination node $d(d^t)$ in period $t$. The $d_r^t$ is calculated based on equation (5.5) proposed in Section 5. After calculating the total amount of demand for each train in each period $(d_r^t)$, the origin/destination nodes for demand $d$ in period $t$ is determined based on available information from customer demand for each city and is placed in $o(d^t)$ and $d(d^t)$, respectively.

$$\sum_{d\in D:o(d)=y_1}\ \sum_{r\in R:o(r)=y_1}\ \sum_{t\in T}X_{dr}^{t}\leq\tau_{y}\qquad\forall y_1\in y. \tag{5.3}$$

Constraint (5.3) assures that the number of cars that can be assembled in the origin yard should not exceed the yard capacity. The number of cars can be assembled only considered in the origin yard because the arrangement of passenger trains does not change along the way.

$$\sum_{d\in D}\sum_{r\in R}\sum_{t\in T}d_{r}^{t}X_{dr}^{t}\leq\sum_{t\in T}\sum_{r\in R}u_{r}Y_{r}^{t}. \tag{5.4}$$

Constraint (5.4) states that the total flow on the train cannot exceed its capacity. They also ensure that no passengers are allowed on a train unless the fixed cost is paid.

$$\sum_{r\in R}\sum_{t\in T}d_{r}^{t}=\sum_{r\in R}\sum_{t\in T}Ad_{r}^{t}-\gamma\left(\sum_{r\in R}\sum_{t\in T}p_{r}^{t}\right)-\sum_{r'\in R'}\sum_{t\in T}\delta_{rr'}\left|\sum_{t\in T}p_{r}^{t}\sum_{t\in T}p_{r'}^{t}\right| \tag{5.5}$$

$$X_{dr}^{t}\geq 0\quad\forall d,r,t;\ \ Y_{r}^{t}\in\{0,1\}\quad\forall r,t. \tag{5.6}$$

Constraints (5.5) calculate the proposed demand function. In general, demand is a function of various factors, but in this study, price is considered as the main factor affecting demand. Accordingly, demand decreases when the price increases. The vector $\gamma$ shows price sensitivity. Also, the number of customers switching from the train $r$ to the train $r'$ per unit increase as a result of the price difference between $p_r$ and $p_{r'}$ considered by using vector $\delta$. The actual demand and the vectors $\gamma$ and $\delta$ achieved based on forecasting the previous period.

## 6. Robust fuzzy stochastic programming approach

Stochastic programming is one of the methods used in the situation the data probability distributions are known or can be estimated, is an appropriate approach to deal with the uncertainty of parameters. Also, using stochastic programming is logical only when an action can be repeated several times [37]. Due to the repeated characteristics of traveling by train, usually, there is enough historical data to model uncertain parameters within each scenario. Despite the repetition of some events, due to the high sensitivity of some parameters, relying on historical data for decision making is illogical and it is necessary to benefit from experts' experiences and their professional opinions. In practice, when there is not a sufficient amount of historical data or it is necessary to rely on expert opinion to provide reasonable estimations for imprecise parameters [9], fuzzy mathematical programming is used. This method uses fuzzy numbers to formulate parameters with epistemic uncertainty through the possibility theory.

In the context of passenger transportation in the railway system, usually, there is historical data about passenger demand to estimate the probability distribution. But the expert's opinion expresses the existence of inherent impreciseness in the scenario-dependent data. The existence of such inherent uncertainties in demand can significantly influence the performance of railway planning. As such, since we are dealing with a parameter with a mixture of random and possibilistic nature, it is necessary to consider a framework to cope with the hybrid uncertainty simultaneously. Recently, new methods have been developed to model hybrid uncertainty and handle the ambiguity of parameters.

In this paper, we used the method described in Fazli-Khalaf *et al.* [10] named the Robust Fuzzy Stochastic Programming approach that is suitable to cope with the uncertainty of passenger demand with a mixture of random and possibilistic nature. The logic of the RFSP is based on considering the uncertainty parameter as a scenario-based parameter and using the experts' knowledge to estimate the range in each scenario. In order to express the RFSP method, we consider the following compact model:

$$\begin{aligned}
&\text{Max } Z = fy + \widetilde{c}x\\
&\text{s.t.}\\
&\qquad \widetilde{b} \leq Ax\\
&\qquad Sx = \widetilde{N}y\\
&\qquad x \geq 0 \; y \in \{0,1\}.
\end{aligned} \tag{6.1}$$

$Z$ is considered as the variable of the above model. The vectors $f$, $c$, and $b$ represented the parameters of the model. The matrices $A$, $S$ and, $N$ are coefficient matrices of the constraints. Also, all binary decision variables are included in vector $y$ and all the continuous decision variables are included in vector $x$. It is assumed that vectors $c$ and $b$ and the coefficient matrix $N$ are the uncertainty parameters of this issue. To develop the RFSP approach, the method stated in Leung *et al.* [19] is used to consider the stochastic programming aspect of the model. Based on this method, the robust stochastic approach for the Model (6.1) is defined as follows:

$$\begin{aligned}
&\text{Max} = \sum_{\theta} \pi_\theta Z_\theta - \lambda \sum_{\theta} \pi_\theta \left( Z_\theta + \sum_{\theta'} \pi_{\theta'} Z_{\theta'} + 2U_\theta \right) - \sum_{\theta} \pi_\theta \varepsilon_\theta w_\theta\\
&\text{s.t.}\\
&\quad Z_\theta = fy + C_\theta x_\theta && \forall \theta\\
&\quad b_\theta - \varepsilon_\theta \leq Ax_\theta && \forall \theta\\
&\quad S_\theta x_\theta - N_\theta y_\theta + \varepsilon_\theta = 0 && \forall \theta\\
&\quad Z_\theta - \sum_{\theta'} \pi_{\theta'} Z_{\theta'} + U_\theta \geq 0 && \forall \theta\\
&\quad U_\theta, x_\theta \geq 0; y \in \{0,1\}.
\end{aligned} \tag{6.2}$$

$\theta$ shows the possible scenarios while $\pi_\theta$ denotes the probability of scenario $\theta$. The first term in the objective function is the expected value of $Z$, maximizing the total profit of the system. The second term, attempts to minimize the deviation of the total profit, and the third term, try to minimize the penalty violation of constraints.

In order to add fuzziness characteristics to the robust stochastic model and develop the RFSP approach, the possibilistic programming developed by Pishvaee *et al.* [27] is used. By considering trapezoidal possibility disibutions for modeling imprecise parameters and credibility measure ($Cr$), the possibilistic programming approach for the model (6.1) is defined as follows:

$$
\begin{aligned}
&\text{Max } Z = fy + \left(\frac{c^{(1)}+c^{(2)}+c^{(3)}+c^{(4)}}{4}\right)x\\
&\text{s.t.}\\
&\quad Ax \geq (2-2\alpha)d^{(3)} + (2\alpha-1)d^{(4)}\\
&\quad Bx \geq \left(1-\frac{\beta}{2}\right)e^{(2)} + \frac{\beta}{2}e^{(3)}\\
&\quad Bx \leq \left(1-\frac{\beta}{2}\right)e^{(3)} + \frac{\beta}{2}e^{(2)}\\
&\quad Sx \leq \left[(2-2\gamma)N^{(2)} + (2\gamma-1)N^{(1)}\right]y\\
&\quad x \geq 0; y \in \{0,1\}; 0 \leq \alpha,\beta,\gamma \leq 0.5.
\end{aligned}
\tag{6.3}
$$

It should be noted that there are different approaches to deal with the uncertainty of equality constraints. In other words, the expression used to control the feasibility robustness of the equal constraints depends on the defined deterministic and uncertainty parts in the equal constraint. Here, we used the approach proved in Yousefi and Pishvaee [53] to handle the uncertainty of equality constraints. By considering the method extended in Fazli-Khalaf *et al.* [10] to combine the methods stated in Leung *et al.* [19] and Pishvaee *et al.* [27], the RFSP approach for the model (6.1) is defined as follows:

$$
\begin{aligned}
\text{Max} = &\sum_\theta \pi_\theta\left(fy + \left(\frac{c_\theta^{(1)}+c_\theta^{(2)}+c_\theta^{(3)}+c_\theta^{(4)}}{4}\right)x\right) - \sum_\theta \pi_\theta\varepsilon_\theta w_\theta\\
&-\lambda\sum_\theta \pi_\theta\left[\left(fy + \left(\frac{c_\theta^{(1)}+c_\theta^{(2)}+c_\theta^{(3)}+c_\theta^{(4)}}{4}\right)x\right)\right.\\
&\left.+\sum_{\theta'}\pi_{\theta'}\left(fy + \left(\frac{c_{\theta'}^{(1)}+c_{\theta'}^{(2)}+c_{\theta'}^{(3)}+c_{\theta'}^{(4)}}{4}\right)x\right) + 2U_\theta\right]\\
\text{s.t.}&\\
&Ax \geq (2-2\alpha)d_\theta^{(3)} + (2\alpha-1)d_\theta^{(4)}\\
&Bx \geq \left(1-\frac{\beta}{2}\right)e_\theta^{(2)} + \frac{\beta}{2}e_\theta^{(3)}\\
&Bx \leq \left(1-\frac{\beta}{2}\right)e_\theta^{(3)} + \frac{\beta}{2}e_\theta^{(2)}\\
&Sx \leq \left[(2-2\gamma)N_\theta^{(2)} + (2\gamma-1)N_\theta^{(1)}\right]y\\
&\left(fy + \left(\frac{c_\theta^{(1)}+c_\theta^{(2)}+c_\theta^{(3)}+c_\theta^{(4)}}{4}\right)x\right) - \sum_{\theta'}\pi_{\theta'}\left(fy + \left(\frac{c_{\theta'}^{(1)}+c_{\theta'}^{(2)}+c_{\theta'}^{(3)}+c_{\theta'}^{(4)}}{4}\right)x\right) + U_\theta \geq 0 \quad \forall\theta\\
&U_\theta, x_\theta \geq 0; y \in \{0,1\}; 0 \leq \alpha,\beta,\gamma \leq 0.5.
\end{aligned}
\tag{6.4}
$$

### 6.1. Robust fuzzy stochastic programming model

As mentioned previously, in this research we use the robust fuzzy stochastic programming model to deal with the uncertainty of demand. Due to the incompatibility of customer behavior and higher demand on specific routes during seasons of holidays and festivals [18], demand is a parameter with uncertain nature with significant fluctuation. Despite the availability of sufficient and reliable information regarding the demand, fluctuation in

Table 4. Fix cost of each train (Million Toman).

| | Time | | | |
|---|---|---|---|---|
| Fix cost | $t = 1$ | $t = 2$ | $t = 3$ | $t = 4$ |
| Train type 1 | 5 | 6.5 | 7 | 6 |
| Train type 2 | 6 | 7 | 6.5 | 5.5 |
| Train type 3 | 6.5 | 6.5 | 7 | 6.5 |
| Train type 4 | 6.5 | 7 | 6 | 6.5 |
| Train type 5 | 5.5 | 6.5 | 6.5 | 7 |

demand as a result of customer behavior, makes a complex and unpredictable tendency in this parameter. In this situation, hybrid methods generate reliable output that could be desirable. In order to estimate the possibility distribution function for demand, we used the expert's opinions. Based on their comments, demand is defined based on a trapezoidal distribution and credibility measure. Based on the RFSP approach, the joint pricing and train formation problem under demand uncertainty is defined as follows:

### Sets

In addition to the sets defined in Section 5, the following set is added.

$\Omega$ Index related to potential scenarios $\Omega = 1, \ldots, \Omega$
$i$ Index related to some constraint $i = 1, \ldots, I$

### Parameters

Some parameters defined in Section 5 are modified as the following parameters; others are the same as defined before.

$d^t_{r\theta}$ Demand for train $\boldsymbol{r}$ in period $\boldsymbol{t}$ under scenario $\boldsymbol{\theta}$
$\gamma_\theta$ Price sensitivity under scenario $\boldsymbol{\theta}$
$\delta_{rr'\theta}$ Number of customers switching from the train $\boldsymbol{r}$ to the train $\boldsymbol{r'}$ per unit increase in the price difference between $\boldsymbol{p_r}$ and $\boldsymbol{p_{r'}}$ under scenario $\boldsymbol{\theta}$
$p^t_{r\theta}$ Ticket prices on train $\boldsymbol{r}$ in period $\boldsymbol{t}$ under scenario $\boldsymbol{\theta}$

### Variables

As with the parameters, some variables are modified and the others are the same as defined before.

$X^t_{dr\theta}$ Amount of flow of demand $\boldsymbol{d}$ on train $\boldsymbol{r}$ in period $\boldsymbol{t}$ under scenario $\boldsymbol{\theta}$

Based on the sets, parameters, and variables defined in this section, we rewrote the model presented in Section 5 based on the provided descriptions in Section 6.

## 7. Computational results

In this section, our model is applied to an instance problem. As mentioned previously, the real data collected from a railway company in Iran belonging to the stations shown in Figure 7 is used in this research. The scale of the problem is as follows: the number of yards is 6; the type of train is 5, and the number of time periods is 10. Table 4 provides the fixed cost of the train in 4 periods, while the passenger demand is given in Table 5.

The ILOG CPLEX 12.6 optimization software is employed to solve the model of the research. All the experiments are carried out by a Pentium dual-core 3.9 GHz computer with 8 GB of RAM. In order to prepare the data and fitting the model, we used R and Python. A summary of the results in terms of total income obtained during a month, corresponding revenue across the different months, and corresponding revenue across different trains are provided in Figures 6–8, respectively.

TABLE 5. Demand for each train (Passenger).

| | Time | | | |
|---|---|---|---|---|
| Demand | $t = 1$ | $t = 2$ | $t = 3$ | $t = 4$ |
| Train type 1 | 6700 | 7200 | 7800 | 8000 |
| Train type 2 | 7200 | 8000 | 6700 | 6700 |
| Train type 3 | 6700 | 7100 | 6700 | 7200 |
| Train type 4 | 8000 | 6700 | 7400 | 8300 |
| Train type 5 | 6400 | 8000 | 6700 | 8000 |

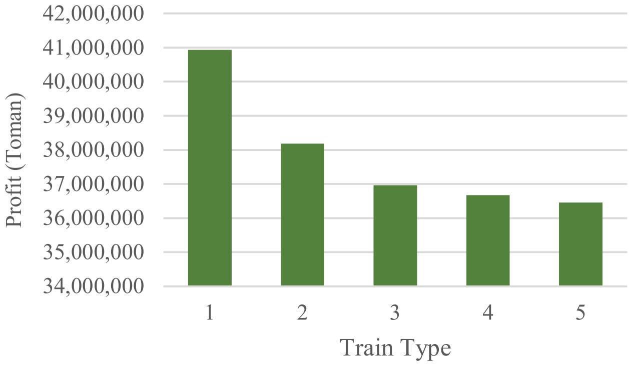


FIGURE 6. The profit obtained from each train during a month.

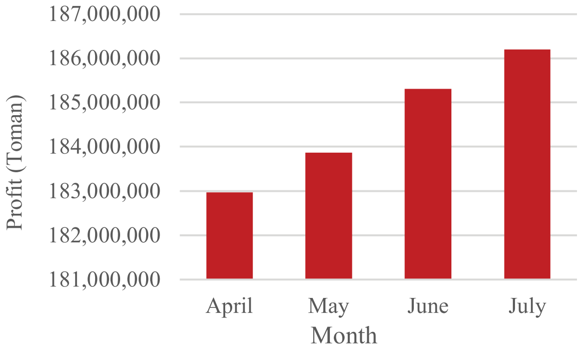


FIGURE 7. Total income obtained during a month.

Based on Figure 6, train type 1, by allocating over 50% of demand, is the most profitable train in comparison to others. Also, the profit obtained from trains types 3, 4, and 5 is almost equal. Although the price of the Type 1 train is lower than others, it can be considered as a profitable train for the company. In other words, not only is train type 1 a profitable train for the company, but it also is a desirable train for customers. The effect of travel history on profit is demonstrated in Figure 7. From mid-June, summer holidays begin in Iran, and train travel increases considerably. Since the Tehran–Mashhad route has high demand, with the start of the holidays, the volume of passengers increases significantly compared to previous months.

Another analysis is worth mentioning is related to the total income obtained from each OD pair (Fig. 8). In our proposed route, a significant percentage of passengers are traveled between Tehran and Mashhad without

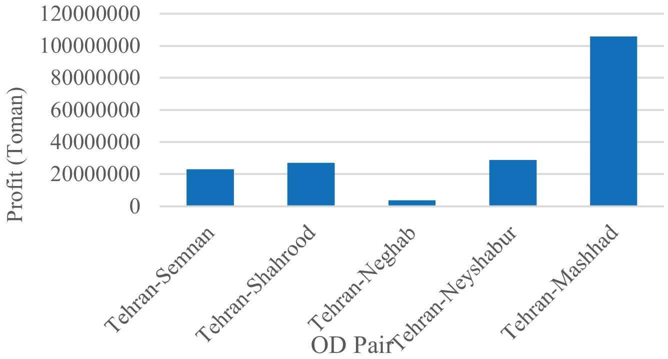


Figure 8. Profit obtained from each OD pair during a month.

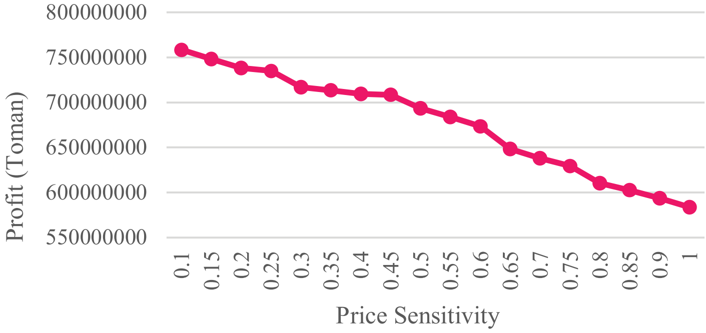


Figure 9. The profit changes as a result of variation the price sensitivity.

Table 6. The accepted demand (Person).

| | Train | | | | |
|---|---|---|---|---|---|
| Period | 1 | 2 | 3 | 4 | 5 |
| 1 | 8390 | 7873 | 7583 | 6734 | 6563 |
| 2 | 8393 | 7834 | 7564 | 6735 | 6486 |
| 3 | 8273 | 7764 | 7543 | 6543 | 6485 |
| 4 | 8317 | 7864 | 7485 | 6497 | 6452 |
| 5 | 8162 | 7654 | 7464 | 6422 | 6399 |
| 6 | 8134 | 7682 | 7564 | 6523 | 6453 |
| 7 | 8312 | 7964 | 7563 | 6742 | 6397 |
| 8 | 8374 | 7936 | 7612 | 6572 | 6563 |
| 9 | 8364 | 7756 | 7601 | 6653 | 6573 |
| 10 | 8389 | 7846 | 7536 | 6734 | 6582 |

stopping at middle stations. Accordingly, the revenue from selling tickets on this route is mostly due to non-stopping passengers. Because of this significant revenue, the railway company can form only non-stopping trains.

The other result obtained from our model is investigating the price sensitivity effect on the income fluctuation. Figure 9 shows the profit variation as a result of changing the coefficient related to price sensitivity. According to this figure, from 0.5 onwards, the profit has fallen more steeply, which indicates that the passenger with high

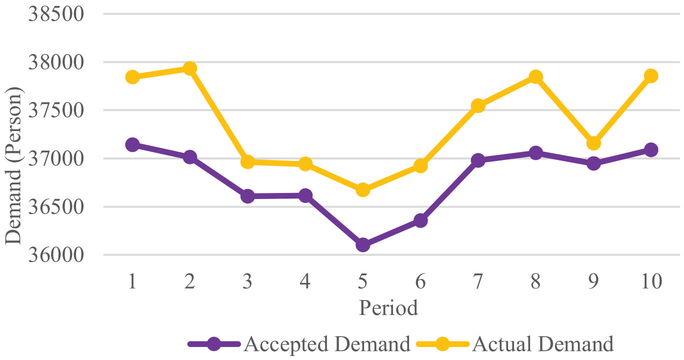


FIGURE 10. The difference between actual demand and accepted demand.

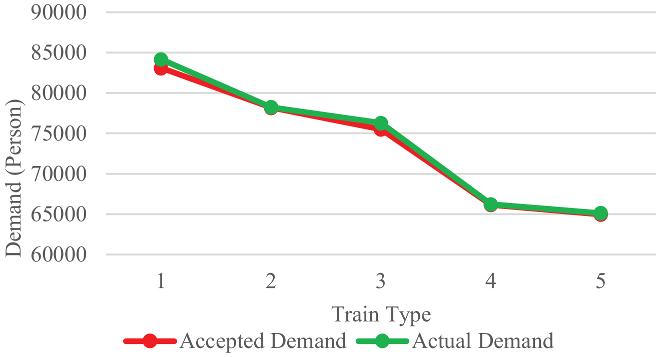


FIGURE 11. The difference between actual demand and accepted demand per train.

sensitivity of price has a significant impact on profit. Therefore, if a significant percentage of passengers are price sensitive, the company's profits will be greatly reduced as a result of price increases. Hence, it is necessary to consider a solution, which minimizes the impact of passengers' price sensitivity on the final profit.

Table 6 indicated the accepted demand $(X^t_{dr})$ for a ten-day period. Based on these results, the average demand for this route is approximately 37 000 people per day, most of which are related to the train type 1. In order to examine the demand more clearly, we demonstrated the difference between the actual and the accepted demand in Figure 10. Accordingly, the average distance between actual demand and accepted demand is about 1000 people. This difference may be due to the inability of the company to satisfy the demand or the significant impact of price-sensitive passengers. In order to determine the company's ability to satisfy the demand, we examined the difference between the actual demand and the accepted demand for each train in Figure 11. Since the difference between them is negligible, the number of price-sensitive passengers seems to be considerable, which led to a significant reduction in demand.

Based on the above analysis, it could be concluded that the company can allocate a significant percentage of its train schedule to train type 1 because of its notable profitability. Also, due to the considerable volume of passengers traveling between Tehran and Mashhad without stopping on middle stations, the company can use special facilities for these passengers or can plan specifically for this route.

### 7.1. Robust fuzzy stochastic programming model performance evaluation

In order to evaluate the performance of the deterministic and RFSP model, several numerical experiments are implemented and the related results are reported in this section. The four prominent values of trapezoidal

Table 7. Fuzzy parameters.

| Parameters | Trapezoidal fuzzy number |
|---|---|
| | $t=1$ |
| $p_{r=1\,\theta=1}$ | $(3050.24, 3100.55, 3250.92, 332.64)$ |
| $d_{r=1\,\theta=1}$ | $(1284, 1394, 1632, 1743)$ |
| $\gamma$ | $(0.15, 0.2, 0.26, 0.3)$ |
| $\delta_{r=1\,r'=1}$ | $(0.11, 0.14, 0.15, 0.17)$ |

fuzzy numbers used to indicate the imprecise parameters are randomly generated between the two extreme points of the corresponding possibility distribution function (*i.e.*, $\text{price}_{\text{real}} \sim [p_1 . p_4]$). Table 7 shows some of the fuzzy numbers generated for uncertain parameters. The weight of each term that is represented in the objective function is set based on the opinion of the experts and decision-makers (DMs). In this paper, we assume that the DM is non-conservative. Also, in order to determine the value of each term, we used the expert's opinions. The values of the coefficients are set as follows:

$$\psi = 0.5\,\gamma = \delta = 0.3. \tag{7.1}$$

Several numerical tests are performed to evaluate the performance of developed models and the corresponding results are reported below. To this aim, firstly, all the models are solved under nominal data. Then, to show the desirability and robustness of the derived solutions, ten random realizations are generated uniformly. Afterward, the obtained solutions under nominal data are replaced in the realization model similar to the model (7.2).

$$\begin{aligned} &\text{Max } Z = ay^* + b_{\text{real}}\,x^* - \sum_i \pi_i\left(\Psi_i^+ + \Psi_i^-\right) \\ &\text{s.t.} \\ &\qquad Cx^* \geq D_{\text{real}}\,y^* + \Psi_1^+ - \Psi_1^- \\ &\qquad Ex^* = F_{\text{real}}\,y^* + \Psi_2^+ - \Psi_2^- \\ &\qquad Gx^* \leq H_{\text{real}}\,y^* + \Psi_3^+ - \Psi_3^- \\ &\qquad x, y \geq 0. \end{aligned} \tag{7.2}$$

Table 8. The performance of the proposed models under realized data.

| No. of realization | RFSP | Deterministic model |
|---|---|---|
| 1 | 738 341 746 | 727 474 784 |
| 2 | 727 307 546 | 721 838 383 |
| 3 | 749 274 623 | 744 127 828 |
| 4 | 731 263 817 | 728 373 783 |
| 5 | 751 836 474 | 748 282 892 |
| 6 | 729 817 463 | 727 348 949 |
| 7 | 712 937 839 | 709 346 746 |
| 8 | 727 348 473 | 718 373 784 |
| 9 | 738 936 736 | 732 937 456 |
| 10 | 751 274 848 | 747 405 283 |
| Average | 735 833 956.5 | 730 550 988.8 |
| Semi variance | 11 350 166.73 | 12 193 373.04 |

In the realization model, we use the objective function of the definite model. Also, we used the constraints have the uncertainty parameters. In this model, $\Psi_i^+$ and $\Psi_i^-$ are the only decision variables that determine the violation of chance constraints under realization (see [27]). The $\pi$ represents the violation fine of constraints that apply when $\Psi^+$ takes a value. In the above model, $D_{\text{real}}$ represents the realization value of the parameter. Values such as $y^*$ marked with * represent the solution obtained by the models under nominal data. Other factors in the above model such as $C$ show the parameters of the problem.

The results obtained from Table 8 demonstrate the efficiency of the RFSP model and the advantage of using it. Based on Table 6, for all experiments, the values of the objective function (profit) are improved in the RFSP model in comparison to the deterministic one while semi-variance has decreased in the RFSP model. Therefore, using RFSP to deal with the uncertainty of demand improves the results. Given the above results, considering the RFSP model to deal with the uncertainty of sensitive parameters such as demand, obtained realistic and reliable results.

## 8. Conclusions

Given the importance of pricing in the railway system and its effect of it on passenger demand, this paper has addressed a framework that focuses on ticket pricing to maximize the revenue of selling the ticket. Due to the complexity of traditional pricing methods, we used regression-based machine learning approaches to obtain the optimal price. The result demonstrates that the accuracy of the fitted model is appropriate to predict the price and is able to combine with the optimization problem. After using a regression-based machine learning model to forecast the ticket pricing of the passenger railway, we applied the obtained price in the train formation problem. The results of the model's analysis show the efficiency of the developed framework in pricing with less complexity and proportionality of it for real and large problems. Because of the imprecise nature of the demand and weaknesses of traditional methods to obtain reliable output decisions, we used the hybrid uncertainty method named the robust fuzzy stochastic programming to handle the inherent uncertainty of this parameter. Using the real data inspired by the real case study, both the non-deterministic and deterministic models are solved and compared to each other. The results showed the efficiency of the result obtained from a robust fuzzy stochastic programming model.

As future works, in terms of the machine learning methods, taking into account the others technique such as neural networks could be considered in the railway ticket price prediction. Also, considering other variables such as the sales impact of other railway companies and government effect could be considered as an influential variable. In terms of the optimization methods, developing competitive models for considering the competition between railway companies, and using other alternative methods to maximize the companies' profit may be considered in the train formation problem. Create a balance between the customer's desired price and the company's satisfying price is also an important issue that should be considered in future studies.

## References

[1] J.A. Abdella, N.M. Zaki, K. Shuaib and F. Khan, Airline ticket price and demand prediction: a survey. *J. King Saud Univ. – Comput. Inf. Sci.* **33** (2021) 375–391.

[2] A.A. Ali, J. Warg and J. Eliasson, Pricing commercial train path requests based on societal costs. *Transp. Res. Part A* **132** (2020) 452–464.

[3] I. Belošević, S. Milinković, M. Ivić and P. Marton, Advanced evaluation of simultaneous train formation methods based on fuzzy compromise programing. *E3S Web Conf.* **135** (2019) 02026.

[4] P. Beria, S. Tolentino, A. Bertolin and G. Filippini, Long-distance rail prices in a competitive market. Evidence from head-on competition in Italy. *J. Rail Transp. Planning Manage.* **12** 100144.

[5] F. Branda, F. Marozzo and D. Talia, Ticket sales prediction and dynamic pricing strategies in public transport. *Big Data Cong. Comput.* **4** (2020) 36.

[6] T. But'ko and A. Prokhorchenko, Investigation into train flow system on Ukraine's railways with methods of complex network analysis. *Am. J. Ind. Eng.* **1** (2013) 41–45.

[7] C. Chen, T. Dollevoet and J. Zhao, One-block train formation in large-scale railway networks: an exact model and a tree-based decomposition algorithm. *Transp. Res. Part B Methodol.* **118** (2018) 1–30.

[8] L. Deng, Q. Zeng, W. Zhou and F. Shi, The effect of train formation length and service frequency on the determination of train schedules. *Proc. Inst. Mech. Eng. Part F J. Rail Rapid Transit* **228** (2014) 378–388.

[9] M. Farrokh, A. Azar, G. Jandaghi and E. Ahmadi, A novel robust fuzzy stochastic programming for closed loop supply chain network design under hybrid uncertainty. *Fuzzy Sets Syst.* **341** (2018) 69–91.

[10] M. FazliKhalaf, A. Mirzazadeh and M.S. Pishvaee, A robust fuzzy stochastic programming model for the design of a reliable green closed-loop supply chain network. *Human Ecol. Risk Assess. Int. J.* **23** (2017) 2119–2149.

[11] X. Gong, H. Wang and J. Zhu, Suboptimal pricing model and analysis of high-speed railway. *J. Interdisciplinary Math.* **20** (2017) 1203–1222.

[12] X. Gong, H. Wang and J. Zhu, Sub-time pricing model and effect analysis of high-speed railway. *J. Discrete Math. Sci. Cryptography* **20** (2017) 971–990.

[13] C. Gremm, The effect of intermodal competition on the pricing behavior of a railway company: evidence from the German case. *Res. Transp. Econ.* **72** (2018) 49–64.

[14] P. Hetrakul and C. Cirillo, A latent class choice-based model system for railway optimal pricing and seat allocation. *Transp. Res. Part E* **61** (2014) 68–83.

[15] Z. Jin, L. Yu-jing and L. Yan-Yang, Pricing model of train passenger transport based on the value of travel time and bi-level programming. In: 20th International Conference on Management Science & Engineering. Harbin, P.R. China (2013) 393–399.

[16] I. Kalathas and M. Papoutsidakis, Predictive maintenance using machine learning and data mining: a pioneer method implemented to Greek railways. *Designs* **5** (2021) 5.

[17] D. Kozachenko, V. Bobrovskiy, B. Gera, I. Skovron and A. Gorbova, An optimization method of the multi-group train formation at flat yards. *Int. J. Rail Transp.* **9** (2021) 61–78

[18] A. Kumar, A. Gupta and A. Mehra, A bi-level programming model for operative decisions on special trains: an Indian railways perspective. *J. Rail Transp. Planning Manage.* **8** (2018) 184–206.

[19] S.C.H. Leung, S.O.S. Tsang, W.L. Ng and Y. Wu, A robust optimization model for multi-site production planning problem in an uncertain environment. *Eur. J. Oper. Res.* **181** (2007) 224–238.

[20] B. Lin, Integrating car path optimization with train formation plan: a non-linear binary programming model and simulated annealing-based heuristics. *Optim. Control.* Preprint arXiv:1707.08326 (2017).

[21] B. Lin and Y. Zhao, The systematic optimization of train formation in loading stations. *Symmetry* **11** (2019) 1238.

[22] D.Y. Lin, J.H. Fang and L.K. Huang, Passenger assignment and pricing strategy for a passenger railway transportation system. *Transp. Lett. Int. J. Transp. Res.* **11** (2019) 320–331.

[23] B. Lin, F. Yang, S. Zuo, C. Liu, Y. Zhao and M. Yang, An optimization approach to the low-frequency entire train formation at the loading area. Sustainability **11** (2019) 5500.

[24] B.L. Lin, Y.N. Zhao, R.X. Lin and C. Liu, Integrating traffic routing optimization and train formation plan using simulated annealing algorithm. *Appl. Math. Modell.* **93** (2021) 811–830.

[25] Z. Mingbao, C. Ying, Z. Ning and Z. Xiaojun, Pricing of urban rail transit for different operation stages based on game theory. In: 2th IEEE International Conference on Information and Financial Engineering. Chongqing, China (2010) 17–19.

[26] N. Noordin and Mohd Ali Amran N.S. (2018) Optimizing Efficiency of Electric Train Service (ETS) ticket pricing. In: Proceedings of the Second International Conference on the Future of ASEAN (ICoFA). Singapore (2018) 381–391.

[27] M.S. Pishvaee, J. Razmi and S.A. Torabi, Robust possibilistic programming for socially responsible supply chain network design: a new approach. *Fuzzy Sets Syst.* **206** (2012) 1–20.

[28] M.S. Pishvaee, S.A. Torabi and J. Razmi, Credibility-based fuzzy mathematical programming model for green logistics design under uncertainty. *Comput. Ind. Eng.* **62** (2012) 624–32.

[29] F.R. Pratikto, A practical approach to revenue management in passenger train services: a case study of the Indonesian railways Argo Parahyangan. *J. Rail Transp. Planning Manage.* **13** (2020) 100161.

[30] M. Qin, Y. Li and G. Che, Railway passenger ticket pricing policy portfolio. In: International Conference on Logistics, Informatics and Service Sciences (LISS). Sydney, NSW, Australia (2016) 24–27.

[31] J. Qin, W. Qu, X. Wu and Y. Zeng, Deferential pricing strategies of high-speed railway based on prospect theory: an empirical study from China. *Sustainability* **11** (2019) 3804.

[32] D.B. Rubin, Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econ. Stat.* **4** (1986) 87–94.

[33] K. Sato and K. Sawaki, Dynamic pricing of high-speed rail with transport competition. *J. Revenue Pricing Manage.* **11** (2012) 548–559.

[34] C. Shearer, The CRISP-DM Model: the new blueprint for data mining. *J. Data Warehousing* **5** (2000) 13–22.

[35] M. Su, W. Luan and T. Sun, Effect of high-speed rail competition on airlines' intertemporal price strategies. *J. Air Transp. Manage.* **80** (2019) 101694.

[36] K.T. Talluri and G.J. Van Ryzin, The Theory and Practice of Revenue Management. Boston, Springer (2005).

[37] S. Tofighi, S.A. Torabi and S.A. Mansouri, Humanitarian logistics network design under mixed uncertainty. *Eur. J. Oper. Res.* **250** (2016) 239–250.

[38] V.A.C. Van den Berg and E.T. Verhoef, Congestion pricing in a road and rail network with heterogeneous values of time and schedule delay. *Transp. A Transp. Sci.* **10** (2014) 377–400.

[39] A. Vigren, Competition in Swedish passenger railway: entry in an open access market and its effect on prices. *Econ. Transp.* **11** (2017) 49–59.

[40] Y. Wang, Dynamic pricing considering strategic customers. In: 2016 International Conference on Logistics, Informatics and Service Sciences (LISS). Sydney, NSW (2016) 1–5.

[41] W. Wang, W. Xia, A. Zhang and Q. Zhang, Effects of train speed on airline demand and price: theory and empirical evidence from a natural experiment. *Transp. Res. Part B* **114** (2018) 99–130.

[42] H. Wu, J. Qin, W. Qu, Y. Zeng and S. Yang, Collaborative optimization of dynamic pricing and seat allocation for high-speed railways: an empirical study from China. *IEEE Access* **7** (2019) 139409–139419.

[43] J. Xiao and B. Lin, Comprehensive optimization of the one-block and two-block train formation plan. *J. Rail Transp. Planning Manage.* **6** (2016) 218–236.

[44] J. Xiao, B. Lin and J. Wang, Solving the train formation plan network problem of the single-block train and two-block train using a hybrid algorithm of genetic algorithm and tabu search. *Transp. Res. Part C: Emerg. Technol.* **86** (2018) 124–146.

[45] Z. Xiaoqiang, M. Lang and Z. Jin, Dynamic pricing for passenger groups of high-speed rail transportation. *J. Rail Transp. Planning Manage.* **6** (2017) 346–356.

[46] Z. Xueyu and Y. Jiaqi, Research on the bi-level programming model for ticket fare pricing of urban rail transit based on particle swarm optimization algorithm. *Proc. Soc. Behav. Sci.* **96** (2013) 633–642.

[47] M. Yaghini, M. Momani and M. Sarmadi, An improved local branching approach for train formation planning. *Appl. Math. Modell.* **37** (2013) 2300–2307.

[48] M. Yaghini, M. Momani and M. Sarmadi, Solving train formation problem using simulated annealing algorithm in a simplex framework. *J. Adv. Transp.* **48** (2014) 402–416.

[49] M. Yaghini, M. Momani and M. Sarmadi, A hybrid solution method for fuzzy train formation planning. *Appl. Soft Comput.* **31** (2015) 257–265.

[50] Z.Y. Yan, X.J. Li and B.M. Han, Collaborative optimization of resource capacity allocation and fare rate for high-speed railway passenger transport. *J. Rail Transp. Planning Manage.* **10** (2020) 23–33.

[51] C.W. Yang and C.C. Chang, Applying price and time differentiation to modeling cabin choice in high-speed rail. *Transp. Res. Part E* **47** (2011) 73–84.

[52] H. Yang and A. Zhang, Effects of high-speed rail and air transport competition on prices, profits and welfare. *Transp. Res. Part B* **46** (2012) 1322–1333.

[53] A. Yousefi and M.S. Pishvaee, A fuzzy optimization approach to integration of physical and financial flows in a global supply chain under exchange rate uncertainty. *Int. J. Fuzzy Syst.* **20** (2018) 2415–2439.

[54] M. Zamir Khan and F. Naheed Khan, Estimating the demand for rail freight transport in Pakistan: a time series analysis. *J. Rail Transp. Planning Manage.* **14** (2020) 100176.

[55] X. Zhang and L. Li, An integrated planning/pricing decision model for rail container transportation. *Int. J. Civil Eng.* **17** (2019) 1537–1546.

[56] R. Zhang, D. Johnson, W. Zhao and C. Nash, Competition of airline and high-speed rail in terms of price and frequency: empirical study from China. *Transp. Policy* **78** (2019) 8–18.

[57] Y.Q. Zhao, D.W. Li, Y.H. Yin, X.L. Dong and S.L. Zhang, Integrated optimization of train formation plan and rolling stock scheduling with multiple turnaround operations under uneven demand in an urban rail transit line. In: 23rd International Conference on Intelligent Transportation Systems (ITSC). Rhodes, Greece (2020) 1–6.