Queueing Models and
Service Management
Vol. 7, No. 2, page 57–73, 2024

QMSM

A Quantitative Model for Staffing Problems in Inpatient Units with Multi-type Patients

Siping Su[1,2*], Shey-Huei Sheu[1] and Kuo-Hsiung Wang[1]

[1]Department of Business Administration, School of Management
Asia University, Wufeng, Taichung 41354, Taiwan

[2]Department of Decision Sciences
Western Washington University, Bellingham, WA98225, USA

(*Received September 2023; accepted July 2024*)

Abstract: This paper addresses the staffing problem for inpatient units treating diverse patient types in a hospital. We develop a queueing model to determine the optimal staffing levels within resource constraints. This quantitative approach offers a valuable analytical tool for hospital managers, enabling them to make more informed and efficient staffing decisions for inpatient units. To ensure the robustness of the results from our analytical model in real-world scenarios, we also employ a simulation model. Numerical examples are provided to illustrate the procedure and generate practical insights for healthcare practitioners.

Keywords: Arena simulations, Erlang-B formula, inpatient units staffing, optimal policy, queueing model.

1. Introduction

The importance of setting appropriate nurse staffing levels in acute hospital wards is widely recognized. However, effective quantitative methods for determining the optimal staffing policy remain scarce. As Hossny [6] points out, nurses are a critical resource in healthcare settings, accounting for 60% of the healthcare system. In these settings, nurses are expected to provide proficient, patient-centered, and cost-effective care.

Human resources in healthcare can be categorized into generalists and specialists. Nurses fall into the former category, while specialist doctors, such as cardiologists or urologists, belong to the latter. This classification implies that nurses provide services to a diverse range of patients, each requiring different types of care. These varied services entail different costs and times. Additionally, nursing service systems are characterized by the randomness of service times and the inter-arrival times of patients to inpatient units in a hospital. These characteristics suggest the use of stochastic models to analyze the nurse staffing problem.

In general, nurse resources are limited by budget constraints, yet it is essential to maintain a certain level of service for patients. Therefore, the nurse staffing problem can be formulated as an optimization problem aimed at minimizing costs while meeting service level

* Corresponding author
Email: Sue.Su@wwu.edu

constraints. In this paper, we present a continuous-time Markov chain (CTMC) model to address this problem. The CTMC model, also known as a stochastic knapsack with random arrivals, involves N resource units to which customers of K mutually independent types arrive. Arrivals of type i, $i = 1, \cdots, K$, are governed by a random process. If an arriving type i customer is admitted to the knapsack, they occupy n_i resource units for a random holding time. The knapsack follows a complete sharing policy for admitting different types of arrivals (see Ross [9]). The capacity constraint can be expressed as $\sum_{j=1}^{K} X_i(t)n_i \leq N$, where $X_i(t)$ is the number of type i customers in the system at time t. This stochastic knapsack model aligns well with the nurse staffing situation. Consequently, we develop an expected cost function for operating inpatient units in a hospital and determine the optimal nurse staffing level to minimize this cost function.

The paper is organized as follows. Section 2 reviews the related literature and positions our work. Section 3 presents a general model for the nurse staffing problem based on the stochastic knapsack. Then, the expected cost function is developed specifically for staffing nurses in inpatient units with two types of patients. In Section 4, some numerical examples are presented to illustrate the results obtained. To test the robustness of the results, we also build Arena simulation model to test more realistic non-exponentially distributed random variables. In Section 5. we discuss the approximation method to extend the model to treat the case with more than two types of patients. Finally, Section 6 concludes the paper with a summary.

2. Literature Review

Our work relates to two streams of studies. The first stream focuses on nursing workload and methodologies and tools for nurse staffing. The second stream encompasses stochastic models that address resource allocation in service systems with random factors. Many of these models are inspired by optimal control and design problems in computer networks. For instance, as noted in the introduction, the stochastic knapsack model, primarily developed for analyzing telecommunication systems, serves as a key reference. We provide a brief review of the literature in these two areas.

Griffiths et al. [5] conducted a systematic scoping review on nursing workload, nurse staffing methodologies, and tools. To determine the appropriate nurse staffing level, hospital administrators or schedulers need the following key information: (i) the major required nursing activities in inpatient units for each type of patient (see Hossny [6]); (ii) the time spent on each of these activities for each patient (see Abbey [1]); and (iii) the hiring cost of registered nurses. While the third piece of information is generally straightforward to obtain, the first two are often difficult to quantify, necessitating further research. In this paper, we assume that all three pieces of information are available. Given this information, the optimal nurse staffing level that minimizes the overall cost of operating the inpatient units can be determined, assuming the staffing level influences service capacity.

The determination of appropriate nurse staffing levels and workload measurement was first studied by Lewinski-Corwin [8]. Since then, numerous studies and reviews have fo-

cused on methods for determining nurse staffing levels, as summarized in Hossny [6]. These methods, developed over the years, can be categorized into several approaches: ”professional judgment,” ”benchmarking,” ”volume-based,” ”patient prototype,” ”multi-factorial indicator,” and ”time-task” approaches, all detailed in Hossny [6].

Regarding the quantitative nature of nurse staffing decisions, researchers have proposed various methods based on operational research, as surveyed by Saville et al. [11]. They identified 27 papers employing methodological approaches from operational research, including optimization (24/27 papers), simulation (6/27 papers), queuing theory (3/27 papers), and forecasting (1/27 paper). For more details on these approaches and additional references, readers are referred to Saville et al. [11].

Another related stream of research concerns resource allocation in service networks, primarily motivated by the management of telecommunication and computer networks. A notable example in this area is the work by Sarangan et al. [10], who examined a tele-traffic system based on the internet and World Wide Web (WWW), where arrivals are bursty, modeling it as a stochastic knapsack problem constrained by bandwidth. Although the primary focus of Sarangan et al.'s study is not on staffing level decisions in service systems, the general structure aligns well with the problem we address in this paper. This research field remains active due to the continuous evolution of the Internet and WWW. Chen and Ross [3] and Arlotto and Xie [2] are representative studies that reflect recent developments in this area. For a concise overview, we refer readers to these works and the references therein for the latest advancements.

Although these two streams of research are related to our work, there appears to be a gap between them. No existing study seems to effectively integrate these approaches. Our aim is to bridge this gap by proposing a stochastic knapsack model to address the nurse staffing issue for inpatient units in hospitals.

3. The Model and Cost Function

In this section, we first present some preliminaries for the model formulation and analysis and then develop a general stochastic nurse staffing model based on the stochastic knapsack.

3.1. Preliminaries

The main focus of analyzing a stochastic nurse-staffing model is to figure out the stationary distribution of the system. Such a analysis is based on the time reversibility of the stochastic process. We start with the definition of this property. At time t, let $X(t)$ be the stochastic process which represents the state of the system that is continuously observed, for all $t \in (-\infty, \infty)$. If stochastic process $\{X(t), -\infty < t < \infty\}$ is stochastically identical to the process $\{X(\tau - t), -\infty < t < \infty\}$ for all $\tau \in (-\infty, \infty)$, then $\{X(t), -\infty < t < \infty\}$ is a reversible process. The process $\{X(\tau - t), -\infty < t < \infty\}$ for any $\tau \in (-\infty, \infty)$ is known as the reversed process at τ. Although this is the definition of a reversed process, it is usually hard to show a CTMC is reversible based on that directly. Instead we resort to

one of the properties of reversible processes that are especially applied to CTMCs. The first step is to check and see if a CTMC is not reversible. In the rate diagram if there is an arc from node i to node j of the CTMC, then there must be an arc from j to i as well for the CTMC to be reversible. This is straightforward because only if you can go from j to i in the forward video, you can go from i to j in the reversed video. Note that this is necessary but not sufficient as we need to verify if the same probability law is followed by both the process and its reversed process (the requirement of being stochastically identical). For an ergodic CTMC that has reached the steady state, we study the probability structure of the reverse process. Tracing the process, denoted by $X(t)$, going backward in time, we first look at the time spent in each state. Given that the CTMC is in state i at some time t, the probability that the reverse process has been in this state for an amount of time greater than s is just $e^{-v_i s}$. This is because

$$\begin{aligned} P(X(\tau) = i, \tau \in [t-s, t] | X(t) = i) &= \frac{P(X(\tau) = i, \tau \in [t-s, t])}{P(X(t) = i)} \\ &= \frac{P(X(t-s) = i) e^{-v_i s}}{P(X(t) = i)} = e^{-v_i s}, \end{aligned}$$

where $P(X(t-s) = i) = P(X(t) = i)$ due to the steady state. Thus, going backward in time, the amount of time spent in state i follows the same exponential distribution as that in the original process. Next, we need to find the condition under which the jump-to probability distribution of the reverse process is the same as that in the original process. Again, we assume that the CTMC has reached the steady state and consider a sequence of state transition instants going backward in time. That is, starting at time instant n (nth transition instant), consider the sequence of states reached at these instants, denoted by $X_n, X_{n-1}, X_{n-2}, ...$ It turns out that this sequence of states is itself a Markov chain process with the transition probabilities, denoted by Q_{ij}. According to the definition, we have

$$\begin{aligned} Q_{ij} = P(X_m = j | X_{m+1} = i) &= \frac{P(X_m = j, X_{m+1} = i)}{P(X_{m+1} = i)} \\ &= \frac{P(X_m = j) P(X_{m+1} = i | X_m = j)}{P(X_{m+1} = i)} = \frac{p_j P_{ji}}{p_i}. \end{aligned}$$

To prove that the reversed process is indeed a Markov chain, we must verify that

$$P(X_m = j | X_{m+1} = i, X_{m+2}, X_{m+3}, ...) = P(X_m = j | X_{m+1} = i).$$

To confirm this property, we can use the fact that the Markov property implies the conditional independence between the past and future given the current state and the independence is a symmetric relationship. Thus, to ensure that the reverse process has the same probability structure as the original process, we need $Q_{ij} = P_{ij}$ which results in the condition for the reversible Markov process. That is $p_j P_{ji} = p_i P_{ij}$. Such a relation can simplify solving the balance equations for stationary distributions. The next two properties will enhance the power of using the time reversibility to find stationary distributions of more complex

CTMCs. We present them briefly and use them to find the stationary distribution of our model. For more details on these two properties, we refer to Kelly [7].

Property 1: Joint processes of independent reversible processes are reversible.

Suppose that we have n independent reversible processes, denoted by $\{X_1(t), -\infty < t < \infty\}$,$\{X_2(t), -\infty < t < \infty\}$,...,$\{X_n(t), -\infty < t < \infty\}$. If we are interested in the joint process $\{X_1(t), X_2(t), ..., X_n(t), -\infty < t < \infty\}$, then it is also reversible and its steady-state probabilities would just be the product of those of the corresponding states of the individual reversible processes. As an example, if each process is a one-dimensional BDP, then the joint process is an n-dimensional BDP which is also reversible. You can verify the reversible condition in terms of the stationary probabilities and transition rates. A truncated process is the process of a finite-state space that results from cutting off part of the infinite state space.

Property 2: Truncated processes of reversible processes are reversible.

Consider a reversible and ergodic CTMC $\{X(t), -\infty < t < \infty\}$ with infinitesimal generator $Q = [q_{ij}]$ defined on state space S and steady-state probabilities p_j that the CTMC is in state j for all $j \in S$. Now consider another CTMC $\{Y(t), -\infty < t < \infty\}$ which is a truncated version of $\{X(t), -\infty < t < \infty\}$ defined on state space A such that $A \in S$. By truncation, we can keep the inter-state transition rates of the truncated process $Y(t)$ the same as those in the original process $X(t)$ and adjust the diagonal elements of the infinitesimal generator of $Y(t)$ by letting its negative value equal the sum of the off-diagonal transition rates in the row. Next we present a stochastic nurse-staffing model.

3.2. A Nurse-staffing model

Consider an inpatient unit that can admit N types of patients. A type i patient, arriving at the inpatient unit according to an independent Poisson process with rate λ_i, requires $c_i > 0$ unit of nurse resource per time unit with $i = 1, 2, ...N$. For example, if $c_1 = 0.7$ and $c_2 = 2$, it means that each type 1 patient needs 0.7 nurse per hour and each type 2 patient needs 2 nurses per hour if time unit is one hour. Let C be the total number of nurses staffed for the planning horizon (e.g. one shift). Let $X_i(t)$ be the number of type i patients in the unit at time t. The staffing capacity constraint can be written as

$$c_1X_1(t) + c_2X_2(t) + \cdots + c_NX_N(t) \leq C.$$

This constraint makes the state space finite as it controls the admission of patients of N types. That is whenever the maximum capacity C is fully utilized, new arrivals are denied. For example, if a type i patient arrives and $c_i + c_1X_1(t) + c_2X_2(t) + \cdots + c_NX_N(t) > C$ holds at that instant (i.e., the maximum capacity is exceeded), this patient is rejected. The type i patient's stay time in the inpatient unit is assumed to be exponentially distributed with rate μ_i. During the entire stay time, each class i patient uses c_i unit of nurse.

Due to the finite state space, the multi-dimensional CTMC, $\{X_1(t), X_2(t), ..., X_N(t)\}$, has the stationary distribution denoted by

$$p_{x_1,x_2,...,x_N} = \lim_{t\to\infty} P(X_1(t) = x_1, X_2(t) = x_2, ..., X_N(t) = x_N).$$

To obtain this stationary distribution, we consider an N-dimensional CTMC with the same state variables and $C = \infty$. Such a CTMC has the state space $S = \mathbf{Z}_+^N$. Obviously, the CTMC of our model is a truncated process of this unconstrained CTMC due to the constraint. Note that $C = \infty$ implies that the system can be considered as N independent $M/M/\infty$ queues since the arrival processes are independent. Each process is reversible and the joint process is also reversible according to Property 1 presented above. Moreover, the stationary probability of the number of type i patients in the system, denoted by $p_i^\infty(j), j = 0, 1, ...,$ has a closed-form expression (i.e. Poisson process with parameter λ_i/μ_i) as follows:

$$p_i^\infty(j) = e^{-\lambda_i/\mu_i} \left(\frac{\lambda_i}{\mu_i}\right)^j \frac{1}{j!},$$

where the superscript ∞ indicates $C = \infty$ (omitting this superscript implies $C < \infty$, the truncated or constrained case). Clearly, the stationary distribution for the joint process is given by

$$\begin{aligned}&\lim_{t\to\infty} P(X_1^\infty(t) = x_1, X_2^\infty(t) = x_2, ..., X_N^\infty(t) = x_N) \\ &\quad = p_1^\infty(x_1)p_2^\infty(x_2)\cdots p_N^\infty(x_N) = \left(e^{-\sum_{i=1}^N \lambda_i/\mu_i}\right) \prod_{i=1}^N \left(\frac{\lambda_i}{\mu_i}\right)^{x_i} \frac{1}{x_i!}.\end{aligned}$$

It follows from Property 2 presented above that the stationary probability for our system subject to the capacity constraint, denoted by $p_{x_1,x_2,...,x_N}$, can be written as

$$p_{x_1,x_2,...,x_N} = M \prod_{i=1}^N \left(\frac{\lambda_i}{\mu_i}\right)^{x_i} \frac{1}{x_i!} \tag{1}$$

subject to $c_1x_1 + c_2x_2 + \cdots + c_Nx_N \le C$. Here M is a constant and can be determined by the normalization condition as

$$M = \left[\sum_{x_1,x_2,...,x_N : c_1x_1+c_2x_2+\cdots+c_Nx_N \le C} \prod_{i=1}^N \left(\frac{\lambda_i}{\mu_i}\right)^{x_i} \frac{1}{x_i!}\right]^{-1}.$$

Although we have obtained the expression for M, it is computationally expensive when the number of patient types N is getting larger. We outline one approach based on the enumeration of all feasible states via a tree diagram as shown in Figure 1 for an inpatient unit with three types of patients. With such a tree, we can figure out all feasible states under the constraint of C nurses. There are $\lfloor C/c_1 \rfloor$+1 trees with 2 stages from the largest tree of $k_1 = 0$ (having most branches) to the smallest tree of $k_1 = \lfloor C/c_1 \rfloor$ (having only one branch which is the case of $x_1 = \lfloor C/c_1 \rfloor, x_2 = x_3 = 0$. This is because as x_1 increases, the number of feasible values for x_2 decreases (i.e., the number of branches in stage 1 reduces); similarly, as x_1 or x_2 increase, the number of feasible values for x_3 decreases (i.e., the number of branches in stage 2 reduces). Then, we can use the fact that these feasible state probabilities sum up

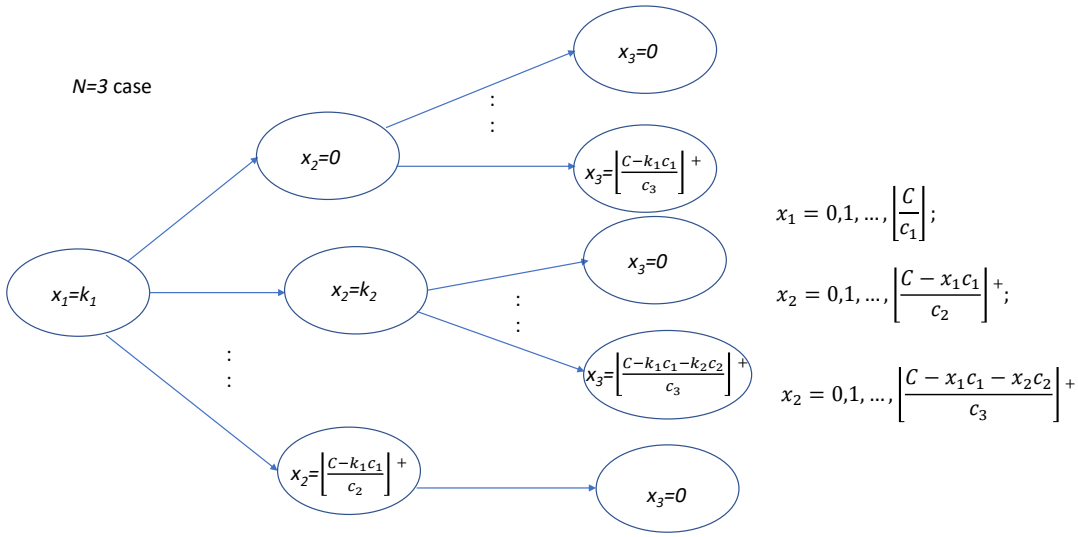


Figure 1. Tree Diagram.

to 1 to determine the parameter M (the normalization condition). Such an approach can be extended to a general N type customer case with the feasible x_i, for $i = 1, 2, ...N$ given by

$$
\begin{aligned}
&x_1 = 0, 1, 2, \cdots, \lfloor \frac{C}{c_1} \rfloor; \\
&x_2 = 0, 1, 2, \cdots, \lfloor \frac{C - x_1 c_1}{c_2} \rfloor^+; \\
&x_3 = 0, 1, 2, \cdots, \lfloor \frac{C - x_1 c_1 - x_2 c_2}{c_3} \rfloor^+; \\
&\quad \vdots \\
&x_j = 0, 1, 2, \cdots, \lfloor \frac{C - \sum_{i=1}^{j-1} x_i c_i}{c_j} \rfloor^+; \quad \text{for } \; j = 4, 5, \cdots, N.
\end{aligned}
$$

Denote by $\bar{x}_j(x_1, ..., x_{j-1}) = \lfloor \frac{C - \sum_{i=1}^{j-1} x_i c_i}{c_j} \rfloor^+$ the upper bound of x_j value, which depends on the set of x_i where $i = 1, ..., j-1$. Then, we can compute M as follows:

$$
M = \left[\sum_{x_1=0}^{\lfloor \frac{C}{c_1} \rfloor} \sum_{x_2=0}^{\bar{x}_2(x_1)} \cdots \sum_{x_N=0}^{\bar{x}_N(x_1,\ldots,x_{N-1})} \prod_{i=1}^{N} \left(\frac{\lambda_i}{\mu_i} \right)^{x_i} \frac{1}{x_i!} \right]^{-1}.
$$

Remark: (1) The computational complexity becomes higher when the number of patient types is getting larger. Fortunately, in a practical inpatient unit, the number of patient types is low. This makes our approach applicable in real situations. (2) Since we consider the system as a queueing system with multiple servers without a waiting buffer (i.e., $M/M/s/s$), the

model presented can be called an Erlang-B based. If we keep a wait list for patients instead of rejecting patients when all beds are occupied, we may consider the $M/M/s/K$ with $K > s$ based model (i.e., we can keep a waitlist up to $K - s$ patients). Such a model, called "inpatient unit with waitlist", can be analyzed similarly.

With the stationary distribution, we can develop the expected cost function of the nurse-staffing level and determine the optimal staffing level that minimizes the cost function. The expected cost function may consist of the "expected cost of rejecting patients" plus the "nurse hiring cost" for inpatient units without waitlist or the "expected waiting cost" plus the "nurse hiring cost" for inpatient units with waitlist. In the next section, we will demonstrate the determination of the optimal staffing level for an inpatient unit serving two types of patients.

It is worth noting that this model is general enough to analyze a variety of stochastic service systems besides the nurse-staffing problem of our interest. For example, it can be utilized to analyze the multi-media traffic problem in the internet.

4. Numerical Illustrations

For numerical illustrations, we consider an inpatient unit that serves two types of patients without waitlist. Inpatient units in a hospital includes intensive care patients, surgery patients, and rehab patients. Consider a cardiology/General Internal Medicine Inpatient Unit that admits two types of patients, cardiology acute care (CAC) called type 1, and general internal medicine (GIM) called type 2. A key constraint for admitting a patient is the number of acute care nurse practitioners (denoted by RN - registered nurses with advanced training) available. Let C be the total number of RNs available. Each CAC patient admitted requires c_1 RNs and each GIM patient admitted requires c_2 RNs. Assume that (a) patients of type n arrive at the unit according to an independent Poisson process with rate λ_n; and (b) each admitted patient type i spends an exponentially distributed time with rate μ_n for $n = 1, 2$. Let $X_1(t)$ and $X_2(t)$ be the number of type-1 and type-2 patients at time t. Then the two-dimensional CTMC $\{(X_1(t), X_2(t)), t = 0\}$ will reach steady-state.

We now write down the stationary probabilities and related performance measures of this model based on the structure of the solution for the general model. Due to the linear capacity constraint $c_1X_1(t) + c_2X_2(t) \leq C$, the state space is finite and the boundary states can be determined in two directions. These two directions will determine two types of boundary states, called X_1 boundary and X_2 boundary states. For a given $x_2 = j$ with $j = 1, ..., \lfloor \frac{C}{c_2} \rfloor$, the feasible state is (x_1, j) where $x_1 = 0, 1, ..., \lfloor \frac{C-jc_2}{c_1} \rfloor^+$. Here $\lfloor x \rfloor$ is the lower floor function that gives the greatest integer less than or equal to x and $\lfloor x \rfloor^+ = max(\lfloor x \rfloor, 0)$. Thus, X_1 boundary state is $(\lfloor \frac{C-jc_2}{c_1} \rfloor^+, j)$ where $j = 0, 1, ..., \lfloor \frac{C}{c_2} \rfloor$. This boundary state set is denoted by $S^B_{X_1}$. Similarly, For a given $x_1 = i$, the feasible state is (i, x_2) where $x_2 = 0, 1, ..., \lfloor \frac{C-ic_1}{c_2} \rfloor^+$. Thus, X_2 boundary state is $(i, \lfloor \frac{C-ic_1}{c_2} \rfloor)^+$ where $i = 0, 1, ..., \lfloor \frac{C}{c_1} \rfloor$. This boundary state set is denoted by $S^B_{X_2}$. Then the double boundary state set is $S^B_{X_1,X_2} = \{(X_1(t) = i, X_2(t) = j : (i, j) \in S^B_{X_1} \cap S^B_{X_2}\}$; X_1 only boundary state set is $\{(X_1(t) = i, X_2(t) = j : (i, j) \in S^B_{X_1,X_2} \; / \; S^B_{X_2}\}$; and X_2 only boundary state set is $\{(X_1(t) = i, X_2(t) = j : (i, j) \in S^B_{X_1,X_2} \; / S^B_{X_1}\}$.

According to the solution of the general model, we have the stationary probabilities as

$$p_{ij} = \lim_{t\to\infty} P(X_1(t) = i, X_2(t) = j) = M \left(\frac{\lambda_1}{\mu_1}\right)^i \left(\frac{\lambda_2}{\mu_2}\right)^j \frac{1}{i!j!},$$
$$i = 0, 1, ..., \lfloor \frac{C}{c_1} \rfloor; j = 0, 1, ..., \lfloor \frac{C - ic_1}{c_2} \rfloor^+,$$

where

$$M = \left[\sum_{i=0}^{\lfloor \frac{C}{c_1} \rfloor} \sum_{j=0}^{\lfloor \frac{C-ic_1}{c_2} \rfloor^+} p_{ij} \right]^{-1}.$$

The performance measures, such as the expected number of patients in the system and the blocking probability of each type, can be obtained based on the stationary probabilities. For example, under a certain cost structure, we can construct a cost function for the system. Assume that there are two types of costs that are critical from the customer service and operating cost perspectives. The first type is the cost of rejecting a patient of either type due to the system blocking (i.e. no room for admitting a patient). Let L_n be the cost of rejecting a type n patient, $n = 1, 2.$ and let h be the cost of hiring one RN per time unit. Then the total expected cost per time unit, denoted by g, is given by

$$g = \lambda_1 L_1 \sum_{j=0}^{\lfloor \frac{C}{c_2} \rfloor} p_{(\lfloor \frac{C-jc_2}{c_1} \rfloor, j)} + \lambda_2 L_2 \sum_{i=0}^{\lfloor \frac{C}{c_1} \rfloor} p_{(i, \lfloor \frac{C-ic_1}{c_2} \rfloor)} + hC.$$

Clearly the first two terms are decreasing in C and the last term is increasing in C.

Numerically, we can determine the optimal C that minimizes the expected cost rate g. As a numerical example, consider a system with $c_1 = 2, c_2 = 1, \lambda_1 = 2, \lambda_2 = 3, \mu_1 = 0.5,$ and $\mu_2 = 1$ and varying C values. The normalization constant is

$$M = \left[\sum_{i=0}^{C/2} \sum_{j=0}^{C-2i} \left(\frac{\lambda_1}{\mu_1}\right)^i \left(\frac{\lambda_2}{\mu_2}\right)^j \frac{1}{i!j!} \right]^{-1}.$$

Figure 2 shows g as a function of C for $h = 16, L_1 = 100$ and $L_2 = 20$. The optimal number of RNs in this example is 8. Based on this performance measure and Erlang-B formulas, we can also evaluate the benefit of completely sharing admission policy compared with the two dedicated systems for the two types of patients.

Robustness test via Arena Simulation

To examine the robustness of the results, we consider the same model with more general distributed service times (or the length of stay - LOS). All other parameters such as arrival rates, service rates, and the number of nurses required per hour for each patient are the same as those above. An Arena model is developed with the system configuration as as shown in Figure 3. However, instead of exponential distributed LOS (or service times), we consider

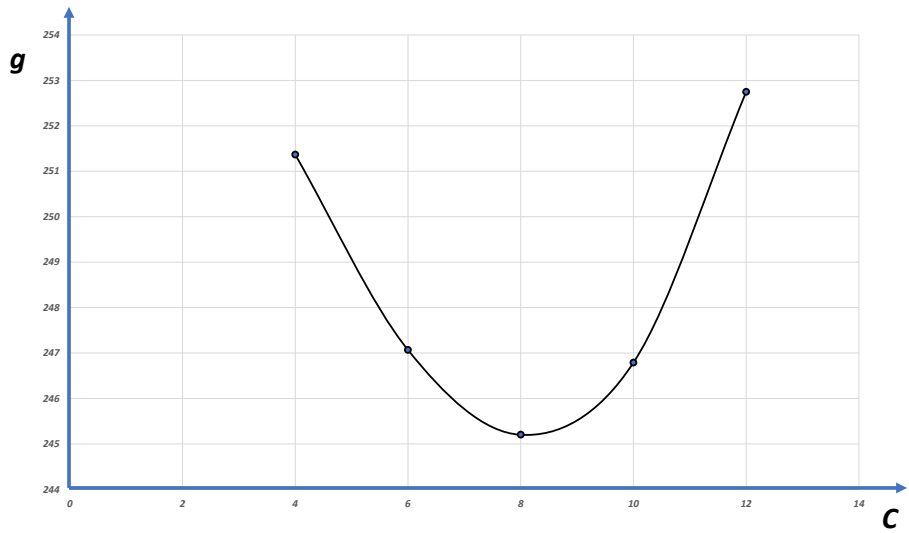


Figure 2. Expected total cost per unit time v.s. total number of RNs available.

the more flexible distributions for the LOS that have been confirmed in several empirical studies (see Dehouche et al. [4]). The first LOS distribution is Gamma distribution with shape parameter, denoted by α, and scale parameter, denoted by β. Then the mean is $\alpha\beta$ and the variance is $\alpha\beta^2$. For each type of patients, we assume that the mean service time remains the same as that in the example of Chapter 3. Namely, the mean service time for type 1 (2) patients is 2 (1). That is for LOS of type 1, $\alpha = 1, \beta = 2$ and for LOS for type 2, $\alpha = 0.5, \beta = 2$.

We run the simulation model for 10 replications with 8 hours as the length of each replication. Figure 4 shows the cost behavior of such a system based on the simulation results. We have observed a similar total expected cost function in this Gamma distributed LOS case as that in the exponentially distributed LOS case presented in Figure 2.

Next, we test the heavy-tailed LOS case. It is well-known that typical heavy-tailed distributions include Lognormal, Pareto, Cauchy, and Weibull distributions. These heavy-tailed distributions have specific characteristics and applications. While Parato distribution and Lognormal distribution may be applicable in modeling financial data with heavy tails, Weibull distributions are commonly utilized in modeling patient LOS in hospital. Therefore, we examine the nursing staffing problem with Weibull distributed LOS. Again, we keep all other parameters the same as before except for the LOS. We first give a brief introduction to Weibul distribution for modeling hospital LOS.

Weibull Distribution Overview: The Weibull distribution is characterized by two parameters: - Shape parameter (k): Determines the shape of the distribution. - Scale parameter (λ): Stretches or compresses the distribution along the horizontal axis.

The probability density function (PDF) for the Weibull distribution is given by:

$$f(x; k, \lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$$

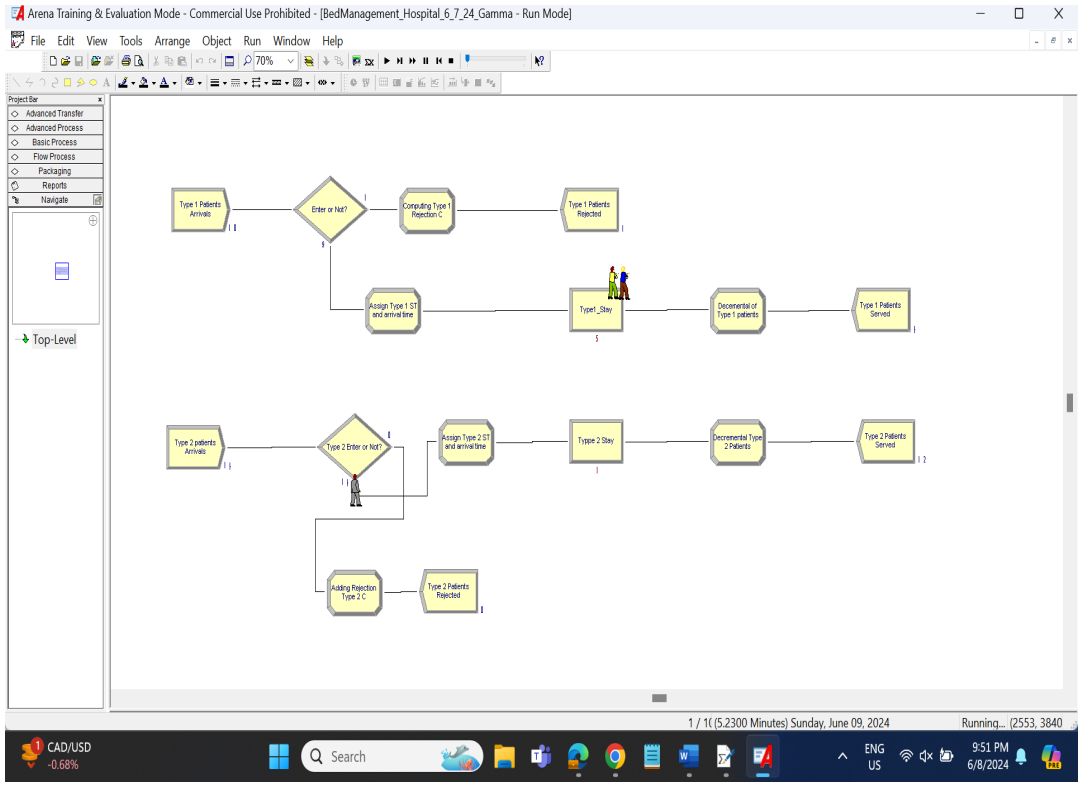


Figure 3. Arena Simulation Model for the Nursing Staffing Problem with Gamma distributed LOS

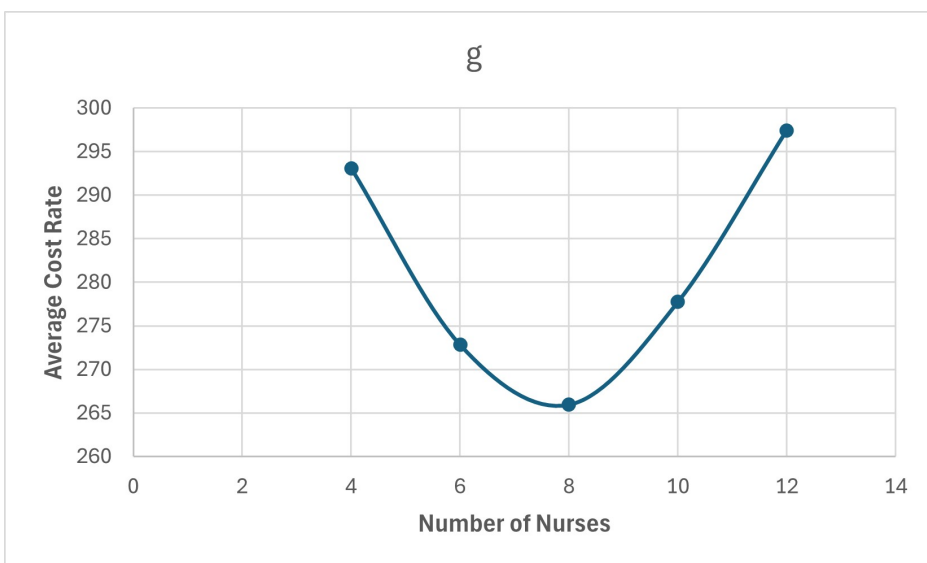


Figure 4. Cost function for the Nursing Staffing Problem with Gamma distributed LOS

Parameter Ranges for Heavy-Tailed Weibull: To model a heavy-tailed length of stay in a hospital using a Weibull distribution, we focus on the shape parameter k being less than 1. This makes the distribution heavy-tailed, implying a relatively high probability of long stays.

In our nursing staffing problem, to keep the mean service times the same as those used in the previous examples, we choose the following parameters:

For type 1 patients: - Shape parameter (k): $0.5 \leq k < 1$. Let $k = 0.7$ - Scale parameter (λ): λ. Let $\lambda = 2$

For type 2 partients: - Shape parameter (k): $0.5 \leq k < 1$. Let $k = 0.7$ - Scale parameter (λ): λ. Let $\lambda = 1$

Using the Weibull distributed LOS, we can model the LOS that often shows a heavy-tailed pattern, where a significant portion of patients have extended stays due to complications or recovery times. Knowing the tail behavior helps hospitals plan resources and manage risks associated with prolonged stays. The exact values for k and λ should be estimated based on historical data using techniques such as maximum likelihood estimation (MLE). These parameter values will depend on the specific hospital or patient population. Larger λ values might be used in hospitals dealing with more severe cases (e.g. longer LOS cases)

This example provides a practical framework for modeling hospital length of stay with a heavy-tailed Weibull distribution. Adjusting the shape and scale parameters allows flexibility to fit the distribution to observed data. As illustrated in Figure 5, compared with the exponential and Gamma distributed LOS cases, this heavy-tailed LOS case has a higher average cost rate due to higher rejection cost. However, the optimal staffing level is reduced from 8 to 6. This is an interesting and counter-intuitive result.

Overall, these simulation results confirm the general cost behavior in nurse staffing problem discovered in our queueing (quantitative) model.

5. "Grouping 2" Method

As illustrated in the numerical section, the problem with two types of patients is easier to compute. This observation motivates us to propose an approximation method to solve a large scale nursing-staffing problem with more than two types of patients. The main idea is to divide N-classes of patients into $N/2$ groups of two types. For each group of two types, we can utilize the procedure developed to determine the optimal staffing level C by minimizing its operating cost. To make it simple, we assume that N is even (for the odd N, just figure out the number of groupings of 2 for $N+1$ and know that there is one group with only one type of patients). Denote by C_n^N the number of combinations if taking n objects out of N distinguished objects. Then, there will be

$$k = \frac{\binom{N}{2}\binom{N-2}{2}\binom{N-4}{2}\cdots\binom{2}{2}}{\left(\frac{N}{2}\right)!} \tag{2}$$

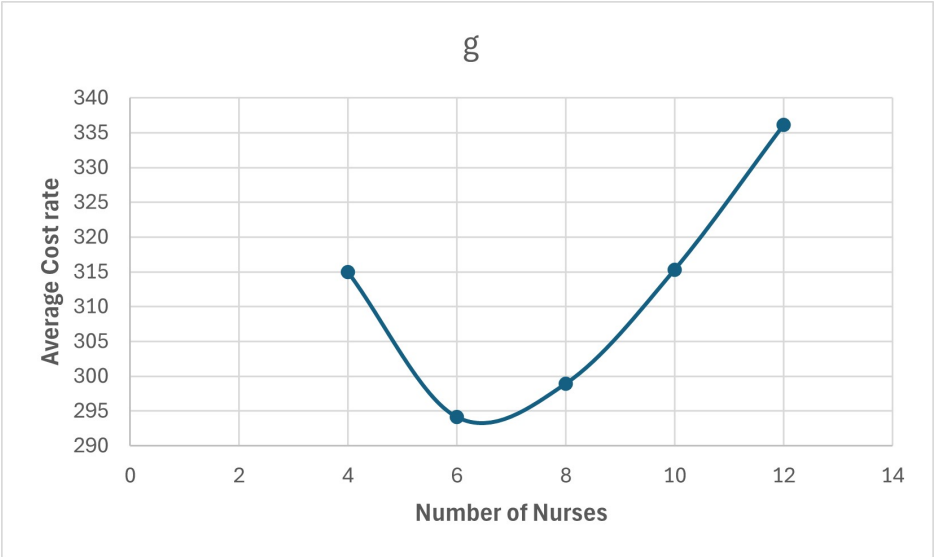


Figure 5. Cost function for the Nursing Staffing Problem with Weibull distributed LOS with heavy-tail

different possible groupings of two. For each possible groupings of two, indexed by j with $j = 1, ..., k$, we can figure out the optimal staffing level C_i^j and its associate cost g_i^j, where $i = 1, 2, ..., N/2$ for each group i. Then, compute the total staffing level $C^j = \sum_{i=1}^{N/2} C_i^j$ and total cost $g^j = \sum_i^{N/2} g_i^j$. The optimal staffing level for cost minimization would be C^{j^*} where $j^* = \arg\min_{j \in 1,2,...,k} g^j$. Here we present an example of $N = 4$. With the hiring cost $h = 16$, the other parameters of the system are summarized in the table below.

Table 1. Parameters for the system with four types of patients with $i = A, B, C, D$.

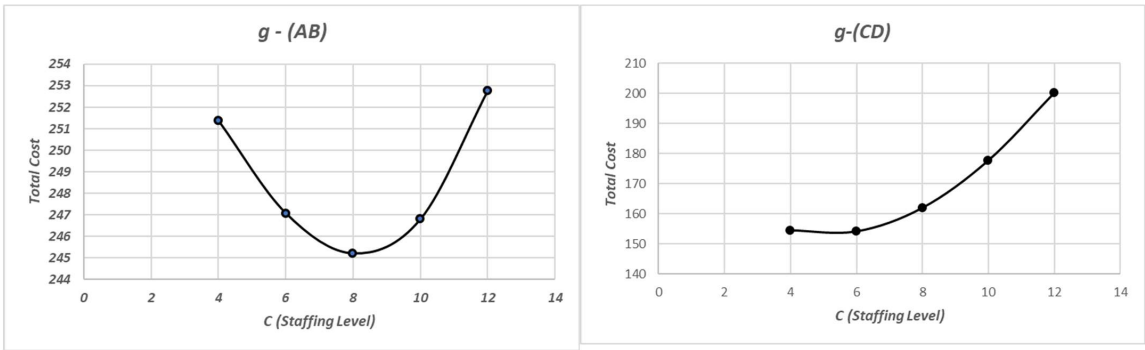
Parameters for type i patients

Patient Type	A	B	C	D
c_i	2	1	2	1
λ_i	2	3	2	1
μ_i	0.5	1	0.8	1.2
L_i	100	20	60	60

Based on (2), the number of possible grouping 2 would be $k = (\binom{4}{2}\binom{2}{2})/(4/2)! = 3$. The optimal staffing levels and their associated costs are given in Table 2.

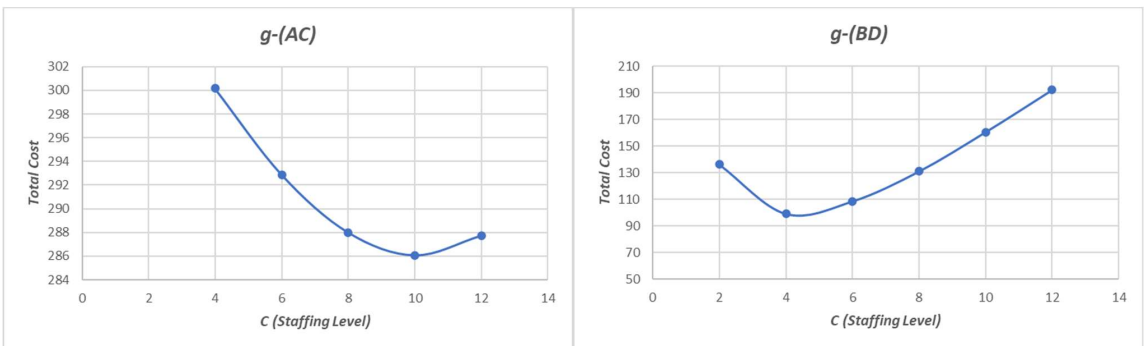
Table 2. Optimal solutions for three grouping scenarios.

j	Groups	C_1^*	g_1^*	C_2^*	g_2^*	$C^* = C_1^* + C_2^*$	$g^* = g_1^* + g_2^*$
1	$(AB)(CD)$	8	245.20	6	154.13	14	399.33
2	$(AC)(BD)$	10	286.05	4	99.32	14	385.37
3	$(AD)(BC)$	8	217.54	4	177.74	12	395.28



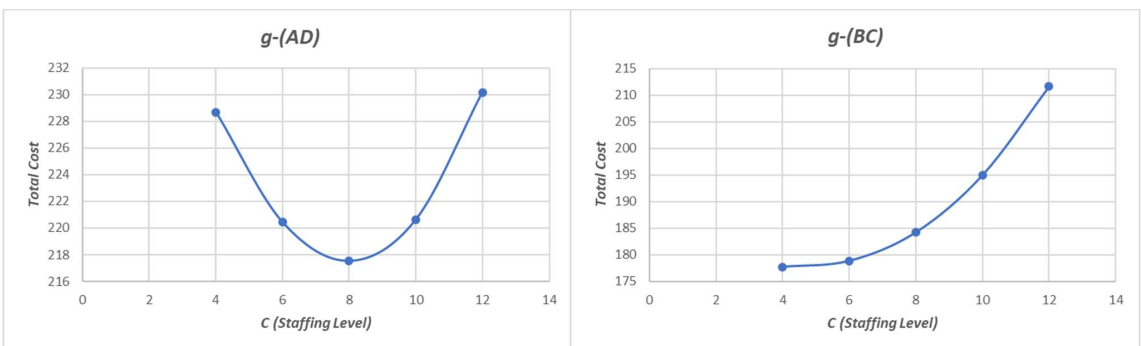
Total cost of grouping 2 as a function of staffing level for (*AB*) and (*CD*) scenario.

Figure 6. Optimal staffing levels and their associate total costs for (AB) and (CD), respectively.



Total cost of grouping 2 as a function of staffing level for (*AC*) and (*BD*) scenario.

Figure 7. Optimal staffing levels and their associate total costs for (AC) and (BD), respectively.



Total cost of grouping 2 as a function of staffing level for (*AC*) and (*BD*) scenario.

Figure 8. Optimal staffing levels and their associate total costs for (AD) and (BC), respectively.

Among the three possible grouping 2 scenarios, the optimal scenario is $(AC), (BD)$ with a minimum cost of 385.37 dollars and staffing level at 14. The average total cost rates for these combination scenarios are shown in Figures 6, 7, and 8. To get an idea of how much cost reduction the grouping 2 approach can result, we compute the total cost of a non-pooling system with the proportional allocation of a fixed service capacity (e.g., 14 nurses) based on $L_i\alpha_i$ rule, where L_i is the cost of losing a type i customer and $\alpha_i = \lambda_i/\mu_i$, the loading ratio. Assume that the fixed total service capacity is C. Under the $L_i\alpha_i$ rule, the staffing level dedicated for type i customers, denoted by C_i, where $i = 1, 2, ..., N$, is determined by

$$\begin{aligned} \frac{C_1}{L_1\alpha_1} = \frac{C_2}{L_2\alpha_2} = \cdots = \frac{C_N}{L_N\alpha_N}, \\ C_1 + C_2 + \cdots + C_N = C. \end{aligned} \tag{3}$$

In our numerical example, we have $N = 4$ (four types of patients). Solving (3) with $L_A\alpha_A = 400, L_B\alpha_B = 60, L_C\alpha_C = 150$ and $L_D\alpha_D = 50$ as

$$\begin{aligned} \frac{C_A}{400} = \frac{C_B}{60} = \frac{C_C}{150} = \frac{C_D}{50}, \\ C_A + C_B + C_C + C_D = 14, \end{aligned}$$

we obtain the solution $C_A = 8.48, C_B = 1.27, C_C = 3.18$ and $C_D = 1.06$,, which suggests an integer solution of $C_A = 9, C_B = 1, C_C = 3$, and $C_D = 1$ with the total of 14. For each dedicated system, the probability of losing a customer is the Erlang B formula for the blocking probability given by

$$P(\text{losing a type } i \text{ patient}) = \frac{\frac{\alpha_i^{\lfloor C_i/c_i \rfloor}}{\lfloor C_i/c_i \rfloor!}}{1 + \alpha + \frac{\alpha^2}{2!} + \cdots + \frac{\alpha^{\lfloor C_i/c_i \rfloor}}{\lfloor C_i/c_i \rfloor!}},$$

where $i = A, B, C, D$. Then the cost of losing patients for system i would be

$$\lambda_i L_i P(\text{losing a type } i \text{ customer}), \quad \text{where } \ i = A, B, C, D.$$

It is easy to find that the total cost for this setting (cost of losing patients plus nurse hiring cost) is given by

$$g_{ABCD} = \sum_{i=A,B,C,D} \lambda_i L_i P(\text{losing a type } i \text{ customer}) + h \sum_{i=A,B,C,D} C_i. \tag{4}$$

Substituting the numerical values for the parameters in (4) yields $g = 476.35$ dollars, which is about 24 percent more than the optimal grouping 2 scenario ($(AC)(BD)$) cost of 385.37 dollars at the same total staffing level of 14. Note that other possible integer solutions can be also evaluated. For example, rounding down all non-integer capacities gives an integer solution of $C_A = 8, C_B = 1, C_C = 3$ and $C_D = 1$, which has a total of 13 nurses hired. In contrast, rounding up all non-integer capacities gives another integer solution of $C_A = 9, C_B = 2, C_C = 4$ and $C_D = 2$, which has a total of 17 nurses hired. Evaluating the total costs of these solutions shows the advantages of polling customer service.

6. Conclusions

In this paper, we have addressed the issue of determining the optimal nurse staffing level for inpatient units in hospitals. Considering the stochastic nature of arrival and service processes, we modeled the service system as a stochastic knapsack problem. By leveraging the time reversibility of a continuous-time Markov chain (CTMC), we obtained the stationary distribution of the number of patients in the system. This distribution allows us to compute various performance measures for the system. Under a defined cost structure, we determine the optimal nurse staffing level that minimizes the long-term average cost. To test the robustness of the model, we developed Arena simulation models to evaluate more realistic lengths of stay (LOS) for healthcare facilities, finding results consistent with those predicted by the analytical CTMC model.

A key feature of our model is its ability to handle staffing problems involving multiple patient types. However, the computational complexity can become an issue with a large number of patient types. Fortunately, in practical inpatient units, the number of patient types is typically small, making our model applicable in real decision-making situations.

A potential future direction for this research is to extend the model to incorporate scenarios with patient waitlists. Another extension is to consider the time-varying arrivals to inpatient units, which leads to a special class of time-varying queues. For the past research on time-varying queues, we refer to Whitt [12].

Acknowledgements

The authors are very grateful to the editor and anonymous referees for the careful reading of the paper and for their comments and constructive suggestions which helped us to improve significantly this paper.

References

[1] Abbey, M., Chaboyer, W., & Mitchell, M. (2012). Understanding the work of intensive care nurses: A time and motion study. *Australian Critical Care*, 25(1), 13–22.

[2] Arlotto, A., & Xie, X. (2020). Logarithmic Regret in the Dynamic and Stochastic Knapsack Problem with Equal Rewards. *Stochastic Systems*, 10(2), 170–-191.

[3] Chen, K., & Ross, S. M. (2014). An adaptive stochastic knapsack problem. *European Journal of Operational Research*, 239(3), 625–635.

[4] Dehouche, N., Viravan, S., Santawat, U., Torsuwan, N., Taijan, S., & Intharakosum A, et al. (2023). Hospital length of stay: A cross-specialty analysis and Beta-geometric model. *PLoS ONE*, 18(7), e0288239. https://doi.org/10.1371/journal.pone.0288239

[5] Griffiths, P., Saville, C., Ball, J., Jones, J., Pattison, N., & Monks, T. (2020). Nursing workload, nurse staffing methodologies and tools: A systematic scoping review and discussion. *International Journal of Nursing Studies*, 103, 103487.

[6] Hossny, E. K. (2022). Studying nursing activities in inpatient units: a road to sustainability for hospitals. *BMC Nursing*, 21(148), 1–11.

[7] Kelly, F. P. (1979). Reversibility and Stochastic Networks. Chichester, UK, Wiley.

[8] Lewinski-Corwin, E. H. (1922). The hospital nursing situation. *American Journal of Nursing*, 22(8), 603–-606.

[9] Ross, K. W. (1995). Multiservice Loss Models for Broadband Telecommunication Networks. New York, Springer-Verlag.

[10] Sarangan, V., Ghosh, D., Gautam, N., & AcharyaSteady, R. (2005). State Distribution for Stochastic Knapsack with Bursty Arrivals. *IEEE Communication Letters*, 9(2), 187–189.

[11] Saville, C. E., Griffiths, P., Ball, J. E., & Monks, T. (2019). How many nurses do we need? A review and discussion of operational research techniques applied to nurse staffing. *International Journal of Nursing Studies*, 97, 7–-13.

[12] Whitt, W. (2018). Time-Varying queues. *Queueing Models and Service Management*, 1(2), 79–164.