Computers and Operations Research 98 (2018) 13–23

Contents lists available at ScienceDirect

Computers and Operations Research

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

An analytical approach to determine the window fill rate in a repair shop with cannibalization

Michael Dreyfuss, Yahel Giat*, Alan Stulman

Jerusalem College of Technology, Jerusalem, Israel

ARTICLE INFO

Article history:
Received 15 August 2017
Revised 21 March 2018
Accepted 17 May 2018
Available online 18 May 2018

Keywords:
Inventory
Cannibalization
Window fill rate
Exchangeable-item repair system
$M/G/\infty$
Spares allocation problem

ABSTRACT

We consider a repair shop in which each unit comprises multiple component types and cannibalization is allowed. The shop's managers have a budget for purchasing spare components and need to decide how many spares of each component type to purchase. Customers arrive to the shop with a single unit of which at least one of its components has failed and expect to be served within a tolerable waiting time. Accordingly, the shop's performance measure is the window fill rate, that is, the fraction of customers who are served within the tolerable wait. In our analysis, we develop exact formulas for the window fill rate that comprise multiple dependent Skellam random variables. We overcome the practical complexity of evaluating these formulas by using simulation to evaluate only the random elements. We discuss run-time considerations for solving the spares allocation problem and demonstrate how the optimal solution and the window fill rate depend on the tolerable wait, budget and the customer base using an illustrative numerical example.

1. Introduction

Many advanced and complex systems, such as aircrafts and marine ships, comprise exchangeable components that are very expensive and therefore whenever they fail they are repaired in lieu of being discarded. One way to ensure the availability of such operable replacements is to purchase sufficiently many spare items. The inventory system just described is commonly referred to as an exchangeable-item repair system with spares. Such inventory systems and its many variants are in common practice in military and civil settings, such as the aviation and marine industries, nuclear power plants and high-end electronic machines (Tsai and Liu, 2015). Purchasing spare items for these systems is usually very costly, and therefore, it is common that managers use an inventory practice called cannibalization. With cannibalization, operable *components* are removed (i.e., cannibalized) from non-functioning *units*, thus, in effect, generating additional spare components free of cost.

In this paper, we consider an exchangeable-item single-location repair system, in which the inventory unit comprises multiple components and in which cannibalization is permitted. Given a budget for spares, the managers solve the spares allocation problem, i.e., deciding how many spare components of each type to purchase. For the performance measure we use a generalization of the fill rate, the window fill rate. Whereas the fill rate measures the fraction of customers whose demand is met *upon* their arrival, the window fill rate, in contrast, measures the fraction of customers whose demand is met *during* a given time window. The window fill rate captures two important features in the relationship between customers and service providers. Firstly, most service agreements allow for service to be rendered within some window of time (Caggiano et al., 2009). Secondly, even absent a specific agreement, customers usually tolerate a certain wait and will not penalize the service providers if they are served within this tolerable wait (Durrande-Moreau, 1990).

This paper makes a number of contributions. We derive an exact formula for the window fill rate in a multi-component setting with cannibalization. Similarly to Dreyfuss and Giat (2017a), we track supply and demand equations for each of the component types. As a result, we make repeated use of counting the number of components arriving to the system and exiting it. Customer arrivals are modelled as a Poisson process and as a result the window fill rate formula comprises multiple dependent Skellam random variables. A Skellam random variable is the difference between two Poisson random variables (Skellam, 1946). Unfortunately, the dependencies between these Skellam random variables make the evaluation of the window fill rate computationally cumbersome. As a result, the paper's second contribution is a hybrid algorithm in which we separate between the random and the non-random elements of the window fill rate and evaluate the random

* Corresponding author.
E-mail address: yahel@jct.ac.il (Y. Giat).

https://doi.org/10.1016/j.cor.2018.05.018
0305-0548/

elements using Monte Carlo simulation. Thirdly, we propose two heuristics to solve the spares allocation problem. Fourthly, we use a numerical example to illustrate the importance of assessing correctly the customers' tolerable wait. We use the numerical example to investigate how the optimal solution changes with the tolerable wait, spares budget and the number of customers served by the repair shop. One of the insights gleaned from the numerical analysis is the existence of a wide range of budgets that do not affect the window fill rate. As a consequence, within these ranges it is likely that some of the budget allocated to spares will be "wasted" in that its effect on the window fill rate is negligible.

2. Literature review

Our model is a single-location version of the METRIC model (Multi-Echelon Technique for Recoverable Item Control) described in Sherbrooke (1968) seminal paper whose many extensions are described in Sherbrooke (2004) and Muckstadt (2005) and surveyed in Basten and van Houtum (2014). Among the main assumptions of METRIC are that demand follows a Compound Poisson distribution, "first come first serve" (FCFS) service policy, ample repair facilities and a "continuous one for one" $(S-1, S)$ review policy. Models that depart from these assumptions include Díaz and Fu (1997) who consider limited repair capacity, Levner et al. (2011) and Park and Lee (2014) who consider a non-Poisson arrival, and Park and Lee (2011, 2014) who investigate an (*S, Q*) policy.

The fill rate, i.e., the fraction of customers who receive a replacement item immediately upon arrival, is one of the commonly used performance measure of such inventory systems (e.g, Shtub and Simon, 1994, Caggiano et al., 2007, Lien et al., 2014). The *window fill rate* for time *t* is a generalization of the fill rate and defined as the proportion of customers who are served within the time window *t*. This measure is appropriate when customers tolerate a certain wait (Durrande-Moreau, 1990; see also Katz et al., 1991 who use the term "reasonable duration"), or when the inventory system's contracts are such that service must be rendered within a predetermined time frame (Caggiano et al., 2009). Papers with a similar approach include Dreyfuss and Giat (2017a), Song (1998) and Caggiano et al. (2007). Dreyfuss and Giat (2017a) develop exact formulas for the window fill rate and develop an algorithm for finding a near-optimal solution to the spares allocation problem. Song (1998) and Caggiano et al. (2007) use approximation or bounding techniques to estimate the fill rate when service must be rendered within a time window. Caggiano et al. (2007) uses these approximation formulas and simulation to solve the spares allocation problem.

Most METRIC-related papers examine items that comprise multiple types of components. Smidt-Destombes et al. (2011) use the METRIC framework to consider the case in which there are redundant components. With redundancy, an item may still be considered operable when some of its components have failed (see more in Cochran and Lewis, 2002). Another related inventory practice is removing functional components from units that are being phased out, a practice reported as far back as Mendershausen (1958) and Geisler (1959). Recently, Block et al. (2014) proposed a model to optimally manage spares from an aircraft once it is decided to be phased-out. Similarly, Wijk et al. (2017) combine simulation and a genetic algorithm to optimize the spares management of aircraft maintenance being phasing out.

Cannibalization, which we consider in this paper, is a common practice in the aviation industry and is investigated in a number of contexts. Our modelling is similar to Sherbrooke (2004, Chapter 8) who assumes complete and instantaneous cannibalization. Fisher and Brennan (1986) use simulation to compare between eight cannibalization schemes ranging from no cannibalization to complete cannibalization and Dreyfuss and Stulman (2018) develop cannibalization policies depending on the number of customers in the queue and the components' assembly times. Salman et al. (2007) examine the benefits of cannibalization in the presence of maintenance-induced damage and in Sheng and Prescott (2016) the demand for parts arises from preventive and corrective maintenance needs. Except for Sherbrooke (2004) the aforementioned papers use simulation to evaluate the system's performance. Unfortunately, simulation is too time consuming if many evaluations are needed such as in the case of finding the optimal spares allocation. In contrast to these simulation-oriented papers, Sherbrooke (2004) develops formulas for the expected back-order amount and the aircraft's availability. In our study, the system's performance measure is the window fill rate.

In this paper, we make repeated use of the Skellam distribution, also known as the Poisson Difference distribution, which is a discrete distribution that describes the difference between two independent Poisson random variables. Formally, $S \sim Skellam(\lambda_1, \lambda_2)$ if $S = P_1 - P_2$ and $P_i \sim Poisson(\lambda_i), i = 1, 2$ are independent. Mathematical treatment of the Skellam distribution goes back to Irwin (1937) and Skellam (1946). This distribution is used mainly when difference of counts are needed, and has been applied in a broad range of settings such as sports simulation (e.g., Karlis and Ntzoufras, 2009), biology (e.g., Jiang et al., 2014) medical treatment (e.g., Karlis and Ntzoufras, 2006) and signal imaging (e.g., Hirakawa et al., 2009). In the context of this paper, we use multiple Skellam variables to count the difference between the number of failed components brought by customers and the number of functioning components given to the customers. With the exception of Dreyfuss and Giat (2017a), we are unaware of any other inventory model that makes use of Skellam random variables in its analysis.

3. The model

Customers arrive to a repair shop with a non-functioning unit that comprises *K* distinct components. Units have no redundant components and therefore, a unit is non-functioning if one or more of its components have failed. Consequently, to render it functional, each of the failed components must be either repaired or exchanged with an operable component of the same type. Some of the item's components may be repairable e.g., a battery is chargeable, an engine refurbished, etc. Non-repairable component types are ordered from a supplier or central warehouse. In the context of this paper, we treat the various replenishment operations (actual repair or ordering) as repair and therefore assume that all component types are repairable.

Customers are served based on a FCFS policy. Let $\hat{R}_k(\cdot)$ denote the cumulative repair times distribution function of type-*k* component, $k = 1, \ldots, K$. As with most METRIC models, we assume that there are ample servers, namely, that the repair facility in the shop is sufficiently equipped that once a failed component is removed there is always a repair station available and therefore the component's repair commences immediately. After repair the component's condition is as new. The repair times of each server is independent and follows $\hat{R}_k(\cdot)$.

In this model, customers are differentiated only by which of their components have failed. Throughout the paper we use the following notation. Let *j* be an integer number between 1 and $2^K - 1$. The binary representation of *j* is a string of *K* binary bits where the *k*th bit (from the right) is 1 if the *k*th component is non-functioning and 0 if the *k*th component is functional. We say that a customer is of type *j* if for each $k = 1, \ldots, K$:

$$j^{(k)} = 1 \quad \Leftrightarrow \quad \text{the } k\text{th component of type-}j \text{ customer is non-functional.}$$