The following publication Wang, S., & Yan, R. (2023). Fundamental challenge and solution methods in prescriptive analytics for freight transportation. Transportation Research Part E: Logistics and Transportation Review, 169, 102966 is available at https://doi.org/10.1016/j.tre.2022.102966.

Fundamental challenge and solution methods in prescriptive analytics for freight transportation

Abstract

Prescriptive analytics, in which some parameters are predicted using statistical or machine learning models and then input into an optimization model, is often used to prescribe recommended solutions to freight transportation problems. The effectiveness of the optimal decision prescribed by prescriptive analytics is typically evaluated through a comparison with the results of the current decision model using predicted data. However, such comparisons are often flawed because of insufficient and uncertain data. We use four freight transport examples to illustrate this fundamental challenge in prescriptive analytics modeling. Furthermore, we propose three solutions to fully or partially overcome this challenge and fairly compare the optimal decisions generated by prescriptive analytics and the current approach. The three solutions involve using sufficient historical data, constructing new test sets, and generating synthetic data. We show how these solutions address the challenges in the four examples and are suitable for different problems considering data availability. The proposed solutions allow for a more comprehensive, accurate, and fair comparison of the optimal decisions to validate those generated by prescriptive analytics. This improves the effectiveness of the prescriptive analytics paradigm and can promote its application in freight transport and other disciplines.

Keywords

Prescriptive analytics; freight transportation; prediction; optimization; fundamental challenge

1. Introduction

As the quantity and diversity of available data are growing rapidly, modern business analytics is increasingly using historical data to optimize freight transportation decisions (Ng., 2015; Sun and Zheng, 2016; Wang et al., 2018; Qi et al., 2022; Hu et al., 2022). Addressing practical problems using business analytics involves three main types of analysis with various levels of difficulty, value, and data: descriptive analytics, predictive analytics, and prescriptive analytics (Lepenioti et al., 2020). Descriptive analytics uses statistical methods to explore what has happened and why it happened as well as what is happening and why it is happening. Predictive analytics uses statistical modeling and machine learning (ML) to predict what will happen and why. Prescriptive analytics, one of the most significant and sophisticated emerging technologies for business analytics, combines descriptive analytics, predictive analytics, and operations research to prescribe the optimal decision to maximize business value. The core of prescriptive analytics is identifying potential outcomes (decisions) using accurate predictions that incorporate artificial intelligence, operations research, and expert systems.

In freight transportation, descriptive analytics is the most basic and commonly used method to make sense of raw data using rudimentary mathematics. It uses data to describe, analyze, and summarize past and current situations. As a classic business intelligence tool, descriptive analytics is used to analyze shipments, freight generation and freight trip generation, accident/damage reports, customer behavior and feedback, and impacts on freight transportation and logistics processes in existing literature. For example, De Jong and Ben-Akiva (2007) proposed a new logistics model to analyze the national freight model system in Norway and Sweden on the firm-to-firm level which also simulated the choice of shipment size. Sakai et al. (2020) used empirical modelling approaches to identifying the effects of factors and heterogeneity on the shipment size selection mechanism. Holguín-Veras et al. (2014) summarized freight generation and freight trip generation models in existing studies, and the models included time series, trip rates, input-output, ordinary least squares, and spatial regression. Kuran et al. (2022) applied aggregated Accimaps to analyze whether adaptive nonconform behavior was uncovered in heavy goods transportation accident investigations conducted by the Norwegian Safety Investigation Authority. Esmaeeli et al. (2022) applied root cause analysis and bow tie analysis to train derailment database of Canada from 2007 to 2017 to identify the main causes and consequences of train derailment accidents. Oliveire et al. (2022) evaluated the influence of implementing a training and feedback procedure with event data recorder systems to promote the behavior of professional drivers using data envelopment analysis and concluded that such training could improve economic and operational results in road freight transport. The impact of off-peak delivery on greenhouse gases and air pollutant emissions in the Greater Toronto Area was analyzed by Saleh et al. (2022) in various scenarios, and it was found that off-peak delivery could improve network congestion and travel times but increase vehicle kilometers travelled.

Going a step further, predictive analytics leverages historical data to predict outcomes and effects, providing business leaders a proactive and data-driven approach to decision-making. Predictive analytics is widely used in freight transportation. Common examples on the strategy level include traffic assignment, modal shift analysis, supply chain network design, and scheduling and route design, and on the operations level, freight demand forecasting, travel time and delay prediction, failure and accident prediction, and energy efficiency and emission prediction and optimization (Wang et al., 2016). For example, on the strategy level, Hwang (2021) studied the procedures for freight truck shipment demand network assignment in the entire U.S. highway network, and a traffic assignment model was proposed to solve the freight truck shipment assignment problem. Wang et al. (2022) proposed an air traffic assignment framework for urban air mobility vehicle operations in 3D air transport networks by developing optimization models. Rosell et al. (2022) developed a non-linear integer programming model that jointly evaluated modal split and railway freight flows to analyze modal shift from road to rail in Europe. There are a number of existing studies on supply chain network design using operations research models, and some of them are combined with machine learning models. For example, Aboytes-Ojeda et al. (2022) proposed a bi-objective two-stage stochastic optimization model to design a biofuel supply chain network to decide the location of biorefineries and the flow between suppliers and biorefineries. Xiao et al. (2022) proposed a data-driven metaheuristic framework based on the combination of optimization and machine learning approaches, where the predictive power of machine learning models is leveraged by the optimization model to address a production system design problem. Similarly, there are many studies addressing the scheduling of freight transport services and designing of freight transport networks, which can be found on freight aircrafts (Bombelli and Fazi, 2022), electric commercial vehicles (Raeesi and Zografos, 2020), and sea-going cargo ships (El Noshokaty, 2020).

At the operational level, Yang (2015) proposed several regional freight transportation demand prediction models based on linear regression. Mahdavian et al. (2021) proposed a universal data-driven framework for truck traffic volume prediction using different machine learning models. Barbour et al. (2018) estimated the time of arrival (ETA) of individual freight trains using support vector regression. Yu et al. (2018) predicted vessels' ETAs to a container port and explored the value of such prediction on container terminal operation. Akgüngör and Doğan (2009) developed an artificial neural network model with genetic algorithm to predict traffic accident regarding the number, fatalities, and injuries of accidents in Ankara, Turkey. As green transportation becomes a main concern of various stakeholders in freight transport in recent years, there have been an increasing number of studies on energy efficiency prediction and optimization for all freight transportation modes, including but not limited to unmanned aerial vehicles in air transport (An et al., 2022), hydrogen electric multiple units trains in rail transport (Li et al., 2022), mining dump trucks in road transport (Siami-Irdemoosa and Dindarloo, 2015), and cruise ships in maritime transport (Wang et al., 2016). Based on the predictions of fuel consumption or energy efficiency, the corresponding emissions can be calculated, and the navigation statuses of the vehicles can be optimized to increase their energy efficiency.

Prescriptive analytics builds on descriptive and predictive analytics with the aim of going from providing a good prediction to providing a good decision, enhancing business value. The best decision is prescribed based on predictions. Therefore, prescriptive analytics includes both prediction and optimization models. In words, a prediction model (usually powered by ML techniques) is used to predict the unknown parameters in the optimization model, which is used to generate optimal decisions. Many freight transportation problems are solved by prescriptive analytics in everyday practice. For example, Liu et al. (2020) proposed a framework integrating travel-time predictors with order-assignment optimization, in which a prediction model predicts the total travel time given a set of orders, and the predictions are then used to optimize order-assignment decisions. Prescriptive analytics is also commonly used to prescribe daily decisions in other logistics processes. For example, Ferreira et al. (2016) studied how a retailer can use historical sales data to price a product it has never sold before. They used ML models to predict historical lost sales and future demand for a new product and then optimize the pricing decision using those predictions. In cooperation with Zara (a clothing chain store), Gallien et al. (2015) developed and tested a decision support system containing a data-driven model of demand forecast updating and a dynamic optimization formulation to allocate limited stock by location.

One challenge of prescriptive analytics is to rigorously show that the optimal decision prescribed is indeed better than that derived using the current approach (i.e., to make a fair comparison). Ideally, if decision makers are willing to try innovative prescriptive analytics approaches, the comparison is much easier. For example, each approach can be used in turn to solve the same problem for a certain period (e.g., the current approach is applied in weeks 1, 4, 5, 8, 9, ..., and the prescriptive analytics approach is applied in weeks 2, 3, 6, 7, 10, 11, ...), and then their average performances are compared. Alternatively, each approach can be applied for a future period of two similar business scenarios (e.g., two similar branches of a company or two sister ships serving the same route) for which the same strategies have been used. Then, changes in projected profits or costs compared with historic profits or costs can be used to verify the superiority of the prescriptive analytics approach. If the decisions derived via the learned prescriptive analytics are indeed better, decision makers should adopt those instead of the decisions suggested by the current approach. However, a fair comparison can only be made if decision makers are willing to try the learned prescriptive analytics for long enough to gather sufficient data. In reality, as the loss from a failed trial can be significant, conservative decision makers may not be willing to try prescriptive analytics unless they are sure that it is superior to the current approach.

Thus, strong evidence showing that the learned prescriptive analytics is indeed better than the current approach is a prerequisite for its adoption. In academic studies, this is mainly achieved by comparing the average predicted benefit/cost of the optimal decision given by prescriptive analytics and the average actual cost of the decision given by the current approach. However, a lack of sufficient data and data uncertainties result in unfair comparisons, and thus the results lack persuasiveness. Consequently, a dead cycle occurs: researchers cannot provide strong evidence that prescriptive analytics is better than the current approach, so industry practitioners do not adopt it; because industry practitioners do not adopt prescriptive analytics, it is difficult to show its superiority.

The contribution of this study is that it is the pioneer that innovatively defines the difficulty in comparing the optimal decisions given by prescriptive analytics and the current approach as a fundamental challenge and describes it within the context of freight transport using four concrete examples in a systematic and rigorous way. Then, a total of three solutions, where one is a full solution by finding sufficient historical data and the other two are partial solutions based on formulating test sets by sampling and generating synthetic data, are proposed to overcome the fundamental challenge. Case study and computational experiments are conducted to validate the existence of the fundamental challenge and the proposed solution approaches, verifying the applicability and effectiveness of these solutions. In addition to the four examples in freight transport, the proposed solutions are expected to have a wider range of applications in other problems in freight transport as long as both prediction and optimization are needed. If the fundamental challenge in the application of prescriptive analytics can be addressed by the proposed solutions, it can be expected that decision makers in freight transport will be more willing to accept and use them.

2. Problem description

Consider a freight transportation decision problem that aims to minimize total costs and is repeated periodically, e.g., every day. On day t, the problem can be formulated as follows:

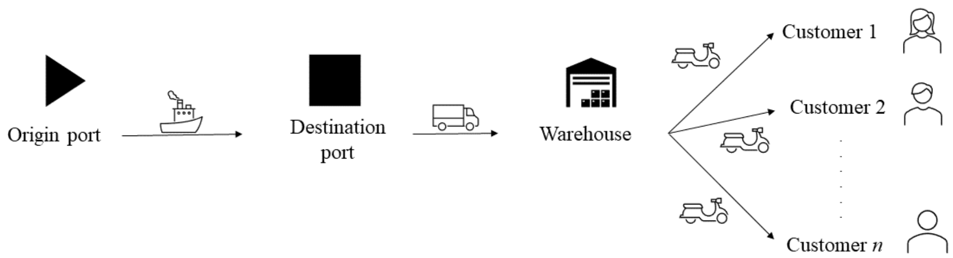
$$\min_{x\in X^{(t)}} \sum_{k=1}^{K} c_k(x_k, u_k^{(t)}) \tag{1}$$

where x_k is the decision variable for the kth component of total costs, $k=1,...,K$, $x=(x_1,...,x_K)$; $X^{(t)}$ accounts for the domain of x that is not influenced by uncertainty; and $c_k(x_k, u_k^{(t)})$ is the cost function of decision x_k and its associated features $u_k^{(t)}$ (e.g., $u_k^{(t)}$ can include information such as whether day t is a workday and the weather on day t). A challenge frequently encountered in solving model (1) is that $c_k(x_k, u_k^{(t)})$ is unknown a priori (e.g., a decision regarding a future period) and thus is predicted using an ML model. Specifically, suppose that decisions are made on day $T+1$. Thus, the decision problem is encountered on days $1,...,T$. On each day $t=1,...,T$, the decision maker records the features of $u_k^{(t)}$, makes decision $x^{\#(t)}$ using an existing method (e.g., based on experience or heuristic rules), and observes the actual costs $c_k(x_k^{\#(t)}, u_k^{(t)})$, $k=1,...,K$. We can thus take advantage of the dataset

$(x_k^{\#(t)}, u_k^{(t)}, c_k(x_k^{\#(t)}, u_k^{(t)}), k = 1, \ldots, K)$, $t = 1, \ldots, T$, to develop an ML model that predicts the cost for any $(x_k, u_k^{(t)})$, which we denote it by $\hat{c}_k(x_k, u_k^{(t)})$ for all $x_k \in X^{(t)}$, and all possible $u_k^{(t)}$ of concern. Then, to make a decision on day $T+1$, we solve the following model:

$$\min_{x \in X^{(T+1)}} \sum_{k=1}^{K} \hat{c}_k(x_k, u_k^{(T+1)}), \tag{2}$$

where $X^{(T+1)}$ accounts for the decision for day $T+1$ and $u_k^{(T+1)}$ is the associated features. This decision problem is quite common in freight transportation. Suppose that a product is transported from a port in the production country to a customer in another country. This decision problem might be solved several times in the multimodal freight transport process as illustrated in Figure 1 and elaborated below.



Problem	Vessel sailing speed optimization	Ship inspection by port state control	Truck route optimization problem	Last-mile delivery
Process	From the origin port to the destination port	At the destination port	From the destination port to the warehouse	From the warehouse to the customers
Objective	Minimize vessel fuel consumption over the voyage	Maximize the noncompliance found on vessels	Minimize the total travel time	Minimize the total travel distance of the drivers
Features	Sea and weather conditions	Ship characteristics and historical inspection results	Temperature, rainfall, and road condition, etc.	The number of locations, the joint effects of visiting area and the number of locations, and the distances between the depot and customer locations, etc.
Decision variables	Sailing speed between each two consecutive ports	The set of visiting ships to inspect	The route from the port to the warehouse	How to assign the customers to the drivers

Figure 1. An illustration of the decision problems in the multimodal freight transport process

Example 1. Vessel sailing speed optimization

Suppose that a product is transported by a bulk carrier ship from the origin port to the destination port where the whole voyage can be divided into K segments. The period of each segment is one day, and thus the sea and weather conditions can be seen as identical in one segment. Vessel sailing speed x_k on each segment $k = 1, \ldots, K$ is to be optimized (Wang and Meng, 2012; Yan et al., 2020). Domain $X^{(t)}$ contains information on the maximum and minimum ship speeds and the total sailing time available. Fuel consumption $c_k(x_k, u_k^{(t)})$ on segment k is related to speed x_k and the surrounding sea and weather conditions $u_k^{(t)}$, but the exact functional form is unknown because of the complexity of hydrodynamics (Zis et al., 2015; Psaraftis et al., 2019; Zis et al., 2020; Zis and Cullinane, 2020; Tillig et al., 2020).

To be more specific, for its $(T+1)$th voyage, the bulk carrier ship starts a voyage consisting of K segments from the origin port to the destination port and the sailing speed on each of the K segments needs to be planned. The noon report of the ship for a past period is available, which contains a total of T voyages and each voyage has a total of K_t records. One record in ship noon report corresponds to one day's ship sailing information, including the ship's geographic location at the recording time (usually at noon), distance travelled since last report, total sailing time, average sailing speed, hourly/daily fuel consumption rate, total hold cargo, etc., and the surrounding sea and weather conditions such as sea and swell directions, sea current type, wind force and direction, and sea water temperature (Zis and Psaraftis, 2021). Then, we denote $(x_k^{\#(t)}, u_k^{(t)}, c_k(x_k^{\#(t)}, u_k^{(t)}), k=1,...,K_t)$, $t=1,...,T$ by the historical dataset on ship sailing, where $x_k^{\#(t)}$ is the average sailing speed on segment k of voyage t, $u_k^{(t)}$ is the sea and weather conditions on segment k of voyage t, and $c_k(x_k^{\#(t)}, u_k^{(t)})$ is the total fuel consumption consumed on segment k of voyage t. The total number of historical noon report records is $K_1+...+K_T$. Then, these records are used as the training set to develop an ML model to predict the ship's hourly fuel consumption rate based on ship sailing behavior and the sea and weather conditions on each segment for voyage $T+1$, which can then be converted into the total fuel consumption over the segment by multiplying the hourly fuel consumption rate by the sailing time on the segment and is presented by $\hat{c}_k(x_k, u_k^{(T+1)})$, $k=1,...,K$.

Further denote the length of segment k by L_k (unit: nautical mile [nm]), the earliest and latest allowable arrival time to the destination port by $T_{\min}$ and $T_{\max}$ (unit: hour), the minimum and maximum allowable sailing speeds on segment k by $x_{k,u_k}^{\min}$ and $x_{k,u_k}^{\max}$ (unit: knot), and the ship sailing speed before departure by x_0 (unit: knot). The main decision variable is x_k, which is the sailing speed on segment k (unit: knot), and the auxiliary decision variable is t_k, which is the arrival time to the beginning of segment k (unit: hour), $k=1,\dots,K$. Then, the sailing speed optimization problem for the next voyage containing K segments for voyage $T+1$ can be formulated by mathematical model M1 as follows:

[M1] $$\min \sum_{k=1}^{K} \hat{c}_k(x_k, u_k^{(T+1)}) \tag{3}$$

s.t.

$$t_{k+1} = t_k + \frac{L_k}{x_k},\ k=1,\dots,K \tag{4}$$

$$T_{\min} \le t_{K+1} \le T_{\max} \tag{5}$$

$$x_0 = 0 \tag{6}$$

$$x_{k,u_k}^{\min} \le x_k \le x_{k,u_k}^{\max},\ k=1,\dots,K \tag{7}$$

$$t_k \ge 0,\ k=1,\dots,K+1. \tag{8}$$

Objective function (3) minimizes the fuel consumption over the voyage, where $\hat{c}_k(x_k, u_k^{(T+1)})$ is the predicted fuel consumption on segment k. Constraints (4) specify the arrival time to each segment, and constraint (5) limits the arrival time to the destination port is no later than $T_{\max}$ and is no earlier than $T_{\min}$. Constraints (6) and (7) impose limits on the sailing speeds on each segment. Constraints (8) guarantee that the arrival time to the beginning of each segment is nonnegative.

Example 2. Ship inspection by port state control (PSC)

When the bulk carrier ship arrives at the destination port, it might be subject to a PSC inspection, the purpose of which is to ensure maritime safety and protect the marine environment by inspecting foreign visiting ships and identifying their noncompliance with international regulations and conventions conducted by the national port authority. Such noncompliance is recorded as a deficiency. A critical problem is whether to inspect a foreign visiting ship given the port authority's limited inspection resources (Yan et al., 2021). Suppose that at most K ships are candidates for inspection (i.e., K is the total number of foreign visiting ships to the port) each day. On day t, $M^{(t)}$ candidate ships are subject to inspection (i.e., $M^{(t)}$ is the number of ships that can be inspected according to the port authority's ship selection rules); $M^{(t)} \le K$. $N^{(t)}$ ships are inspected (because of the limited inspection resources of the port authority); $N^{(t)} \le M^{(t)}$. Then, x_k is a binary decision variable that is one if ship k is inspected and zero otherwise; domain $X^{(t)}$ requires that $N^{(t)}$ ships are inspected and that ships $M^{(t)}+1,...,K$ are not inspected as they do not exist; $c_k(x_k,u_k^{(t)})$ accounts for some condition of ship k (e.g., a negative value of the number of deficiencies) when $x_k=1$ and is equal to zero when $x_k=0$, where $u_k^{(t)}$ is the set of features used to predict a ship's condition, such as ship characteristics and historical inspection results. Note that in this example, we only need one ML model (instead of K models) to predict one function: the condition of the ship given its features.

To be more specific, suppose our goal is to maximize the total number of deficiencies that can be detected from the inspected ships, which is also equivalent to minimizing the negative value of the total number of deficiencies detected. At a port authority, the process of high-risk ship selection and inspection is conducted on each working day, and a certain number of inspection records for the past period can be accumulated. Each inspection record includes the inspected ship's characteristics, historical inspection performance, performance of management parties, and its performance (i.e., deficiency and detention conditions) in the current inspection. Then, an ML model can be constructed using ship related information from the inspection records as the input and the negative ship deficiency number as the output. Then, on day $T+1$, the total number of foreign visiting ships is K. $M^{(T+1)}$ ships are candidate for inspection, and the available inspection resources at port allow up to $N^{(T+1)}$ ships to be inspected, $N^{(T+1)} \le M^{(T+1)}$. Denote the predicted negative deficiency number of ship k by the ML model by $\hat{e}_k$, and thus $\hat{c}_k(x_k,u_k^{(t)}) = \hat{e}_k \times x_k$. One day $T+1$, the ship selection planning problem for the port authority can be formulated by mathematical model M2 as follows:

[M2] $$\min \sum_{k=1}^{K} \hat{c}_k(x_k, u_k^{(T+1)}) \quad (9)$$

s.t.

$$\sum_{k=1}^{K} x_k \le N^{(T+1)} \quad (10)$$

$$x_k \in \{0,1\},\ k=1,...,K. \quad (11)$$

Objective function (9) minimizes the sum of the negative total number of deficiencies detected from the inspected ships. Constraint (10) guarantees that at most

$N^{(T+1)}$ ships are inspected by the port. Constraints (11) ensure the domain of the decision variable.

Example 3. Truck route optimization problem

After an imported product is unloaded from the ship onto a truck, the truck transports the product to a warehouse. There are several candidate routes for the truck to choose from to minimize the total travel time. Suppose that the number of links (i.e., road segments) in the road network is K. Then, x_k is a binary decision variable that equals one if link k, $k=1,...,K$, is traversed by the truck and zero otherwise; domain $X^{(t)}$ requires that x_ks define a route from the port to the warehouse; $c_k(x_k, u_k^{(t)})$ is the travel time on link k when $x_k = 1$ and zero when $x_k = 0$, where $u_k^{(t)}$ is the set of factors that influence the travel time, such as temperature, precipitation, and road conditions.

For the sake of simplicity, further denote the set of nodes in the road network by $N := \{1, 2, ..., n\}$ where 1 is the origin and n is the destination. Denote the set of links in the road network by $A := \{(i, j), i \in N, j \in N, i \neq j\}$, whose travel time is $c_{ij}(x_{ij}, u_{ij}^{(t)})$, $t = 1,...,T$ when $x_{ij} = 1$ and zero when $x_{ij} = 0$, and $u_{ij}^{(t)}$ accounts for the condition of the link as well as the surrounding environment. Then, on day $T+1$, the predicted travel time on segment $(i, j) \in A$ is $\hat{c}_{ij}(x_{ij}, u_{ij}^{(T+1)})$, and the truck route optimization problem aiming to minimize the total travel time can be formulated by mathematical model M3 as follows:

[M3]

$$\min \sum_{(i,j)\in A} \hat{c}_{ij}(x_{ij}, u_{ij}^{(T+1)}) \tag{12}$$

s.t.

$$\sum_{j\in N\setminus\{1\}} x_{1j} = 1 \tag{13}$$

$$\sum_{i\in N\setminus\{n\}} x_{in} = 1 \tag{14}$$

$$\sum_{j\in N\setminus\{i\}} x_{ij} = \sum_{j\in N\setminus\{i\}} x_{ji},\ i \in N \setminus \{1, n\} \tag{15}$$

$$x_{ij} \in \{0,1\},\ (i, j) \in A. \tag{16}$$

Objective function (12) minimizes the predicted total travel time of the truck. Constraints (13) to (15) are the flow conservation of the origin node, destination node, and internal nodes, respectively. Constraints (16) specify the domain of decision variable x_{ij}.

Example 4. Last-mile delivery

When customers order the product, it must be delivered to them from the warehouse (Pani et al., 2022). This is basically a vehicle routing problem (VRP). Using actual delivery data, Liu et al. (2021) found that drivers do not follow the sequence obtained by solving the VRP; instead, they visit customers in a sequence derived from their experience, and as a result, their actual travel distance differs from what the VRP model calculates[1], jeopardizing the efficiency of the customer assignments provided by solving the VRP model. Here, x_k refers to the set of customers to be visited by vehicle $k=1,...,K$; domain $X^{(t)}$ requires the maximum number of customers in one set and that each customer is visited by exactly one vehicle; we do not consider the effect of $u_k^{(t)}$, and thus we set $u_k^{(t)}=0$. $c_k(x_k,u_k^{(t)}=0)$ is the actual travel distance of driver k to serve the set of customers x_k assigned to the driver. As we do not know how driver k determines the sequence of customers to visit, we do not know $c_k(x_k,0)$. To develop an ML model $\hat{c}_k(x_k,0)$ to estimate $c_k(x_k,0)$, we use factors derived from x_k such as the number of locations, the joint effect of visiting area and the number of locations, and the average/shortest/longest distance between the depot and customer locations, and the actual total travel distance, over each of the past T days. With $\hat{c}_k(x_k,0)$ as the predicted travel distance of vehicle $k=1,...,K$, we can optimize the assignment of customers to drivers using a model that differs from the traditional VRP model.

To be more specific, denote the number of customers by I and one customer by $i=1,...,I$. At most C customers can be assigned to each vehicle. Decision variable $x_k^i\in\{0,1\}$ is further introduced which is one if customer i is assigned to vehicle k

[1] Liu et al. (2021) examines the travel time instead of travel distance. To avoid confusion with Example 3, we use travel distance in Example 4.

and zero, otherwise. The last-mile delivery problem can be formulated by mathematical model M4 as follows:

[M4]

$$\min \sum_{k=1}^{K} \hat{c}_k(x_k, 0) \tag{17}$$

$$\sum_{i=1}^{I} x_k^i \le C,\ k = 1,\dots,K \tag{18}$$

$$\sum_{k=1}^{K} x_k^i = 1,\ i = 1,\dots,I \tag{19}$$

$$x_k^i \in \{0,1\},\ i = 1,\dots,I,\ k = 1,\dots,K\,. \tag{20}$$

Objective function (17) minimizes the predicted total travel time of all the K vehicles. Constraints (18) guarantee that the total number of customers assigned to vehicle k is no more than C. Constraints (19) ensure that each customer i is served by exactly one vehicle. Constraints (20) specify the domain of decision variable x_k^i.

3. Fundamental challenge in prescriptive analytics

To compare the performance of the decisions given by the current approach and by prescriptive analytics, the following method is commonly adopted in the literature. Divide the past T days' data into two sets: the first T_1 days' data are the training set, and the last T_2 days' data are the test set (e.g., $T_1 = 0.8T$ and $T_2 = 0.2T$). Using the training set, an ML model $\hat{c}_k(x_k, u_k)$, is developed to predict $c_k(x_k, u_k)$. Then, the following surrogate models are solved:

$$\min_{x \in X^{(t)}} \sum_{k=1}^{K} \hat{c}_k(x_k, u_k^{(t)}),\ t = T_1 + 1,\dots,T\,, \tag{21}$$

and the resulting solution for day $t = T_1 + 1,\dots,T$ is $x_k^{*(t)}$. That is, $x^{*(t)}$ is the solution if prescriptive analytics were used for making decisions on day $t = T_1 + 1,\dots,T$ (recall that $x^{\#(t)}$ is the actual decision made on day $t = T_1 + 1,\dots,T$). The prescriptive analytics paradigm is considered superior to the current decision-making approach if

$$\frac{1}{T - T_1} \sum_{t=T_1+1}^{T} \sum_{k=1}^{K} \hat{c}_k(x_k^{*(t)}, u_k^{(t)}) < \frac{1}{T - T_1} \sum_{t=T_1+1}^{T} \sum_{k=1}^{K} c_k(x_k^{\#(t)}, u_k^{(t)}) \tag{22}$$

where the left side is the average daily predicted cost if the decisions given by prescriptive analytics were implemented, and the right side is the average daily actual cost of the decisions given by the current approach.

However, this comparison can be flawed, as illustrated by a simple example. Suppose that the actual but unknown function is $c_k(x_k, u_k^{(t)}) = |x_k|$, $-10 \le x_k \le 10$, and $k = 1 = K$. $u_k^{(t)}$ has no effect, and the current decision-making approach is to randomly choose an x_k from $\{0,1\}$. The training dataset has two such examples $(x_k^{\#(1)}, u_k^{(1)}, c_k(x_k^{\#(1)}, u_k^{(1)})) = (0,0,0)$ and $(x_k^{\#(2)}, u_k^{(2)}, c_k(x_k^{\#(2)}, u_k^{(2)})) = (1,0,1)$, and the test dataset has one such example $(x_k^{\#(3)}, u_k^{(3)}, c_k(x_k^{\#(3)}, u_k^{(3)})) = (0,0,0)$. Further suppose that the ML model is a linear regression model. Then, $\hat{c}_k(x_k, u_k^{(t)}) = x_k$, and using prescriptive analytics for day 3, we solve $\min_{-10 \le x_k \le 10} \hat{c}_k(x_k, u_k^{(3)})$ and have $x_k^{*(3)} = -10$. As a result, the left side of Eq. (4) is -10, the right side is 0, and we wrongly conclude that the prescriptive analytics paradigm is superior to the current decision-making approach. An intuitive explanation for the error is that only the real realization corresponding to the decision $x_k^{\#(3)} = 0$ is revealed, while its counterfactuals, i.e., when $x_k^{\#(3)} \ne 0$, cannot be observed. Then, $x^{*(t)}$ can only be derived using the (inaccurate) predicted function $\hat{c}_k(x_k, u_k^{(t)})$ which can be distinct from the real cost function, and of course, it performs well on $\hat{c}_k(x_k, u_k^{(t)})$; however, it may not work as well on the actual (but unknown) function $c_k(x_k, u_k^{(t)})$. Hence, the comparison in Eq. (4) is unfair. We consider this a fundamental challenge in prescriptive analytics, which is formally presented as follows.

Fundamental challenge in prescriptive analytics: Whereas in principle $\frac{1}{T_2}\sum_{t=T_1+1}^{T}\sum_{k=1}^{K} c_k(x_k^{*(t)}, u_k^{(t)})$ and $\frac{1}{T_2}\sum_{t=T_1+1}^{T}\sum_{k=1}^{K} c_k(x_k^{\#(t)}, u_k^{(t)})$ should be compared, in many cases $\frac{1}{T_2}\sum_{t=T_1+1}^{T}\sum_{k=1}^{K} \hat{c}_k(x_k^{*(t)}, u_k^{(t)})$ and $\frac{1}{T_2}\sum_{t=T_1+1}^{T}\sum_{k=1}^{K} c_k(x_k^{\#(t)}, u_k^{(t)})$ are actually compared because the value of $c_k(x_k^{*(t)}, u_k^{(t)})$ is unrevealed, $k = 1,...,K$, $t = T_1+1,...,T$, resulting in an unfair comparison.

This fundamental challenge is frequently encountered in practice. For example, in the ship fuel consumption prediction and speed optimization problem in Example 1, the "optimal" sailing speeds that minimize total fuel consumption are generated using an imperfect ML model, resulting in an inaccurate total fuel consumption figure, while the ground-truth fuel consumption given the "optimal" sailing speed (i.e., the counterfactuals of the real sailing speeds) is unknown. As no information on actual fuel consumption of the speed after optimization is available, the comparison with actual fuel consumption is unfair, and the results are invalid.

4. Solutions to the fundamental challenge

We propose three solutions that fully or partially address the fundamental challenge by, respectively, searching sufficient historical data, constructing new test sets, and generating synthetic data considering the differences in the accessibility and the costs of obtaining historical data. Detailed descriptions and examples of each solution are provided in Table 1. Each solution only addresses a subset of the problems related to prescriptive analytics.

Table 1. Summary of the solutions to the fundamental challenge

Solution	Characteristics
Solution 1	Finding sufficient historical data if they are accessible: We can find a way to know $c_k(x_k, u_k^{(t)})$, $k = 1,...,K$, for any $(x_1,...,x_K) \in X^{(t)}$, for days $t = 1,...,T$, but we still do not know $c_k(x_k, u_k^{(T+1)})$. Then, we can fairly compare prescriptive analytics and the current decision-making approach to overcome the fundamental challenge.
Solution 2	Constructing test sets where the costs of all the solutions to the examples are available using historical data: Using the test data obtained over days $T_1 + 1,...,T$, we can find a way to construct T_2 new test sets such that the actual cost of any solution is known for each new test example. Then, we can compare the results of prescriptive analytics and the current decision-making approach for the T_2 new test examples, partially overcoming the fundamental challenge.
Solution 3	Formulating synthetic test set where the costs of all the solutions to the examples are available based on domain knowledge: We can find the known function $d_k(x_k, u_k^{(t)})$ (which complies with practical decision making process) that can approximate $c_k(x_k, u_k^{(t)})$, $k = 1,...,K$. Using $d_k(x_k, u_k^{(t)})$ in place of $c_k(x_k, u_k^{(t)})$, we then generate synthetic data over days $1,...,T$. Using the synthetic data on days $1,...,T_1$, we train an ML model $\hat{d}_k(x_k, u_k)$ that predicts the value of $d_k(x_k, u_k)$. Then, we can compare the solution obtained from prescriptive analytics using $\hat{d}_k(x_k, u_k)$, whose actual cost is known, with the solution obtained from the current decision-making approach for the T_2 new test examples, partially overcoming the fundamental challenge.

The rationale behind the three solutions is the differences in practical problems regarding their accessibility of sufficient historical data and the costs of acquiring such historical data. If sufficient historical data are available and the costs are affordable, solution 1, which is a full solution based on historical data obtaining, is preferred. Unfortunately, in most cases, it is impossible or too costly to access sufficient historical data, and what we have on hand is a certain (limited) amount of historical data. If the components of the decision to be made for a certain day are not related to each other, then, solution 2 based on sampling approaches can be applied to construct more historical data sets to mimic practical decision-making scenarios, and then the comparison between the decisions recommended by prescriptive analytics and the current decision approach can be fairly compared. Another case is that the historical data are limited, while the decisions to be made for each of the components for a certain day are highly related with each other. For example, in the ship fuel consumption prediction and sailing speed optimization problem, the actual arrival time to the destination port should be within the earliest and the latest allowable arrival time, and thus the sailing speed in each segment is highly related to each other. Furthermore, the sea and weather conditions are distinct on each segment of the voyage, and thus sampling cannot be directly applied, as the original relationship and property of the components may not be preserved after sampling, causing the decision invalid. Then, we propose another solution, i.e., solution 3, based on generating synthetic data by following decision rules adopted by decision makers in practice.

Based on the above explanation, we argue that which solution is more suitable for a certain practical problem in freight transport depends on the availability and the costs of acquiring historical data: If sufficient historical data can be accessed and the cost is affordable, solution 1 might be the most suitable; if certain historical data can be accessed and the components of the decision on a certain day are not correlated, solution 2 could be applied; alternatively, if certain historical data are accessible while the components of the decision to be made on a certain day are highly correlated with other, solution 3 could be applied. The detailed description and examples given to the three solutions are presented in the following paragraphs.

Solution 1 (full solution): If sufficient historical data on the cost of the candidate decisions can be obtained, the fundamental challenge can be overcome. For instance, in Example 3, even if many links in the road network are not traversed by the truck or other trucks belonging to the logistics company, their actual travel time in the past T days can be obtained from a third party, such as Google Maps or Uber, and then the value of $c_k(x_k^{*(t)}, u_k^{(t)})$ is known, $k = 1,...,K$, $t = T_1 + 1,...,T$, allowing for a fair comparison between prescriptive analytics and the current decision-making approach. For example, given any pair of origin and destination, Google Maps can use big historical data collected from individual phones to estimate the movement and speed of traffic under different scenarios (e.g., in peak or non-peak periods and on rainy or sunny days) and consider posted speed limits and historical traffic patterns to estimate the travel time between the origin and destination (Verizon Connect, 2020). For example, the Distance Matrix API provided by the Google Maps Platform gives the travelling duration and distance of each pair of origin and destination for different travel modes, together with the request distance data in different units such as kilometers and miles and the estimated travel time in traffic[2]. In this way, accurate and sufficient historical data can be obtained at a low cost to generate the full solution.

Solution 2 (partial solution): In most cases, it is impossible (because of policy provisions and confidentiality agreements) or too costly to obtain the actual cost of each candidate decision. In words, the costs are known for only a fraction of the candidate decisions. Consequently, Solution 1 is not applicable. Alternatively, new test sets could be formulated by adopting sampling approaches that simulate actual operations, such that the actual cost of each candidate decision is known. For example, in Example 2, it is impossible to inspect all $M^{(t)}$ foreign ships on day t because of the high inspection costs (the average inspection without a deficiency costs the port US$506, and if a deficiency is found, the average cost is US$759; Knapp, 2007) and limited inspection resources (there are usually no more than three inspectors on duty each day, and an inspection takes at least 1.5 hours, not including the travel time between ships). Consequently, $N^{(t)}$ ships are inspected by the port on day t, $t = 1,...,T$, and the actual conditions of the ships are available. Suppose that an ML model for ship condition prediction is constructed using inspection records for the first T_1 days, and we have $\sum_{t=T_1+1}^{T} N^{(t)}$ inspected ships for the next T_2 days for testing purposes. Suppose further that there are $M^{(t)}$ candidate ships subject to inspection each day $t = T_1 + 1,...,T$, and the total number of candidate ships during the next T_2 days is

[2] https://developers.google.com/maps/documentation/distance-matrix/overview

$\sum_{t=T_1+1}^{T} M^{(t)}$ and $\sum_{t=T_1+1}^{T} N^{(t)} < \sum_{t=T_1+1}^{T} M^{(t)}$. Then, we do not apply prescriptive analytics to the selection of $N^{(t)}$ ships from the $M^{(t)}$ candidate ships each day $t = T_1+1,...,T$ (because, e.g., if $N^{(t)} = 2$, the current decision-making approach inspects ships D and B and prescriptive analytics recommends inspecting ships D and F, then we cannot fairly compare the two methods as we do not know the actual condition of ship F [which is a counterfactual of the actual decision]). Instead, we construct new test data as follows. Of the $\sum_{t=T_1+1}^{T} N^{(t)}$ ships inspected on days $T_1+1,...,T$, we randomly group $M^{(t)}$ ships in one sample (i.e., a new test set) and perform the sampling process T_2 times (to construct T_2 new test sets). We assume that on day $t = T_1+1,...,T$, ships in the $(t-T_1)$ th sample are candidates for inspection. We apply both prescriptive analytics and the current decision-making approach to the problem on day $t = T_1+1,...,T$ on the $(t-T_1)$ th sample. In this way, the conditions of all ships are known. Then, $\frac{1}{T_2}\sum_{t=T_1+1}^{T}\sum_{k=1}^{K} c_k(x_k^{*(t)}, u_k^{(t)})$ and $\frac{1}{T_2}\sum_{t=T_1+1}^{T}\sum_{k=1}^{K} c_k(x_k^{\#(t)}, u_k^{(t)})$ can be fairly compared.

We further use a concrete Example 2 with a data set consisting of 2,000 real inspection records at the Hong Kong port to illustrate the detailed steps of solution 2 to address the ship selection problem. Suppose that the first 1,800 inspected ships from the first T_1 days serve as the training set $D_{train} = (x_k^{\#(t)}, u_k^{(t)}, c_k(x_k^{\#(t)}, u_k^{(t)}), k = 1,...,K_t)$, $t = 1,...,T_1$, where $x_k^{\#(t)} = 1$, $u_k^{(t)}$ is ship characteristics (including ship type, age, flag/recognized organization/company performance, month interval between the current inspection and the last initial inspection in the Tokyo MoU, whether it is inspected in the last 36 months, and the average deficiency number and detention rate in the last 36 months (Zis, 2021)), and $c_k(x_k^{\#(t)}, u_k^{(t)})$ is the negative value of the number of deficiencies of ship k coming to the port on day t. A random forest model is developed for ship condition prediction using D_{train}. Based on five-fold cross validation and grid search, hyperparameters in the random forest model are set as follows: the number of trees contained is 200, the fraction of features considered for node splitting is 0.4, the maximum tree depth is 6, and the minimum number of examples contained in a leaf node is 5.

For the last T_2 days, the number of candidate ships to be inspected on day t is $M^{(t)} = 30$, the number of ships that can be inspected on day t is $N^{(t)} = 10$, and the total number of inspected ships is 200. Then, instead of applying prescriptive analytics to select $N^{(t)} = 10$ from $M^{(t)} = 30$ for day t, we randomly select 30 ships from the 200 ships in the whole test set without replacement to form a sample of $M^{(t)}$ and we denote it by $M^{(t)'}$, and the above process is repeated 20 times ($T_2 = 200 \div 10 = 20$ days). Then, in each $M^{(t)'}$, prescriptive analytics is applied to select $N^{(t)}$ ships with the smallest negative values of deficiency number for inspection, and denote the optimal decision (inspect or not inspect) given by Eq. (21) for ship k on day t by $x_k^{*(t)}$. Recall that the current inspection decision is $x_k^{\#(t)}$. As the conditions of all the ships in each $M^{(t)'}$ are known, decisions given by prescriptive analytics $\frac{1}{T_2}\sum_{t=T_1+1}^{T}\sum_{k=1}^{K} c_k(x_k^{*(t)}, u_k^{(t)})$ and the current approach $\frac{1}{T_2}\sum_{t=T_1+1}^{T}\sum_{k=1}^{K} c_k(x_k^{\#(t)}, u_k^{(t)})$ can be compared in a fair way, as both values are calculated based on real costs.

In our example, ship inspection performances (i.e., the negative value of the total number of ship deficiencies identified for each day) in the prescriptive analytics model, the current decision approach, and the perfect-forecast policy is shown in Table 2. In particular, random ship selection for inspection is the current decision approach, and its performance is the average of 100 times of random ship sampling (i.e., randomly choose 10 ships out of the 30 candidate ships for inspection as one time of sampling). The perfect-forecast policy is the decision made based on knowing the actual number of deficiencies of the 30 candidate ships for each day, which cannot be achieved in practice.

Table 2. Comparison between prescriptive analytics, current decision approach, and perfect-forecast policy for ship selection in PSC

t	Prescriptive analytics	Current decision approach	Perfect-forecast policy
1	-37	-31.30	-65
2	-89	-48.58	-110
3	-42	-27.02	-63
4	-53	-35.37	-82
5	-68	-48.68	-93
6	-45	-28.06	-60
7	-61	-46.59	-94
8	-86	-52.26	-106
9	-48	-40.27	-83
10	-42	-33.97	-67
11	-53	-49.02	-90
12	-67	-49.96	-98
13	-57	-44.81	-88
14	-74	-46.44	-94
15	-65	-48.54	-97
16	-75	-44.38	-102
17	-43	-27.09	-55
18	-69	-46.08	-94
19	-50	-36.67	-73
20	-69	-38.24	-83
Average performance	-59.65	-41.17	-84.85

It can be seen from Table 2 that ship inspection performance of the decision given by prescriptive analytics can improve the current decision approach by 44.9%, and the gap between the prescriptive analytics and the perfect-policy forecast is 29.7% and the gap between the current decision approach and the perfect-policy forecast is 51.5%. The performance of the three models is compared fairly by using the actual ship deficiency condition. Therefore, it can be safely concluded that ship selection decision given by the prescriptive analytics method (with average costs -59.65 in the last T_2 days) is indeed better than the current decision approach (with average costs -41.17 in the last T_2 days).

Solution 3 (partial solution): Solution 2 is suitable when a new test set can be constructed for which the actual cost of each candidate decision is known. However, in some situations, we cannot construct such new test sets, such as in Examples 1 and 4, because the K components are highly related with each other, and thus random sampling is not applicable. Then, synthetic data that are analogous to actual operations can be constructed. Specifically, we first design known function $d_k(x_k, u_k^{(t)})$ that approximates $c_k(x_k, u_k^{(t)})$. For example, in Example 4, we cannot fairly compare prescriptive analytics and drivers' current decision-making approach. Taking a step back, we ask a highly relevant question: if each driver k adopts the nearest unvisited customer heuristic (Kendall and Li, 2013) to determine the sequence of customers to visit, will prescriptive analytics outperform the VRP model? To answer this, the actual total travel distance given x_k, $k = 1,...,K$, can be calculated, and we denote the function by $d_k(x_k, u_k)$ (recall that u_k has no effect in this example). The new training dataset is $(x_k^{\#(t)}, u_k^{(t)}, d_k(x_k^{\#(t)}, u_k^{(t)}), k = 1,...,K)$, $t = 1,...,T_1$, which is then used to train ML model $\hat{d}_k(x_k, u_k)$ that predicts the total travel distance of driver $k = 1,...,K$ when customer set x_k is assigned to the driver. Then, the following surrogate models are solved:

$$\min_{x \in X^{(t)}} \sum_{k=1}^{K} \hat{d}_k(x_k, u_k^{(t)}),\ t = T_1 + 1,...,T, \tag{23}$$

and the resulting solution for day $t = T_1 + 1,...,T$ is $x^{*(t)}$. The prescriptive analytics paradigm is superior to the current decision-making approach if

$$\frac{1}{T_2} \sum_{t=T_1+1}^{T} \sum_{k=1}^{K} d_k(x_k^{*(t)}, u^{(t)}) < \frac{1}{T_2} \sum_{t=T_1+1}^{T} \sum_{k=1}^{K} d_k(x_k^{\#(t)}, u^{(t)}) \tag{24}$$

where the value of $d_k(x_k^{*(t)}, u^{(t)})$ is known. In such a case, prescriptive analytics and the current decision-making approach can be compared relatively fairly.

Similarly, in Example 1, over one leg, only the fuel consumption rate under the actual sailing speed can be observed, while those under other candidate sailing speeds, i.e., counterfactuals of the real sailing speed, are unknown. For example, suppose that the allowable sailing speed ranges from 8 to 18 knots, and this range is further discretized with 0.01 knots as the interval. If the actual sailing speed is 12 knots, only the real fuel consumption rate when sailing at 12 knots is available, while those of the other 1,000 speed values are unknown. Meanwhile, in practice, white-box models based on hydrodynamics and physical laws (Wang et al., 2021) as well as the well-known cubic law (Adland et al., 2020) are widely used to predict ship fuel consumption rates as they are good approximations of actual values. By utilizing either method as $d_k(x_k, u_k^{(t)})$, e.g., $d_k(x_k, u_k^{(t)}) = 10(x_k)^{2.2}$ where the values 10 and 2.2 should be calibrated using actual data (Wang and Meng, 2012), fuel consumption on leg k under all candidate sailing speeds can be estimated, which is the synthetic data in this example.

We can now ask a highly relevant question: for each k, if the actual fuel consumption function is $d_k(x_k, u_k^{(t)})$, will prescriptive analytics outperform the current decision-making approach? To answer this, the new training dataset is $(x_k^{\#(t)}, u_k^{(t)}, d_k(x_k^{\#(t)}, u_k^{(t)}), k=1,...,K)$, $t=1,...,T_1$, which is then used to train ML model $\hat{d}_k(x_k, u_k)$ that predicts the fuel consumption on leg $k=1,...,K$ at speed x_k. Then, the following surrogate models are solved:

$$\min_{x \in X^{(t)}} \sum_{k=1}^{K} \hat{d}_k(x_k, u_k^{(t)}),\ t = T_1+1,...,T, \tag{25}$$

and the resulting solution for period $t=T_1+1,...,T$ is $x^{*(t)}$. The prescriptive analytics paradigm is superior to the current decision-making approach if

$$\frac{1}{T_2}\sum_{t=T_1+1}^{T}\sum_{k=1}^{K} d_k(x_k^{*(t)}, u_k^{(t)}) < \frac{1}{T_2}\sum_{t=T_1+1}^{T}\sum_{k=1}^{K} d_k(x_k^{\#(t)}, u_k^{(t)}) \tag{26}$$

where the value of $d_k(x_k^{*(t)}, u_k^{(t)})$ is known.

We use Example 1 to show the specific steps of solution 3 based on generating synthetic data. Noon report of this bulk carrier is provided by its company, which contains a total of 333 records. We select two 8-day voyages as the test set (16 records), and the remaining 317 records from 23 voyages as the training set. Then, the training set can be represented by $D_{train} = (x_k^{\#(t)}, u_k^{(t)}, c_k(x_k^{\#(t)}, u_k^{(t)}), k=1,...,K_t)$, $t=1,...,23$, and the test set is $D_{test} = (x_k^{\#(t)}, u_k^{(t)}, c_k(x_k^{\#(t)}, u_k^{(t)}), k=1,...,K_t)$, $t=24,25$. The actual sailing speeds, sailing time, length, and the total fuel consumption of each segment of the two voyages are shown in Table 3.

Table 3. The original sailing speed, sailing time, length, and fuel consumption of each segment in voyage 1 and voyage 2

Segment	Sailing speed	Sailing time	Length	Fuel consumption in noon report	Fuel consumption calculated by cubic law
Original information of voyage 1					
1	11.35	15.5	176	20.90	19.08
2	11.00	24	264	32.50	26.90
3	11.17	24	268	32.60	28.16
4	11.21	24	269	32.50	28.47
5	11.46	24	275	32.50	30.41
6	10.46	24	251	32.45	23.13
7	11.33	24	272	33.00	29.39
8	9.21	24	221	32.50	15.79
Sum	\	183.5	1,996	248.95	201.32
Original information of voyage 2					
Segment	Sailing speed	Sailing time	Length	Fuel consumption in noon report	Fuel consumption calculated by cubic law
1	12.73	14.3	182	19.37	24.84
2	12.73	24.5	312	33.15	42.56
3	12.64	25	316	33.84	42.51
4	11.96	24	287	32.50	34.57
5	12.38	24	297	32.50	38.34
6	12.67	24	304	32.50	41.10
7	11.84	25	296	33.84	34.94
8	12.04	24	289	32.50	35.27
Sum	\	184.8	2,283	250.20	294.13

Suppose that cubic law is adopted by the operator of the bulk carrier to predict its hourly fuel consumption rate in daily vessel operation for speed planning. The function for hourly fuel consumption rate prediction is $d'_k(x_k, u_k^{(t)}) = \alpha(x_k)^\beta$, where α and β are parameters to be estimated from D_{train}, and the final form is $d'_k(x_k, u_k^{(t)}) = 0.000842(x_k)^3$. Based on this, a new training dataset $D^{'}_{train} = (x_k^{\#(t)}, u_k^{(t)}, d_k(x_k^{\#(t)}, u_k^{(t)}), k = 1,...,K)$, $t = 1,...,23$ can be formulated, where $x_k^{\#(t)}$ and $u_k^{(t)}$ are the same as those in D_{train} for segment k in voyage t, while the total fuel consumption over the segment is changed from $c_k(x_k^{\#(t)}, u_k^{(t)})$, which is the original fuel consumption of the segment recorded in the noon report, to $d_k(x_k^{\#(t)}, u_k^{(t)})$, which is the product of the cubic law for hourly fuel consumption rate prediction $d'_k(x_k, u_k^{(t)}) = 0.000842(x_k)^3$ and the corresponding sailing hours. Then, suppose we have an ML model based on artificial neural network constructed on the new training set $D^{'}_{train}$, which has one hidden layer with 8 neurons with activation function Rectified Linear Unit (ReLU), learning rate 0.01, and epoch 100. The ML model developed is then used to predict the hourly fuel consumption rate of each segment in test set D_{test}. Then, the following surrogate model M1' of model M1 is solved:

[M1']

$$\min \sum_{k=1}^{K_t} \hat{d}_k (x_k^{(t)}, u_k^{(t)}),\ t = 24, 25 \tag{27}$$

s.t.

Constraints Eq. (4) to Eq. (8).

Denote the resulted optimal sailing speed for segment k of voyage t by $x_k^{*(t)}$. We further require that the voyage sailing time after optimization is within 5% of the real voyage sailing time, i.e., $T_{\min}$ is no less than 95% of the real total sailing time, and $T_{\max}$ is no more than 105% of the real total sailing time. The sailing speeds, sailing time, and the total fuel consumption of each segment of the two voyages after speed optimization are shown in Table 4.

Table 4. The sailing speed, sailing time, and fuel consumption of each segment after optimization in voyage 1 and voyage 2

Information of voyage 1 after speed optimization				
Segment	Sailing speed	Sailing time	Fuel consumption based on prediction	Fuel consumption calculated by cubic law
1	15.53	11.33	11.14	35.74
2	11.80	22.37	21.56	30.95
3	8.01	33.46	32.10	14.48
4	16.77	16.04	15.55	63.70
5	17.95	15.32	14.85	74.61
6	17.70	14.18	13.76	66.21
7	8.00	34.00	31.14	14.66
8	8.00	27.63	26.34	11.91
Sum	\	174.33	166.43	312.25
Information of voyage 2 after speed optimization				
Segment	Sailing speed	Sailing time	Fuel consumption based on prediction	Fuel consumption calculated by cubic law
1	17.62	10.33	20.25	47.58
2	12.57	24.82	41.10	41.51
3	18.00	17.56	36.20	86.21
4	18.00	15.94	32.43	78.30
5	18.00	16.50	35.31	81.02
6	8.03	37.86	52.26	16.51
7	18.00	16.44	33.80	80.75
8	8.00	36.13	43.09	15.57
Sum	\	175.58	294.43	447.44

The fundamental challenge of prescriptive analytics occurs if the fuel consumption recorded in the noon report before speed optimization and the fuel consumption predicted by the ANN model after speed optimization is directly compared. In voyage

1, the total fuel consumption recorded in the noon report obtained before speed optimization is 248.95 tons, and is 166.43 tons after speed optimization according to the estimation given by the ANN model, which seems to be heavily reduced to a very large extend. However, according to the fuel consumption given by the cubic law, the original total fuel consumption is 201.32 tons, and it creases to 312.25 tons after speed optimization, which means that if the optimized speeds based on ANN are adopted, actually much more fuel would be used to accomplish the voyage. Such increase in fuel consumption is because the ANN model cannot accurately capture the monotonic and convex relationship between ship sailing speed and hourly fuel consumption rate. Consequently, the hourly fuel consumption rate cannot be accurately predicted, and thus leads to improper consequent sailing speed decisions. This shows that the fundamental challenge exists in the comparison process between prescriptive analytics and current decision approach if Eq. (22) is used. In voyage 2, the total fuel consumption before and after speed optimization is 250.20 tons and 294.43 tons, respectively, showing that the performance of prescriptive analytics is worse than the original ship sailing behavior. This result is echoed in the comparison based on cubic law, where 294.13 tons of fuel is used before speed optimization and 447.44 tons is used after speed optimization.

According to the partial solution, Eq. (22) is used to compare the current decision-making approach and prescriptive analytics. Especially, the left-hand side of Eq. (22) is the average fuel consumption over all voyages in the test set using the optimal sailing speeds recommended by model M1', which is 379.85 tons, and the right-hand side of Eq. (22) is the average fuel consumption over each voyage in the test set calculated by applying the real sailing speeds on the synthetic dataset, which is 247.73 tons. Then, it can be fairly concluded that the current decision-making approach is better than the prescriptive analytics approach, as the hourly fuel consumption given by the ANN model is highly inaccurate and biased.

5. Discussion and conclusion

In the era of big data where smart devices, clouding computing, SaaS business and company models are widely adopted, an increasing number of historical data are produced and collected, and analytics has advanced from providing elementary description and summarizing what happened/what is happening (i.e., descriptive analytics), to forecasting what will happen in the future (i.e., predictive analytics), and to prescribing actions to achieve desired outcome given past and current events (i.e., prescriptive analytics). Especially, as a state-of-the-art analytics approach, prescriptive analytics is becoming more and more popular in freight transport planning as it is expected to mitigate risk in strategic decision making and increase customer satisfaction, and hence improve profitability and competitiveness for business organizations in the ever-changing business environments.

One of the key challenges in prescriptive analytics is the lack of sufficient data and data uncertainties that results in an unfair comparison between the optimal decisions by prescriptive analytics and by current decision approaches. Consequently, a dead circle occurs in the adoption of prescriptive analytics models in prescribing business decisions: researchers cannot provide strong evidence that prescriptive analytics is better than the current approach, so industry practitioners do not adopt it; because industry practitioners do not adopt prescriptive analytics, it is difficult to show its superiority as there are insufficient historical data. This study refers to this challenge as the fundamental challenge in prescriptive analytics and illustrates it using four practical examples in freight transportation. Then, three possible solutions that fully or partially address the fundamental challenge by searching sufficient historical data, constructing new test sets, and generating synthetic data are proposed.

The solutions differ from each other regarding their application scenarios considering the accessibility of historical data: solutions 1 requires that sufficient historical data could be acquired with reasonable costs; Solution 2 based on data sampling is suitable when only certain historical inspection records are available while the components of a decision are uncorrelated to each other; Solution 3 based on generating synthetic data is suitable when only certain historical inspection records are available, and it can address the situations when the components of a decision are correlated. Considering the availability and costs of data acquisition, different solutions might be adopted to address different problems: when sufficient and affordable historical data are available, solution 1 is preferred; when there are insufficient historical data or the cost of data acquisition is too high, solution 2 can be used when the components of a decision are uncorrelated and solution 3 can be used when those components are correlated. Considering further the fact that the solutions only consider the characteristics of historical data, the proposed solutions can promote not only in freight transport but also in other disciplines by demonstrating the effectiveness and efficiency of the prescriptive analytics paradigm.

One should also be aware that in addition to the fundamental challenge mentioned in this article, there are several other challenges of applying prescriptive analytics to address practical freight transport problems. One critical challenge is the incompleteness and noise in raw data, which can be addressed by adequate data preprocessing. For example, as discussed by Yan et al. (2021), typical methods to deal with missing values in datasets include re-obtaining of missing data, discarding features with too many missing values, data imputation, and missing value prediction. Common methods to detect outliers in data include z-score, interquartile range, and using unsupervised machine learning models for outlier detection. Another critical challenge is the uncertainties in prediction where the predicted conditional mean is highly likely not to equal the actual condition mean and the uncertainties are inevitably brought to the following optimization model. To deal with the uncertainties in the prediction, when the objective function of the optimization model is nonlinear in the unknown parameter, the parameter should be modeled as a random variable, where local weight functions based on predictive methods such as k-nearest-neighbors regression and linear regression (Bertsimas and Kallus, 2020) as well as global methods considering prediction errors in the training set (Wang and Yan, 2022) and based on quantile regression (Yan and Wang, 2022) can be applied to estimate the conditional distribution of the unknown parameter.

It is worth mentioning that in addition to the four typical examples covered in this article, there are a wider range of application scenarios of the proposed solutions to compare prescriptive analytics and current decision approaches in freight transport for different decision makers, including drayage operators who plan and schedule vehicles between terminals and shippers and receivers (e.g., predicting travel time or cost and scheduling vehicles to minimize delay or the total cost), terminal operators who management transshipment among different transportation modes (e.g., predicting ship arrival time to port and allocating resources for berthing and cargo operating to minimize the total operational time or cost), and network operators who are responsible to plan the construction and operation of transport infrastructure (e.g., predicting traffic amount and assigning and scheduling vehicles to minimize the cost or maximize the benefit).

Meanwhile, it should also be noted that the foundation of all the three solutions is that historical data (no matter sufficient or insufficient) are available. If they are unavailable, prescriptive analytics models themselves cannot be developed, not to mention the solutions to tackle the fundamental challenge. Therefore, we would suggest the stakeholders and decision makers in freight transport properly collecting and storing critical data in daily operations and using such data to develop prescriptive analytics models to prescribe informed decisions. In addition, the solutions proposed in this article can be further applied to compare the current decision-making approaches and the prescriptive analytics in a reasonable way. Furthermore, we would encourage the stakeholders to share their data and information to establish an information sharing platform, where joint decisions in the whole intermodal freight transport network can be optimized.

References

Aboytes-Ojeda, M., Castillo-Villar, K. K., Cardona-Valdés, Y., 2022. Bi-objective stochastic model for the design of biofuel supply chains incorporating risk. Expert Systems with Applications 202, 117285.

Adland, R., Cariou, P., Wolff, F. C., 2020. Optimal ship speed and the cubic law revisited: Empirical evidence from an oil tanker fleet. Transportation Research Part E: Logistics and Transportation Review 140, 101972.

Akgüngör, A. P., Doğan, E., 2009. An artificial intelligent approach to traffic accident estimation: Model development and application. Transport 24(2), 135–142.

An, J. H., Kwon, D. Y., Jeon, K. S., Tyan, M., Lee, J. W., 2022. Advanced sizing methodology for a multi-Mode eVTOL UAV powered by a hydrogen fuel cell and battery. Aerospace 9(2), 1–26.

Barbour, W., Mori, J. C. M., Kuppa, S., Work, D. B., 2018. Prediction of arrival times of freight traffic on US railroads using support vector regression. Transportation Research Part C: Emerging Technologies 93, 211–227.

Bertsimas, D., Kallus, N., 2020. From predictive to prescriptive analytics. Management Science 66(3), 1025–1044.

Bombelli, A., Fazi, S., 2022. The ground handler dock capacitated pickup and delivery problem with time windows: A collaborative framework for air cargo operations. Transportation Research Part E: Logistics and Transportation Review 159, 102603.

De Jong, G., Ben-Akiva, M., 2007. A micro-simulation model of shipment size and transport chain choice. Transportation Research Part B: Methodological 41(9), 950–965.

De Oliveira, L. P., Alonso, F. J., da Silva, M. A. V., de Gomes Garcia, B. T., Lopes, D. M. M., 2020. Analysis of the influence of training and feedback based on event data recorder information to improve safety, operational and economic performance of road freight transport in Brazil. Sustainability 12(19), 8139.

Du, Y., Meng, Q., Wang, S., Kuang, H., 2019. Two-phase optimal solutions for ship speed and trim optimization over a voyage using voyage report data. Transportation Research Part B: Methodological 122, 88–114.

El Noshokaty, S., 2020. Ship routing and scheduling systems: forecasting, upscaling and viability. Maritime Business Review 6(1), 95–112.

Elmachtoub, A. N., Grigas, P., 2022. Smart “predict, then optimize”. Management Science 68(1), 9–26.

Esmaeeli, N., Sattari, F., Lefsrud, L., Macciotta, R., 2022. Critical analysis of train derailments in Canada through process safety techniques and insights into enhanced safety management systems. Transportation research record 2676(4), 603–625.

Ferreira, K. J., Lee, B. H. A., Simchi-Levi, D., 2016. Analytics for an online retailer: Demand forecasting and price optimization. Manufacturing & Service Operations Management 18(1), 69–88.

Frazzetto, D., Nielsen, T. D., Pedersen, T. B., Šikšnys, L., 2019. Prescriptive analytics: a survey of emerging trends and technologies. The VLDB Journal 28(4), 575–595.

Gallien, J., Mersereau, A. J., Garro, A., Mora, A. D., Vidal, M. N., 2015. Initial shipment decisions for new products at Zara. Operations Research 63(2), 269–286.

Holguín-Veras, J., Jaller, M., Sánchez-Díaz, I., Campbell, S., Lawson, C. T., 2014. Freight generation and freight trip generation models. Modelling Freight Transport, 43–63.

Hu, Q., Gu, W., Wang, S., 2022. Optimal subsidy scheme design for promoting intermodal freight transport. Transportation Research Part E: Logistics and Transportation Review 157, 102561.

Hwang, T., 2021. Assignment of freight truck shipment on the US highway network. Sustainability 13(11), 6369.

Jiang, Y., Ding, Z., Zhou, J., Wu, P., Chen, B., 2022. Estimation of traffic emissions in a polycentric urban city based on a macroscopic approach. Physica A: Statistical Mechanics and its Applications, 127391.

Kendall, G., Li, J., 2013. Competitive travelling salesmen problem: A hyper-heuristic approach. Journal of the Operational Research Society 64(2), 208–216.

Knapp S., 2007. The econometrics of maritime safety: recommendations to enhance safety at sea. PhD Thesis in Erasmus University Rotterdam.

Kuran, C. H. A., Newnam, S., Beanland, V., 2022. Adaptive non-conform behaviour in accident investigations in the road based heavy goods transport sector. Safety Science 146, 105539.

Li, Q., Liu, P., Meng, X., Zhang, G., Ai, Y., Chen, W., 2022. Model prediction control-based energy management combining self-trending prediction and subset-searching algorithm for hydrogen electric multiple unit train. IEEE Transactions on Transportation Electrification 8(2), 2249–2260.

Lepenioti, K., Bousdekis, A., Apostolou, D., Mentzas, G., 2020. Prescriptive analytics: Literature review and research challenges. International Journal of Information Management 50, 57–70.

Mahdavian, A., Shojaei, A., Salem, M., Laman, H., Eluru, N., Oloufa, A. A., 2021. A universal automated data-driven modeling framework for truck traffic volume prediction. IEEE Access 9, 105341–105356.

Ng, M.W., 2015. Container vessel fleet deployment for liner shipping with stochastic dependencies in shipping demand. Transportation Research Part B: Methodological 74, 79–87.

Pani, A., Mishra, S., Sahu, P., 2022. Developing multi-vehicle freight trip generation models quantifying the relationship between logistics outsourcing and insourcing decisions. Transportation Research Part E: Logistics and Transportation Review 159, 102632.

Psaraftis, H. N., 2019. Decarbonization of maritime transport: to be or not to be? Maritime Economics & Logistics 21(3), 353–371.

Qi, Y., Harrod, S., Psaraftis, H. N., Lang, M., 2022. Transport service selection and routing with carbon emissions and inventory costs consideration in the context of the Belt and Road Initiative. Transportation Research Part E: Logistics and Transportation Review 159, 102630.

Raeesi, R., Zografos, K. G., 2020. The electric vehicle routing problem with time windows and synchronised mobile battery swapping. Transportation Research Part B: Methodological 140, 101–129.

Rosell, F., Codina, E., Montero, L., 2022. A combined and robust modal-split/traffic assignment model for rail and road freight transport. European Journal of Operational Research, in press.

Sakai, T., Alho, A., Hyodo, T., Ben-Akiva, M., 2020. Empirical shipment size model for urban freight and its implications. Transportation Research Record 2674(5), 12–21.

Saleh, M., Chowdhury, T., Vaughan, J., Wang, A., Mousavi, K., Roorda, M. J., Hatzopoulou, M., 2022. Climate and air quality impacts of off-peak commercial deliveries. Transportation Research Part D: Transport and Environment 109, 103360.

Siami-Irdemoosa, E., Dindarloo, S. R., 2015. Prediction of fuel consumption of mining dump trucks: A neural networks approach. Applied Energy 151, 77–84.

Sun, Z., Zheng, J., 2016. Finding potential hub locations for liner shipping. Transportation Research Part B: Methodological 93, 750–761.
Taslimi, B., Sarijaloo, F. B., Liu, H., Pardalos, P. M., 2022. A novel mixed integer programming model for freight train travel time estimation. European Journal of Operational Research 300(2), 676–688.
Tillig, F., Ringsberg, J. W., Psaraftis, H. N., Zis, T., 2020. Reduced environmental impact of marine transport through speed reduction and wind assisted propulsion. Transportation Research Part D: Transport and Environment 83, 102380.
Verizon Connect, 2020. Traffic Data, Route Planning, and ETA: How Google Maps Predicts Travel Time. Accessed 28 July 2022. https://www.verizonconnect.com/resources/article/google-maps-travel-time/.
Wang, G., Gunasekaran, A., Ngai, E. W., Papadopoulos, T, 2016. Big data analytics in logistics and supply chain management: certain investigations for research and applications. International Journal of Production Economics 176, 98–110.
Wang, H., Lang, X., Mao, W., 2021. Voyage optimization combining genetic algorithm and dynamic programming for fuel/emissions reduction. Transportation Research Part D: Transport and Environment 90, 102670.
Wang, K., Yan, X., Yuan, Y., Li, F., 2016. Real-time optimization of ship energy efficiency based on the prediction technology of working condition. Transportation Research Part D: Transport and Environment 46, 81–93.
Wang, T., Wang, X., Meng, Q., 2018. Joint berth allocation and quay crane assignment under different carbon taxation policies. Transportation Research Part B: Methodological 117, 18–36.
Wang, S., Meng, Q., 2012. Sailing speed optimization for container ships in a liner shipping network. Transportation Research Part E: Logistics and Transportation Review 48(3), 701–714.
Wang, S., Yan, R., 2022. A global method from predictive to prescriptive analytics considering prediction error for "Predict, then optimize" with an example of low-carbon logistics. Cleaner Logistics and Supply Chain, 100062.
Wang, Z., Delahaye, D., Farges, J. L., Alam, S., 2022. Complexity optimal air traffic assignment in multi-layer transport network for urban air mobility operations. Transportation Research Part C: Emerging Technologies 142, 103776.

Xiao, Z., Zhi, J., Keskin, B. B., 2022. Towards a machine learning-aided metaheuristic framework for a production/distribution system design problem. Computers & Operations Research, 105897.

Yan, R., Wang, S., 2022. Integrating prediction with optimization: Models and applications in transportation management. Multimodal Transportation 1(3), 100018.

Yan, R., Wang, S., Cao, J., Sun, D., 2021. Shipping domain knowledge informed prediction and optimization in port state control. Transportation Research Part B: Methodological 149, 52–78.

Yan, R., Wang, S., Zhen, L., Laporte, G., 2021. Emerging approaches applied to maritime transport research: Past and future. Communications in Transportation Research 1, 100011.

Yang, Y., 2015. Development of the regional freight transportation demand prediction models based on the regression analysis methods. Neurocomputing 158, 42–47.

Yu, J., Tang, G., Song, X., Yu, X., Qi, Y., Li, D., Zhang, Y., 2018. Ship arrival prediction and its value on daily container terminal operation. Ocean Engineering 157, 73–86.

Zis, T. P., 2021. A game theoretic approach on improving sulphur compliance. Transport Policy 114, 127–137.

Zis, T. P., Cullinane, K., 2020. The desulphurisation of shipping: Past, present and the future under a global cap. Transportation Research Part D: Transport and Environment 82, 102316.

Zis, T., North, R. J., Angeloudis, P., Ochieng, W. Y., Bell, M. G., 2015. Environmental balance of shipping emissions reduction strategies. Transportation Research Record 2479(1), 25–33.

Zis, T. P., Psaraftis, H. N., 2021. Impacts of short-term measures to decarbonize maritime transport on perishable cargoes. Maritime Economics & Logistics, 1–28.

Zis, T. P., Psaraftis, H. N., Ding, L., 2020. Ship weather routing: A taxonomy and survey. Ocean Engineering 213, 107697.