ORIGINAL PAPER

The impact of customer behavior models on revenue management systems

**Shadi Sharif Azadeh · M. Hosseinalifam ·
G. Savard**

Abstract Revenue management (RM) can be considered an application of operations
research in the transportation industry. For these service companies, it is a difficult
task to adjust supply and demand. In order to maximize revenue, RM systems display
demand behavior by using historical data. Usually, parametric methods are applied to
estimate the probability of choosing a product at a given time. However, parameter
estimation becomes challenging when we need to deal with constrained data. In this
research, we evaluate the performance of a revenue management system when a non-
parametric method for choice probability estimation is chosen. The outcomes of this
method have been compared to the total expected revenue using synthetic data.

1 Introduction

Revenue management systems rely on the expected demand of each fare class. There-
fore, the accuracy of demand models can affect revenue significantly. Most research
in the literature has focused on the optimization methods that booking limits and fare
products are based upon. In 2005, van Ryzin shifted the focus from traditional prod-
uct demand models to the analysis of customer behaviour in revenue management

Sh. Sharif Azadeh (✉) · M. Hosseinalifam · G. Savard
Polytechnique Montreal, 2900 Boulevard Edouard-Montpetit, Montréal, QC H3T 1J4, Canada
e-mail: shadi.sharifazadeh@polymtl.ca

M. Hosseinalifam
e-mail: morad.hosseinalifam@polymtl.ca

G. Savard
e-mail: gilles.savard@polymtl.ca

systems, based on the theory of discrete choice models (random utility) (Ryzin 2005). This change of paradigm has made it possible to blend the concept of revenue maximization with customer behavior analysis in recent research. Cooper et al. (2006) has also shown that ignoring customer behavior in RM systems results in loss of revenue.

In real case studies, the problem of demand modeling is twofold: (1) lack of information about the characteristics of customers and products (or alternatives) (2) size of the historical data. In this paper, we address the first issue. Most research in the area of revenue management solve the problem from parametric point of view by using the maximum likelihood method of parameter estimation. These types of studies require information gathering about customers' characteristics, their arrival rates, and many other factors that affect customers' preferences and consequently their choice probabilities. In this paper however, we consider the cases in which the only available recourses include registered transactions. We propose a new method to tackle this problem by using synthetic data.

An RM system has to decide whether to accept or reject a request from an arriving customer for a given product. Usually, products that are purchased in advance belong to the price-sensitive customer segments. However, the higher fare products are more likely to be bought right before departure. Therefore, availability status of different products at a given time before departure has a direct impact on customer behavior. It is assumed that an arriving customer from each segment has a *consideration set* (a set of alternatives among which the customer selects his choice) and he is willing to purchase the most attractive one based on a *preference vector* (Liu and Ryzin 2008).

This vector is expressed in terms of product-based utilities according to which the probability of choosing an alternative by an arriving customer is determined (e.g., based on a multinomial logit model). In traditional revenue maximization problems, utility and consequently preference vector estimations are performed as an offline study. Based on the Utility Maximization Theory, there are many factors that contribute to finding the utility of a given product. Defining which factors play a greater role in product utilities by taking customer characteristics into account is usually a time consuming procedure.

In the revenue maximization literature, the demand of a given product is often assumed to be independent from the others. That is, every client chooses a product independently from other ones (Cooper et al. 2006; Talluri and Ryzin 2004; Weatherford 2000). One of these revenue maximization techniques is deterministic linear programming (DLP) which was first introduced by Simpson (1989). In their research, the expected demand has been defined by using the mean forecast value. Afterwards, a linear program was suggested to define the optimal demand based on the capacity constraints for a given time period.

A more advanced model was proposed by Gallego et al. (2004), Liu and Ryzin (2008). They have suggested a choice-based deterministic linear program (CDLP) to maximize revenue by defining at a given time which product should be offered to the arriving customers from different segments to maximize revenue. However, in reality, the demand has a stochastic nature and usually the only information available in transportation companies is the registered booking of products during different time periods. Therefore, we need to extract customer behavior based on historical registered data.

In this research, we study a choice-based revenue management network for which we estimate customer preferences by directly using historical data via a new non-parametric algorithm. The revenue performance of this model has been compared with an upper bound resulting from a modified CDLP model (MCDLP) and the outcome of the expected revenue of a simulation model. Our numerical experiments show that the proposed method of preference vector approximation performs as well as a parametric method with less computational cost.

The remainder of the paper is organized as follows. The main problem and the concept of customer preferences are described in Sect. 2. In Sect. 3, a modified CDLP problem is reviewed, which is used to generate synthetic data and an upper bound to the revenue in the comparison scheme. In Sect. 4, a non-parametric mathematical model is represented . In Sect. 5, numerical results are presented that suggest the non-parametric method of preference vector approximation can produce results close to those obtained with the original model. Avenues for further research are outlined in the concluding Sect. 6.

2 Problem description

In order to forecast the expected demand of a given product at a given time we need to have an accurate estimation about choice probabilities (i.e. the probability of choosing a product at a given time). In this research, each choice is made based on two basic rules: (1) each customer can choose only one of the available products, (2) only myopic customers are considered (i.e. people who make their final decision at the time of arrival).

An arriving customer has a personal preference when purchasing a product. A subset of available offered products, considered by a client, is called a *consideration set*. Each customer selects an available alternative from his consideration set based on his preferences. Figure 1 illustrates the registered booking of a given product for different time periods.

Preference vector illustrates the vector of weights of all available alternatives. As soon as one of these products is no longer available the probability of choosing a substitute changes which suggests a conditional probability for choosing another product.

Based on a discrete choice model framework, the probability of choosing a given product by an arriving customer from a given segment is calculated by product utilities.
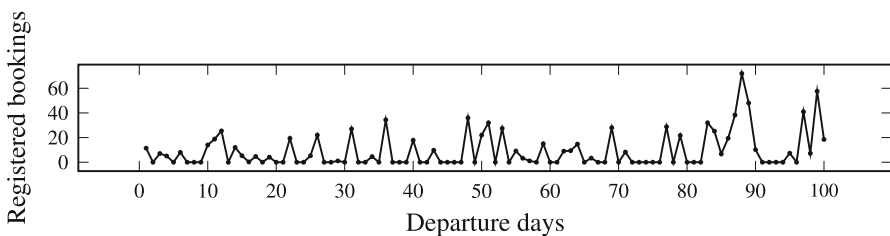

Fig. 1 Registered bookings of a given product for successive departure days

In general, utilities are defined by parametric combination of a deterministic and a stochastic term that are related to the features of each product (McFadden 2001; Train 2009). Parameter estimation becomes more challenging when we have to deal with registered bookings with censored data as a result of the unavailability of some products at different time intervals. In this research, no information about product characteristics is available; therefore, utilities are defined as product-based standalone variables in the mathematical model. Hence, the term non parametric for our proposed demand modeling approach seems appropriate. We use multi-nomial logit (MNL) as our choice model (which will be introduced in more detail in Sect. 3).

According to the above mentioned definitions, choice probabilities are not constant even under the assumption of having time independent product utilities. When the set of available products changes from one time interval to another, the choice probabilities related to available products are adjusted accordingly. We present the mathematical model in Sect. 4. Then, we propose a learning process that helps us to estimate choice probabilities and consequently customer preferences in a more realistic way.

3 Modified CDLP model

In this section, we present a modified version of a choice-based linear program introduced by Bront et al. (2009), Liu and Ryzin (2008). Consider a network with m resources (legs) providing n products. The set of products is expressed by $N = \{1, 2, \ldots, n\}$ and vector $r = (r_1, r_2, \ldots, r_n)$ denotes associated revenue to the products. Vector $c = (c_1, c_2, \ldots, c_m)$ shows the initial capacities of resources. A given product can use more than one resource. The usage of each unit of capacity related to each product is described with an incidence matrix $A = [a_{ij}] \in B^{m \times n}$. The matrix entries are defined by:

$$a_{ij} = \begin{cases} 1, & \text{if resource } i \text{ is used by product } j \\ 0, & \text{otherwise} \end{cases}$$

Time is expressed in discrete periods. The total number of periods is defined by τ, $t = 1, 2, \ldots, \tau$. A customer arrival rate, λ, is considered for a given time interval. We suppose that at most one customer arrives during each period of time and he can buy only a single product or decide not to purchase at all.

Customers are divided into $l = \{1, \ldots, L\}$ different segments with corresponding consideration set C_l. If we have one arrival, p_l represents the probability that an arriving customer belongs to segment l with $\sum_{l=1}^{L} p_l = 1$. We consider a Poisson process of arriving streams of customers from segment l with rate $\lambda_l = \lambda p_l$ and total arriving rate of $\lambda = \sum_{l=1}^{L} \lambda_l$.

In each period of time t, the firm should decide about the offer set (*i.e.* a subset of products, $\mathscr{S} \subseteq N$, that the firm makes available to arriving customers). If set $\mathscr{S}$ is offered, the deterministic quantity $P_j(\mathscr{S})$ indicates the probability of choosing product $j \in \mathscr{S}$ and $P_j(\mathscr{S}) = 0$ if $j \notin \mathscr{S}$. We have $\sum_{j \in S} P_j(\mathscr{S}) + P_0(\mathscr{S}) = 1$, where $P_0(\mathscr{S})$ indicates the no-purchase probability.

Customers' choice probabilities are derived from a multi-nomial logit (MNL) model which is one of the most commonly used models to study how customers make their choices. In the MNL choice model, $v_l \geq 0$ represents a customer's preference vector for available products in consideration set C_l and v_{l0} represents the no-purchase preference. We let $P_{lj}(\mathscr{S})$ denote the probability of selling product $j \in C_l \cap \mathscr{S}$ to a customer from segment l when set $\mathscr{S}$ is offered. According to the MNL choice model, this choice probability can be expressed as follows:

$$P_{lj}(\mathscr{S}) = \frac{v_{lj}}{\sum_{h \in C_l \cap \mathscr{S}} (v_{lh} + v_{l0})} \tag{1}$$

Note that in (1), $P_{lj}(\mathscr{S}) = 0$ if $v_{lj} = 0$ which can be a result of $j \notin C_l$ or $j \notin C_l \cap \mathscr{S}$. That is, if a given product j does not exist in the consideration set of an arriving customer from segment l, the utility associated to that product approaches minus infinity. We assume that $v_{l0} > 0$ for all segments, $l = 1, 2, \ldots, L$.

In the more general case, as a firm cannot recognize the corresponding segment of an arrival in advance, the probability that the firm sells product j to an arriving customer is described as follows,

$$P_j(\mathscr{S}) = \sum_{l=1}^{L} p_l P_{lj}(\mathscr{S}). \tag{2}$$

Therefore, if a set $\mathscr{S}$ is offered the corresponding expected revenue is given by:

$$R(\mathscr{S}) = \sum_{j \in \mathscr{S}} r_j P_j(\mathscr{S}). \tag{3}$$

Given $P(\mathscr{S}) = (P_1(\mathscr{S}), \ldots, P_n(\mathscr{S}))^\top$ be the vector of purchase probabilities, the vector of capacity consumption probabilities $Q(\mathscr{S})$ is denoted by:

$$Q(\mathscr{S}) = A P(\mathscr{S}), \tag{4}$$

where $Q(\mathscr{S}) = (Q_1(\mathscr{S}), \ldots, Q_m(\mathscr{S}))^\top$ and $Q_i(\mathscr{S})$ indicates the probability of using a unit of capacity on leg i, for $i = 1, 2, \ldots, m$.

Let binary variable $X_t(\mathscr{S})$ indicate whether set $\mathscr{S}$ at time t is offered. After obtaining the values of $R(\mathscr{S})$ and $Q_i(\mathscr{S})$, we can embed these functions in the following mathematical programming model to obtain the optimal resource allocation by taking

into account the time and capacity constraints while maximizing revenue.

$$V^{MCDLP} : \max_X \sum_{t=1}^{\tau} \sum_{\mathscr{S}} \lambda R(\mathscr{S}) X_t(\mathscr{S})$$

subject to

$$\sum_t \sum_{\mathscr{S}} \lambda Q_i(\mathscr{S}) X_t(\mathscr{S}) \le c_i \qquad \forall i \qquad\qquad (5)$$

$$\sum_{\mathscr{S}} X_t(\mathscr{S}) \le 1 \qquad\qquad \forall t$$

$$X_t(\mathscr{S}) \in \{0, 1\} \qquad\qquad \forall t, S$$

Model (5) is a modified formulation of the customer choice-based deterministic linear programming model (MCDLP) in which the decision variable $X_t(\mathscr{S})$ indicates offering set $\mathscr{S}$ at booking period t instead of the total time periods during which $\mathscr{S}$ is offered (Bront et al. 2009; Kunnumkal and Topaloglu 2008).

4 Non-parametric approach

In this section, we introduce a mixed integer nonlinear formulation based on which we can estimate customers preference vectors by using registered historical data O_{jt}. The objective function in (6) minimizes the *prediction error* (the difference between estimated demand of each product at a given time and the related registered booking). In this model, we have two sets: (1) set of time intervals $t = 1, 2, \ldots, \tau$, and (2) set of products $j = 1, \ldots, J$. Moreover, S_t shows the choice set or the set of available products at time interval t and p_{jt} presents the probability of choosing product j at time t. As can be observed in this model, we drop the index l that was considered in model (5) for choice probabilities. The reason for this is that when using historical data we do not know which segment a customer came from when choosing the product.

The binary parameter A_{jt} indicates the availability status of product j at time period t. The variables of the model are: u_j that shows the utility related to product j and d_t which presents overall potential demand at time interval t. In this paper, the utilities are assumed to be positive. Also, we define a constant value for the utility of no-purchase in the model that is presented by u_0.

$$\text{MINLP} : \quad \min_{d_t, u_j} \sum_{t \in \tau} \sum_{j \in S_t} (p_{jt} d_t A_{jt} - O_{jt})^2$$

$$\text{subject to} \, p_{jt} = \frac{\exp(u_j)}{\sum_{k \in S_t} \exp(u_k) + \exp(u_0)} \quad j \in S_t \qquad (6)$$

In the above mentioned model, there are two main sources of nonconvexities: first, there is a product of two variables; choice probabilities p_{jt} and potential demand of

each time period d_t in the objective function, and second, there is a fractional term that expresses MNL choice probabilities.

In order to improve the quality of solutions (estimated utilities and potential demand) in (6) without loss of generality, we set some conditions that need to be verified while solving the problem. First, we define a lower bound on d_t based on the cumulative registered bookings of different products at a given time period;

$$d_t \geq D_t^L = \sum_j O_{jt} \qquad t \in \tau. \tag{7}$$

As an upper bound to d_t, we arbitrarily choose a large number. Product utilities are positive and as mentioned above, no-purchase utility is fixed to a constant value.

In addition, we propose a set of valid inequalities based on the logical relations between choice probabilities that improve the performance of our resolution approach.

Valid inequality: It is a property of MNL model that, if the choice set of time interval t is a subset of the choice set of time interval t', that is $S_t \subseteq S_{t'}$, then we have

$$p_{jt} = \frac{\exp(u_j)}{\sum_{k \in S_t} \exp(u_k)a_{kt} + \exp(u_0)} \geq \frac{\exp(u_j)}{\sum_{k \in S_{t'}} \exp(u_k)a_{kt'} + \exp(u_0)} = p_{jt'} \tag{8}$$

Algorithm 1 Valid inequalities on choice probabilities

Input: Registered bookings
Output: Valid inequalities
1: **For all** time periods ($t \in \tau$)
2:     **For all** days ($t' \in \tau$)
       Compare the choice sets
3:     **If** $S_t \subset S_{t'}$ **then**
4:       **If** $S_t = S_{t'}$ **then**
5:         **For all** products ($j \in S_t, j \in S_{t'}$)
6:           Write $P_{jt}(S) = P_{jt'}(S)$
7:       **Else**
8:         **For all** products ($j \in S_t, j \in S_{t'}$)
9:           Write $P_{jt}(S) > P_{jt'}(S)$

These logical relations are extracted from historical data using Algorithm 1.

We add the proposed set of valid inequalities to (6) in order to improve the run time and the quality of solutions of interior point algorithm that is used in the nonlinear solver (KNITRO). After estimating the utility of each product, we obtain the preference vectors based on a multi-nomial logit model. In the next section, we present the computational results in order to compare the outcome of our proposed preference estimation method on revenue with the revenue obtained from the MCDLP model.

5 Computational results

5.1 Data instances

In this research, synthetic data is used to show the impact of our proposed demand model on revenue performance. Twenty-four generated instances are distinguishable based on three main elements: number of products, J (6,8), number of booking intervals, T (7,14,21,28), and the number of customer segments, L (3,4).

The main idea of how to generate synthetic data comes from the benchmark example provided by Liu and Ryzin (2008). In order to generate the historical data, customers are divided for example into three segments (based on their interest towards low fare, medium fare or high fare products) with disjoint consideration sets. Customer segmentations are defined by their fare-class preferences and origin-destination markets. Products are determined according to different combinations of three origin-destinations and their related fare-classes. Moreover, similar to the example presented in the paper of Liu et al., preference vectors (product weights) for different customer segments are also considered as given or parameters of model (5). Usually these values are obtained during an offline study about customer's characteristics using several observable and latent variables. In this study however, we suggest an online study on the preference vectors by using only registered bookings.

For a given number of products, number of segments and number of booking intervals, 3 different random instances have been produced. In order to generate registered historical data, O_{jt}, we use MCDLP via the model presented in (5) to first derive the optimal offer sets, $X_t(\mathcal{S})$, which provide us the choice sets of each time interval, S_t, in the MINLP model.

A total population of around 150–200 customers is considered. Arrival rates are fixed parameters and are defined a priori based on a uniformly distributed variable. The computational results have been carried out on a computer with 2.4 GHz CPU and 8 GB of RAM and 4 cores. We have used Knitro 8, a nonlinear optimizer to solve the problem presented in (6) and we have utilized FICO Xpress-Mosel 7.2 to solve the problem MCDLP (column generation approach).

5.2 Numerical results

Figure 2 shows the resolution framework. First, based on the MCDLP model (5) the efficient offer sets and upper bound to revenue are calculated. In order to avoid conducting the experiment with the same dataset that was used to calibrate the model, we perturb the preference values of different customer segments. Preference vectors are perturbed by adding an error term to the utility values. These added noise terms to the utilities follow a Gumbel (double exponential) distribution with a location parameter zero and scale parameter one (Chaneton and Vulcano 2011). This predefined scale parameter (i.e. dispersion measure or variance) is large enough to present heterogeneity in the out of sample example. Then, the expected revenue of this simulated model is calculated. Using estimated efficient offersets and randomly arriving customers, synthetic historical data is generated. The reason to support this decision is that we
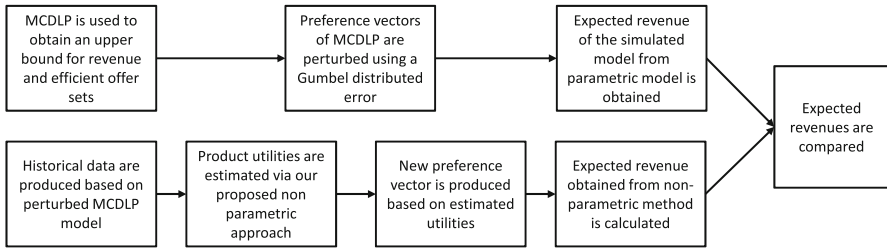
Fig. 2 Non parametric method of preference estimation and its impact on revenue

do not have access to customer information to provide an out of sample example; therefore, we generate it synthetically to test the consistency of the model.

Now, by directly using generated synthetic data, we estimate product utilities with our proposed nonparametric approach. According to obtained estimated utilities, we can predict preference vector of arriving customers. The revenue resulting from both methods are compared to see if changing the method of preference estimation affects revenue performance.

Table 1 presents a comparative study on these 24 instances. As mentioned, these examples are generated based on random customer arrival rates from different segments. Column "Book.Int" shows the number of booking periods. "UB" represents the revenue resulting from MCDLP that provides us an Upper Bound on the expected revenue. Columns "ER(MCDLP)" and "ER(MINLP)" respectively show the Expected Revenues resulting from simulation (with perturbed preference vector) and our non-parametric approach represented in the model (6).

The gap, "Gap(%)", between "ER(MCDLP)" and "ER(MINLP)" is small. Even though there are cases where the non-parametric approach has slightly outperformed the simulated model, this is not surprising. The reason is that while using a non-parametric method, the degree of freedom of the model increases, which can result in slight outperformance. The numerical results suggest that the non-parametric method of preference vector approximation can produce revenues close to those obtained from the original MCDLP .

6 Discussion and conclusion

From a practical point of view, this method can be helpful to revenue management systems in transportation companies or any other service company such as the hotel or rental industries, where there is limited information about customers and product characteristics. Unlike traditional choice models, the main purpose in this research is not to identify the main factors that contribute in the utility function. In this study, we define product utilities as variables of the model that are estimated based on only historical registered bookings. By directly using historical data, we can take product availabilities into account, which brings more dynamism to the decision making process in order to find customer preferences (using MNL choice model).

Table 1 Revenue comparison

	Product	Segment	Book.Int	Instance #	UB	ER(MCDLP)	ER(MINLP)	Gap(%)
1				1	3,954	3,439	3,417	0.64
2			7	2	3,954	3,439	3,431	0.23
3				3	3,954	3,439	3,448	−0.26
4				1	7,500	6,581	6,472	1.66
5			14	2	7,500	6,581	6,743	−2.46
6	6	3		3	7,500	6,581	6,503	1.19
7				1	12,080	10,570	10,305	2.51
8			21	2	12,080	10,570	10,619	−0.46
9				3	12,080	10,570	10,437	1.26
10				1	17,371	15,822	15,062	4.80
11			28	2	17,371	15,822	15,801	0.13
12				3	17,371	15,822	15,796	0.16
13				1	4,817	4,389	4,310	1.80
14			7	2	4,817	4,389	4,347	0.96
15				3	4,817	4,389	4,371	0.41
16				1	9,634	9,234	9,246	−0.13
17			14	2	9,634	9,234	9,194	0.43
18	8	4		3	9,634	9,234	9,282	−0.52
19				1	14,064	13,262	13,264	−0.02
20			21	2	14,064	13,262	13,238	0.18
21				3	14,064	13,262	13,125	1.03
22				1	16,681	15,783	15,597	1.18
23			28	2	16,681	15,783	15,742	0.26
24				3	16,681	15,783	15,663	0.76

In fact, we avoid using other offline studies to capture customer behavior by gathering information about customers' characteristics such as income, purpose of travel and comfort preferences, which can be time consuming and costly.

The results testify to remarkable revenue performance. The gap between these two methods is slight; moreover, there are cases where our model outperforms the simulated revenue of MCDLP. As future work, we want to study the impact of a choice model using more general networks for larger sample sizes.

References

Bront J, Méndez-Díaz I, Vulcano G (2009) A column generation algorithm for choice-based network revenue management. Oper Res 57(3):769–784

Chaneton JM, Vulcano G (2011) Computing bid prices for revenue management under customer choice behavior. Manuf Serv Oper Manag 13(4):452–470

Cooper W, Homem-de Mello T, Kleywegt A (2006) Models of the spiral-down effect in revenue management. Oper Res 54(5):968–987

Gallego G, Iyengar G, Phillips R, Dubey A (2004) Managing flexible products on a network. Working paper, Columbia University, New York

Kunnumkal S, Topaloglu H (2008) A refined deterministic linear program for the network revenue management problem with customer choice behavior. Nav Res Logist (NRL) 55(6):563–580

Liu Q, van Ryzin G (2008) On the choice-based linear programming model for network revenue management. Manuf Serv Oper Manag 10(2):288–310

McFadden D (2001) Disaggregate behavioral travel demand's RUM side. In: Travel behaviour research. Pergamon Press, Oxford

Simpson RW (1989) Using network flow techniques to find shadow prices for market and seat inventory control. In: Proceedings of technical report memorandum, M89–1, Flight Transportation Laboratory, MIT, Cambridge

Talluri K, van Ryzin G (2004) Revenue management under a general discrete choice model of consumer behavior. Manag Sci 50(1):15–33

Train K (2009) Discrete choice methods with simulation, 2nd edn. Cambridge university press, Cambridge

van Ryzin G (2005) Future of revenue management: models of demand. J Revenue Pricing Manag 4(2):204–210

Weatherford L (2000) Unconstraining methods. In: AGIFORS Reservations and Yield Management Study Group