

Estimating True Demand in Airline's Revenue Management Systems using Observed Sales

Alireza Nikseresht
Computer Science and Engineering
Shiraz University
Shiraz, Iran

Koorush Ziarati
Computer Science and Engineering
Shiraz University
Shiraz, Iran

Abstract—Forecasting accuracy is very important in revenue management. Improved forecast accuracy, improves the decision made about inventory and this lead to a greater revenue. In the airline's revenue management systems, the inventory is controlled by changing the product availability. As a consequence of changing availability, the recorded sales become a censored observation of underlying demand, so could not depict the true demand, and the accuracy of forecasting is affected by this censored data. This paper proposed a method to estimate true demand from censored data. In the literature, this process is referred to as unconstraining or uncensoring. Multinomial Logit model is used to model the customer choice behaviour. A simple algorithm is proposed to estimate the parameters (customers' preference) of the model by using historical sales data, product availability info and the market share. The proposed method is evaluated using different simulated datasets and the results are compared with three benchmark models that are used commonly in airline revenue management practice. The experiments show that proposed method outperforms the others in terms of execution time and accuracy. A 47.64% improvement is reported in root mean square error between simulated and estimated demand in contrast to the benchmark models.

Keywords—Demand estimation; demand modelling; forecasting; revenue management; inventory control; unconstraining; uncensoring

I. INTRODUCTION

After airline deregulation in 1978, many low cost airlines were born. They started to offer much cheaper tickets than the major airlines, nevertheless they had benefit, because they had lower operation cost [1]. The major airlines were not able to price their seats below or at least near to the new born airlines. Although they did not lose all of their customers, because some customers are price sensitive while the others are sensitive to quality of service and brands, they loosed a significant amount of revenue. In result, they offered their seats with different prices and features to capture all types of customers, and the problem changed to how to optimally price the seats and controlling seat inventory? To find the optimal solution for this problem, airlines should know the exact amount of demands for each offered seat. Knowing the exact amount of demand is impossible because it is something that will happen in future, so by looking at historical sales data, they try to estimate demand.

In the airline's revenue management systems, the inventory is controlled by means of setting booking limits for each

product. The goal of setting these booking limits is to protect a specific amount of seat for higher profit customers. Although the booking limits cause more revenue, it censors the true demand, which is needed for accurate demand forecasting. In most cases the observed sales data do not reflect the true demand because no sales are recorded after a product has been sold out, or in other word if a product's booking limit has been reached. Therefore, transactional sales data are usually denoted as censored demand [2]. Using the censored data, as true demand has two consequences: 1) under estimation of true demand, which causes a spiral down effect over time; 2) over estimation of true demands [3]–[5]. In both situations the companies lose revenue. Although ignoring the censorship results in significant reductions in revenue, observing demand after its booking limit reached is impractical. So, unconstraining methods are used to estimate the true demand [6]. In addition to censorship, ignoring the correlation of demand between related products leads to inefficient estimates of the true demand [7], [8]. Empirical studies show that between 45% and 84% of demand can be substituted [9]–[12]. Nevertheless, measuring true demands using the available sales data is not an easy task [13]. If the accuracy of forecasting grows about 20%, the resulting revenue will improve about 1% [14]. In a research conducted by Weatherford, it was found that if there is a negative bias in forecast, up to 3% of the potential profit may be lost [15]. So it is rational to look forward to methods measuring the true demands from the available data and use these data for forecasting and managing seats [16].

This paper proposed an unconstraining algorithm to estimate true demand using censored sales transaction data in a fast, accurate, and very simple manner. The proposed algorithm is an extension of MSEG algorithm which is an iterative demand estimation algorithm for airline sales transactions data [17]. MSEG is actually a minimum square error algorithm which runs iteratively and in each iteration tries to reduce the amount of error between estimated and observed sales data, until a minimum error is reached or a certain iteration is elapsed. Gradient descent is used to minimize the error function, and in each iteration, to approach the minimum, it moves toward the opposite of the gradient vector with a small step size L .

Unlike MSEG, in the proposed method the objective function is changed to convex form and the optimal solution is founded easily by a simple derivative. As MSEG, our method needs transactional sales data and product availability, plus

market share which were optional in MSEG. In the proposed method, customer arrival is estimated using a simple heuristic instead of assuming a priori distribution, so the risk of misspecification of distribution is resolved and it benefit the advantages of not requiring any assumption on the form of the distribution function. To model the demand, the multinomial logit choice model is employed and the proposed algorithm estimates the choice model parameters. The proposed method is evaluated using two types of datasets: 1) dataset with full observation of demand occurrences; and 2) dataset with censored observation of demands. The dataset 1 which contains unconstrained observation is served as a benchmark to evaluate our work. The results were compared with the other unconstraining methods which are applied commonly in revenue management practice, such as projection de truncation (PD), expectation maximization (EM) and another version of EM proposed by Vulcano et al. in 2012 [18]. The rest of the paper is structured as follows: In Section 2, similar works have been reviewed and Section 3 discusses the problem description. Section 4 describes how the demand is modelled. Section 5 is dedicated to the solution and algorithm. In Section 6, the simulation process and the datasets are described. The results are demonstrated in Section 7, and finally, Section 8 concludes the paper.

II. LITERATURE REVIEW

Since 1990 in the context of demand unconstraining, there is a vast literature that address the issue of estimating true demand using historical sales data [19]. Choice modelling attempts to model the decision process of customers. The theory of choice modelling assume customers are rational agents who intelligently make decisions when, what, and how much to purchase to achieve the maximum benefit. An important aspect of this rationality assumption is that customer behaviour can be predicted [1]. Multinomial logit model (MNL) is the most popular approaches to choice modelling. The MNL is a discrete customer choice model which presumes customers are rational utility maximizers, and predicts the customer's behaviour [20].

Andersson, Algiers, and Beser, presented a customer choice based method for optimal seat allocation and the parameters of choice model are captured from interview with experts and historical sales data. In the proposed model they considered recapture and buy-up [21], [22]. Ratliff devised a customer choice model-based heuristic to estimate demand, spill and recapture. His heuristic needs sales data and market share to do unconstraining. It is capable of computing recapture alongside different flights and products, and its main advantage is that by taking into account the portion of substitution between products or flights, due to the fact that some products are not available, it prevents double counting of demands [23].

Many researchers in their works, employed expectation maximization (EM) to find the customer choice model parameters. Talluri and Van Ryzin applied a discrete choice model to model customer behaviour and found the related parameters using EM method. In their proposed method, they assumed a distribution function for demand and then tried to find its parameters from the observed data, indicating that their methods perform well [1]. In other work, they developed an

estimation method based on the expectation maximization which is able to jointly estimate arrival rates and choice model parameters when no-purchase outcomes are unobservable [24]. Haensel and Koole utilized the idea of customer choice set to model the buying behaviour of customers. They applied EM algorithm to unconstrain the censored data, not assuming a fixed arrival rate for customers as regularly done. Instead, they estimated a demand function for each group of customers by analysing observed data [2], [25]. Newman et al. presented a parameter estimation method for multinomial logit model in which one alternative is never observed. Their method is based on decomposing the log-likelihood function into marginal and conditional components. They showed that their proposed method is computationally efficient and provides consistent parameter estimates. Simulations based on industry data set demonstrate their method computationally outperform the other alternative estimation methods [26]. Vulcano et al. developed a maximum likelihood estimation algorithm that uses a variation of the EM method to account for unobservable data. With simulation study, they showed that revenue improves about 1%–5% using choice based revenue management [27]. Vulcano et al. proposed an EM-based method estimating spilled and substitute demands. Their method only needs the observed sales data, product availability, and company market share. Their main idea is to consider the problem as primary demand or customer first choice. They supposed that each customer has a set of choices with a primary or first choice. Each customer could buy his/her first choice, and if the first choice is not available, the next choice is substituted or leaves the system without purchasing anything. Then they tried to estimate primary demands, substitute demands, and no purchase count, using EM [18]. Agrawal et al. modeled the consumer choice behavior using the multinomial logit model for assortment selection problem and dynamically estimated the model parameters [28].

Modelling customer choice is a high dimensional problem and it is difficult to dealing with it, so to deal with its difficulty, many researchers prefer to assume a priori distribution and parametric model which they think is able to adequately capture choice model behaviour [29]. The side effects of all parametric approaches are misspecification of the model and overestimation or underestimation true demand. Van Ryzin and Vulcano proposed a nonparametric choice model approach to estimate, customer preferences for a set of substitutable products. With the numerical experiments on a real dataset, they showed that their method perform well and improves root mean square error between predicted and observed sales about 67% [19]. Farias et al. presented an approach to predict the expected sales from historical data. They used a nonparametric approach to model the customer choice. With empirical study using simulated and real dataset, they showed that their method is able to produce accurate revenue forecast without over/under fitting [29]. For a detail review on choice modelling you can refer to: [1], [30], [31]. A good categorization and review of the majority of demand unconstraining methods is presented in [32].

In the context of demand estimation and forecasting in airline revenue management systems, our paper brings three main contributions to the literature. First, this work contributes

to the literature by introducing an easy-to-implement non-parametric estimation algorithm for estimating multinomial logit choice model parameters in a reasonable computation time and accuracy. Second, the objective function is perturbed to a convex form by devising a simple heuristic which, shortens the execution time and guarantees the convergence. Third, a simple method proposed to estimate the customer arrival and no purchase in each period, without considering any a priori distribution, and just by the means of looking at the aggregated sale and market share in each period.

III. PROBLEM DESCRIPTION

Airlines start to sell their flights about a year before departure. To analyse the customer behaviour this long selling time horizon is divided to some periods and discretized. These periods may differ in length, for example a period may be a week while the other may be a month. Then the aggregated sales of each offered product (seats with different conditions) are observed and recorded in each period. So the available dataset, consisting of aggregated sales for each product in each period. Table 1, shows a sample of such a dataset which is from an airline that offers 4 classes (C1 to C4) in 10 periods before flights. As you can see in this table, the aggregated sales of each product are recorded in each period.

TABLE I. A SAMPLE OF AGGREGATED RECORDED SALES IN DIFFERENT PERIODS BEFORE FLIGHT. N MEANS THE PRODUCT IS NOT AVAILABLE IN THAT PERIOD. PERIOD 0 IS THE DEPARTURE DAY

Products	9	8	7	6	5	4	3	2	1	0
C1	13	15	13	N	N	N	N	N	N	N
C2	11	6	7	9	10	N	N	N	N	N
C3	1	8	3	5	8	4	9	N	N	N
C4	4	9	4	7	7	3	9	6	7	5

TABLE II. CLASSES AND THEIR FEATURES OFFERED BY AN AIRLINE

Classes	Miles Earned	Change	Cancellation	Price \$
C1	100%	Charge 0%	Charge 0%	600
C2	100%	Charge 10%	Charge 20%	400
C3	50%	Charge 40%	Charge 50%	300
C4	30%	NO	NO	240

As it alluded to above, based on the theory of choice model the customers are rational utility maximizer and when they face with a choice list, they try to choose the choice which has the most benefit for them. If their first choice was not available, they may choose the second choice to buy in lieu to their first choice (this action is called substitution) or leave the system without purchase. Table 2 presents a sample of offered classes by an airline, along with their features.

To illustrate the problem, suppose that an airline offers 4 classes, C1 to C4 (see Table 2) in an aircraft with the capacity of 100, and the nested booking limits are set to 100, 80, 45 and 15. This could be translated to protecting 20 seats for class C1, 55 seats for classes C1 and C2 and 85 seats for classes C1, C2 and C3. There is no protection for class C4 as like as there is no limitation to sale class C1 except maximum aircraft

capacity. Suppose that 35 customers arrive. The observed sales for product C1 to C4 is (0, 0, 10, 15). The current information systems just register the successful sales and do not record any information about unsuccessful requests. Therefore, in such cases, the sales data could not show the true demands; for example, the sales data here shows 15 requests for class C4 (the observed demand), but due to C4 booking limit of 15, if the requests for this class were more than 15, system just able to record 15 successful sales and the other requests is not recorded. So this number of sales does not show the true amount of customer's request for C4. The observed sales for product C3 are 10. Although it is lower than its booking limit, this number does not show the true demand, because some of customers, whose first choice was C4, prefer to buy C3 instead of leaving the system without purchase. So this number of sales for class C3 is the mixture of customers whose their first choice are C3 and those whom their first choice are C4 but substitute to C3.

Here if these recorded sale is used as true demand, it is obvious that the true demand of C4 is under estimated while the true demand of C3 is over estimated. If these data are used as true demands and are fed into the forecasting module, a spiral down effect will occur in estimation, and the total revenue begins to decrease [3]. In this paper our goal is to find that how the customers prioritize the offered products and choose among them. Hence it is important to model the customer's buying behaviour. In this paper, it is supposed that the customers behave like in the customer choice model. The next section describes the approach to model the demand based on a discrete customer choice model.

IV. DEMAND MODEL AND PROBLEM FORMULATION

Discrete choice model is the theoretical basis of customer behaviour. In 2000, the Nobel prize in economics has been awarded to American economists James Heckman and Daniel McFadden. Daniel McFadden proposed the discrete choice model, which is the basis for the development of the customer choice behaviour [33], [34]. The MNL model of McFadden was the primary basis for analysis of multinomial choice for many years. MNL assumes consumers have homogeneous tastes for observed product attributes, and that the random (unobserved) part of utility is iid. In other word all customers use the following equation to calculate the product j's utility with the same parameters.

$$i. u_j = \beta^T x_j + \epsilon_j \tag{1}$$

Here in (1), β is a vector of parameters and x_j is a vector of attribute values for product j which could include factors such as prices, rewards, time of departure, length of flight, cancellation and change policy and etc. ϵ_j is the random component. In the MNL model, the probability that an alternative j is chosen from a set of offered products S, which contains product j is given by:

$$P_j(S) = \frac{e^{u_j}}{\sum_{i \in S} e^{u_i}} \tag{2}$$

The MNL possesses a restrictive property known as the independence from irrelevant alternatives (IIA) property [1]. For a famous example which describes IIA (red bus/blue bus paradox) please refer to [35]. As a result of IIA, the MNL model must be used with caution. It should be restricted to choice sets that contain alternatives that are dissimilar. Despite this deficiency, the MNL model is widely used in estimating travel demand. The popularity of MNL goes back to being analytically tractable, relatively accurate, and can be estimated easily using standard statistical techniques. The product utility is needed to calculate the probability of selecting an item from choice set. The product utility could be estimated from the sales data. For simplicity, the product preference weight is defined same as in [18]:

$$v_j = e^{u_j} \quad (3)$$

And the probability of choosing product j from choice set S is become:

$$P_j(S) = \frac{v_j}{\sum_{i \in S} v_i} \quad (4)$$

Now, assume that an airline offers n classes to its customers. The selling time horizon before departure, is divided into T periods: $t=1 \dots T$. In each period t, a set of classes (C_t) is offered. The number of observed sales of class j in the period t is o_{jt} , and O_t is the observed sale vector in interval t: $O_t = (o_{1t}, o_{2t}, \dots, o_{nt})$. The customers may decide to not purchase anything when facing with the offered choices, so c_0 is added to the choice set as no purchase alternative. The utility of no purchase is zero, so, the preference weight of no purchase is $v_0 = 1$. Using these preference weights the MNL choice probability is written as in [36], the probability of choosing class j in period t:

$$P(j, t) = \frac{v_j}{\sum_{i \in C_t} v_i + v_0} \quad (5)$$

The probability of choosing class j which is not available in period t is: $P(j, t) = 0$.

Based on no purchase preference weight or $v_0 = 1$ the no purchase probability is defined as:

$$P(0, t) = \frac{v_0}{\sum_{i \in C_t} v_i + v_0} = \frac{1}{\sum_{i \in C_t} v_i + 1} \quad (6)$$

Suppose that the number of arrivals in each period t is known to be A_t , and if all classes were available in period t then, the probable number of sales of class j in the period t or true demand, is:

$$d_{jt} = P(j, t) \cdot A_t = \frac{v_j}{\sum_{i=1}^n v_i + v_0} A_t \quad (7)$$

The total sale in period t is:

$$E_t = \sum_{j \in C_t} d_{jt} = \frac{\sum v_j}{\sum v_j + v_0} A_t \quad (8)$$

But in the real world application the number of arrival is not visible, so it should be estimated. Suppose the preference weights are known. As an estimate, consider that the aggregated number of sales in period t (q_t) is equal to the aggregated number of estimated sales in period t (E_t), so the number of arrival could be found using:

$$q_t \approx E_t \Rightarrow q_t \approx \sum_{j=0}^n d_{jt} = \frac{\sum v_j}{\sum v_j + v_0} A_t \Rightarrow A_t \approx q_t \cdot \frac{\sum v_j + v_0}{\sum v_j} \quad (9)$$

The number of customers who could not buy their first choice in period t is estimated by using:

$$s_t = \sum_{j \notin C_t} d_{jt} \quad (10)$$

The number of requests which choose class k as a substitute in period t, could be estimated using:

$$r_{kt} = s_t \cdot P(k, t) = s_t \cdot \frac{v_k}{\sum_{j \in C_t} v_j + v_0} \quad k \in C_t \quad (11)$$

The number of observed sales of class j in period t or o_{jt} consists of two components: 1) the number of customers interested in class j as their first choice (true demand or d_{jt}); 2) the number of customers buying j because their first choice was not available (recapture or r_{jt}) [21], [37]. Thus, the demand mass balance equation is as follows:

$$o_{jt} = d_{jt} + r_{jt} \quad (12)$$

It supposed that the preference weights is known, but the fact is that, the values of preference vector is not known, so at first the preference vector V is needed to compute arrival in each period, spill and recapture. Therefore, first the preference vector is estimated, and then by using it and applying foregoing equations the number of true demands, recaptures, and spills are estimated.

V. ALGORITHM

The Airline's true demand estimation algorithm is an iterative algorithm which starts with an initial value for preference vector and tries to make a better estimate for preference vector in each iteration by minimizing an error function which is the gap between estimated and observed sales in each period:

$$e_{jt} = o_{jt} - (d_{jt} + r_{jt}) \quad (13)$$

The objective function which should be minimized is a quadratic error function (Least Square Error):

$$\min_{v_j} (E = \sum_{t=1}^T \sum_{j=1}^N \frac{1}{2} \cdot e_{jt}^2) \quad (14)$$

A. Analysis of Objective Function

To analysis the objective function, the selling time horizon is divided into two parts: 1) periods which all classes are available; and 2) periods which some classes are not available. In the first part, since all classes are available the spill and recapture is zero, so the observed sales are equal to true demand. Using (7), (9) and (14) the objective function is rewritten as:

$$\min_{v_j}(E = \sum_{t=1}^T \sum_{j=1}^N \frac{1}{2} \left(o_{jt} - q_t \cdot \frac{v_j}{\sum_{i=1}^n v_i} \right) \quad (15)$$

Because of the fractional term, the objective function is not convex, also the Hessian matrix is not positive definite. To illustrate its non-convexity, assume a company offers two classes A and B in 10 periods. At the first 6 periods both of the classes were available, but in the rest of periods, just class B was available. Based on (15) the error function in the first 6 periods which all classes were available (so no spill and recapture exist) becomes:

$$e = \left(x[t, 1] - v1 * \frac{\sum_{i=1}^2 x[t,i]}{v1+v2} \right)^2 + \left(x[t, 2] - v2 * \frac{\sum_{i=1}^2 x[t,i]}{v1+v2} \right)^2 \quad (16)$$

The 3D plot of (16) is shown in Fig. 1. In the second part, which some classes are not available, some of the arrived customers prefer to buy substitute classes (recapture) or leave the system without purchase (spill). Using (7), (9), (13) and (14) the objective function is rewritten as:

$$\min_{v_j}(E = \sum_{t=1}^T \sum_{j=1}^N \frac{1}{2} \left(o_{jt} - q_t * \frac{\sum_{i \in C_t} v_i + v_0}{\sum_{i \in C_t} v_i} * \frac{v_j}{\sum_{i=1}^n v_i + 1} - s_t * \frac{v_j}{\sum_{i \in C_t} v_i + v_0} \right) \quad (17)$$

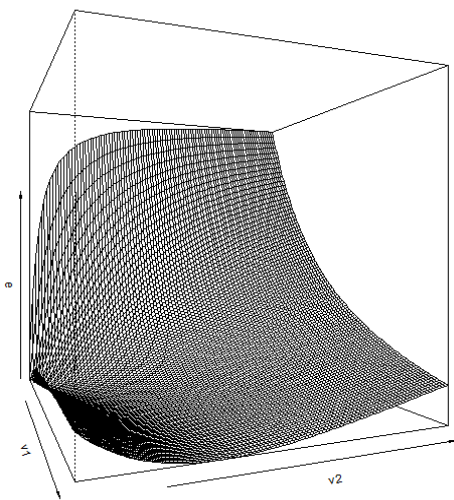


Fig. 1. The 3D plot of error function in the periods that all products are available, for a company with two offered classes A and B in 10 periods.

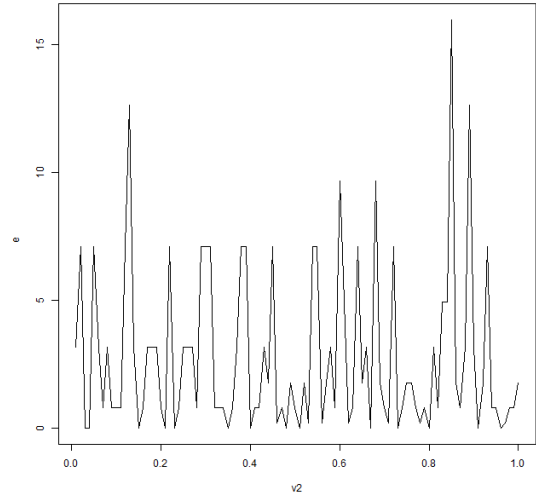


Fig. 2. Plot of error function in the second part of periods that some products are not available, for a company with two offered products A and B in 10 periods.

In above foregoing airline company example, class A is not available in the last 4 periods. So, because of spilled and recaptured demands, the error function is like:

$$e = \left(x[t, 2] - (v2 + 1) * \frac{x[t, 2]}{v1 + v2 + 1} - v1 * \frac{x[t, 2]}{v1+v2+1} \right)^2 \quad (18)$$

The value of v1 is not updated as class A is not available, so it assumed that, the value of v1 is set to its optimal value, then the error function with fixing v1 at its optimal value is plotted. The plot is shown in Fig. 2 and obviously is ill structured and not convex.

B. Convexification Process

Because of the fractional term in the demand (7), arrival (9) and the recapture (11), the objective function is not convex. A simple heuristic is devised to make that function convex. As it said before, the nonconvexity is because of the existence of the fractional term in the demand function, so if the denominator ($\sum v_j + v_0$) is replaced with a fixed value, the objective function becomes convex. This value could be extracted from market share. The market share (M) is considered same as that in [38]:

$$M = \frac{\sum v_j}{\sum v_j + v_0} \xrightarrow{v_0=1} \sum v_j = M/(1 - M) \quad (19)$$

Fig. 3 shows the error function after convexification.

The proposed method initiated the values of preference weights equally using (19). For example, the product j's initial preference weight is:

$$v_j = \left(\frac{M}{1 - M} \right) / (\text{number of products}) \quad (20)$$

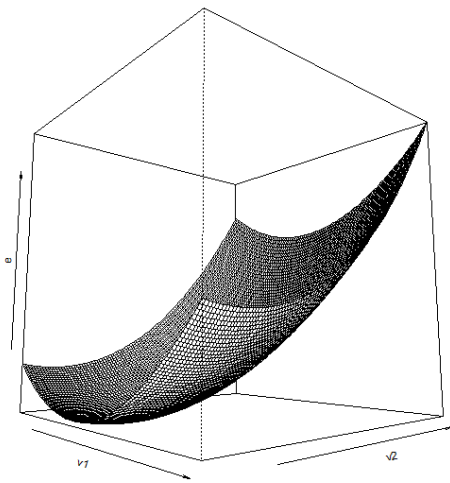


Fig. 3. 3D plot of error function after convexification.

Pseudo Code of the algorithm:

1. Ct: Set of available classes in period t.
2. V: Preference weight vector.
3. At: The number of arrival in period t.
4. Ot: The vector of purchases in period t.
5. M: The market Share
6. n: the number of offered classes
7. Initialize :
 - Estimating v_j
 - For each class j in [1..n]

$$v_j = \left(\frac{M}{1-M} \right) / n$$
8. For each period t in [1..T]
 - Compute At based on V_j s: $A_t = q_t \cdot \frac{\sum v_j + v_0}{\sum v_j}$
 - Compute no purchase: $d_{0t} = A_t - q_t$
 - For each class j in [1..n]
 - $d_{jt} = \frac{v_j}{\sum v_j + v_0} A_t$
 - if (cjt not available)
 - $V_{j_new} = V_{j_old}$
 - $S = S + d_j$
 - else
 - $r_j = s \cdot \frac{v_j}{\sum_{j \in Ct} v_j + v_0}$
 - $V_{j_new} = \text{mean} \left(\frac{O_{jt}}{\left(\frac{A_t}{\sum v_j} \right) + \left(\frac{S}{\sum_{j \in Ct} v_j + v_0} \right)} \right)$

VI. DATA AND SIMULATION PROCESS

This section describes the simulation process and the data which is used for evaluating proposed algorithm. To have a dataset of customer sales, 6 types of information are needed: 1) offered product set; 2) product availability information during the selling time horizon; 3) the number of periods before flight; 4) the customer arrival process; 5) the customer buying behaviour; 6) the market share of the company. So the simulation process is as follows:

1) It is considered that an airline company offers a set of products in each flight. So for example, it is assumed that the company offers 4 classes in 10 periods. Then 100 flight data are generated.

2) In the simulation process, the product availability info is set exogenously. The product availability info is in matrix format, with the number of columns equal to the number of periods and the number of rows is equal to the number of products. For example $AVL_{i,j} = 1$ means that the product i is available in period j and vice versa. This information also could be generated randomly.

3) The number of periods which the aggregated sales of each class are observed is considered different for each dataset. For example, assume 10 periods for selling the seats before the flight.

4) In the simulation process, it assumed that, customers arrive based on a poison process. The mean of poison distribution is different for each dataset. Here for 10 periods, 4 classes example, a poison arrival process with the mean of 60 is considered.

5) To simulate the customer's buying behaviour, a customer choice model is used. This is a predefined model with predefined products and their preference weights. This model may pick from real-world data or synthetic datasets. For example, consider a model with offered classes, c1 to c4 and the preference vector of $V = (0.85, 0.68, 0.33, 0.14)$.

6) The market share of the company is considered different for each dataset. In our example, the market share is considered 67%.

7) To simulate the demands, the selected model is used to simulate the customer's request to buy the products in each period. By having the simulated number of arrivals in each period, the number of sales for each product and the number of no purchase could be simulated using foregoing equations. This generated dataset, which contains the unconstrained demands is served as a benchmark to evaluate the accuracy and performance of the proposed method.

8) By applying the product availability info on simulated sales from previous step a censored observation of the dataset is achieved. For example, if the product i was not available on period j (i.e. $AVL_{i,j}=0$) the simulated sales of that product on foregoing period reset to zero (ie. $d_{ij} = 0$). This dataset is used as input to our unconstraining algorithm.

9) By executing the proposed algorithm on the generated dataset in step viii, the preference vector is estimated.

10) Using estimated preference vector, the true demand, spill and recapture could be calculated.

A sample of simulated dataset is available in Table 3. The number of arrivals and no purchase is not visible in real-world application; thus, these values are separated in two rows entitled Hidden Data.

Each dataset contains 100 instances of flights, which each instance contains a set of classes and the number of sales of each class on every period. For example, in the 4X10 dataset which contains a set of 4 offered classes and 10 intervals, 00 instances of flight exists. This means totally $100 \cdot 10$ intervals that 4 products is offered so 4000 sales data records is generated in this dataset.

TABLE III. A SAMPLE OF SIMULATED SALES DURING PERIODS BEFORE DEPARTURE. N MEANS THE CLASS IS NOT AVAILABLE IN THAT PERIOD

Classes	9	8	7	6	5	4	3	2	1	0
C1	13	15	13	N	N	N	N	N	N	N
C2	11	6	7	9	10	N	N	N	N	N
C3	1	8	3	5	8	4	9	N	N	N
C4	4	9	4	7	7	3	9	6	7	5
↓ Hidden Data ↓										
No Purchases	47	42	46	72	58	76	75	68	83	9
#Arrivals	76	80	73	93	83	83	93	74	90	4

VII. RESULTS

To evaluate the performance and the accuracy of the proposed algorithm, the algorithm is executed on the observed sales of all simulated datasets. As a result of applying the proposed algorithm on simulated dataset, preference values are estimated for each dataset. Table 4 shows a sample of estimated preference vector in contrast to the true value of preference vector which is used to generate the simulated dataset. There are two favourable attributes for an estimator: accuracy and precision. Accuracy is lack of bias and precision is small variance. If an estimator is unbiased, then its variance is investigated. If it is biased, it is good to look at the mean squared error. As the Table 4 shows, the proposed estimator has a negative bias on estimated values and this is because of perturbing the objective function to a convex form and the new convex function is always below the true values.

Table 5 shows the mean, variance, mean square error, accuracy and precision of the proposed algorithm for estimated values of preference vector. As Table 5 shows the values of MSE and Variance are almost equal and this is because of the mean of the estimated preference vector is almost equal to the true value of preference vector. Here in Fig. 4, the box plot of estimated preference values for 100 executions of the algorithm over simulated datasets is shown. The line in the middle of the box is the median. The box itself represents the middle 50% of the data. The box edges are the 25th and 75th percentiles. There is some disk visible below or above of some boxes which is shown in some cases the simulated sales were far from the regular values. So, the estimated preference of that simulated dataset is far from the median.

TABLE IV. ESTIMATED PREFERENCE VECTOR

Classes	Preference Vector	True Value	Estimated Value	Estimator Bias
C1	V1	0.85	0.822	-0.028
C2	V2	0.68	0.659	-0.020
C3	V3	0.33	0.327	-0.002
C4	V4	0.14	0.139	-0.0009

TABLE V. MEAN, VARIANCE AND MEAN SQUARE ERROR OF DIFFERENCE BETWEEN ESTIMATED AND TRUE VALUES OF PREFERENCES.

Classes	Mean	Variance	MSE	Accuracy%	Precision%
C1	-0.028	0.0023	0.0031	97.16	99.7
C2	-0.020	0.0020	0.0024	97.91	99.8
C3	-0.002	0.0014	0.0014	99.73	99.8
C4	-0.0009	0.0004	0.0004	99.90	99.9

TABLE VI. ROOT MEAN SQUARE ERROR BETWEEN ESTIMATED TRUE DEMAND AND SIMULATED TRUE DEMAND IN CONTRAST TO EM AND PD.

Methods	C1	C2	C3	C4
Our Method				
RMSE	3.39	2.49	1.29	0.59
EM				
RMSE	2.34	3.68	3.9	2.34
PD				
RMSE	2.20	3.59	3.78	2.34
Vulcano EM				
RMSE	3.57	2.61	1.39	0.66

Figure 5 shows the cumulative sum of simulated demand, estimated demand, observed sales and estimated sales over 30 periods for products c1 to c4. It is obvious that always estimated sales is close, but a bit upper than observed sales. One of the important aspect of the simulation is that even the hidden data which are not observable in the real world are generated, hence it is easy to measure the accuracy of the estimation process. Here, the estimated true demand is compared with simulated true demand data. In this way, we have the ability to show how much accurate the forecasting is, and this will indicate the accuracy of estimated preference vector. To have a better insight about the results and the accuracy of proposed method, the results are compared to three other methods which are usually applied in practice: 1-Expectation Maximization or EM, 2-Projection Detruncation or PD methods and 3-EM method that proposed by Vulcano et al in 2012 [18],[39]–[41]. To compare the results, three above methods are applied on a dataset for 1000 times, then the mean of root mean square error for each method is calculated. The result of the comparison is depicted in table VI. The product C4 which is the most expensive product, is always available in all periods. The EM and PD methods, assume that the observed sales of product 4 is equal to true demand because these methods do not consider the substitution. The value of 2.34 which is seen in the table VI for the root mean square error of product c4 in EM and PD methods shows that the true demand is differ from observed sales. Here the observed sales are a mixture of true demand and recaptured demand, hence, the observed sales are greater than the true demand. Our proposed method and the EM method which is presented by Vulcano et al. are able to compute the spill and recapture from observed sales data and so, they are more accurate than EM and PD. If there is no substitution, the EM and PD methods are also accurate, even more accurate than our method or the Vulcano’s EM method.

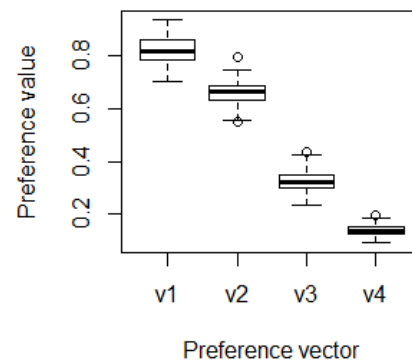


Fig. 4. The Boxplot of preference values.

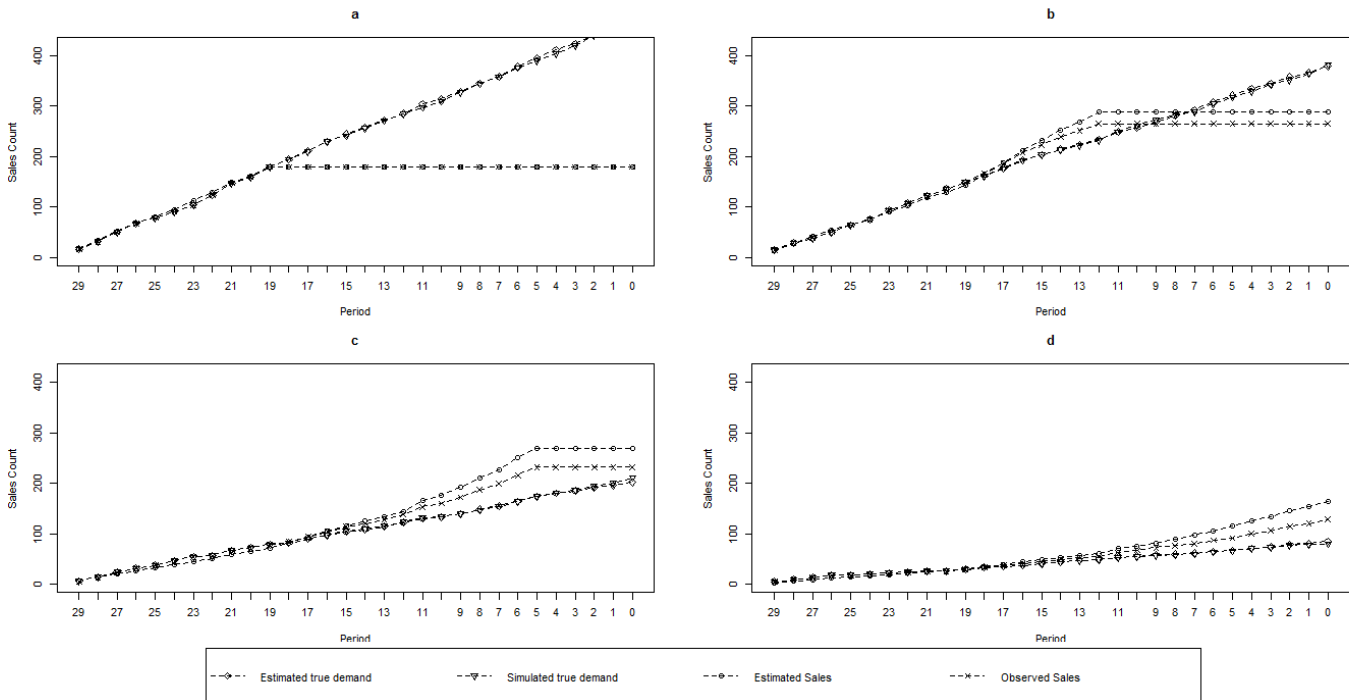


Fig. 5. The cumulative sum of simulated and estimated demand, observed sales and estimated sales over 30 periods for classes c1(a), c2(b), c3(c) and c4(d)

TABLE. VII. EXECUTION TIME FOR DIFFERENT DATASETS

Dataset	Execution Time for 100 flights (sec)
5X15	2.01
4X30	3.61
5X30	3.72
10X30	4.70
10X60	9.12
20X60	12.77

This could be deduced by looking at column C1 of Table 6. As you can see in this column the root mean square error of EM and PD are smaller than the others, because product C1 is the least expensive and more favourable product in the choice set and in the periods that this product is available, all of the other classes are available too. So the observed sales of product C1 are always equal to its true demand. The results show that our proposed method out performs the other three, and also has up to 47.64% improvement in root mean square error in contrast to the other mentioned methods.

Different size datasets used to show the efficiency and speed of the proposed algorithm. At first it is evaluated with a rather small dataset with 5 products and 15 periods, in this case the algorithm lasts 2.01 seconds to estimate 100 instances of flights. The second dataset is composed of observed sales from a flight with 4 classes and 30 periods. The third dataset is for a flight having 5 classes sold in 30 intervals. Finally, in the last dataset which is our biggest dataset, there exist 20 classes in a flight with 60 time intervals. For the last datasets algorithm takes about 12.77 seconds to solve 100 instances, which is very good for this size of dataset in contrast to current demand

unconstraining methods. Table 7 shows the execution time for all six datasets. R revision 3.2.2 used to implement the algorithm on a computer with an Intel core2 Quad Q8300 and 4 GB of internal memory. Code execution time is measured by `proc.time()` in R[42]. Column 1 of Table 7 shows the datasets and column 2 is the mean convergence time.

VIII. CONCLUSION

In this paper, a new approach for demand unconstraining is proposed. Multinomial logit model is used to model the customer choice behaviour. Transactional sales data, product availability and market share are the only data used to estimate the choice model parameters. Spilled and recaptured demands are estimated alongside with true demand. The problem of double counting demands in available classes is resolved by considering spill and recapture. The customer arrival is estimated using a simple heuristic. In this heuristic method, the customer arrival count is assumed to be a fraction of aggregated sales in each period (see (9)).

Unlike the proposed method, most of current in practice choice based methods, assume a priori distribution to estimate, customer arrival rate in each period, so suffer from misspecification of the distribution function and its parameters. Numerical experiments prove that proposed method performs well in terms of the speed and the accuracy of estimation. Based on the simulation results, our method improved root mean square error between simulated and estimated demands by 47.64%, in contrast to other methods such as PD, EM and another version of EM proposed by Vulcano et al. It is observed that if there is no substitution in the datasets, the EM and PD methods perform well and even better than the other methods. But in most cases the substitution exists between classes and this affects the true demand. So the EM and PD

methods are not able to perform well in these kind of datasets. For a rather large dataset of size 20 products and 60 periods, it takes 0.127 seconds to estimate the true demand, spill and recapture, which is fast in this context.

Based on the experiment's result, we believe that the proposed method would be a good replacement for current demand unconstraining methods in airline revenue management systems.

REFERENCES

- [1] K. T. Talluri and G. J. Van Ryzin, *The Theory and Practice of Revenue Management*, vol. 1. 2005.
- [2] A. Haensel and G. Koole, "Estimating unconstrained demand rate functions using customer choice sets," *J. Revenue Pricing Manag.*, vol. 10, no. February, pp. 438–454, 2011.
- [3] W. L. Cooper, T. Homem-de-Mello, and A. J. Kleywegt, "Models of the Spiral-Down Effect in Revenue Management," *Oper. Res.*, vol. 54, no. December 2014, pp. 968–987, 2006.
- [4] R. Saleh, "Estimating lost demand with imperfect availability indicators," in *AGIFORS Reservations and Yield Management Study Group*, 1997.
- [5] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*. 2002.
- [6] A. M. Fouad and A. F. Atiya, "A New Unconstraining Method for Demand Forecasting," pp. 16–19, 2012.
- [7] C. Stefanescu, "Multivariate Demand: Modeling and Estimation from Censored Sales," pp. 1–33, 2009.
- [8] B. Tan and S. Karabati, "Retail inventory management with stock-out based dynamic demand substitution," *Int. J. Prod. Econ.*, vol. 145, no. 1, pp. 78–87, 2013.
- [9] T. Gruen, D. Corsten, and S. Bharadwaj, "Retail Out-of-Stocks: A Worldwide Examination of Extent, Causes and Consumer Responses," 2002.
- [10] K. Campo, E. Gijsbrechts, and P. Nisol, "The impact of retailer stockouts on whether, how much, and what to buy," *Int. J. Res. Mark.*, vol. 20, no. 3, pp. 273–286, 2003.
- [11] J. Aastrup and H. Kotzab, "Analyzing out-of-stock in independent grocery stores: an empirical study," *Int. J. Retail Distrib. Manag.*, vol. 37, no. 9, pp. 765–789, 2009.
- [12] G. Xin, P. R. Messinger, and J. Li, "Influence of soldout products on consumer choice," *J. Retail.*, vol. 85, no. 3, pp. 274–287, 2009.
- [13] R. R. Wickham, "Evaluation of forecasting techniques for short-term demand of air transportation," MIT, 1995.
- [14] K. T. Talluri, G. J. van Ryzin, I. Z. Karaesmen, and G. J. Vulcano, "Revenue management: Models and methods," in *2008 Winter Simulation Conference*, 2008, pp. 145–156.
- [15] L. R. Weatherford, "A review of optimization modeling assumptions in revenue management situations," in *AGIFORS Reservations and Yield Management Study Group*, 1997.
- [16] A. Nikseresht and K. Ziarati, "Review on the Newest Revenue Management Demand Forecasting Methods," in *International Conference on Management, Economics and Industrial Engineering*, 2015, vol. 1, no. 1.
- [17] A. Nikseresht and K. Ziarati, "A Non-Parametric Customer Choice-Based Method to Unconstrain Airline Sales Transaction Data," *J. Air Transp. Manag.*, 2016.
- [18] G. Vulcano, G. van Ryzin, and R. Ratliff, "Estimating Primary Demand for Substitutable Products from Sales Transaction Data," *Oper. Res.*, vol. 60, no. 2, pp. 313–334, 2012.
- [19] G. van Ryzin and G. Vulcano, "A Market Discovery Algorithm to Estimate a General Class of Nonparametric Choice Models," *Manage. Sci.*, vol. 61, no. 2, pp. 281–300, 2015.
- [20] A. Hübner, H. Kuhn, and S. Kühn, "An efficient algorithm for capacitated assortment planning with stochastic demand and substitution," *Eur. J. Oper. Res.*, vol. 250, no. 2, pp. 505–520, Apr. 2016.
- [21] S.-E. Andersson, "Passenger choice analysis for seat capacity control: A pilot project in Scandinavian Airlines," *Int. Trans. Oper. Res.*, vol. 5, no. 6, pp. 471–486, 1998.
- [22] M. B. Staffan Algers, "Modelling choice of flight and booking class - a study using Stated Preference and Revealed Preference data," *Int. J. Serv. Technol. Manag.*, vol. 2, no. 1/2, 2001.
- [23] R. M. Ratliff, B. Venkateshwara Rao, C. P. Narayan, and K. Yellepeddi, "A multi-flight recapture heuristic for estimating unconstrained demand from airline bookings," *J. Revenue Pricing Manag.*, vol. 7, no. 2, pp. 153–171, 2008.
- [24] K. Talluri and G. van Ryzin, "Revenue Management Under a General Discrete Choice Model of Consumer Behavior," *Manage. Sci.*, vol. 50, no. 1, pp. 15–33, Jan. 2004.
- [25] A. Haensel, G. Koole, and J. Erdman, "Estimating unconstrained customer choice set demand: A case study on airline reservation data," *J. Choice Model.*, vol. 4, no. 3, pp. 75–87, 2011.
- [26] J. P. Newman, M. E. Ferguson, L. A. Garrow, and T. L. Jacobs, "Estimation of choice-based models using sales data from a single firm," *Manuf. Serv. Oper. Manag.*, vol. 16, no. 2, pp. 184–197, 2014.
- [27] G. Vulcano, G. van Ryzin, and W. Chaar, "OM Practice--Choice-Based Revenue Management: An Empirical Study of Estimation and Optimization," *Manuf. Serv. Oper. Manag.*, vol. 12, no. 3, pp. 371–392, 2010.
- [28] S. Agrawal, V. Avadhanula, V. Goyal, and A. Zeevi, "A near-optimal exploration exploitation approach for assortment selection," in *EC 2016 - Proceedings of the 2016 ACM Conference on Economics and Computation*, 2016, pp. 599–600.
- [29] V. F. Farias, S. Jagabathula, and D. Shah, "A Nonparametric Approach to Modeling Choice with Limited Data," *Manage. Sci.*, vol. 59, no. March 2015, p. 44, 2013.
- [30] B. Wierenga, *Handbook of marketing decision models*. Springer, 2008.
- [31] S. R. Chandukala, J. Kim, T. Otter, P. E. Rossi, and G. M. Allenby, "Choice Models in Marketing: Economic Assumptions, Challenges and Trends," *Found. Trends Mark.*, vol. 2, no. 2, pp. 97–184, 2007.
- [32] S. Sharif Azadeh, P. Marcotte, and G. Savard, "A taxonomy of demand uncensoring methods in revenue management," *J Revenue Pricing Manag.*, vol. 13, no. 6, pp. 440–456, 2014.
- [33] W. Qi, X. Luo, and X. Liu, "Discrete choice model of customer behavior and empirical study," in *2016 Chinese Control and Decision Conference (CCDC)*, 2016, pp. 5677–5682.
- [34] D. L. McFadden and P. Zarembka, "Frontiers in econometrics," *Cond. logit Anal. Qual. choice Behav.*, p. 105, 1974.
- [35] G. Debreu, "Review of R. D. Luce, individual choice behavior: A theoretical analysis," *Am. Econ. Rev.*, vol. 50, pp. 186–188, 1960.
- [36] M. N. Lee G. Cooper, *Market-Share Analysis*. Boston Dordrecht London: Kluwer Academic Publishers, 2010.
- [37] S. Ja, S., Rao, B. V. Chandler, "Passenger recapture estimation in airline RM," in *AGIFORS 41st Annual Symposium*, 2001.
- [38] T. S. G. and D. Sudharshan, "Equilibrium Characteristics of Multinomial Logit Market Share Models," *Mark. Res.*, vol. 28, no. 4, pp. 480–482, 1991.
- [39] R. H. Zeni, "Improved Forecast Accuracy In Revenue Management By Unconstraining Demand Estimates From Censored Data," *The State University of New Jersey*, 2001.
- [40] P. Guo, B. Xiao, and J. Li, "Unconstraining methods in revenue management systems: Research overview and prospects," *Adv. Oper. Res.*, vol. 2012, 2012.
- [41] T. Abdallah and G. Vulcano, "Demand Estimation under the Multinomial Logit Model from Sales Transaction Data," no. May, 2016.
- [42] S. L. Niël J le Roux, *A Step-by-Step R Tutorial: An introduction into R applications and programming*, 1st ed. 2015.