Data-Driven Frequency-Based Airline Profit Maximization

BO AN and HAIPENG CHEN, Nanyang Technological University
NOSEONG PARK, University of North Carolina, Charlotte
V. S. SUBRAHMANIAN, University of Maryland

Although numerous traditional models predict market share and demand along airline routes, the prediction of existing models is not precise enough, and to the best of our knowledge, there is no use of data mining–based forecasting techniques for improving airline profitability. We propose the maximizing airline profits (MAP) architecture designed to help airlines and make two key contributions in airline market share and route demand prediction and prediction-based airline profit optimization. Compared to past methods used to forecast market share and demand along airline routes, we introduce a novel ensemble forecasting (MAP-EF) approach considering two new classes of features: (i) features derived from clusters of similar routes and (ii) features based on equilibrium pricing. We show that MAP-EF achieves much better Pearson correlation coefficients (greater than 0.95 vs. 0.82 for market share, 0.98 vs. 0.77 for demand) and R^2-values compared to three state-of-the-art works for forecasting market share and demand while showing much lower variance. Using the results of MAP-EF, we develop MAP–bilevel branch and bound (MAP-BBB) and MAP-greedy (MAP-G) algorithms to optimally allocate flight frequencies over multiple routes to maximize an airline's profit. We also study two extensions of the profit maximization problem considering frequency constraints and long-term profits. Furthermore, we develop algorithms for computing Nash equilibrium frequencies when there are multiple strategic airlines. Experimental results show that airlines can increase profits by a significant margin. All experiments were conducted with data aggregated from four sources: the U.S. Bureau of Transportation Statistics (BTS), the U.S. Bureau of Economic Analysis (BEA), the National Transportation Safety Board (NTSB), and the U.S. Census Bureau (CB).

1. INTRODUCTION

Since the deregulation of U.S. airlines in 1978, there has been intense competition among airlines for market share and, eventually, profitability. Although there is

considerable work on predicting market share and demand along airline routes, the prediction of existing models is not precise enough, and to the best of our knowledge, there is no use of data mining–based forecasting techniques for improving airline profitability. In this article, we formally define the maximizing airline profits (MAP) problem by allowing airlines to decide the flight frequencies on their flying routes to maximize their profits, subject to cost constraints. For instance, if there is a set $\mathcal{R}$ of routes in the world, and the airline can fly up to n routes where $n \leq |\mathcal{R}|$, then our MAP problem would capture both route selection (which routes to fly) and frequency.[1]

The first key contribution of this article is a precise route-specific prediction of both total demand of an origin-destination market and an airline's market share. Existing related works only use simple regression methods and focus on a limited number of variables. Our studies of three major route market share [Hansen 1990; Suzuki 2000; Wei and Hansen 2005] and demand prediction [Alperovich and Machnes 1994; Bhadra and Kee 2008; Bhadra and Wells 2005] methods show that the Pearson correlation coefficients between the predicted values and the actual values are a max of 0.82 (for one market share method) and 0.77 (for one demand model). Other existing models predict much lower numbers. Although one of these regression methods [Hansen 1990] outputs satisfactory prediction results in our experiments, there is still a gap between the prediction and the real value. The relatively "small" prediction gap can lead to a huge revenue loss if airlines make decisions based on prediction. Worse still, due to the inflexibility of existing models, the prediction gap can be large for some routes. We propose a new ensemble-based prediction method for forecasting total demand and market share, which uses an extensive number of features and several state-of-the-art clustering and regression algorithms that have never been adopted for MAP. The new ensemble method builds on existing models, but also collects several new features, together with novel clustering and game theoretic methods. We are the first to propose a prediction method, for a broad set of routes (around 700; past works stopped at 200), that considers the predictions of both total demand (demand generation) of a route and market share (demand allocation) of each airline operating in that route (13 airlines in total).

The second key contribution is two novel algorithms for solving MAP, which becomes computationally intractable with brute-force search when the number of routes is large. This is because (1) the solution space is exponential with respect to the number of routes, and (2) the profit-frequency function, which is generated by our proposed prediction method, is neither convex nor concave (and thus not linear). We show in Section 2 that MAP belongs to the hardest subclass of knapsack problems (KPs) and resource allocation problems (RAPs), and despite a vast amount of existing works [Bretthauer and Shetty 2002; Caprara et al. 2000; Pferschy 1999; Pisinger 2000] for both KPs and RAPs, all previous algorithms fail to solve it efficiently. Based on predictions made by a novel ensemble forecasting (MAP-EF), we come up with two optimization algorithms to solve the raised profit maximization problem. We present an exact algorithm to solve MAP based on a novel bilevel branch and bound (MAP-BBB) approach that computes the true optimal solution, as well as a MAP greedy (MAP-G) algorithm that more quickly computes suboptimal solutions.

As our third key contribution, we study two practical extensions of the basic frequency-based airline profit maximization problem and develop algorithms to compute Nash equilibrium strategies when there are multiple strategic airlines. First, airlines are usually conservative about changing their flight frequencies. Furthermore, they are concerned about the effect of the current frequency strategies on future

---

[1]Because the airline industry is very complex, it is difficult to model all other sources of profit that might be available (e.g., in-flight sales, baggage fees) to the airline, especially as such data is not freely available.

Table I. Summary of Notations

$\mathcal{A}$	Set of airlines	$N_i(f_i)$	Profit when A operates f_i flights on route R_i
A_i	i^{th} airline of $\mathcal{A}$	$\langle f_i \rangle$	Frequency strategy for all routes in $\mathcal{R}$ of airline A
$\mathcal{R}$	Set of routes	$\langle f_i^* \rangle$	Optimal frequency strategy of airline A
R_i	i^{th} route of $\mathcal{R}$	$N(\langle f_i \rangle)$	Profit of frequency strategy $\langle f_i \rangle$
Symbols Related to Predict Market Share of Route R		b	Budget limit of the profit maximization problem
X_{ik}	Value of the kth variable (e.g., price) of airline A_i	K	Number of routes to optimize profit of airline
η_k	Weight of X_{ik}	**Symbols Related to MAP-BBB in Algorithm 2**	
V_i	Airline A_i's customer utility in R	n_p	Parent node
m_i	Market share of airline A_i in R	c	Child node
Symbols Related to Predict Demand of Route R		c_l, c_r	Left and right children of a parent node
Y_k	Value of the kth variable (e.g., population)	N^*	Optimal profit sought by the algorithm
λ_k	Exponent of Y_k	$MaxLB$	Bounding threshold
d	Total demand (the number of passengers) of route R	UB_c, LB_c	Lower bound and upper bound profit of node c
Symbols Related to Calculate Nash Equilibrium of Route R		C_c	Lower bound cost of node c
p_i	Average ticket price of A_i for route R	**Symbols Related to MAP-G in Algorithm 3**	
c_i^{psg}	Unit passenger-related cost of A_i in route R	g	Group of route
$N(A_i)$	Utility (profit) of A_i in route R	q	Number of routes in a group g
Symbols Related to Profit Maximization of Airline A		$\mathcal{G}$	Set of route groups, $\mathcal{R}$ is divided into $|\mathcal{G}|$ groups
C_i	Per flight cost of route R_i for a given airline A	δ	Budget allocation unit
f_i	Flight frequency of route R_i for a given airline A	$\langle f_g^* \rangle$	Optimal frequency strategy of group g
$r_i(f_i)$	Revenue when A operates f_i flights on route R_i		

"potential" profits. As a result, we introduce bounded frequencies and a future profit effect factor in the basic profit maximization problem as two extensions of the basic problem and study their effects on airlines' profits. Second, we consider multiple strategic airlines. In the basic profit maximization problem, we assume that only one airline is strategic in deciding its frequencies. We release this assumption and study Nash equilibrium frequency strategies with multiple strategic airlines, each of which is smart enough to reason about the optimal behavior of the other airlines.

As our final contribution, we conduct extensive experimental evaluations to compare MAP's prediction results with those of past works. We show that past prediction models are significantly "beaten" by MAP, which increases these predictive accuracies to greater than 0.95 while significantly reducing the variance in our prediction error compared to past works. We also compare both the optimality and scalability of our proposed profit maximization algorithms to several benchmarks. The result shows that our predictions are far superior to past efforts. Moreover, by using MAP, an airline can increase its profit by at least 55% under mild conditions.[2]

The rest of this article is organized as follows. Section 2 introduces related work. Section 3 discusses the dataset used in our study. Section 4 shows our proposed prediction framework based on our ensemble method. Section 5 presents our proposed algorithms for the basic profit maximization problem. Section 6 introduces two practical extensions of the profit maximization problem. Section 7 studies equilibrium frequency strategies with multiple strategic airlines. Section 8 conducts extensive experimental evaluations. We conclude in Section 9. For notational convenience, Table I summarizes all notations used in this article.

2. RELATED WORK

Existing airline market share and demand prediction models try to write down math formulas for market share/demand and then use regression to find values of the parameters in these formulas that minimize the sum of the squared errors between the predicted values and the true values. We therefore call these *math models* to indicate that the structural form of these models is written down a priori without reference to

---

[2]Like all past models, our analysis of airline profits is only based on publicly available data, so the true number might be smaller. Nevertheless, even a 10% improvement in the real world would be substantial.

any data and that the parameter values that minimize error are then computed using available data.

Airline market share prediction. Most works on airline market share prediction use a multinomial logit (MNL) regression model [Hansen 1990; Suzuki 2000; Wei and Hansen 2005]. For a set $\mathcal{A}$ of airlines in a given route, an airline A_i's market share m_i is modeled as

$$m_i = \frac{e^{V_i}}{\sum_{A_j \in \mathcal{A}} e^{V_j}}. \tag{1}$$

Here, $V_i = \sum_k \eta_k X_{ik}$ is the customer's utility for choosing airline A_i's service, X_{ik} is the value of the kth variable for airline A_i, and η_k is its corresponding weight to be learned from data. Various variables have been studied. Market1 [Hansen 1990] considers a route's price, frequency, and number of stops. Based on the prediction, the competition among airlines is analyzed. Market2 [Suzuki 2000] considers an airline's frequency, delay, and safety. Market3 [Wei and Hansen 2005] studies the effect of aircraft size and seat availability on market share and considers other variables, such as price and frequency.

Total demand prediction. For total demand of a route, a multiplicative [Alperovich and Machnes 1994; Bhadra and Kee 2008; Bhadra and Wells 2005] regression model is the most frequently adopted. Formally, a route's total demand is represented as

$$d = \prod_k Y_k^{\lambda_k}, \tag{2}$$

where Y_k is a value of the kth variable and λ_k is a parameter to be learned from data that measures the importance of Y_k. Demand1 [Alperovich and Machnes 1994] considers wealth-related variables such as price, income, and the consumer price index (CPI). Demand2 [Bhadra and Wells 2005] studies the influence of demographic variables such as price, income, and population, as well as hub status (i.e., large, medium, small, or nonhub) on passenger demand. Demand3 [Bhadra and Kee 2008] focuses on variables such as price, income, population, and origin-destination distance.

In this article, our prediction models build on all of these methods but use an ensemble approach to achieve not only much higher predictive accuracy but also much greater robustness.

KPs and RAPs. Airline profit maximization through optimal frequency allocation over multiple routes, given a certain budget constraint, belongs to the family of KPs and RAPs, which have been extensively studied ever since the work of Dantzig [1957]. Although all variants of KPs are known to be NP-hard, MAP, which has bounded integer variables, and a nonconcave (arbitrary) objective function, is notoriously hard [Bretthauer and Shetty 2002].

There are some polynomial-time approximation schemes (PTAS) and even fully-PTAS to solve bounded knapsack problems (BKPs), using greedy algorithms [Keller et al. 2004], dynamic programming [Pisinger 2000], and branch and bound (BB) [Martello and Toth 1977]. The first two methods assume linear objective functions, whereas the last approach uses linear relaxation. Several algorithms have been proposed to solve KPs/RAPs with nonlinear objective functions, especially for a concave and nondecreasing objective function [Bretthauer and Shetty 2002]. The common idea is to decompose the original objective function into a set of approximated piecewise linear functions. Unfortunately, none of these works for our problem. This is because the profit-frequency function in our problem is neither concave nor convex (and thus not linear), varies from one route to another, and cannot be described by a simple
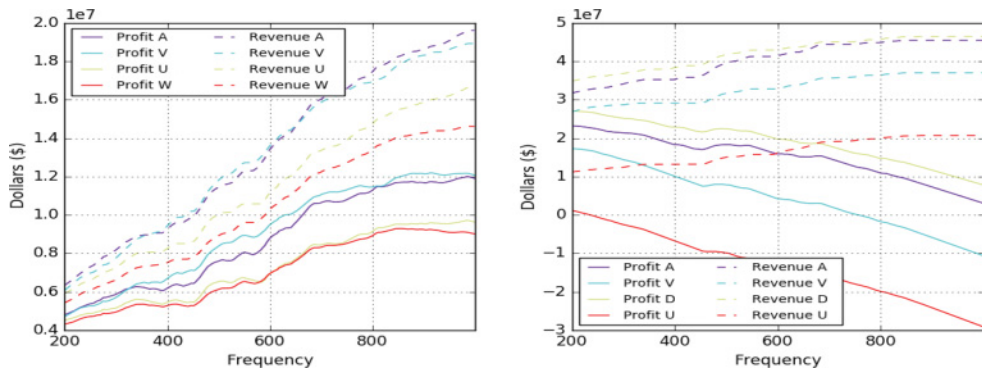
Fig. 1. Quarterly revenue (dashed lines) and profit (solid lines) in 10 million dollars generated by MAP-EF. Revenue is defined as the average ticket price multiplied by the number of passengers. Cost is the per flight cost multiplied by frequency. Note that the profit-frequency curve (revenue minus cost) is very irregular even though the revenue-frequency curve is monotonically increasing. The shapes of the curves are quite different.

math formula but by a series of clustering, regression, and other types of functions.[3] As shown in Figure 1, where the *x*-axis is the frequency and the *y*-axis is the profit (or revenue), the two graphs are for two routes (SFO-LAX route on the left and JFK-LAX route on the right) and four major U.S. airlines. We see that the shapes of the curve on the left (SFO-LAX) is quite different from another on the right (JFK-LAX). In fact, across the 700 routes we studied, there was huge variation in the actual profit/revenue curves. Thus, any method specially devised for a certain form of objective functions (e.g., linear, concave, or convex) cannot be applied.

Optimal decision making based on data mining. Data mining–based optimization has been implemented by many existing KDD works, including optimal bidding [Zhang et al. 2014; Zhu et al. 2015], smart pacing [Xu et al. 2015], and optimal recommendation systems [Lu et al. 2014]. In the first step of these works, a predicted objective function is learned using regression analysis techniques. Based on the prediction results, an optimization problem is solved, either analytically [Lu et al. 2014; Zhang et al. 2014] or via optimization algorithms [Xu et al. 2015; Zhu et al. 2015]. Our MAP framework follows this general scheme; however, it differs in many ways, including the fact that the predicted objective function is in a very complicated linear algebra form.

3. THE MAP DATASET

We have created an integrated MAP dataset by aggregating information from four publicly available data sources: the U.S. Bureau of Transportation Statistics (BTS) [BTS 2016], the U.S. Bureau of Economic Analysis (BEA) [BEA 2016], the National Transportation Safety Board (NTSB) [NTSB 2016], and the U.S. Census Bureau (CB) [CB 2016]. The entire dataset amounts to 539 million rows, 80% of which are from the BTS dataset. The BTS provides almost all aspects of airline market information, such as ticket price,[4] frequency, number of stops, delay records, aircraft size, and available seats. The BEA and the CB release regional income, economic state, and population information, whereas the NTSB provides all safety- and accident-related information.

---

[3]We will revisit this point after Definition 5.2.
[4]The BTS releases a 10% sample of route and airline specific ticket sales logs, from which the average ticket price is obtained.

Table II. Variables for Predictions in Related Works

Market Share	
ticket price	Average price of tickets sold by airline
frequency	Number of flights operated by airline
delay time	Average delay time of all flights operated by airline
delay ratio	Ratio of delayed flights to total number of flights by airline
cancel ratio	Ratio of canceled flights to total number of flights by airline
stop	Average number of stops of all flights by airline
safety	Number of casualties caused by past accidents of airline
craftsize	Average number of seats for a flight of airline
total seat	Total number of seats provided by airline during a quarter
Total Demand	
avg. price	Average of all airlines' ticket prices
population	Geometric mean of origin and destinations' population
income	Geometric mean of origin and destinations' income
CPI	Geometric mean of origin and destinations' CPI
distance	Distance between origin and destination

ALGORITHM 1: ITER

1 Randomly assign values to $\mathbf{p}_{t-1}$ within $[p_i^{min}, p_i^{max}]$
2 **repeat**
3   **for** $A_i \in \mathcal{A}$ **do**
4   *counter* $\leftarrow 0$
5        $p_{i,t} \leftarrow$ Solution of Equations (3) through (5)
6   **if** $|p_{i,t} - p_{i,t-1}| \leq \varepsilon$ **then** *counter* $++$
7     $\mathbf{p}_{t-1} \leftarrow \mathbf{p}_t$
8 **until** *counter* $= n$
9 **return** $\mathbf{p}_t$
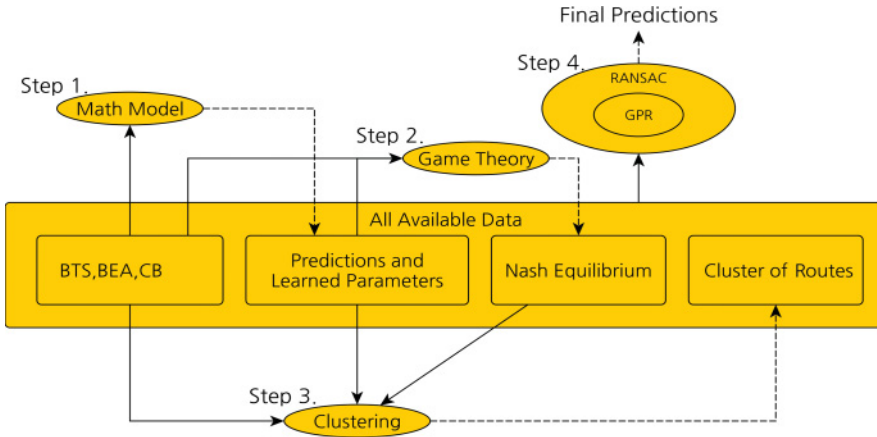
Fig. 2. Structure of the MAP-EF prediction model. Rectangles mean data and ellipses mean algorithms. Each algorithm is marked with a processing sequence number. Solid lines mean inputs and dashed lines mean outputs.

The BTS provides a quarterly dataset. A detailed description of the variables we consider is presented in Table II. We plan to publicly release our tools and dataset in 2016.

4. ENSEMBLE PREDICTION MODEL

In this section, we design an ensemble-based predictor, MAP-EF, which produces highly accurate route-specific predictions.

Overall architecture. Figure 2 shows the architecture of our ensemble model with four major steps:

—*Step 1: Past models with regularization.* In the first step, we use the MAP dataset in conjunction with existing market share models (Market1 [Hansen 1990], Market2 [Suzuki 2000], Market3 [Wei and Hansen 2005]) and demand models (Demand1 [Alperovich and Machnes 1994], Demand2 [Bhadra and Wells 2005], Demand3 [Bhadra and Kee 2008]) to generate the predictions made by models developed

previously in the literature. These predictions are used as additional features. In addition, we augment past models with regularization [Bishop 2006].[5] Because this step is straightforwardly built on top of past works, we do not go into it in further detail.

—*Step 2: Nash equilibrium*. Ticket pricing clearly influences both market share and demand. Moreover, the competitive pricing behavior of airlines would have further influence. Thus, we include equilibrium pricing calculated based on past models as an additional feature in MAP-EF. After this step, each route has one feature table for market share prediction and another for demand prediction. Each table[6] includes a set of features collected from the MAP dataset and from Steps 1 and 2.

—*Step 3: Clustering of similar routes*. We cluster all routes into groups of similar routes with the features of the quarter we predict. For each group of routes, we concatenate all individual route tables constructed in Step 2 for all routes in that group. The merged tables are then used for training within MAP-EF. We describe this step in further detail.

—*Step 4: Iterative Gaussian process regression (GPR) and random sample consensus (RANSAC)*. After Step 3, we have a comprehensive MAP feature table. We use a combination of the well-known methods of GPR [Rasmussen and Williams 2006] and RANSAC [Fischler and Bolles 1981]. To make market share/demand predictions for a route R, this step trains only on data of the group that contains route R, removes outliers with RANSAC, and then uses GPR to make the final prediction.

Nash equilibrium price feature. In a competitive airline market, the pricing behaviors of airlines are strategic. To capture the competitive nature of the market, we include equilibrium price as an additional feature.[7] The rationale behind this is that when an airline's quoted price is higher or lower than its equilibrium price, the customer might lose or gain utility, which has influence on the customer's choice and thus further affects an airline's market share. Therefore, we use the gap between equilibrium price and real price as an additional feature. The utility (net income) $N(A_i)$ of an airline A_i with respect to a specific route is $N(A_i) = (p_i - c_i^{psg}) \cdot d \cdot m_i$, where p_i is the ticket price and c_i^{psg} is the passenger-related per ticket cost obtained from the BTS dataset; d and m_i are a route's total demand and airline A_i's market share as described in Section 2. Recall that Nash equilibrium is an equilibrium situation in a multiplayer noncooperative game, where each player is assumed to know the equilibrium strategies of the other players, and no player has anything to gain by changing only its own strategy. Following this definition, we formulate the computation of Nash equilibrium strategy of each airline $A_i \in \mathcal{A}$ with respect to a given route:

$$\max_{p_i} \quad N(A_i), \tag{3}$$

$$\text{subject to} \quad p_i^{min} \le p_i \le p_i^{max}, \tag{4}$$

$$d \cdot m_i \le Q_i. \tag{5}$$

---

[5]In past works for market share prediction, the authors minimize sum squared error (SSE)—that is, they try to find parameter values for the math models that minimize $\Sigma_{A_i}(m_i - \hat{m_i})^2$, where m_i (respectively $\hat{m_i}$) is the true (respectively predicted) market share. In addition to this, regularization minimizes $\sum_{A_i}(m_i - \hat{m_i})^2 + ||W||$, where W is a vector that consists of parameters η_k (or λ_k for demand prediction) and $||W||$ is the norm of W. $||W||$ can be defined using either the L^1 or L^2 norm. Thus, we have three options: SSE, SSE+L^1 regularization, and SSE+L^2 regularization. To decide the best model among these, we use grid search and cross validation. For cross validation, we predict a quarter of the training data after training with the remaining quarters regardless of temporal sequence, which creates n-fold cross validation, where n is the number of quarters in the training data.

[6]Please refer to the appendix for detailed examples of the merged table.

[7]Note that when computing the equilibrium price, the market share and total demand is the math formula obtained in Step 1, and the competition is only within a route.

Equation (4) specifies the feasible section of price, and Equation (5) indicates that the total number of passengers cannot exceed its capacity Q_i. To compute the equilibrium price on a route, we propose the iterative method for nash equilibrium (ITER) computation in Algorithm 1, where $\mathbf{p}_{t-1}$ and $\mathbf{p}_t$ denote the optimal price vectors of two subsequent "virtual" periods. The algorithm starts with an initialization of $\mathbf{p}_{t-1}$ within $[p_i^{min}, p_i^{max}]$, where p_i^{min} and p_i^{max} are the minimum and maximum possible prices quoted by airline A_i. After this, iteration proceeds. In each repeat loop (lines 2 through 8), the algorithm updates optimal prices of all airlines. Within each repeat loop, there is a for loop (lines 3 through 7). In the for loop, each airline alternately updates its optimal price $p_{i,t}$ based on the previous price vector $\mathbf{p}_{-i,t-1}$ of the other airlines in the last virtual period. It iterates until for each airline $A_i \in \mathcal{A}$, the price is not updating in two subsequent virtual periods (i.e., the difference between the two prices is no larger than a small threshold $|p_{i,t} - p_{i,t-1}| \leq \varepsilon$), which means that a Nash equilibrium is obtained. Note that the existence and uniqueness of Nash equilibrium prices are dependent on the utility function $N(A_i)$ (e.g., whether it is concave with respect to price). Due to the complex form of the utility function, it is intractable to prove it analytically. However, the following theorem guarantees that if there exists an equilibrium price, ITER always returns it, and our empirical results show that ITER always returns a solution after a few iterations, which means that Nash equilibrium prices always exist.

THEOREM 4.1. *If ITER returns a solution, then that solution is the Nash equilibrium price.*

PROOF. In each repeat loop, ITER finds the best response of all players (airlines) given the strategies of the last virtual period. When the best response strategies of all players in the current virtual period are equal to the best response strategies of the last virtual period, a Nash equilibrium is found according to the definition of Nash equilibrium. □

Clustering similar routes. Some routes are more similar than others. For example, LAX-JFK may show more similar market attributes to SFO-JFK than LAX-SFO because the first two are long-haul routes from west to east and the last one is a short-haul local route, and estimating market share/demand would probably benefit from examining all similar routes. To cluster routes, we use all features gathered in the MAP dataset, Steps 1 and 2. As we do not know the "ground truth" definition of whether a cluster is correct or not, we evaluate cluster quality via a *silhouette score*[8] [Rousseeuw 1987]. The silhouette score tends to keep increasing (with some glitches) as the number of clusters increases. The elbow method [Thorndike 1953] chooses the optimal number of clusters when silhouette scores stabilize. We tested several clustering methods, including k-means[++] [Arthur and Vassilvitskii 2007], and connectivity-based and density-based clustering algorithms, such as DBSCAN [Ester et al. 1996]. Of these, we achieved the best result with k-means[++].

RANSAC/GPR regression. Training with similar routes does not always contribute positively to predictive accuracy. In several cases, it incurs additional error. We tested many linear regression algorithms, such as Ridge, Lasso, and Lars [Bishop 2006]. None of them produces stable performance across multiple routes. We then tested a specially devised structure that consists of RANSAC and GPR. RANSAC removes outliers from

---

[8]The silhouette score is to calculate the consistency of clusters. As more similar routes are contained by a cluster, its silhouette score is higher. If $\mathbf{G}$ is a set of clusters, dis is some distance (dissimilarity) measure, and e is some data item, the silhouette score $SS(e) = \frac{b(e) - a(e)}{\max\{a(e), b(e)\}}$, where $a(e)$ is the average distance between e and the data items of the cluster to which it belongs, and $b(e)$ is the minimum average distance between e and the data items of clusters to which it does not belong.

training data, and GPR makes the final prediction only with inliers. GPR is known to be robust because it is much less parametric than other methods. GPR assumes that priors follow Gaussian processes, and the prior mean values are set to the same means as the training data means. The prior's covariance is calculated based on a user-input covariance function (i.e., kernel function), and the hyperparameters of the kernel function are optimized by maximizing the log marginal likelihood. The final predictions for test cases are calculated using the covariance matrix of all training and testing cases. The entire process can be described by a very complicated linear algebra form. RANSAC regression is known to be robust with respect to noise. In the first step, RANSAC performs regression with a small random subset of training data called *hypothetical inliers*. In the second step, all training data that do not fit the model constructed in the first step (in terms of a user-specified loss function) are classified as outliers and others as inliers. These two processes iterate over time to refine the set of inliers called the *consensus set* until the size of consensus set is large enough. Thus, MAP-EF uses RANSAC with GPR as a base estimator. We show prediction results of other regression techniques in the appendix.

Time overheads. In comparison with past models, MAP-EF should spend additional time for clustering and more advanced regression methods, which takes cubic time of training data size.

5. AIRLINE PROFIT MAXIMIZATION VIA FREQUENCY ALLOCATION

In this section, we suggest how to maximize an airlin's profit by optimally allocating quarterly frequencies to routes.

Definition 5.1. A *frequency strategy* of an airline A for a given set $\mathcal{R} = \{R_1, \ldots, R_n\}$ of routes is a vector $\langle f_i \rangle$ of nonnegative integers.

Intuitively, the ith element f_i of the frequency strategy denotes the number of times the airline should fly route R_i in a quarter. For a given airline A, a set $\mathcal{R} = \{R_1, \ldots, R_n\}$, per flight cost C_i for each route R_i, and total operation budget b, the profit maximization problem finds the frequency allocation $\langle f_i \rangle$ of A assuming that its competitors' behaviors are fixed. We also conduct experiments to test the robustness of this approach when competitors change their strategies simultaneously.

Definition 5.2. The *frequency-based profit maximization problem* can be encoded as the optimization problem:

$$\max_{\langle f_i \rangle} \quad \sum_{R_i \in \mathcal{R}} N_i(f_i) \tag{6}$$

$$\text{subject to} \quad 0 \leq f_i \leq f_i^{max}, \ f_i \in \mathbb{N}, \ \forall R_i \in \mathcal{R} \tag{7}$$

$$\sum_{R_i \in \mathcal{R}} C_i f_i \leq b. \tag{8}$$

$N_i(f_i)$ is the profit of route R_i given frequency f_i, which is ticket revenue $r_i(f_i)$ minus total operation cost $C_i f_i$ and can be calculated by the market share and total demand predicted by MAP-EF. The ticket revenue is defined as the price per ticket multiplied by airline A's total passenger demand on route R_i. $\sum_{R_i \in \mathcal{R}} N_i(f_i)$ is the sum of profits for all routes. Because routes (e.g., the SFO-LAX and JFK-LAX routes shown in Figure 1) can differ dramatically in profitability, and as these numbers are basically estimated by our market share and demand prediction algorithms, it follows that the individual terms $N_i(f_i)$ in our objective function can vary dramatically in form from one route to another. This is what makes using a single form of objective function difficult. Constraint (7)

specifies the bound of frequency for each route as f_i^{max} and that the frequencies are nonnegative integers. The frequency bound $f_i^{max} = \min(\lfloor \frac{b}{C_i} \rfloor, f^{BTS})$, where f^{BTS} is the maximum frequency discovered in the BTS dataset, so that f_i does not exceed the feasible frequency bound. Constraint (8) indicates the budget constraint. As discussed in Section 2, this optimization problem belongs to the hardest class of KPs and RAPs, mainly for two reasons: (1) the solution space is exponentially large as $\Pi_i f_i^{max}$, and (2) the objective function is neither convex nor concave (thus not linear).

Before presenting the complexity analysis and the algorithms to solve MAP, we introduce two important assumptions:

—*Independence of routes*: We assume that varying the frequency of a route does not affect other routes' total demand and market share. This assumption is valid for most cases where the passenger's traveling pattern on one route is independent of the other routes.

—*Monotonicity of revenue and cost*: We assume that revenue (respectively cost) is a monotonically increasing nonlinear (respectively linear) function of frequency. As shown in Figure 1, this assumption is intuitive. Increasing frequency leads to larger market share (and thus higher revenue) and higher operation cost. Therefore, the monotonicity of revenue and cost does not imply the monotonicity of profit. This prevents us from applying many well-developed algorithms for KPs and RAPs.

We show the computational hardness of the MAP problem with the following theorem.

THEOREM 5.3. *The decision version of MAP (D-MAP) that answers the question* "Is there a frequency strategy whose profit exceeds a threshold P without violating budget b?" *is NP-complete.*

PROOF. We first show that there is a polynomial reduction from the KP to D-MAP. In the KP, there are n items (each has p_i profit and w_i cost), budget limit W, a threshold V, and the objective is to find a set of items whose profit is no smaller than V without violating budget W. This a special instance of D-MAP where $b = W$, $P = V$, $C_i = w_i$, and the profit-frequency function of route R_i is linear with the slope of p_i, and the reduction from the KP to profit maximization can be done in polynomial time.

Next, we show its correctness—that is, Y is a *yes* instance of the KP $\Longleftrightarrow X$ reduced from Y is a *yes* instance for D-MAP.

"If" direction. Assume that there exists a solution s for a KP instance. We can create a frequency strategy $\langle f_i \rangle$, where f_i is the number of ith items in the solution s; s can achieve a profit of at least V without violating budget W, and so is $\langle f_i \rangle$ because $b = W$, $P = V$, $C_i = w_i$, and the same linear profit function.

"Only if" direction. Assume that a frequency strategy $\langle f_i \rangle$ achieves a profit of at least P without violating budget b. We can create a knapsack solution s to select f_i items for the ith item; s is a valid solution of the KP because $b = W$, $P = V$, $C_i = w_i$, and the same linear profit function.

Finally, D-MAP is in $\mathcal{NP}$ because the verification process to check whether a frequency strategy's profit exceeds a threshold requires $|\mathcal{R}|$ oracle queries. Note that each oracle query takes a constant time for a given route and airline. Thus, the verification process takes polynomial time. □

Due to the computational complexity, it is intractable to find an optimal solution of the MAP problem by performing brute-force search. As a result, we propose two new approaches: (1) MAP-BBB is a bilevel BB method, and (2) MAP-G is a greedy algorithm operating over groups of routes. Section 8 shows that these two algorithms have complementary performance in terms of runtime (scalability) and optimality.
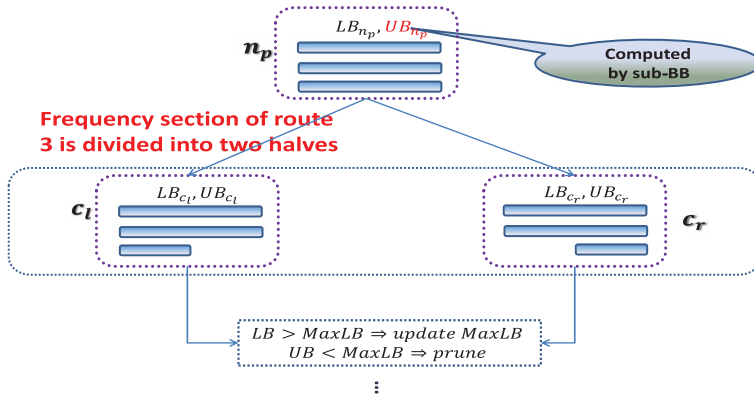
Fig. 3. An example BB process for MAP-BBB with three routes. Each blue bar of node represents a frequency range. Lower BB is used to calculate an upper bound profit of a node. Here, c_l and c_r are two children of the parent node n_p.

5.1. Bilevel BB (MAP-BBB)

BB is a popular search algorithm for combinatorial optimization to efficiently enumerate the entire solution space and find the optimal solution. The solution space of BB can be represented by a rooted tree, and each child of a parent node considers a subset of solution space of its parent node. BB consists of a series of repeated branching and bounding processes. In the branching process, the feasible solution space is divided into a number of smaller subsets, whereas in the bounding process, an upper and a lower bound of the objective (airline profit in this problem) is derived for each subset of solution space. The largest lower bound is updated, and when the upper bound of a certain subset is smaller than the current largest lower bound, that subset is pruned since it does not contain the optimal solution. As we can see, the key of BB lies in finding a tight upper and lower bound for the bounding process, especially for the upper bound. Although most of the BB algorithms for KPs find upper bounds through linear relaxation [Martello and Toth 1977], the complicated form of our prediction model prevents us from implementing such upper bounds. As a result, we propose the MAP-BBB algorithm, which uses another sub-BB search to find an upper bound for the main BB process.

Algorithm 2 shows the MAP-BBB algorithm that builds on classical BB methods. The search for an optimal solution explores a tree whose nodes are labeled with n intervals, one for each of the routes $R_1, \ldots, R_n$ under consideration. Intuitively, if a node is labeled with the interval $[\ell, u]$ for a particular route R_i, it means that we are currently restricting the frequency in the interval $[\ell, u]$ for that route. MAP-BBB also maintains a priority queue Q of all nodes where the head has the largest upper bound profit. The root node has the full range of $[0, f_i^{max}]$ for each route $R_i \in \mathcal{R}$. As we descend the tree, these intervals get smaller. In Figure 3, for example, the parent node n_p is divided into two children nodes c_l and c_r, where the frequency range of the last route of n_p is divided into two halves.

For the bounding process of child node $c \in \{c_l, c_r\}$, upper and lower bound profits of the optimal strategy that can be defined with the frequency range of c are calculated (to be further described later). If its lower bound LB_c is larger than the temporary maximum lower bound, denoted as $MaxLB$ in Algorithm 2, and its lower bound cost C_c of the optimal strategy does not violate the budget b, we update $MaxLB$ (line 8). If its upper bound UB_c is no larger than $MaxLB$ or its lower bound cost C_c violates budget b, we prune this node (line 9). Once we find a node that consists of only nondivisible

---

ALGORITHM 2: MAP–Bilevel Branch and Bound (MAP-BBB)

---

1 $N^* \leftarrow 0, \langle f_i^* \rangle \leftarrow \mathbf{0}$; Create *root*; $MaxLB \leftarrow LB_{root}$;
2 PriorityQueue $Q \leftarrow \emptyset$; $Q.enque(root)$
3 **while** $|Q| > 0$ **do**
4 | $n_p \leftarrow Q.poll()$
5 | **if** n_p *is divisible* **then**
6 | | Create c_l, c_r
7 | | **for** $c \in \{c_l, c_r\}$ **do**
8 | | | **if** $LB_c > MaxLB$ and $C_c \leq b$ **then** $MaxLB \leftarrow LB_c$
9 | | | **if** $UB_c \geq MaxLB$ and $C_c \leq b$ **then** $Q.enque(c)$
10 | **else if** $N^* < N(n_p)$ and $C_{n_p} \leq b$ **then**
11 | | $N^* \leftarrow N(n_p); \langle f_i^* \rangle \leftarrow$ frequency strategy of n_p
12 **return** $\langle f_i^* \rangle, N^*$

---

ranges (i.e., min and max of ranges are all the same) and its cost is within budget b, it is a candidate of the optimal frequency strategy. The optimal strategy is selected among them (line 10). N^* and $\langle f_i^* \rangle$ stand for the obtained optimal profit and optimal frequency strategy, respectively.

Lower bound profit and lower bound cost. Calculating a lower bound of profit LB_c and cost C_c of node c is a key to the MAP-BBB algorithm. We choose the min frequency value for each route's range and calculate its lower bound profit and cost with MAP-EF. This is valid because the optimal profit and the associated cost of node c are always no smaller than that.

Upper bound. As we cannot employ linear relaxation, the upper bound calculation is more challenging. For this, we use another BB process, which is why the proposed algorithm is called a *bilevel branch and bound*. For each range of R_i in node c, we perform the following sub-BB process to find its optimal frequency within the range without a budget constraint. This is good because the optimal profit without budget limit is definitely no smaller than the optimal profit with budget constraint, and it is much faster than the upper bound with budget check across multiple routes. We omit its pseudocode because it is more or less the same as Algorithm 2, except the sub-BB does not check budget violation. To distinguish two BBs, we use different terms: *sub* and *main*. The sub root node is initialized with the specified range of R_i. During the sub branching process, we equally divide a sub node interval into two halves and create two sub children nodes. In the sub bounding process of a sub node, the heuristic for computing a lower bound of the sub-BB is defined as $r_i(f_{random}) - C_i f_{random}$, where f_{random} is a random frequency in the frequency range. This holds since any feasible frequency generates a lower bound of the optimal profit. The heuristic for computing an upper bound of the sub-BB is $r_i(f_{max}) - C_i f_{min}$, where f_{max} and f_{min} are the maximum and minimum frequency of the sub node, which utilizes the monotonicity of revenue and cost. The sum of maximum profits calculated for each route with the sub-BB without budget constraint is always no smaller than that with a budget, and thus sub-BB provides an upper bound profit of the main node. With the lower and upper bounds derived earlier, MAP-BBB eliminates only nonoptimal frequency strategies of routes, which proves correctness and optimality of MAP-BBB. We show the best-case complexity analysis of MAP-BBB in the following theorem.

THEOREM 5.4. *The best-case runtime complexity of MAP-BBB is $O((\log(\max\{f_i^{max}\}))^2 \times n)$, where n is the number of routes.*
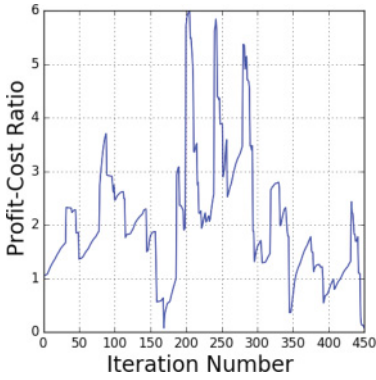
Fig. 4. Best PCR among all routes in each iteration of the naive greedy knapsack algorithm.

ALGORITHM 3: MAP-G
1 $N^* \leftarrow 0, \langle f_i^* \rangle \leftarrow \mathbf{0}$
2 $\mathcal{G} \leftarrow$ divide $\mathcal{R}$ into $\lfloor \frac{n}{k} \rfloor$ groups
3 $b_g \leftarrow \delta, \; \forall g \in \mathcal{G}$;
4 **while** $\sum_{g \in G} b_g \leq b$ **do**
5   **for** $g \in \mathcal{G}$ **do**
6      $\langle f_g^* \rangle \leftarrow$ solve MAP of g with budget b_g
7    $g^* \leftarrow$ the best PCR group
8    $\langle f_i \rangle \leftarrow merge(\{\langle f_g^* \rangle | g \in \mathcal{G}\})$
9    $b_{g^*} \leftarrow b_{g^*} + \delta$;
10   **if** $N^* < N(\langle f_i \rangle)$ **then** $\langle f_i^* \rangle \leftarrow \langle f_i \rangle$
11 **return** $\langle f_i^* \rangle, N^*$

PROOF. The best-case complexity of the BBB algorithm happens when in each branch and bound process, one of the children nodes is pruned. For each route R_i, its maximum frequency is f_i^{max}, and it takes $\log(f_i^{max})$ branches until the frequency in this route is nondivisible; for all routes, it takes $\max\{f_i^{max}\} \times n$ time. However, for each branch, it needs at least $O(\log(\max\{f_i^{max}\}))$ time to solve one sub-BB. Thus, the total complexity is $O((\log(\max\{f_i^{max}\}))^2 \times n)$. □

The worst-case complexity happens when there is no pruning in each BB process. However, the experimental results in Section 8 show that this never happens and that the average runtime of MAP-BBB is promising.

5.2. Greedy Algorithm (MAP-G)

As MAP is NP-hard, it is clear that the exact algorithm MAP-BBB may not be scalable. In this section, we propose a greedy algorithm to solve MAP more efficiently while sacrificing optimality. A naive greedy algorithm is to increase the frequency of the route with the largest marginal profit/cost ratio (PCR), which is known to work very well for conventional KPs. However, unlike conventional KPs where the PCR is fixed (i.e., linear objective function), the PCR in our problem fluctuates as can be seen from the example in Figure 4. According to our experiments, this naive greedy approach on average achieves only 0.66 of the optimal profit.

To address the fluctuating PCRs, we suggest an improved "group-greedy" algorithm (MAP-G). The basic idea of MAP-G is that instead of greedily raising the frequency of a single route with the largest PCR, MAP-G first divides routes into groups and then compares the PCRs of different groups and raises frequencies of the group of routes with the largest PCR. As shown in Algorithm 3, it randomly divides n routes into $\lfloor \frac{n}{q} \rfloor$ groups (line 2), each consisting of q routes.[9] Initially, each group is assigned small budget δ (line 3), and we independently solve MAP within each group using exact algorithms such as MAP-BBB as a subroutine (line 6). We merge all groups' most recent frequency strategy to create a complete frequency strategy of all routes (line 8), and its profit is compared to the temporary maximum profit (line 10). The budget of the group that creates the largest marginal profit density is increased by δ (line 9).

---

[9]The actual number of routes in a group may slightly vary due to the floor operation.

This process iterates until the sum of groups' budgets equals the total budget b. This method is robust with respect to the fluctuating PCR because we calculate the optimal strategy for each group in each iteration. On the other hand, calculating the optimal frequency strategy for a group of routes is much faster than that of all routes because the solution space is much smaller. We set $\delta = b \times \frac{1}{h}$, where $h > 1$ is an integer. The quality of the solution increases as q increases—if $q = n$ and $h = 1$, it finds the optimal solution. The benefit of a small q is the short runtime, although it sacrifices optimality. By adjusting q, MAP-G can trade off between runtime and optimality.

6. EXTENSIONS OF THE FREQUENCY-BASED PROFIT MAXIMIZATION PROBLEM

In the preceding section, we formalized a basic frequency-based profit maximization problem and proposed two search algorithms to solve the problem. In this section, we introduce two extensions of the basic problem: bounded frequencies and future profits.

6.1. Bounded Frequency

As shown in Figure 1, the profit function of an airline is rather irregular. For some routes, it is even monotonically decreasing (respectively increasing) with respect to frequency. For these routes, we find out that the optimal frequency values computed by the proposed algorithms are rather small (respectively large), with an abrupt decrease (respectively increase) compared to the flight frequency strategy of the last quarter. Although in practice even if it is more profitable to change the frequency by a large number, airlines tend to be conservative and make only bounded changes to the frequencies. Interestingly, the BTS dataset shows that in some extreme cases airlines maintain their flight frequency on a route even when profit is negative. As a result, we vary the constraint in Equation (7) and introduce tighter frequency bounds:

$$f_i^0(1 - \alpha) \leq f_i \leq f_i^0(1 + \alpha), \tag{9}$$

where f_i^0 is the frequency on a route R_i in the last quarter and $\alpha \in [0, 1]$ is a bounding factor that indicates the "conservativeness" of an airline. With the new bounded frequency constraint, in the adapted MAP-BBB (MAP-BBB1), the root node's frequency range should be altered $[f_i^0(1 - \alpha), f_i^0(1 + \alpha)]$ and the remaining steps of MAP-BBB are the same.

COROLLARY 6.1. *The best-case runtime complexity of the adapted algorithm MAP-BBB1 is $O((\log(2\alpha \max\{f_i^0\}))^2 \times n)$, where n is the number of routes.*

PROOF. According to Theorem 5.4, the runtime runtime complexity of searching each route R_i is $O((\log(\max\{f_i^{max} - f_i^{min}\}))^2 \times n)$. In MAP-BBB1, $f_i^{max} = f_i^0(1 + \alpha)$, $f_i^{min} = f_i^0(1 - \alpha)$. As a result, the best-case runtime complexity of MAP-BBB1 is $O((\log(2\alpha \max\{f_i^0\}))^2 \times n)$. □

6.2. "Potential" Long-Term Profits

Another interpretation of airlines' conservativeness in changing (decreasing in particular) their frequency strategies is that for some highly competitive routes, these airlines (especially large airlines) are hoping to maintain their market share with relatively large frequency numbers. The maintained market, which may not be profitable in the current stage, will "drive off" the other competing airlines and bring potential profits in the future. To better capture this property, we add an extra linear term (with respect

to frequency) in the objective:[10]

$$\max_{\langle f_i \rangle} \sum_{R_i \in \mathcal{R}} N_i(f_i) + \beta_i f_i \tag{10}$$

$$\text{subject to} \quad f_i^0(1 - \alpha) \leq f_i \leq f_i^0(1 + \alpha), \ f_i \in \mathbb{N}, \forall R_i \in \mathcal{R} \tag{11}$$

$$\sum_{R_i \in \mathcal{R}} C_i f_i \leq b. \tag{12}$$

$\beta_i > 0$ is a route-dependent factor indicating the extent to which future market share is important to the airline. Note that we also set bounded frequency in this extension (as indicated by the constraints in Equation (11)). With larger frequency, one airline may not have larger immediate profit, whereas in the new objective, the linear term, which increases with respect to frequency, indicates the potential increase of future profits.

It is guaranteed that with positive β_i values, the term $\beta_i f_i$ monotonically increases with respect to f_i. Therefore, by merging this term with the revenue term $r_i(f_i)$, the monotonicity of the combined term $r_i(f_i) + \beta_i f_i$ is inherited. Inspired by this, we replace $r_i(f_i)$ with $r_i(f_i) + \beta_i f_i$ in the sub-BB process, and the upper bound of sub-BB is represented as $r_i(f_i^0(1 + \alpha)) + \beta_i f_i^0(1 + \alpha) - C_i f_i^0(1 - \alpha)$. With this adaptation, sub-BB finds a proper upper bound for the main BB process, and the adapted BBB algorithm (MAP-BBB2) still works.

COROLLARY 6.2. *The best-case runtime complexity of MAP-BBB2 is still* $O((\log(2\alpha \max\{f_i^0\}))^2 \times n)$, *where n is the number of routes.*

PROOF. With the adaptation in the sub-BB process, the runtime for the searching process of each route is the same as MAP-BBB1. □

7. MULTIPLE STRATEGIC AIRLINES AND NASH EQUILIBRIUM FREQUENCY STRATEGIES

Previously, we assumed that only one airline is strategic in deciding its frequencies of a certain set of routes $\mathcal{R}$, which may not be true in practice. In this section, we investigate the case where multiple airlines are strategic and change their frequencies using our proposed frequency-based profit maximization algorithm and optimally respond to the current frequency strategies of the other airlines. The selfish and strategic behaviors of the competing airlines result in a noncooperative game among the airlines, for which Nash equilibrium is a natural solution concept. Recall from the definition of Nash equilibrium in Section 4 that in a Nash equilibrium, the frequency strategies of the each airline $A \in \mathcal{A}$ satisfy

$$\max_{\langle f_{i,A} \rangle} \sum_{R_i \in \mathcal{R}} N_{i,A}(f_{i,A}), \tag{13}$$

$$\text{subject to} \quad f_{i,A}^0(1 - \alpha) \leq f_{i,A} \leq f_{i,A}^0(1 + \alpha), \ f_{i,A} \in \mathbb{N}, \forall R_i \in \mathcal{R} \tag{14}$$

$$\sum_{R_i \in \mathcal{R}} C_{i,A} f_i \leq b_A. \tag{15}$$

We add a subscript $_A$ in f_i, f_i^0, C_i, and b to distinguish different airlines. We adapt Algorithm 1 and propose ITER-Freq (ITER with respect to frequency strategies) to

---

[10]Although potential future profits cannot be precisely measured with the simple linear term, to some extent it does reveal an airline's concern for long-term profit maximization. To better study this, a more complicated model (e.g., using MDPs to formulate long-term profits of a series of future quarters or data mining/machine learning methodologies to obtain profit function of a series of future quarters with respect to current frequency) should be applied. We leave the long-term profit maximization problem as future research.

ALGORITHM 4: ITER-Freq

---

1 Randomly assign values to $\langle f_{i,A} \rangle$ within $[f_{i,A}^0(1-\alpha), f_{i,A}^0(1+\alpha)]$
2 **repeat**
3 | **for** $A_i \in \mathcal{A}$ **do**
4 | | $counter \leftarrow 0$
5 | | $f_{i,A}' \leftarrow$ Solution of Equations (13) through (15)
6 | | **if** $|f_{i,A} - f_{i,A}'| \leq \varepsilon$ **then** $counter + +$
7 | $\langle f_{i,A} \rangle \leftarrow \langle f_{i,A}' \rangle$
8 **until** $counter = n$
9 **return** $\langle f_{i,A} \rangle, \forall A \in \mathcal{A}$

---

compute Nash equilibrium frequency strategies, which is shown in Algorithm 4. First, we initialize the frequency of each airline A for each route R_i by randomly assigning a value within a bounded frequency range $[f_{i,A}^0(1-\alpha), f_{i,A}^0(1+\alpha)]$. In the repeat loop (lines 2 through 8), it updates the optimal frequency strategies of each airline with the for loop (lines 3 through 7) and terminates until, for each airline A and for each route R_i, the difference between the updated frequency value $f_{i,A}'$ and the original value $f_{i,A}$ is within a small bound ε. Note that after each iteration, the predicted profit function (i.e., MAP-EF) is updated with the new frequency values of the corresponding airlines. To do this, the features described in Section 4 are also updated.

THEOREM 7.1. *If ITER-Freq returns a solution, then that solution is the Nash equilibrium frequency.*

PROOF. In each repeat loop, ITER-Freq finds the best response of all airlines given their strategies of the last iteration. When the best response strategies of all airlines in the current iteration are equal to the best response strategies of the last iteration, a Nash equilibrium is reached according to its definition. □

The existence and uniqueness of Nash equilibrium heavily depend on the shape (e.g., concave, convex) of the profit functions of the competing airlines. Unfortunately, the profit functions are represented with a very complicated linear algebra form, and thus it is infeasible to analytically study the existence and uniqueness of Nash equilibrium frequencies. At the same time, it is known that the runtime complexity of iteration-based computation of Nash equilibrium is unpredictable, whereas in practice it normally takes a few iterations before convergence of equilibrium (if achieved). As a result, we leave further discussion to Section 8 and study these two aspects experimentally.

8. EXPERIMENTAL RESULTS

In this section, we first compare MAP-EF to several benchmark methods for both route demand and airline market share prediction. We then show the improvement of the optimal profit obtained based on MAP-EF prediction compared to that based on traditional methods. Last, we compare both scalability and optimality of our proposed algorithms to several benchmarks.

8.1. Market Share and Total Demand Prediction

8.1.1. Experimental Environment. For route-specific predictions, we selected the 700 largest routes in terms of the number of passengers. We predicted market share and total demand for the first quarter of 2015, the most recent period for which all data are available, after training with 10 years of data up to the last quarter of 2014. We tested three pre-existing prediction models for both market share and total demand as described in Section 2, and our proposed method, MAP-EF.

For each route, we evaluated with the following three criteria: the correlation coefficient (CC), R^2, mean absolute error (MAE) divided by the maximum market share value of a route or divided by true total demand value. These two metrics are much stricter than pure MAE values. A good predictor will perform well under all three metrics.

8.1.2. Experiment Results.

Market share. Table III (respectively Table IV) summarizes mean values (respectively variances) of the three metrics for all routes (All) and for routes with four or more operating airlines ($|\mathcal{A}| \geq 4$)—as the number of airlines increases, predictions become more challenging. Among existing models, Market1 shows the best performance. Its CC of 0.82 is quite reasonable with a low variance around 0.18, but its poor R^2 of -0.84 indicates some issues. Two metrics revealing contradictory results indicates that the rise or decline of the predicted values may have high error. Thus, Market1 can tell us which airline has a larger market share than another, but its absolute market share value is not reliable. Profit optimization based on the best competitor, Market1, does not make sense because even a small error in predicting market share may lead to a high loss in profit. As shown in Tables III and IV, MAP-EF shows very stable performance in terms of both mean and variance for all metrics, especially when $|\mathcal{A}| \geq 4$. Note that for all tables, ↑ indicates that larger values are preferred and vice versa. The best method is highlighted in yellow.

Total demand. Whereas route-specific market share prediction is a list (one value for each airline), total demand prediction of a route is a scalar value. Thus, we create two lists, one of true demand and one of predicted demand, and perform CC and R^2 analysis. MAE divided by true demand can be measured for each route, and its mean and variance are also summarized in Table V. Again, MAP-EF shows better performance than all other methods.

Effectiveness of additional features in MAP-EF. To evaluate the effectiveness of the equilibrium price feature (Step 2 in Figure 2) and route clustering (Step 3 in Figure 2), we created two simpler prediction models: one without Step 2 and the other without Step 3. On average, the model without the equilibrium feature achieved 88% (respectively 93%) of the full model's CC (respectively R^2), whereas the model without the clustering achieved 90% (respectively 72%). All of these results indicate that profit maximization should be based on the prediction results of MAP-EF, whereas "optimal" strategies obtained with other models are more likely to be far from real optimal strategies. Nevertheless, we note that MAP-EF builds on other preexisting prediction models' findings, and hence we build on top of these prior scientists' work.

8.2. Profit Maximization

8.2.1. Experimental Environment.
We performed experiments on a cluster of 64 machines running Linux with a 2.4GHz Xeon CPU and 24GB RAM. We selected six major airlines and calculated the optimal frequency allocation strategy for their top K biggest markets (a market means an origin-destination pair, i.e., a route). We retrieved real budget information of routes from the BTS dataset such that the budget for a set of routes is the sum of their real budgets. For the scalability test, we chose the number of markets $K = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30\}$. We compare the proposed MAP-BBB and MAP-G with two benchmark algorithms, namely N-Greedy (naive greedy algorithm) and DP (dynamic programming). DP is a well-known approach for KPs and RAPS, and we refer to the appendix for a detailed description of DP. MAP-BBB is the proposed BBB algorithm. We used MAP-BBB as a subroutine for MAP-G to solve the profit maximization subproblem in each group.

Table III. Mean Values of CC, R^2, and MAE/Max for Market Share Predictions

	All			$|\mathcal{A}| \geq 4$		
	CC(↑)	R^2(↑)	MAE/Max(↓)	CC(↑)	R^2(↑)	MAE/Max(↓)
Market1	0.82	-0.84	0.133	0.82	0.18	0.157
Market2	0.48	-3.51	0.188	0.62	-1.92	0.212
Market3	0.34	-6.96	0.27	0.44	-3.99	0.32
MAP-EF	0.96	0.88	0.052	0.95	0.89	0.046

Table IV. Variance of CC, R^2, and MAE/Max of Market Share Predictions

	All			$|\mathcal{A}| \geq 4$		
	CC(↓)	R^2(↓)	MAE/Max(↓)	CC(↓)	R^2(↓)	MAE/Max(↓)
Market1	0.18	132.2	0.06	0.09	4.29	0.08
Market2	0.54	109.7	0.08	0.31	26.77	0.09
Market3	0.55	964.9	0.10	0.35	38.66	0.10
MAP-EF	0.03	0.12	0.01	0.01	0.05	0.01

Table V. CC, R^2, and MAE Divided by True Demand of Total Demand Predictions

	CC(↑)	R^2(↑)	MAE/True (Mean,↓)	MAE/True (Variance,↓)
Demand1	0.77	0.31	0.29	0.06
Demand2	0.49	−1.4	0.57	0.1
Demand3	0.56	−0.8	0.49	0.08
MAP-EF	0.98	0.96	0.07	0.004

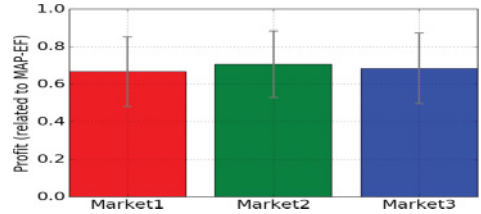*Note*: The variance of CC and R^2 cannot be defined because it predicts a scalar value for each route.

Fig. 5. Profit maximization with different prediction methods (relative to MAP-EF) and 90% confidence interval are shown.

8.2.2. Maximum Profit of Different Predictors. Ground truth is not available for the true maximal profit an airline could possibly make, as their actual quarterly profit may not reflect the total profit that they could have made. In theory, MAP-BBB is guaranteed to produce the maximal profit under the assumptions made in our model. As we see in the following, MAP-BBB works when the number of markets is relatively small. To explore the profits obtained using predictions made by pre-existing market share predictors [Hansen 1990; Suzuki 2000; Wei and Hansen 2005], we compare them to the profits obtained using the MAP-EF predictor. We also obtain the optimal frequency allocation strategies using MAP-BBB for $K \leq 6$ based on different prediction methods and then calculate the corresponding profit of these strategies using MAP-EF (for larger K, MAP-BBB, as an exact algorithm, takes an inordinate amount of time). As shown in Figure 5, the optimal profit obtained by all existing methods has a gap of around 30% compared to MAP-EF, which is a huge amount considering the scale of profit (typically several million dollars per route per quarter).

8.2.3. Comparing Different Optimization Algorithms. In this section, we first compare several algorithms that are targeted at small K (number of markets) values, such as DP, MAP-BBB, and N-Greedy. Using these algorithms as subroutines, we then test the performance of MAP-G with large K values. Average optimized profits of airlines and runtime are summarized in Figures 6 and 7. "Real" stands for the average profit of airlines calculated with real frequency strategies in the BTS dataset. Interestingly, Real shows worse results than even the simple N-Greedy. We see that past market share models can only capture about 75% of the profit generated by MAP-EF. This is a substantial loss in absolute dollars.

Profit and runtime analysis for $K \leq 6$. As we can see from Figure 6(a), since both DP and MAP-BBB are exact algorithms, their maximum profit values are the same and are the largest. However, DP's runtime, shown in Figure 6(b), grows exponentially, and we could not finish it on time for $K \geq 4$. N-Greedy cannot achieve a reasonable profit. When $K = 2$, it has around 70% of the optimal profit calculated with exact algorithms;
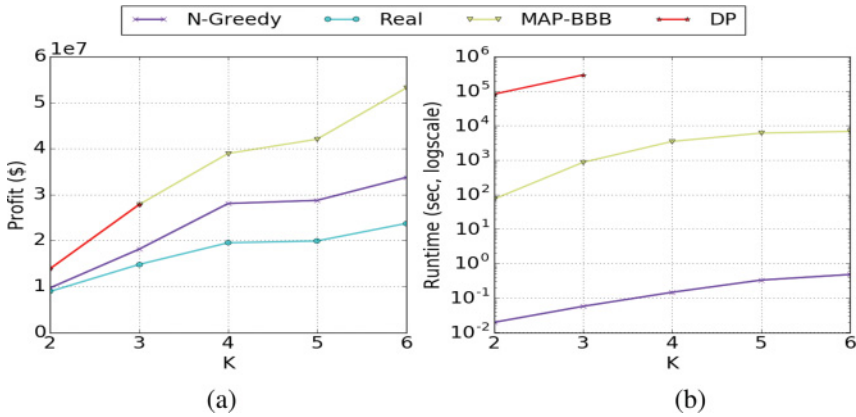
Fig. 6. Optimized profits (a) and runtime for $K \leq 6$ (b).
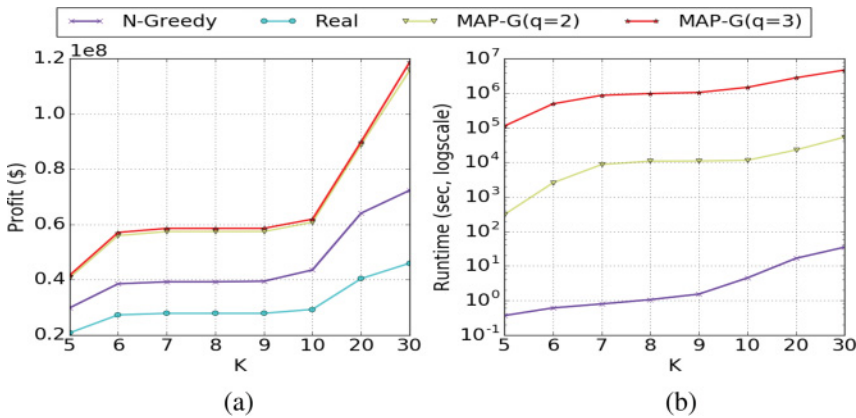
Fig. 7. Optimized profits (a) and runtime for $K \geq 5$ (b), with the number of routes in a group set to $h = 100$.

Table VI. Profit Maximization of MAP-G(BBB) with
$q = 3$ (Relative to the Optimal Profit)

	$K = 4$	$K = 5$	$K = 6$
Profit	0.99	0.97	0.99

as K increases, however, it decreases to only 60%. Before $K = 5$, MAP-BBB can find the optimal profit in a few seconds, whereas DP takes around 10,000 seconds in average. This is due to its pseudopolynomial runtime, which behaves more like exponential runtime than polynomial time. This indicates that MAP-BBB scales really well. The MAP-BBB is more suitable as the subsolver of MAP-G.

Profit and runtime analysis for $K \geq 5$. Table VI shows the optimality of MAP-G when $K = 4$, 5, or 6. The exact maximum profit when $K \geq 7$ cannot be computed, as the exact algorithms either run out of memory or take inordinately long. But in cases when $K \leq 6$, MAP-G achieves 99% of the maximum profit, suggesting that it does very well. As expected, MAP-G is robust with respect to the problem of fluctuating profit and independent processing of each route from which N-Greedy suffers. Figure 7(a) shows that MAP-G is consistently better than N-Greedy and that N-Greedy obtains only about 70% of MAP-G's profit, suggesting that MAP-G is far superior. We confirm
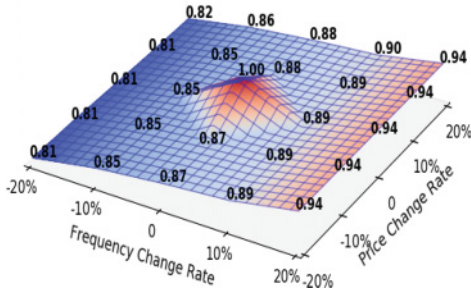
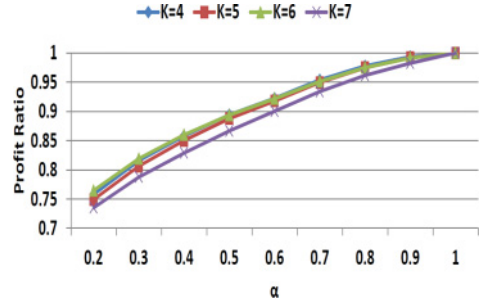Fig. 8. Profits by varying price and frequency of other competitors.

Fig. 9. Ratio of optimal profits for bounded frequency and unbounded frequency.

that $q = 3$ (i.e., groups of similar routes are limited to have three routes in them) leads to an increase in profit than when $q = 2$ for MAP-G—in Figure 7(a), MAP-G with $q = 3$ is consistently better than $q = 2$—which corresponds to our conjecture that optimality will be improved as the size of groups increases.

The profit increase from Real to the optimized one is quite impressive. Airlines can increase their profits by at least 55% by adopting MAP based on our model with available data. But please note that we assume the behaviors of other airlines are fixed, which is not the case in reality. In the following, we will test the robustness of MAP (i.e., when the frequency and pricing strategies of other airlines also vary (and are not known in advance), how much profit can be achieved by the airline to be optimized).

Robustness analysis. In the preceding experiments, we assume that the frequency allocation and pricing strategies of all other airlines are fixed and known in advance. We now test the robustness of our approach—that is, we test the airline profits obtained by our approach when the frequency and pricing strategies of other airlines vary. To do this, we randomly varied price (respectively frequency) by a rate of {[−20%, −10%), [−10%, 0), 0, [0, 10%), [10%, 20%]}. We first compute the optimal strategy by assuming that the other airlines' behaviors are fixed (because they are unknown). We then evaluate this obtained optimal strategy with the prediction model of the changed price and the frequency strategies of other airlines. We computed the ratio of the suboptimal profit (not knowing other airlines' price/frequency changes) to the optimal profit (knowing the changed price/frequency strategies of other airlines), and to calculate the exact optimal profit on time, we tested K up to 6. For each frequency and price change rate combination, we generated 30 scenarios with all different random seeds, and profits are averaged on them. As shown in Figure 8, this profit ratio is always larger than 0.8, which indicates robustness of our MAP approach even with a 40% fluctuation of both price and frequency. Moreover, it is shown that for a fixed frequency change rate, the profit ratio is almost the same for all price change rates, which means that frequency change is more effective in airlines' competition. Last, the largest ratio happens in the case that competitors increase frequencies, which means that MAP is more effective in competitive markets.

8.3. Extensions of Profit Maximization

Bounded frequencies. Figure 9 shows the averaged optimal profit values of airlines operating on different number of routes with varying frequency ranges, where the x-axis represents different values of α from 0 to 0.8 (recall that the frequency range is set as $[f_i^0(1 - \alpha), f_i^0(1 + \alpha)]$), the y-axis represents the optimal profit with bounded frequencies (related to the optimal profit with unbounded frequencies), and the K
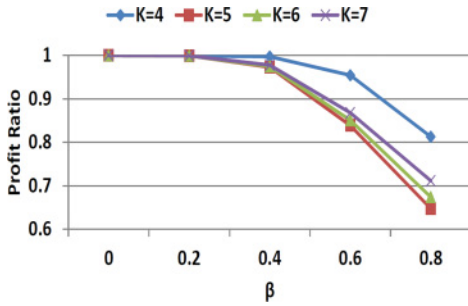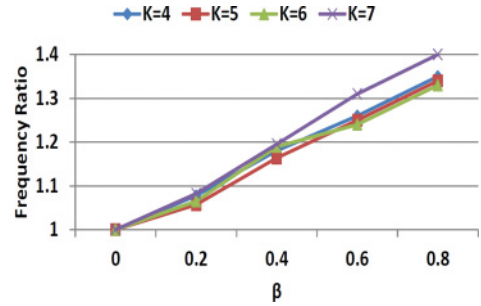
Fig. 10.   Ratio of optimal profits.

Fig. 11.   Ratio of frequency values.

Table VII. Convergence Behavior of Nash Equilibrium Frequency

K	1	2	3
Percentage of convergence	5/10	4/10	5/9
Number of iterations	1.8	2	2.4

value in the legend means the number of routes. With larger α values (i.e., larger frequency ranges), the airlines can obtain larger short-term profits. However, in practice, airlines normally do not make abrupt changes in their flight frequencies—even when larger changes in profits may lead to more profits, which implies the conservativeness of airlines.

Considering potential profits. To evaluate the effect of airlines' concern on potential profits, we first compare the optimal profit obtained in the extended profit optimization problem in Equations (10) through (12) to that computed by the basic profit optimization problem in Equations (6) through (8). To reveal a different extent of importance of potential profits, we set β_i as the airline's per-flight cost of a route R_i multiplied by a varying factor β. As shown in Figure 10, the x-axis denotes different β values from 0.4 to 0.8, the y-axis denotes the ratio of profits obtained by the preceding two optimization problems, and K denotes the number of routes considered.

As we can see, the profit ratio is always smaller than 1, and it decreases with respect to the value of β. Meanwhile, Figure 11 shows the ratio of the averaged optimal frequency values obtained by the extended profit optimization problem and the basic profit optimization problem. For all β values, the frequency ratio is always larger than 1, and it increases with respect to the value of β. Combining these two figures, we can see that (1) when future profit is taken into consideration, airlines do not "rush" for immediate profit, and (2) when airlines put more weight on potential profits, they tend to invest more in flight frequencies even if not more profitable currently.

8.4. Equilibrium Frequencies

Here, we evaluate the convergence behavior and profits obtained with Nash equilibrium frequency. To do this, we select three of the five largest airlines (denoted as A_1, A_2, A_3, A_4, A_5) in the United States, which we assume are strategic and use Nash equilibrium frequency strategies, and then we evaluate when these three airlines are competing over a different number of routes ($K = 1, 2, 3$). Since there are a total of 10 combinations of airlines ($C_3^5 = 10$), for each number of routes, all of the following results are averaged over 10 cases. Table VII shows the percentage of convergence[11]

---

[11]Note that we regard a case as not able to reach Nash equilibrium when the ITER-Freq algorithm does not converge within 10 iterations.

to equilibrium and, if a case converges, the averaged number of iterations required to reach equilibrium (not counting the cases where it does not converge). For the percentage of convergence, the number in the denominator means the number of cases reported. Note that for $K = 3$, this number is not equal to 10, because for one case, not all of the three airlines are operating over the same three routes simultaneously. We note that some cases do not converge to a Nash equilibrium within 10 iterations. However, for those cases that do converge, they converge very quickly—1.8 to 2.4 iterations on average. Our interpretation for the cases that do not converge is that for these routes, the profit function of certain airline(s) obtained with our MAP-EF method is not concave, and the frequency values are restricted to natural number. We prove in the following theorem that when the frequency domain is relaxed to continuous and the profit function of each competing airline is concave, Nash equilibrium always exists.

THEOREM 8.1. *When the frequency domain of airlines (described in Equations (14) and (15)) is relaxed from a natural number (i.e., $f_{i,A} \in \mathbb{N}$) to a positive real number (i.e., $f_{i,A} \in \mathbb{R}^+$), and the profit function $\sum_{R_i \in \mathcal{R}} N_{i,A}(f_{i,A})$ of each airline A is concave with respect to its frequency strategy $\mathbf{f}_A$, Nash equilibrium (with respect to frequency) always exists.*

PROOF. Denote the domain described in Equations (14) and (15) (with linear relaxation) as D. Accordingly, D_A denotes the domain for each airline A, and D_{-A} denotes the domain for all other competing airlines $A \in \mathcal{A}$ except A. For each airline A as a player, given the frequency strategies $\mathbf{f}_{-A}$ of all other competing airlines, it will select an optimal frequency strategy that maximizes its profit function, which is a mapping $\phi_A : D_A \rightarrow D_{-A}$ defined as $\phi_A(\mathbf{f}_{-A}) = \mathbf{f}_A$. Furthermore, according to Equations (14) and (15), both the domain and range of the mapping $\phi_A(\mathbf{f}_{-A})$ are compact (closed and bounded). Combining the compactness of the domain and range of the mapping with the concavity of airline A's profit function $\sum_{R_i \in \mathcal{R}} N_{i,A}(f_{i,A})$, it is guaranteed that the mapping $\phi_A(\mathbf{f}_{-A})$ is nonempty (*existence*). The compactness of D_{-A} along with the linear relaxation also implies that the mapping $\phi_A(\mathbf{f}_{-A})$ is *upper hemicontinuous*.

The existence and upper hemicontinuity of the mapping $\phi_A(\mathbf{f}_{-A})$ are naturally inherited by the mapping $\phi : D \rightarrow D$ defined as $\phi(\mathbf{f}) = \times_{A \in \mathcal{A}} \phi_A(\mathbf{f}_{-A})$, where $\mathbf{f}$ is a matrix that denotes the frequency strategies of all ompeting airlines. In addition, Equations (14) and (15) (with linear relaxation) imply that D is a *simplex* (and closed). Applying Kakutani's fixed point theorem [Kakutani et al. 1941], we conclude that there always exists a fixed point $\mathbf{f}$ in domain D that maps to a same point $\mathbf{f}$ in the range D with the mapping $\phi(\mathbf{f})$. As a result, with the linear relaxation and concavity assumption stated previously, the Nash equilibrium (with respect to frequency) always exists. □

It is worth mentioning that in this problem, the concavity of airlines' profit functions is not a necessary condition (but only sufficient condition) of the existence of the Nash equilibrium. Nonetheless, evidence (e.g., Figure 1, where profit airlines' functions of one route are displayed) shows that the profit functions of airlines are sometimes nonconcave, which we believe is a major reason some cases do not converge to equilibrium within 10 iterations. However, it is still possible that the Nash equilibrium exists when the profit functions of some competing airlines are not globally concave but, for example, are locally concave.

To evaluate the profits obtained by airlines with equilibrium strategy, we compare the profits of the airlines in three situations: (1) airlines use the current frequency strategy, (2) only one airline is strategic, and (3) multiple airlines are strategic. As shown in Table VIII, we show the profits obtained by the five airlines over different numbers of routes. For each cell in the table, the left number in the bracket denotes the ratio of profits in the second and first situations, whereas the right number denotes the ratio of profits for the third and first situations. Note that "N/A" means not applicable. This

Table VIII. Profits Obtained by Airlines
Under Equilibrium

Airline \ K	1	2	3
A_1	[1.27,1.30]	[1.37,1.44]	[1.36,1.43]
A_2	[2.18,2.73]	[2.81,2.94]	[2.30,2.27]
A_3	[1.51,1.56]	[1.58,1.75]	[1.46,1.69]
A_4	[N/A]	[N/A]	[1.34,1.27]
A_5	[1.33,1.67]	[1.14,1.30]	[2.51,2.90]

is because for $K = 1$ and $K = 2$, all cases where airline A_4 is involved do not converge to equilibrium, and thus we do not record the profits for these cases. All profit ratios are averaged among the cases where equilibrium frequencies converge. For example, for airline A_1 and $K = 1$, if in four out of six ($= C_2^4$) cases the equilibrium frequencies converge, both the left and right profit ratio numbers for A_1 and $K = 1$ are averaged over four cases. As we can see, first, for all numbers of routes and all airlines, the profits obtained in both the second (only one airline is strategic) and third (when multiple airlines are strategic) situations are larger than that obtained in the first situation (airlines use current frequency strategy). Second, it shows, somewhat counterintuitively, that with a few exceptions (highlighted in yellow), the profits obtained when multiple airlines are strategic are mostly even larger than those obtained when only one airline is strategic. Note that this result is applicable under the assumption that each airline is only optimizing one-shot profit. When each airline is only concerned about short-term profit, instead of "competing" with each other by increasing frequencies, airlines tend to "concede" and decrease their frequencies for routes that are not profitable. This result, together with Figures 9 through 11, is a vast improvement that provides an enlightenment to the researchers for a more precise profit prediction model that takes into account not only profits in the current period but also profits in the future.

9. CONCLUSION AND FUTURE WORK

In this article, we propose the MAP framework for optimal frequency allocation over multiple routes. We make three key contributions. First, we design a novel ensemble predictor based on past regression-based work, clustering techniques, and game theoretic analysis, which have never been utilized for this purpose. Second, based on the prediction, we design several algorithms to solve the profit maximization problem. Third, we conduct extensive experiments. We show that the prediction performance of MAP-EF is much better than past methods. We also compare runtime and optimality of our proposed optimization algorithms with benchmarks. In addition, we show that MAP increases profitability per route by at least 55%. Even if there are factors that we did not consider on the basis of open source data and these numbers would be reduced substantially if private airline data were used, the increased profitability would still be substantial. Furthermore, we show that considering bounded frequencies and future profits, the one-shot profit might be less compared to that without these concerns, which to some extent reveals the practical implications of the current airline industry—airlines are usually conservative in changing their frequencies and more concerned about long-term profits. Last, we show the optimal frequencies and profits of airlines when there are multiple strategic airlines.

We consider two potential directions for future work. First, our current optimization considers only frequency allocation. In the future, we may also consider joint optimization over frequency and pricing. However, the monotonicity property of revenue and

cost is not guaranteed with ticket pricing,[12] and it is already one of the hardest KPs and RAPs with only frequency being considered. Second, we currently use a simple linear term (with respect to frequency) to evaluate the effects of current frequency strategies on future profits, which has limited veracity to represent long-term profits. To more explicitly model future profits, a more realistic model should be employed, such as MDPs, POMDPs, or more sophisticated data mining/machine learning methodologies.

APPENDIXES

A. FEATURE TABLES GENERATED AFTER STEP 2 OF MAP-EF

Tables IX and X separately present the example tables of variables for market share and total demand prediction, which are generated in Step 2 of MAP-EF. We use Table IX to illustrate. In this table, the first four columns (Route, Quarter, Airline, and Existing Variables) are existing features that are also used in past works. Column 5 is the output in Step 1 of MAP-EF, including the parameter values η_k in the math model and the intermediate predicted market share values. Column 6 is the difference between the Nash equilibrium price and the real price of an airline, which is obtained in Step 2. The highlighted last column (True Market Share) indicates the dependent variable to be predicted. Based on both the market share–related features extracted in Table IX and the demand-related features in Table X, we further employ clustering techniques in Step 3 to find similar routes.

Table IX. Variables to Predict Airline Market Share of R_1 by MAP-EF

Route	Quarter	Airline	Existing Variables	Outputs of Step 1	Equil. Price − Real Price	True Market Share
R_1	2010Q1	A_1	Price, Freq, Delay, $\cdots$	15% (Predicted value), η_k (Parameters)	-50	10%
R_1	2010Q1	A_2	Price, Freq, Delay, $\cdots$	85%, η_k	60	90%
R_2	2010Q1	A_1	Price, Freq, Delay, $\cdots$	11% (Predicted value), η_k (Parameters)	-10	5%
R_2	2010Q1	A_2	Price, Freq, Delay, $\cdots$	89%, η_k	50	95%
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Note: The True Market Share column (highlighted) indicates the dependent variable to be predicted.

Table X. Variables to Predict Total Demand of R_1 by MAP-EF

Route	Quarter	Existing Variables	Outputs of Step 1	Equilibrium Price	True Total Demand
R_1	2010Q1	Avg Price, Population, Income, $\cdots$	20,000 (Predicted value), λ_k (Parameters)	$340	22,000
R_1	2010Q2	Avg Price, Population, Income, $\cdots$	25,000, λ_k	$350	24,000
R_2	2010Q1	Avg Price, Population, Income, $\cdots$	19,000 (Predicted value), λ_k (Parameters)	$360	21,000
R_2	2010Q2	Avg Price, Population, Income, $\cdots$	18,000, λ_k	$250	20,000
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Note: The True Total Demand column (highlighted) indicates the dependent variable to be predicted.

B. PREDICTION RESULTS OF OTHER REGRESSION TECHNIQUES

Tables XI and XII show the market share prediction performance of all tested regression algorithms, including MAP-EF(Ridge), MAP-EF(Lasso), MAP-EF(Lars), and MAP-EF(TheilSen). We refer to Section 8.1 for a description of the experiment settings and the three metrics (CC, R^2, MAE/Max) for measuring prediction performance. As we can see in Table XI, our selected MAP-EF with RANSAC and GPR (herein called MAP-EF for short) shows the best performance in the mean values of all metrics. Although the

---

[12]By decreasing ticket price, market share will increase but revenue may or may not increase due to the change of price.

Table XI. Mean Values of CC, R^2 and MAE/Max for Market Share
Predictions. ↑ Indicates that Larger Values are Preferred and
Vice Versa. The Best Method is Highlighted in Yellow

	All			$|\mathcal{A}| \geq 4$		
	CC(↑)	R^2(↑)	MAE/ Max(↓)	CC(↑)	R^2(↑)	MAE/ Max(↓)
Market1	0.82	−0.84	0.133	0.82	0.18	0.157
Market2	0.48	−3.51	0.188	0.62	−1.92	0.212
Market3	0.34	−6.96	0.27	0.44	−3.99	0.32
MAP-EF(Ridge)	0.92	0.65	0.07	0.91	0.64	0.07
MAP-EF(Lasso)	0.93	0.71	0.06	0.93	0.69	0.06
MAP-EF(Lars)	0.94	0.71	0.06	0.91	0.70	0.06
MAP-EF(TheilSen)	0.82	0.13	0.07	0.81	0.70	0.07
MAP-EF	0.96	0.88	0.052	0.95	0.89	0.046

Table XII. Variance of CC, R^2 and MAE/Max of Market Share Predictions

	All			$|\mathcal{A}| \geq 4$		
	CC(↓)	R^2(↓)	MAE/ Max(↓)	CC(↓)	R^2(↓)	MAE/ Max(↓)
Market1	0.18	132.2	0.06	0.09	4.29	0.08
Market2	0.54	109.7	0.08	0.31	26.77	0.09
Market3	0.55	964.9	0.10	0.35	38.66	0.10
MAP-EF(Ridge)	0.06	2.66	0.02	0.05	3.07	0.03
MAP-EF(Lasso)	0.06	2.31	0.02	0.04	3.2	0.02
MAP-EF(Lars)	0.06	2.33	0.02	0.04	3.1	0.03
MAP-EF(TheilSen)	0.23	18.54	0.02	0.21	0.13	0.02
MAP-EF	0.03	0.12	0.01	0.01	0.05	0.01

Table XIII. CC, R^2 and MAE Divided by True Demand of Total
Demand Predictions

	CC(↑)	R^2(↑)	MAE/True (Mean,↓)	MAE/True (Variance,↓)
Demand1	0.77	0.31	0.29	0.06
Demand2	0.49	−1.4	0.57	0.1
Demand3	0.56	−0.8	0.49	0.08
MAP-EF(Ridge)	0.34	−4.68	0.22	0.07
MAP-EF(Lasso)	0.94	0.92	0.12	0.02
MAP-EF(Lars)	0.95	0.93	0.12	0.02
MAP-EF(TheilSen)	0.95	0.88	0.14	0.01
MAP-EF	0.98	0.96	0.07	0.004

Note: The variance of CC and R^2 cannot be defined because it predicts a
scalar value for each route.

mean metrics values of some regression algorithms are quite close to MAP-EF (e.g., MAP-EF(Lars)), it is shown in Table XII that their variances are much larger than MAP-EF, which means that others regression algorithms are not as stable as MAP-EF. This is the main reason we chose MAP-EF.

Table XIII shows the total demand prediction results for all regression algorithms. As we can see, the mean values of both CC and R^2 of MAP-EF are larger than the other regression algorithms, whereas the mean and variance values of MAE/True are much smaller. This also indicates the superiority of MAP-EF while predicting total demand.

C. DYNAMIC PROGRAMMING APPROACH

The following recurrence formula shows the suggested dynamic programming approach:

$$N^*(i, w) = \max_{k=0}^{f_i^{max}}\{N^*(i - 1, w - k \cdot C_i) + N_i(k)\},$$

where $N^*(i, w)$ is the best profit we can get when operating flights from R_1 to R_i and the budget is w. $N_i(k)$ is the profit of operating k flights on R_i. Hence, $N^*(n, b)$ will output the optimal strategy of the profit maximization.

Example C.1. Assume that there are three routes, R_1, R_2, and R_3, and $C_1 = C_2 = C_3 = 10$ and $b = 100$. The optimal profit is $N^*(3, 100) = \max\{N^*(2, 100), N^*(2, 90) + N(3, 1), N^*(2, 80) + N(3, 2), \cdots\}$.

D. SCATTER PLOTS

Figure 12 shows full details of market share predictions, especially when $|\mathcal{A}| \geq 4$ (the most challenging prediction case). Market share prediction of a route is evaluated with
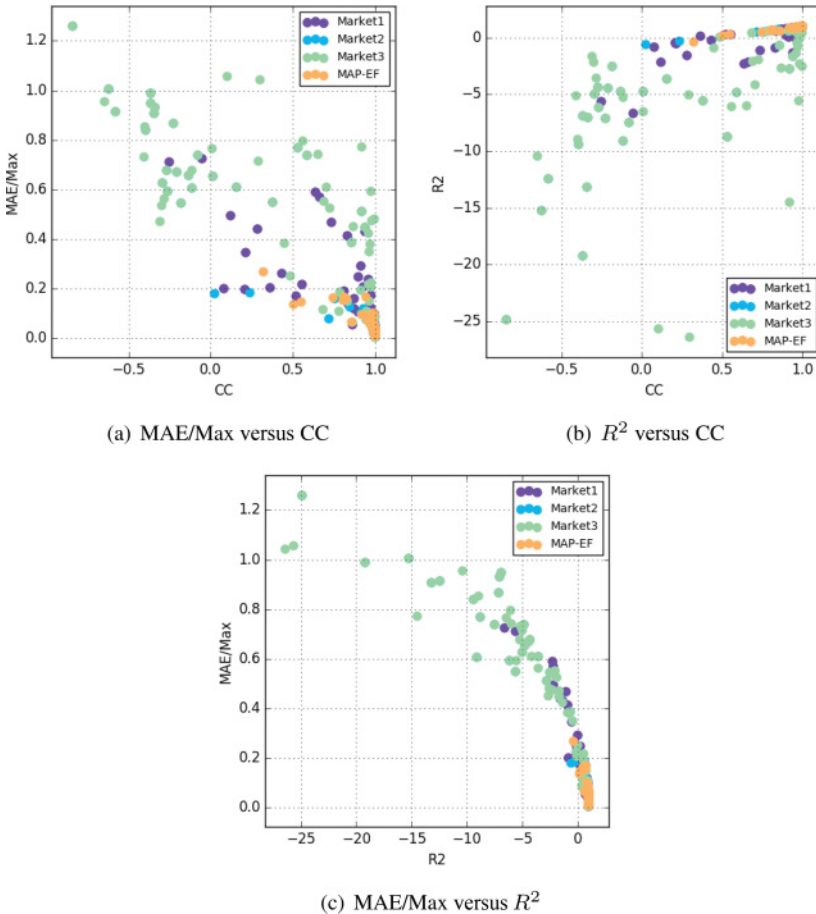

(a) MAE/Max versus CC

(b) R^2 versus CC

(c) MAE/Max versus R^2

Fig. 12.   Scatter plot of market share predictions when $|\mathcal{A}| \geq 4$. Each dot corresponds to a prediction of route.

three metrics: CC, R^2, and MAE/Max. For each pair of possible combinations of these metrics, we show how well each prediction model predicts in the figure. As is shown, MAP-EF's dots are always collected at the "good" corner of the chart. To illustrate, we take Figure 12(a) as an example, which shows the prediction performance of metrics R^2 versus CC. As we know, larger R^2 and CC values indicate better prediction. In this figure, MAP-EF's dots always stay at the top right corner of the chart, which means that it always has large R^2 and CC values. For the dots of existing prediction methods, although the dots of some routes stay at the top right corner, the rest of them do not. Worse still, some dots are far away from the corner. As is pointed out, even a small prediction gap can lead to a large profit loss in reality. These results indicate that market share prediction should be performed by MAP-EF.

REFERENCES

Gershon Alperovich and Yaffa Machnes. 1994. The role of wealth in the demand for international air travel. *Journal of Transport Economics and Policy* 28, 2, 163–173.

B. An, H. Chen, N. Park, and V. S. Subrahmanian. 2016. MAP: Frequency-based maximization of airline profits based on an ensemble forecasting approach. In *Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'16)*. 421–430.

David Arthur and Sergei Vassilvitskii. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'07)*. 1027–1035.

Dipasis Bhadra and Jacqueline Kee. 2008. Structure and dynamics of the core US air travel markets: A basic empirical analysis of domestic passenger demand. *Journal of Air Transport Management* 14, 1, 27–39.

Dipasis Bhadra and Michael Wells. 2005. Air travel by state its determinants and contributions in the United States. *Public Works Management and Policy* 10, 2, 119–137.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, New York, NY.

Kurt M. Bretthauer and Bala Shetty. 2002. The nonlinear knapsack problem—algorithms and applications. *European Journal of Operational Research* 138, 3, 459–472.

Alberto Caprara, Hans Kellerer, Ulrich Pferschy, and David Pisinger. 2000. Approximation algorithms for knapsack problems with cardinality constraints. *European Journal of Operational Research* 123, 2, 333–345.

George B. Dantzig. 1957. Discrete-variable extremum problems. *Operations Research* 5, 2, 266–288.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'96)*. 226–231.

Martin A. Fischler and Robert C. Bolles. 1981. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* 24, 6, 381–395.

Mark Hansen. 1990. Airline competition in a hub-dominated environment: An application of noncooperative game theory. *Transportation Research Part B: Methodological* 24, 1, 27–43.

Shizuo Kakutani et al. 1941. *A Generalization of Brouwer's Fixed Point Theorem*. Duke University Press.

H. Keller, U. Pferschy, and D. Pisinger. 2004. *Knapsack Problems*. Springer-Verlag, Berlin, Germany.

Wei Lu, Stratis Ioannidis, Smriti Bhagat, and Laks V. S. Lakshmanan. 2014. Optimal recommendations under attraction, aversion, and social influence. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'14)*. 811–820.

Silvano Martello and Paolo Toth. 1977. Branch-and-bound algorithms for the solution of the general unidimensional knapsack problem. *Advances in Operations Research* 13, 9, 295–301.

National Transportation Safety Bureau (NTSB). 2016. Aviation Accident Database & Synopses. Retrieved January 30, 2017, from http://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx.

Ulrich Pferschy. 1999. Dynamic programming revisited: Improving knapsack algorithms. *Computing* 63, 4, 419–430.

David Pisinger. 2000. A minimal algorithm for the bounded knapsack problem. *INFORMS Journal on Computing* 12, 1, 75–82.

C. E. Rasmussen and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20, 53–65.

Yoshinori Suzuki. 2000. The relationship between on-time performance and airline market share: A new approach. *Transportation Research Part E: Logistics and Transportation Review* 36, 2, 139–154.

Robert L. Thorndike. 1953. Who belongs in the family. *Psychometrika* 18, 4, 267–276.

U.S. Bureau of Economic Analysis (BEA). 2016. Regional Economic Accounts: Download. Retrieved January 30, 2017, from http://www.bea.gov/regional/downloadzip.cfm.

U.S. Bureau of Transportation Statistics (BTS). 2016. Airlines and Airports—Data and Statistical Resources. Retrieved January 30, 2017, from http://www.rita.dot.gov/bts/data_and_statistics/by_mode/airline_and_airports/index.html.

U.S. Census Bureau (CB). 2016. Metropolitan and Micropolitan. Retrieved January 30, 2017, from https://www.census.gov/population/metro/data/.

Wenbin Wei and Mark Hansen. 2005. Impact of aircraft size and seat availability on airlines demand and market share in duopoly markets. *Transportation Research Part E: Logistics and Transportation Review* 41, 4, 315–327.

Jian Xu, Kuang-Chih Lee, Wentong Li, Hang Qi, and Quan Lu. 2015. Smart pacing for effective online ad campaign optimization. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'15)*. 2217–2226.

Weinan Zhang, Shuai Yuan, and Jun Wang. 2014. Optimal real-time bidding for display advertising. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'14)*. 1077–1086.

Yada Zhu, Hongxia Yang, and Jingrui He. 2015. Co-clustering based dual prediction for cargo pricing optimization. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'15)*. 1583–1592.