MAP: Frequency-Based Maximization of Airline Profits based on an Ensemble Forecasting Approach

Bo An
School of Computer Science and Engineering
Nanyang Technological University
boan@ntu.edu.sg

Haipeng Chen
Interdisciplinary Graduate School
Nanyang Technological University
chen0939@ntu.edu.sg

Noseong Park, V.S. Subrahmanian
UMIACS & Department of Computer Science
University of Maryland
{npark,vs}@cs.umd.edu

ABSTRACT

Though there are numerous traditional models to 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 to improve airline profitability. We propose the MAP (Maximizing Airline Profits) 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 with past methods 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 (over 0.95 vs. 0.82 for market share, 0.98 vs. 0.77 for demand) and R^2-values compared with 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. Experimental results show that airlines can increase profits by a significant margin. All experiments were conducted with data aggregated from four sources: US Bureau of Transportation Statistics (BTS), US Bureau of Economic Analysis (BEA), the National Transportation Safety Board (NTSB), and the US Census Bureau (CB).

Keywords

Ensemble Prediction; Regression; Airline Demand and Market Share Prediction; Airline Profit Maximization;

1. INTRODUCTION

Since the deregulation of US airlines in 1978, there has been intense competition amongst airlines for market share and, eventually, profitability. While 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 to improve airline profitability. In this paper, we formally define the MAP problem by allowing airlines to decide the flight frequencies on their flying routes in order 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 paper 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 [16, 24, 26] and demand prediction [6, 8, 9] 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 [16] 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 the 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 upon 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), which 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 w.r.t. the number of routes, 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 [11, 12, 20, 21] for both KPs and RAPs, all previous algorithms fail to solve it

KDD '16, August 13-17, 2016, San Francisco, CA, USA

DOI: http://dx.doi.org/10.1145/2939672.2939726

[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, etc.) to the airline, especially as such data is not freely available.

efficiently. Based on predictions made by 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 approach (MAP-BBB) that computes the true optimal solution, as well as a Greedy algorithm (MAP-G) that more quickly computes suboptimal solutions.

Third, 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 over 0.95, while significantly reducing the variance in our prediction error compared with past works. We also compare both the optimality and scalability of our proposed profit maximization algorithms with several benchmarks. The result shows that our predictions are far superior to past efforts. Moreover, by using MAP, an airline can averagely increase its profit by at least 55% under mild conditions.[2]

The rest of this paper is organized as follows. Section 2 introduces related works. Section 3 discusses the data set used in our study. Section 4 shows our proposed prediction framework based on our ensemble method. Section 5 presents our proposed algorithms for the profit maximization problem. Section 6 conducts extensive experimental evaluations. *For notational convenience, Table 1 summarizes all the notations used in this paper.*

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 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 [16, 24, 26]. For a set $\mathcal{A}$ of airlines in a given route, an airline A_i's market share m_i is modeled as:

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

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 k^{th} variable for airline A_i and η_k is its corresponding weight to be learnt from data. Various variables have been studied. In Market1 [16], the authors consider a route's price, frequency, and the number of stops. Based on the prediction, they analyze competition among airlines. Market2 [24] considers an airline's frequency, delay, and safety. Market3 [26] 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 [6, 8, 9] regression model is the most frequently adopted. Formally, a route's total demand is represented as:

$$d = \prod_k Y_k^{\lambda_k}, \quad (2)$$

where Y_k is a value of the k^{th} variable and λ_k is a parameter to be learnt from data which measures the importance of Y_k. Demand1 [6] considers wealth related variables such as price, income and CPI (consumer price index). Demand2 [9] studies the influence of demographic variables such as price, income and population, as well as hub status (i.e., large, medium, small or non-hub) on passenger demand. Demand3 [8] focuses on variables such as price, income, population and origin-destination distance.

In this paper, we show that 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.

[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.

Table 1: Summary of Notations

$\mathcal{A}$	a set of airlines
A_i	i^{th} airline of $\mathcal{A}$
$\mathcal{R}$	a set of routes
R_i	i^{th} route of $\mathcal{R}$
Symbols related to predict market share of route R	
X_{ik}	the value of the k^{th} variable, e.g., price, of airline A_i
η_k	the weight of X_{ik}
V_i	airline A_i's customer utility in R
m_i	market share of airline A_i in R
Symbols related to predict demand of route R	
Y_k	the value of the k^{th} variable, e.g., population
λ_k	the exponent of Y_k
d	total demand (the number of passengers) of route R
Symbols related to calculate Nash Equilibrium of route R	
p_i	average ticket price of A_i for route R
c_i^{psg}	unit passenger related cost of A_i in route R
$N(A_i)$	utility (profit) of A_i in route R
Symbols related to Profit Maximization of airline A	
C_i	per flight cost of route R_i for a given airline A
f_i	flight frequency of route R_i for a given airline A
$r_i(f_i)$	revenue when A operates f_i flights on route R_i
$N_i(f_i)$	profit (revenue minus total cost) when A operates f_i flights on route R_i
$\langle f_i \rangle$	frequency strategy for all routes in $\mathcal{R}$ of airline A
$\langle f_i^* \rangle$	optimal frequency strategy of airline A
$N(\langle f_i \rangle)$	profit of frequency strategy $\langle f_i \rangle$
b	budget limit of the Profit Maximization problem
K	the number of routes to optimize profit of airline
Symbols related to MAP-BBB in Algorithm 2	
n_p	parent node
c	child node
c_l, c_r	left and right children of a parent node
N^*	optimal profit sought by the algorithm
$MaxLB$	bounding threshold
UB_c, LB_c	lower bound and upper bound profit of node c
C_c	lower bound cost of node c
Symbols related to MAP-G in Algorithm 3	
g	a group of route
q	the number of routes in a group g
$\mathcal{G}$	a set of route groups, $\mathcal{R}$ is divided into $\lvert\mathcal{G}\rvert$ groups
δ	budget allocation unit
$\langle f_g^* \rangle$	optimal frequency strategy of group g

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 [13] in 1957. While all variants of KPs are known to be *NP-hard*, MAP, which has bounded integer variables, and a non-concave (arbitrary) objective function, is notoriously hard [11].

There are some Polynomial Time Approximate Schemes (PTAS) and even Fully-PTAS to solve bounded knapsack problems (BKPs), using greedy algorithms [17], dynamic programming [21] and branch and bound [19]. The first two methods assume linear objective functions, while the last approach uses linear relaxation. Several algorithms have been proposed to solve KPs/RAPs with non-linear objective functions, especially for a concave and non-decreasing

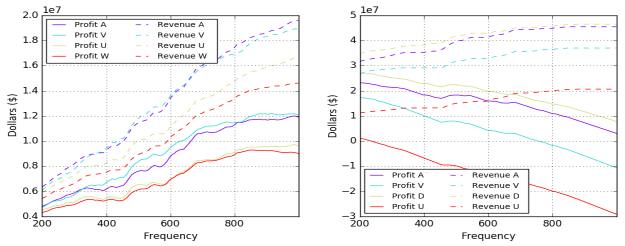


Figure 1: Quarterly revenue (dashed lines) and profit (solid lines) in 10 million dollars generated by MAP-EF. Revenue is defined as average ticket price multiplied by the number of passengers. Cost is per flight cost multiplied by frequency. Note that the profit-frequency curve (revenue minus cost) is very irregular even though the revenue-frequency curve is monotonically increasing. The shapes of the curves are quite different.

objective function [11]. 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 (thus not linear), varies from one route to another, and cannot be described by a simple 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 4 major US 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 papers, including optimal bidding [28, 29], smart pacing [27] and optimal recommendation systems [18]. 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 [18, 28] or via optimization algorithms [27, 29]. While our MAP framework follows this general scheme, 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" data set by aggregating information from 4 publicly available data sources, viz. the Bureau of Transportation Statistics (BTS) [3], the Bureau of Economic Analysis (BEA) [2], the National Transportation Safety Board (NTSB) [4], and the U.S. Census Bureau (Census) [5]. 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 Census release regional income, economic state, and population information while the NTSB provides all safety and accident related information. The BTS provides a quarterly dataset. A detailed description of the variables we consider is presented in Table 2. We plan to publicly release our tools

[3]We will revisit this point after Definition 5.2.

[4]BTS releases a 10% sample of route and airline specific ticket sales logs, from which average ticket price is obtained.

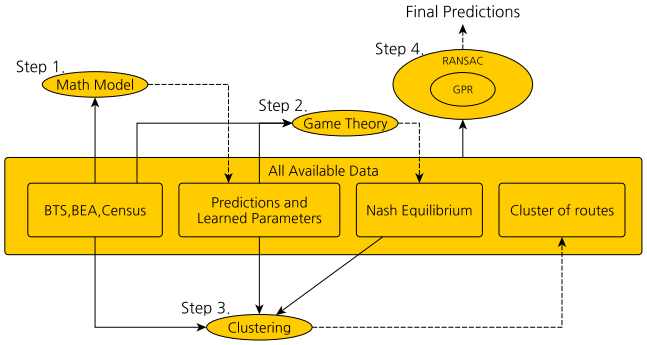


Figure 2: Structure of the MAP-EF prediction model. Rectangles mean data and elipses algorithms. Each algorithm is marked with a processing sequence number. Solid lines mean inputs and dashed lines outputs.

and data set in 2016. Our training data includes 10 years (40 quarters) of data on 13 airlines and 700 routes. Our validation data includes one quarter (all 700 routes and 13 airlines' market share and profits are predicted, leading to a total 9100 predictions in total). This is consistent with most corporate forecasts where forecasts are made for the next quarter[5].

Table 2: Variables used to make predictions by 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

4. ENSEMBLE PREDICTION MODEL

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

Overall architecture. Figure 2 shows the architecture of our ensemble model which runs through 4 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 [16], Market2 [24], Market3 [26]) and demand models (Demand1 [6], Demand2 [9], Demand3 [8]) in order 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 [10].[6] Because this step

[5]These predictions exceed that of related work, e.g., [26], which predicted for 10 airlines with a 10 year history for 200 routes.

[6]In past works for market share prediction, the authors minimize sum-squared-error (SSE), i.e. they try to find parameter values for the math models that minimize $\Sigma_{A_i}(m_i - \hat{m}_i)^2$ where m_i (resp. $\hat{m}_i$) is the true (resp. predicted) market share. In addition to this, regularization minimizes $\sum_{A_i}(m_i - \hat{m}_i)^2 + ||W||$, where W is a

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. At the same time, the competitive pricing behavior among airlines would have further influence. Thus, we include equilibrium pricing calculated based on past models as an additional feature in MAP-EF. At the end of this step, each route has one feature table for market share prediction and another for demand prediction. Each table[7] 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 the 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). At the end of Step 3, we have a comprehensive *MAP Feature Table*. We use a combination of the well-known methods of GPR [22] and RANSAC [15]. To make market share/demand predictions for a route R, this step trains only on data about the group that contains route R, removes outliers using 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[8]. The rationale behind this is that when an airline's quoted price is higher/lower than its equilibrium price, the customer might lose/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 w.r.t. 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 multi-player non-cooperative 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 $A_i \in \mathcal{A}$ w.r.t. a given route:

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

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

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

Eq.(4) specifies the feasible section of price, and Eq.(5) indicates that the total number of passengers cannot exceed its capacity Q_i.

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. In order 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.

[7]Please refer to the Appendix [1] for detailed examples of the merged table.

[8]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.

Algorithm 1: ITER

```
1 Randomly assign values to p_{t-1} within [p_i^min, p_i^max]
2 repeat
3   for A_i ∈ A do
4     counter ← 0
5     p_{i,t} ← Solution of Eqs.(3)-(5)
6     if |p_{i,t} − p_{i,t−1}| ≤ ε then counter++
7   p_{t−1} ← p_t
8 until counter = n
9 return p_t
```

To compute equilibrium price on a route, we propose ITER (Iterative meThod for nash EquilibRium 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 2-9), the algorithm updates optimal prices of all the airlines. Within each repeat loop, there is a for loop (Lines 3-7). 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 $|p_{i,t} - p_{i,t-1}| \le \varepsilon$), which means a Nash equilibrium is obtained.

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

Proof sketch. 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, e.g., 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 of similar routes. In order 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 *silhouette score*[9] [23]. The silhouette score tends to keep increasing (with some glitches) as the number of clusters increases. The elbow method [25] chooses the optimal number of clusters when silhouette scores stabilize. We tested several clustering methods including K-Means^{++}[7] and connectivity-based and density-based clustering algorithms such as DBSCAN [14]. 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,

[9]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 it belongs to, and $b(e)$ is the minimum average distance between e and the data items of clusters it does not belong to.

it incurs additional error. We tested many linear regression algorithms such as Ridge, Lasso, and Lars [10]. None of them could produce stable performance across multiple routes. We then tested a specially devised structure that consists of RANSAC and GPR. RANSAC removes outliers from training data and GPR makes the final prediction only with inliers. GPR is known to be robust because it is much less parametric than the others. Intuitively, GPR says that if features of two airlines are similar, their market shares should also be similar. It assumes that priors follow a Gaussian distribution. After that, GPR also calculates posteriors using the famous bayes rule, $\mathbb{P}(posterior) \propto \mathbb{P}(likelihood)\mathbb{P}(prior)$ that can be analytically calculated using a covariance matrix fitted to the training data. To predict, GPR, using the learned covariance matrix, calculates similarities between a test case and all training cases, and makes a prediction. The entire process can be described by a very complicated linear algebra form. RANSAC Regression is known to be robust w.r.t. noise. In the first step, RANSAC performs regression with a small random subset of training data. In the second step, all training data that does not fit the model constructed during the first step is classified as outliers and others as inliers. These two processes are iterated over time to refine the set of inliers called the *consensus set*. Thus, MAP-EF uses RANSAC with GPR as a base estimator. We show prediction results of other regression techniques in the Appendix [1].

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 airline's profit by optimally allocating quarterly frequencies to routes.

DEFINITION 5.1 (FREQUENCY STRATEGY). *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 non-negative integers.*

Intuitively, the i'th 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 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:*

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

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

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

$N_i(f_i)$ 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 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 (such as 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 non-negative 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 we prevent f_i from exceeding 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 due to two reasons: i) the solution space is exponentially large as $\Pi_i f_i^{max}$, and ii) 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. First, we assume that varying the frequency of a route does not affect other routes' total demand and market share — *Independence of Routes*. This assumption is valid for most cases where the passenger's traveling pattern on one route is independent of the other routes. Second, we assume revenue (resp. cost) is a monotonically increasing non-linear (resp. linear) function of frequency — *Monotonicity of Revenue and Cost*. As shown in Figure 1, this assumption is intuitive. Increasing frequency leads to larger market share (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) which answers the question* "Is there a frequency strategy whose profit exceeds a threshold P without violating budget b?" *is NP-complete.*

Proof sketch. We first show that there is a polynomial reduction from the Knapsack problem to D-MAP. In the Knapsack problem, 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 Knapsack problem to profit maximization can be done in polynomial time.

Next, we show its correctness, i.e., Y is a *yes* instance of the Knapsack problem $\iff$ X reduced from Y is a *yes* instance for D-MAP.

"if" direction. Assume that there exists a solution s for a Knapsack problem instance. We can create a frequency strategy $\langle f_i \rangle$, where f_i is the number of i^{th} item 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 $\langle f_i \rangle$ can achieve a profit of at least P without violating budget b. We can create a Knapsack solution s to select f_i items for i^{th} item. s is a valid solution of the Knapsack problem 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: i) MAP-BBB is a bi-level branch and bound method, and ii) MAP-G is a greedy algorithm operating over groups of routes. Section 6 shows that these two algorithms have complementary performance in terms of runtime (scalability) and optimality.

5.1 Bilevel Branch and Bound (MAP-BBB)

Branch and bound (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, while 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 upper bound. While most of the BB algorithms for Knapsack problems find upper bounds through linear relaxation [19], the complicated form of our prediction model prevents us from implementing such upper bounds. As a result, we propose a novel *Bilevel branch and bound* (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 branch and bound 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 the 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 below). 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, then we update *MaxLB* (line 11). 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 12). Once we find a node that consists of only non-divisible 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 13). 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.

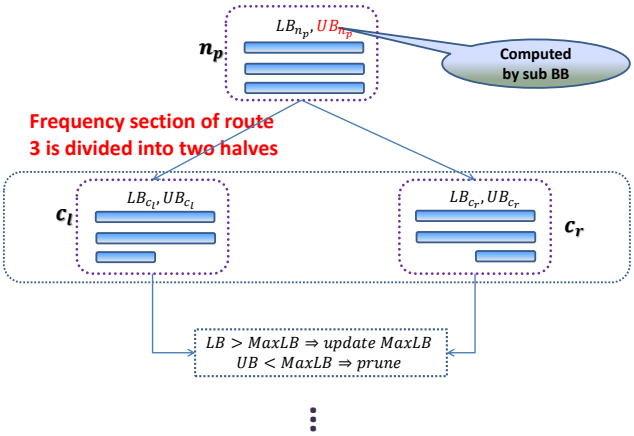


Figure 3: An example branch and bound process for MAP-BBB with three routes. Each blue bar of node represents a frequency range. Lower BB is used to calculate an upper bound profit of a node. c_l and c_r are two children of the parent node n_p.

Algorithm 2: MAP-Bilevel Branch and Bound (MAP-BBB)

```
1  N* ← 0, ⟨f_i*⟩ ← 0
2  Create root
3  MaxLB ← LB_root
4  PriorityQueue Q ← ∅
5  Q.enque(root)
6  while |Q| > 0 do
7      n_p ← Q.poll()
8      if n_p is divisible then
9          Create c_l, c_r
10         for c ∈ {c_l, c_r} do
11             if LB_c > MaxLB and C_c ≤ b then MaxLB ← LB_c
12             if UB_c ≥ MaxLB and C_c ≤ b then Q.enque(c)
13     else if N* < N(n_p) and C_{n_p} ≤ b then
14         N* ← N(n_p)
15         ⟨f_i*⟩ ← frequency strategy of n_p
16 return ⟨f_i*⟩, N*
```

Upper bound. As we cannot employ linear relaxation, upper bound calculation is more challenging. For this, we use another branch and bound 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 branch and bound 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 pseudo code because it is more or less the same as Algorithm 2 except that the sub branch and bound does not check budget violation. In order to distinguish two branch and bounds, 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 branch and bound 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 branch and bound 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 branch-and-bound without budget constraint is always no smaller than that with a budget, thus sub branch-and-bound provides an upper bound profit of the main node. With the lower and upper bounds derived above, MAP-BBB eliminates only non-optimal frequency strategies of routes, which proves correctness and optimality of MAP-BBB. We show the best-case complexity analysis of MAP-BBB:

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.*

Proof sketch. The best-case complexity of the bi-level branch and bound 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. While for each branch it needs at least $O(\log(\max\{f_i^{max}\}))$ time to solve one sub branch and bound. 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 branch and bound process. however, the experimental results in Section 6 show that this never happens and 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 knapsack problems. However, unlike conventional knapsack problems where 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[10]. 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 sub-routine (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 with the temporary maximum profit (line 10). The budget of the group that creates the largest marginal profit density is increased by δ (line 9). This process iterates until the sum of groups' budgets equals to the total budget b. This method is robust w.r.t. the fluctuating PCR because we calculate the optimal strategy for each group in each iteration. On the other hand, calculating 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, while sacrificing optimality. By adjusting q, MAP-G can trade-off between runtime and optimality.

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

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** $\langle f_i^* \rangle, N^*$

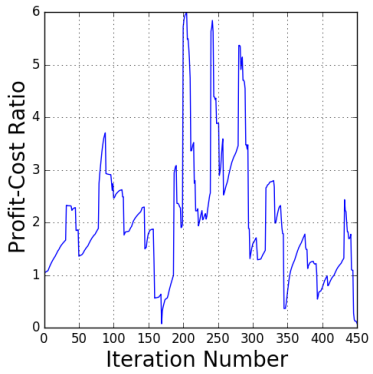


Figure 4: Best Profit-Cost Ratio among all routes in each iteration of the naive greedy Knapsack algorithm.

6. EXPERIMENTAL RESULTS

In this section, we first compare MAP-EF with 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 with that based on traditional methods. Last, we compare both scalability and optimality of our proposed algorithms with several benchmarks.

6.1 Market Share & Total Demand Prediction

6.1.1 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 data. We tested 3 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: i) Correlation Coefficient (CC), ii) R^2, and iii) 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.

6.1.2 Experiment Results

Market Share. Table 3 (resp. Table 4) summarizes mean values (resp. variances) of the three metrics for all routes (All) and

for routes with 4 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 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/decline of the predicted values may have high error. Thus, Market1 can tell us which airline has 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 Table 3, MAP-EF shows very stable performance in terms of both mean and variance for all metrics, especially when $|\mathcal{A}| \geq 4$.

Table 3: 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	0.96	0.88	0.052	0.95	0.89	0.046

Table 4: 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

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 5. Again, MAP-EF shows better performance than all other methods.

All these results indicate that profit maximization should be based on the prediction results of MAP-EF, while "optimal" strategies obtained with other models are more likely to be far from real optimal strategies. Nevertheless, we note that MAP-EF builds upon other pre-existing prediction models' findings and hence we build on top of these prior scientists' work.

6.2 Profit Maximization

6.2.1 Experimental Environment

We performed experiments on a cluster of 64 machines running Linux with 2.4GHz Xeon CPU and 24GB RAMs. 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 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 Appendix [1] for a detailed description of DP. MAP-BBB is the proposed bilevel branch-and-bound algorithm. We used MAP-BBB as subroutine for MAP-G to solve the profit maximization subproblem in each group.

Table 5: CC, R^2 and MAE divided by true demand of Total Demand Predictions. Note that the variance of CC and R^2 cannot be defined because it predicts a scalar value for each route.

	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

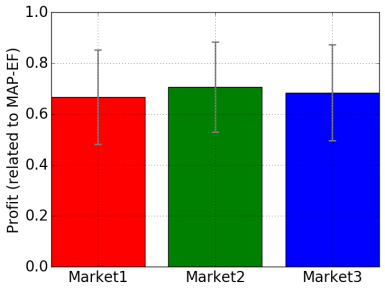


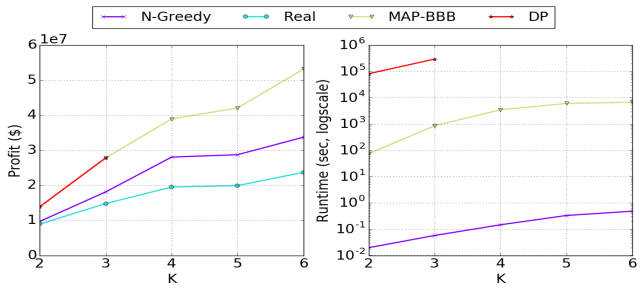
Figure 5: Profit maximization with different prediction methods (relative to MAP-EF) and 90% confidence interval is shown.

6.2.2 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 maade. In theory, MAP-BBB is guaranteed to produce the maximal profit *under the assumptions made in our model*. As we see below, MAP-BBB works when the number of markets is relatively small. In order to explore the profits obtained using predictions made by pre-existing market share predictors [16, 24, 26], we compare them with 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 with MAP-EF, which is a huge amount considering the scale of profit (typically on the order of millions of dollars per route per quarter).

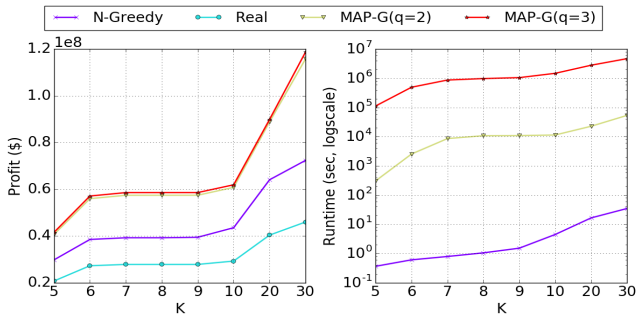
6.2.3 Comparing Different Optimization Algorithms

In this subsection, we first compare several algorithms which are targeted at small K (number of markets) values, such as DP, MAP-BBB, and N-Greedy. Using these algorithms as subroutines, we then tested 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.



(a) (b)

Figure 6: (a) Optimized profits and (b) Runtime for $K \leq 6$. K is the number of routes.



(a) (b)

Figure 7: (a) Optimized profits and (b) Runtime for $K \geq 5$, with the number of routes in a group as $h = 100$.

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; as K increases, however, it decreases to only 60%. Before $K = 5$, MAP-BBB can find the optimal profit in a few seconds while DP takes around 10,000 seconds in average. This is due to its pseudo polynomial 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 6 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 w.r.t. 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 N-Greedy obtains only about 70% of MAP-G's profit, suggesting that MAP-G is far superior. We confirm that $q = 3$ (i.e. groups of similar routes are limited to have 3 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 not known in advance), how much profit can be achieved by the airline to be optimized.

Table 6: Profit maximization of MAP-G(BBB) with $q = 3$ (relative to the optimal profit) when $K = 4$, 5 or 6. K is the total number of routes, q is the number of routes allowed in a group.

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

Robustness Analysis. In the above 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, i.e., 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 (resp. frequency) by a rate in $\{[-20\%, -10\%), [-10\%, 0), 0, [0, 10\%), [10\%, 20\%]\}$. 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 frequency strategies of other airlines. We computed the ratio of the sub-optimal profit (not knowing other airlines' price/frequency changes) to the optimal profit (knowing the changed price/frequency strategies of other airlines) — in order 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 is 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 MAP is more effective at competitive markets.

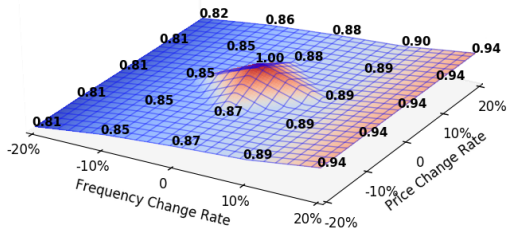


Figure 8: Profits by varying price and frequency of other competitors

7. CONCLUSION AND FUTURE WORK

In this paper, we propose the MAP framework for optimal frequency allocation over multiple routes. We make three key contributions. i) 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; ii) based on the prediction, we design several algorithms to solve the profit maximization problem; iii) 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. We also show that MAP increases profitability per route by at least 55%. Even if there are factors we did not consider on the basis of open source data and these numbers are reduced substantially if private airline data was to be used, the increased profitability would still be substantial.

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[11] and it is already one of the hardest KPs and RAPs with only frequency being considered. Second, this paper assumes the behaviors of all other competing airlines are fixed. When other airlines are also strategic, game theory would be a natural methodology to solve the problem, where instead of a single-airline-centric optimization, a Nash equilibrium will be defined.

Acknowledgements

This research is supported by the National Research Foundation Singapore under its Interactive Digital Media (IDM) Strategic Research Programme.

8. REFERENCES

[1] Appendix. http://all-is-well-everyday.weebly.com or http://www.cs.umd.edu/~npark/kdd2016.pdf.

[2] Bureau of Economic Analysis. http://www.bea.gov/regional/downloadzip.cfm.

[3] Bureau of Transportation Statistics. http://www.rita.dot.gov/bts/data_and_statistics/by_mode/airline_and_airports/index.html.

[4] National Transportation Safety Board. http://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx.

[5] U.S. Census Bureau. https://www.census.gov/population/metro/data/.

[6] G. Alperovich and Y. Machnes. The role of wealth in the demand for international air travel. *Journal of Transport Economics and Policy*, 28(2):163–173, 1994.

[7] D. Arthur and S. Vassilvitskii. K-Means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1027–1035, 2007.

[8] D. Bhadra and J. Kee. 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, 2008.

[9] D. Bhadra and M. Wells. Air travel by state its determinants and contributions in the United States. *Public Works Management & Policy*, 10(2):119–137, 2005.

[10] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

[11] K. M. Bretthauer and B. Shetty. The nonlinear knapsack problem – algorithms and applications. *European Journal of Operational Research*, 138(3):459–472, 2002.

[12] A. Caprara, H. Kellerer, U. Pferschy, and D. Pisinger. Approximation algorithms for knapsack problems with cardinality constraints. *European Journal of Operational Research*, 123(2):333–345, 2000.

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

[13] G. B. Dantzig. Discrete-variable extremum problems. *Operations research*, 5(2):266–288, 1957.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. 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)*, pages 226–231, 1996.

[15] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981.

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

[17] H. Keller, U. Pferschy, and D. Pisinger. *Knapsack problems*. Springer-Verlag Berlin, 2004.

[18] W. Lu, S. Ioannidis, S. Bhagat, and L. V. Lakshmanan. Optimal recommendations under attraction, aversion, and social influence. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 811–820, 2014.

[19] S. Martello and P. Toth. Branch-and-bound algorithms for the solution of the general unidimensional knapsack problem. *Advances in Operations Research*, 13(9):295–301, 1977.

[20] U. Pferschy. Dynamic programming revisited: Improving knapsack algorithms. *Computing*, 63(4):419–430, 1999.

[21] D. Pisinger. A minimal algorithm for the bounded knapsack problem. *INFORMS Journal on Computing*, 12(1):75–82, 2000.

[22] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, 2006.

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

[24] Y. Suzuki. 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, 2000.

[25] R. L. Thorndike. Who belongs in the family. *Psychometrika*, 18(4):267–276, 1953.

[26] W. Wei and M. Hansen. 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, 2005.

[27] J. Xu, K.-c. Lee, W. Li, H. Qi, and Q. Lu. Smart pacing for effective online ad campaign optimization. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 2217–2226, 2015.

[28] W. Zhang, S. Yuan, and J. Wang. Optimal real-time bidding for display advertising. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 1077–1086, 2014.

[29] Y. Zhu, H. Yang, and J. He. Co-clustering based dual prediction for cargo pricing optimization. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 1583–1592, 2015.