European Journal of Operational Research 254 (2016) 113–126

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

Risk measures and their application to staffing nonstationary service systems

Jamol Pender*

School of Operations Research and Information Engineering, Cornell University, 228 Rhodes Hall, United States

ARTICLE INFO

Article history:
Received 28 January 2015
Accepted 9 March 2016
Available online 20 April 2016

Keywords:
Queues and service systems
Risk measures
Healthcare
Time inhomogeneous markov processes
Staffing

ABSTRACT

In this paper, we explore the use of static risk measures from the mathematical finance literature to assess the performance of some standard nonstationary queueing systems. To do this we study two important queueing models, namely the infinite server queue and the multi-server queue with abandonment. We derive exact expressions for the value of many standard risk measures for the $M_t/M/\infty$, $M_t/G/\infty$, and $M_t/M_t/\infty$ queueing models. We also derive Gaussian based approximations for the value of risk measures for the Erlang-A queueing model. Unlike more traditional approaches of performance analysis, risk measures offer the ability to satisfy the unique and specific risk preferences or tolerances of service operations managers. We also show how risk measures can be used for staffing nonstationary systems with different risk preferences and assess the impact of these staffing policies via simulation.

1. Introduction

Time varying queueing models such as the $M_t/G/\infty$ queue and the $M_t/M/k_t + M$ queue are standard models for describing the dynamics of large scale service systems like telecommunication systems, call centers, and healthcare systems like hospitals. To get a good understanding of the wide variety of applications of nonstationary queueing models, see for example (Khudyakov, Feigin, & Mandelbaum, 2010) for applications to call centers with interactive voice response and Yom-Tov and Mandelbaum (2014) for application to healthcare systems. However, staffing these systems appropriately and stabilizing salient performance measures such as the probability of delay and waiting times for these stochastic systems has been a long standing problem in the queueing literature for many years.

One of the first solutions for stabilizing the delay probabilities for multiserver queues without abandonment was developed by Jennings, Mandelbaum, Massey, and Whitt (1996). Jennings et al. (1996) develop a novel square root staffing algorithm that uses the offered load of an infinite server queue and the square root of the offered load for refinements to stabilize the delay probabilities in multi-server queues. In the case of exponential service times, it only requires the solution to a simple ordinary differential equation to find the appropriate staffing level. However, as noted in Feldman, Mandelbaum, Massey, and Whitt (2008) and Liu and Whitt (2012) and Massey and Pender (2013), this algorithm for stabilizing the delay probabilities does not stabilize the abandonment probabilities and other performance measures. Thus, Liu and Whitt developed a new approach that stabilizes the abandonment probabilities and mean delay using the combination of two infinite server queues.

Nonetheless, these algorithms for performance stabilization are only useful for a few performance measures that are well-studied in the queueing literature and are especially tailored for applications in telecommunications where there is no extreme consequence if a customer waits a long time for service. For instance, in a call center it is considered good performance if 99 percent of customers are served within 2 minutes and we might not care about the 1 percent of customers who might have extremely long wait times. However, in a healthcare or emergency care setting, patients with extremely long waiting times can be very costly to the hospital, especially if their health deteriorates while waiting and subsequently they die before being seen, see for example (Castillo, 2014). Consequently, it is not sufficient to just make sure that waiting times are short, but it is also important to make sure that even excessive waiting times are short in the context of healthcare.

To address the difference between application settings like telecommunications and healthcare, in this paper we propose analyzing the new problems in applications like healthcare with new

* Tel.: +1 6464185950; fax: +1 646 418 5950.
E-mail address: jjp274@cornell.edu

http://dx.doi.org/10.1016/j.ejor.2016.03.012
0377-2217/

ideas, namely using static risk measures from the mathematical finance literature. The advantage of using static risk measures over traditional approaches of performance analysis, is that the risk measure approach can be adapted to a manager's risk preferences and the particular application context. The fact that the risk measure approach can be adapted to different applications and in different contexts within a particular application is quite useful for managers of service systems. One example in healthcare is that patients with shortness of breath might be less willing to tolerate long waits than patients with an ankle sprain so a different risk measure should be used for those patients. Thus, this risk measure approach allows the manager of a service center such as hospital to choose his or her own risk preferences for the overall performance of the system as well as the individual parts of the system.

In order to develop this risk measure approach for general service systems, we need to specify a stochastic model for the dynamics of our service systems. In this paper, we begin with the infinite server queueing model. This model is very natural as a start since its dynamics are tractable in the stationary and nonstationary setting. Not only are the mean and variance dynamics tractable, but also the entire distribution is known for the infinite server queue when initialized with a Poisson distribution or at zero. Besides the fact that the infinite server queue is a relatively simple model, it is also an offered load model. Thus, the infinite server dynamics represents the system when an unlimited number of resources are available and serves as an lower bound for the dynamics of finite server systems.

In addition to the infinite server queue, we also analyze the canonical nonstationary Erlang-A queueing model. The nonstationary Erlang-A model assumes the customer arrival process is a nonhomogenous Poisson process with nonstationary arrival rate $\lambda(t)$. We also have k servers with i.i.d. service times that are exponentially distributed with mean $1/\mu$. Finally, all the customers have i.i.d. abandonment times that are also exponentially distributed with mean $1/\beta$. Although the Erlang-A model is a simple model for some complex realities, it is also very hard to find closed form expressions for many of the performance measures of interest in the nonstationary setting. Thus, we must find approximations of the Erlang-A that are accurate and more tractable in terms of providing closed form expressions for performance measures of interest.

One standard method would be to use the fluid and diffusion limits of Mandelbaum, Massey, and Reiman (1998). However, it is well known that for small values of the scaling parameter η, the fluid and diffusion limits are not warranted. Moreover, when the mean queue length is near the number of servers, the fluid and diffusion limits are not Gaussian. Thus, in this work, we use another approximation to accurately estimate the queue length process. This approximations is known as the Gaussian variance approximation (GVA) of Massey and Pender (2011) and uses a Gaussian surrogate distribution to approximate the queue length dynamics. With this approximation for the queue length dynamics, we then approximate various risk measures for the queue length process and illustrate their performance as tools for staffing the system. We are not the first to study staffing issues in queues, see for example (Engblom & Pender, 2014; Pender, 2015; Stolletz, 2008; Tirdad, Grassmann, & Tavakoli, 2016; Yarmand & Down, 2013), however, we are the first to use risk measures in this context.

1.1. Contributions

To the best of our knowledge our contributions in this work are the following.

- We are the first to illustrate how static risk measures from the mathematical finance literature can be used in the context of server staffing and performance analysis in queueing theory.
- We derive explicit approximate staffing schedules for various risk measures that are widely used in the financial community and derive closed form expressions for the values of risk measures under Poisson and Gaussian distributional assumptions.
- We use the risk measures as staffing procedures and assess the results through comparing standard performance measures such as the probability of delay and abandonment probabilities.

1.2. Outline of paper

The rest of the paper is as follows. In Section 2, we introduce the concept of risk measures and provide several examples of risk measures. We also introduce the concept of functional risk measures, which will also be used throughout the rest of the paper. In Section 3, we start with the infinite server queue and derive closed form formulas for several risk measures for the queueing process. In Section 4, we introduce the Erlang-A model and several approximations for it. In Section 5, we use the approximations for the Erlang-A model queueing model and derive closed form expressions for the risk measures of the queueing model. In Section 6, we give numerical results and describe the impact of using the risk measures for staffing the system. We give examples of some extensions and conclude with final remarks in Section 7.

2. Static risk measures

One of the central goals in mathematical finance is to assess the risk of financial positions. The risk of a financial position may be seen as the capital reserves that a bank should hold in response to the risk it exposes itself to. Inspired by this notion of risk as a minimal capital reserve and by the shortcomings of *V@R*, Artzner et al. (1997,1999) introduced an axiomatic approach to coherent risk measures. The goal of a coherent risk measure is to quantify the risk of X by a number $\rho(X)$. It is our goal in this paper to introduce this notion of risk measures into the world of queueing theory where there are analogous notions of risk and reserves. In fact, in the context of queueing theory and staffing, the notions of risk and reservers can be viewed as the number of staff needed to maintain a specific quality of service level. Before we describe how various risk measures are related to various performance quantities in the service systems literature, we give a brief overview of risk measures to make the paper self-contained for the reader's convenience.

Definition 2.1. A mapping $\rho : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ is called a monetary risk measure if $\rho(0)$ is finite and if ρ satisfies the following conditions for X,Y $\in \mathcal{X}$.

- Monotonicity: If X $\leq$ Y , then $\rho(X) \geq \rho(Y)$.
- Cash Invariance: If m $\in \mathbb{R}$, then $\rho(X + m) = \rho(X) - m$

These two conditions are very necessary to define risk measure. It is clear that if X is always smaller than Y under every scenario $(\forall\omega)$, then the risk associated with X should be higher than the risk associated with Y. Moreover, if we add cash to our position, it should reduce the risk of that position because cash is not a risky asset.

Definition 2.2. A monetary risk measure ρ is called a convex or quasi-convex risk measure if ρ satisfies the following condition for X,Y $\in \mathcal{X}$.

- Convex: If $\rho(\lambda X + (1-\lambda)Y) \leq \lambda\rho(X) + (1-\lambda)\rho(Y)$ for all $\lambda \in$ [0,1].
- Quasi-Convex: If $\rho(\lambda X + (1-\lambda)Y) \leq \max\{\rho(X), \rho(Y)\}$ for all $\lambda \in$ [0,1]