Omega 39 (2011) 73–85

Contents lists available at ScienceDirect

Omega

journal homepage: www.elsevier.com/locate/omega

Capacity rationing in rental systems with two customer classes and batch arrivals

Felix Papier, Ulrich W. Thonemann *

Department of Supply Chain Management and Management Science, University of Cologne, Albertus-Magnus-Platz, 50923 Cologne, Germany

ARTICLE INFO

Article history:
Received 19 September 2007
Accepted 4 March 2010
Available online 1 April 2010

Keywords:
Queuing
Optimization
Rail transport
Rental systems
Revenue management

ABSTRACT

Service differentiation is an emerging method to improve profit and to better serve high-priority customers. Such an approach has recently been introduced by one of Europe's leading rail cargo companies. Under this approach, customers can choose between classic and premium services. Premium service is priced above classic service and premium customers receive a service guarantee which classic customers do not receive. The company has to decide under which conditions it should ration its fleet capacity to classic customers in order to increase service of premium customers. We model such a situation as a batch-arrival queuing loss system. We describe the model, solve it optimally, and derive quantities of interest such as service probabilities. We further analyze it by performing numerical experiments based on the data from the company that motivated our research. We show that the potential of capacity rationing can be substantial in situations like the one we analyzed. We also derive conditions under which rationing is especially beneficial, such as under high unit fleet holding costs or in the presence of batch arrivals compared to single arrivals.

1. Introduction

Historically, cargo logistics companies have not differentiated service between customers and have treated all orders equally. With increasing demand for high service quality on the one hand and increasing pressure to cut costs on the other hand, this is about to change. Some customers require high availability, e.g., when their production schedule depends on timely transportation. They are willing to pay a high price for the rental services in return for a guarantee that their order is served. Other customers are less dependent on availability but require service at low cost. Cargo logistics companies are beginning to leverage this difference in customer needs to improve profit.

In this paper, we analyze a rental system that operates in such a situation. Our research has been motivated by a leading European rail cargo company that recently introduced service differentiation by offering two types of rental services: premium service and classic service. Under premium service, customers pay more for the rental service than under classic service but receive a compensation when their demand is not filled. Under classic service, customers do not receive such a compensation. Both services are served by the same fleet of cargo rail cars, so that joint planning is necessary. In strategic planning, the company has to decide on the size of its rental fleet. In operational planning, it must decide how to allocate rail cars to customers.

We model the rental business as a queuing loss system and explicitly take into account service differentiation. We maximize expected profit by optimizing the fleet size and the admission policy. The contribution of our research is threefold: Firstly, we model a rail car rental business with bursty demand such as the one that motivated our research as a queuing loss system with batch arrivals. Secondly, we develop an algorithm to find the joint optimum of the fleet size and the admission policy. Prior research has focused on characterizing the optimal policy for a given fleet size, but no article shows how to find an optimal solution in a possibly infinite solution space. Thirdly, we analyze extensively the impact of system parameters such as order sizes, penalty cost or holding cost on the expected profit. From these analyses, we derive managerial insights that help decision makers of related rental businesses.

The remainder of the paper is structured as follows: In Section 2, we review the relevant literature. In Section 3, we define the model, derive the objective function, and discuss the main assumptions of the model. In Section 4, we calculate the expected profit and customer service probabilities and optimize the decision variables. In Section 5, we perform numerical experiments which are based on fleet data from the company that motivated our research. From the experimental results, we derive implications for the management of rental businesses. In Section 6, we conclude.

2. Literature review

In this paper, we build on previous research on capacity rationing in stochastic queuing and rental systems. As to our best

* Corresponding author. Tel.: +49 221 470 0.
E-mail addresses: felix.papier@gmx.de (F. Papier), ulrich.thonemann@uni-koeln.de (U.W. Thonemann).

0305-0483/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.omega.2010.03.002

knowledge, Miller [1] was the first to analyze rationing in loss systems. He studies a queuing loss system with multiple customer classes, single arrivals, and identical service time distributions for all customers and proves the optimality of threshold policies. Shonick and Jackson [2] extend the model of Miller to a model in which one class of customers waits in a queue and the other one is blocked if the remaining capacity is low. They apply their model to capacity management of hospital beds. Lippman [3] analyzes queuing loss models with a finite decision horizon. Albright [4] studies queuing loss models with multiple customer classes in which some servers are faster than the others. Carrizosa et al. [5] analyze a queuing loss system with multiple customer classes and arbitrary service time distributions in which a quota-based admission policy is used. The authors derive the optimal admission quotas for each class. Ku and Jordan [6,7] analyze admission control in queuing loss systems in which servers are connected sequentially and customers may require several processing steps.

All of the previously described research assumes single arrivals, but admission control has also been extended to batch arrivals. Kawanishi et al. [8] study a queuing system with two customer classes and batched demand, in which one class of customers is blocked and the other one waits in a queue. Ross and Tsang [9] and Altman et al. [10] also study queuing loss models with batch arrivals. They allow for multiple customers classes and assume entire batch blocking. For the model with two customer classes, Altman et al. prove that a threshold policy is optimal. Örmeci and Burnetas [11] analyze a queuing loss system with batch arrivals and multiple customer classes but with partial batch blocking. All customers are assumed to have the same service time distribution. They show the optimality of a threshold policy with threshold levels that are monotone in the class-specific revenues. Örmeci and Burnetas [12] extend this model to class-specific service-time distributions. They conjecture but do not analytically prove the optimality of threshold policies.

Apart from the extension to batch arrivals, other extensions have recently been published. Örmeci and Van Der Wal [13] consider single-arrival loss models with general interarrival times. For such models, the optimality of threshold policies could be shown. Ku et al. [14] extend admission control to loss networks where customers are routed to other servers after they have received service at one server. Gans and Savin [15] study a queuing loss model for rental businesses with two customer classes: contract customers and walk-in customers. Contract customers have fixed prices while walk-in customers are offered an individual price upon their arrival. The authors derive an optimal admission policy for contract customers and optimal prices for walk-in customers.

The idea of rationing has also been analyzed extensively in the context of inventory management. In inventory management, two or more customer classes request items from a common inventory. As with rental systems, rationing can be used to protect high-priority customers. While the idea of rationing in inventory management is identical to the idea of rationing in our research, the underlying methods differ considerably. For an overview on existing research in this context, we refer the interested reader to Zhao et al. [16] or Tempelmeier [17].

Revenue management approaches have also been applied to a variety of different industry settings. For example, Syam and Côté [18] present a revenue management model for health care services, Yao et al. [19] and Tsai and Hung [20] each develop models for multi-channel settings with internet e-tailing channels.

In this research, we analyze a stochastic rental system with two customer classes and compound Poisson distributed demand. Unlike previous research, we also focus on deriving the optimal fleet size in addition to the optimal admission policy. We also study the effect of the system parameters on performance and derive managerial insights from this study. The objective function, the decision variables, and the model assumptions are described in detail in the next section.

3. Model description

We consider a rental company that rents vehicles such as rail cars to commercial customers. The company operates a fleet of c homogeneous vehicles. The company offers two different services to its customers: *premium service* (*class* 1) and *classic service* (*class* 2). The contribution margins are r_1 and r_2, respectively, with $r_1 \geq r_2$. We do not allow switching of customers between classes after they have observed the state. We can make this assumption, because customers do not change the type of service they usually order. This is due to the fact that in the cargo rail business customers subscribe to one type of service by long-term contracts.

Customers arrive at random arrival times and demand a certain number of cars (this number is referred to as order size B). We assume that customer demands arrive according to two independent compound Poisson arrival streams. The arrival rate of class i orders is denoted by λ_i. The order size distribution of an order of class i is given by the probability mass function $f_i(n)$ with mean q_i and $F_i(0)=0$. Order sizes can have an arbitrary distribution.

When a customer arrives, the rental company can decide whether to allocate the customer all requested cars, some cars or no cars at all. The admission decision of a customer is denoted by a. $a=n$ corresponds to an allocation of n cars, while $a=0$ corresponds to a rejection. Obviously, the number of cars allocated must be less than or equal to the number of requested cars (B) and the number of idle cars upon arrival ($c-N$), $a \leq \min(B, c-N)$, where N denotes the number of cars rented out right before the order arrives. Customers that are denied service are lost. Lost class 1 demand receives a compensation p per lost car. We denote a policy which specifies an a_j for each arrival j by $\wp$.

Once customers are allocated cars, they hold them for random periods of time. The duration of a rental service is referred to as the *rental time*. We denote the mean of the rental time by $\overline{S} = \mu^{-1}$. It comprises delivery of the empty cars to the customer, loading operations, delivery of the cars to the destination, unloading operations, and the return to the rental company. After the customer has used a car, she returns it to the rail car company. Rental times are assumed to be i.i.d. random variables with an exponential distribution. This assumption is made by many articles on capacity rationing in rental systems (e.g. cf. [8,11,15]), because class-specific rental time means or other rental time distributions limit the analytical tractability of the model and because this assumption reflects the idea that all rail cars are used in common processes, independent of the class origin.

On a long-term basis, the rental company can decide on the fleet size. On a short-term basis, it can decide on the admission policy. Thus, the decision variables are the fleet size c and the admission policy $\wp$. The objective of the rental company is to maximize expected profit per time unit. Profit consists of regular revenues and operating costs, penalties paid to customers, and holding costs for the fleet. The fleet size c and the admission policy $\wp$ determine the expected amount of served class i demand per time unit, denoted by $V_i(c,\wp)$. The expected revenues are $r_1 V_1(c,\wp)$ for class 1 demand and $r_2 V_2(c,\wp)$ for class 2 demand. The rental company must pay a compensation for each unmet class 1 demand, which amounts, on average, to $p(\lambda_1 q_1 - V_1(c,\wp))$. Holding