European Journal of Operational Research 000 (2017) 1–11

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Production, Manufacturing and Logistics

Optimal spares allocation to an exchangeable-item repair system with tolerable wait

Michael Dreyfuss, Yahel Giat*

Department of Industrial Engineering, Jerusalem College of Technology, Jerusalem, Israel

ARTICLE INFO

Article history:
Received 27 January 2016
Accepted 13 February 2017
Available online xxx

Keywords:
Inventory
Logistics
Truncated waiting time
Window fill rate
Optimization criteria

ABSTRACT

In a multi-location, exchangeable-item repair system with stochastic demand, the expected waiting time and the fill rate measures are oftentimes used as the optimization criteria for the spares allocation problem. These measures, however, do not take into account that customers will tolerate a reasonable delay and therefore, a firm does not incur reputation costs if customers wait less than their tolerable wait. Accordingly, we generalize the expected waiting time and fill rate measures to reflect customer patience. These generalized measures are termed the truncated waiting time and the window fill rate, respectively. We develop efficient algorithms to solve the problem for each of the criteria and demonstrate how incorporating customer patience provides considerable savings and profoundly affects the optimal spares allocation.

1. Introduction

Exchangeable-item repair systems have been investigated and applied in a variety of service, maintenance and inventory settings. In these systems, customers bring a failed item and exchange it for a serviceable item that is available in stock. The failed item itself is repaired on site after which it is returned to stock. To increase the availability of serviceable items in stock, spare items are placed in stock. If the system has multiple locations into which customers may arrive, then in addition to determining the number of spares to be purchased, managers must also decide how to allocate the spares to the various locations with the goal of optimizing the predetermined service measure.

The target level of service is by itself an important managerial consideration and lies at the heart of this paper. Researchers frequently assume that the firm's objective is to minimize the expected number of backorders, which, by Little's Law may be reformulated to minimizing the expected waiting time, (e.g., Van der Heijden, Alvarez, & Schutten, 2013; Wong, van Houtum, Cattrysse, & van Oudheusden, 2006). The reasoning behind this is quite straightforward; longer waiting times are associated with negative customer satisfaction and result in reputation losses to firms. The prevalent use of this measure is further facilitated by the fact that expected waiting time is easy to compute and is a convex function of the number of spares and therefore a simple greedy algorithm can attain the optimum efficiently. In many settings of practical interest, however, expected waiting time is not an accurate proxy for the firm's costs. Frequently, firms are obliged, either through government regulation or by contractual commitment, to reduce the waiting time to below a threshold. From the customers' standpoint, too, there is a certain tolerable or acceptable period of wait, which may depend on their level of patience or expectation. In these cases, the objective should be to minimize the expected wait *beyond* this tolerable wait.

In other situations, the cost to the firm does not depend on the time the customer waits but on *whether* the customer has to wait. In these situations, the fill rate measure, i.e., the portion of customers who are served upon arrival, is the appropriate criterion. Furthermore, if the firm is penalized only for the number of customers who wait beyond the tolerable wait then the fill rate measure should be generalized to reflect the firm's correct objective. Accordingly, given a spares allocation $\vec{n}$, we consider in this paper two measures of service performance to serve as the optimality criteria for the spares allocation problem:

- The truncated waiting time (TWT), $W(\vec{n}, t)$: The expected time waited beyond t units of time. For $t = 0$, this measure is the expected waiting time.
- The window fill rate (WFR), $F(\vec{n}, t)$: The expected fraction of customers who are served within t units of time. For $t = 0$, this measure is the fill rate.

The proposed service measures may be used in a variety of settings. The bulk of exchangeable-item repair system models focus on logistic systems such as maintenance and inventory systems.

* Corresponding author.
E-mail addresses: dreyfuss@jct.ac.il (M. Dreyfuss), yahel@jct.ac.il (Y. Giat).

http://dx.doi.org/10.1016/j.ejor.2017.02.031
0377-2217/

Service contracts in these systems compensate customers if they do not receive service within a certain time window, which makes a compelling case for optimizing the WFR (see also Caggiano, Jackson, Muckstadt, & Rappold (2009)).

Exchangeable-item repair systems have also been applied to service-oriented settings. Such an application is the electric vehicle battery-swapping network (see, for example, (Avci, Girotra, & Netessine, 2014)). Electric vehicles' batteries need to be recharged frequently with inconveniently long recharging time. One suggestion to overcome this problem is to leave battery ownership to a third party that will construct multiple battery-swapping stations in which car owners replace their depleted batteries for recharged ones. In such a system, customers expect a certain wait and the system's reputation costs accrue only after this expected wait. Thus, managers are advised to minimize the TWT.

The use of the expected waiting time or fill rate as a criterion for optimality has been considered repeatedly by researchers and dates back to Sherbrooke (1968, p. 127). There is, however, a difficulty with the treatment of the fill rate. The fill rate measure is concave in the number of spares in each location only if each location has been allocated sufficiently many spares. Therefore, the researchers who considered the fill rate as a criterion for optimality imposed the strict limitation of considering only the region for which the fill rate is concave. The goals of this paper are therefore twofold. The first, to develop an algorithm that efficiently computes the spares allocation that maximizes the fill rate. Secondly, and more importantly, we aspire to understand how taking into account customer patience, (i.e., the tolerable wait), affects the optimal allocation.

To achieve these goals we build on Berg and Posner (1990) who derive a mathematical expression for the WFR. We show that when item arrival is Poisson the WFR is either concave, or initially convex and then concave in the number of spares in each location. Further, we show that for a sufficiently large tolerable wait, the WFR is concave shaped. These observations allow us to make a number of contributions to current theory in exchangeable-item repair systems:

First, we are novel in our generalization of the expected waiting time measure and develop a mathematical expression for the TWT. Since this expression depends on the WFR, we exploit the mathematical properties of the WFR to show that the TWT is decreasing and convex in the number of spares and therefore optimality can be attained efficiently using a greedy algorithm.

Second, we develop an algorithm to efficiently derive a near optimal solution to the problem of optimizing the WFR. Using the properties of the WFR, we define a concave covering function of each of the locations' WFR, and optimize these covering functions efficiently. We find that if there are few spares and the tolerable wait is short, then spares must be allocated to only part of the locations leaving the others with no spares. We characterize the *a priori* bound for the distance from optimum of the suggested algorithm. For the cases that the solution is sub-optimal for the original problem, we characterize the *a posteriori* distance from optimum. We show that the distance from optimum depends on the functional value of only one location and therefore as the scale of the problem increases the *a priori* and *a posteriori* distances from optimum decrease.

Third, we derive formulas for the WFR when demand follows a Compound Poisson process and where partial service is not allowed. Although the convex-concave property of the WFR does not generally hold, we show that the optimization algorithm requires only minor adjustments to achieve a near optimal solution.

Our paper also provides practical insights to managers, which we demonstrate through a large-scale numerical example motivated by the electric vehicle battery swapping problem. This example underscores the importance of defining correctly the criterion of optimality. In particular, we demonstrate how costly it is for companies who neglect to take into account the customers' patience level. Second, we show how the optimal allocation for the TWT changes with the tolerable wait. As the tolerable wait increases, the relative importance of high-arrival locations is mitigated and these locations are allotted fewer spares. Third, we show that the fill rate is more sensitive to customer patience than the expected waiting time. This implies that accurately estimating the tolerable wait is essential particularly when the fill rate is the optimality criterion. Finally, we show that the WFR criteria creates two classes of locations, such that one class receives spares and the other does not. As a consequence, in this case, managers should develop two different policies with respect to their service time to customers.

2. Literature review

Exchangeable-item inventory systems have been extensively researched in various settings (van Houtum, 2014; Muckstadt, 2005; Sherbrooke, 2004). There are two popular service measures for these systems, the fill rate and the expected backorders (e.g., Basten, van der, & Schutten, 2012; Caggiano, Jackson, Muckstadt, & Rappold, 2007; Ghaddar, Sakr, & Asiedu, 2016; Kranenburg & van Houtum, 2009; Sherbrooke, 1968; Tsai & Zheng, 2013; Wong et al., 2006). Berg and Posner (1990) extend the fill rate measure by deriving the waiting time distribution so that one can measure the probability to be served within any time window *t*. In our paper, we denote the waiting time distribution as the WFR. This measure is similar to Song (1998) who compute the order fill rate for a multi-item system with lost sales, Caggiano et al. (2009) who derive approximations for the channel fill rate and Dreyfuss and Posner (2015) who derive the WFR for a system with batch arrivals. The WFR differs from Kutanoglu and Lohiya (2008) lost-sales fill rate, since in our model orders are backlogged.

The second popular service measure is the expected backorder measure, which is proportional to the expected waiting time. The TWT incorporates customer patience into the expected waiting time. To the best of our knowledge, there is no research considering the TWT as a service measure. Therefore, while there are many papers which optimize the spares allocation according to the expected waiting time (e.g., Basten et al., 2012; Kranenburg & van Houtum, 2009; Wong et al., 2006), our paper is novel in that it allows optimization using the TWT.

In spite of the prevalent use of the fill rate as a service measure, only a few papers consider it as a criterion for the optimal spares allocation due to its non-concave form. The few papers who discuss the fill rate as a criterion for optimality, limit the search space to its concave region only (e.g., van Houtum, 2014; Larsen & Thorstenson, 2014; Muckstadt, 2005; Sherbrooke, 1968; Song & Yao, 2002). Caggiano et al. (2007) is the only paper that attempts to *optimize* according to the generalization of the fill rate. Their model is a multi-item, multi-echelon model that allows different delivery times to different locations and items. We differ from their paper in a number of ways. First, we do not assume deterministic lead times. Instead, in our model the lead time or repair time is stochastic with a general distribution. Second, our WFR can be computed for any time window and not only the lead times. Third, our optimization procedure uses exact values of the WFR and not approximations. Finally, Caggiano et al. (2007) assume that each location receives sufficiently many parts so that the fill rates are typically concave. In contrast, we do not limit the spares search space. In fact, we show that limiting the search space may decrease the fill rate, since, it may be optimal that certain locations receive no spares at all.

Our finding that the WFR's functional form is generally S-shaped ties our research to optimization models of additively

Please cite this article as: M. Dreyfuss, Y. Giat, Optimal spares allocation to an exchangeable-item repair system with tolerable wait, European Journal of Operational Research (2017), http://dx.doi.org/10.1016/j.ejor.2017.02.031