DECISION SCIENCES
A JOURNAL OF THE DECISION SCIENCES INSTITUTE

Decision Sciences
Volume 00 Number 0
xxxx 2016

A Queueing Theoretic Approach to Set Staffing Levels in Time-Dependent Dual-Class Service Systems*

J. L. Vile†
School of Mathematics, Cardiff University, Cardiff, UK and NHS Delivery Unit, Bridgend, UK, e-mail: julievile1@gmail.com

J. W. Gillard, P. R. Harper, and V. A. Knight
School of Mathematics, Cardiff University, Cardiff, UK, e-mail: GillardJW@cf.ac.uk, Harper@cf.ac.uk, KnightVA@cf.ac.uk

ABSTRACT

This article addresses the optimal staffing problem for a nonpreemptive priority queue with two customer classes and a time-dependent arrival rate. The problem is related to several important service settings such as call centers and emergency departments where the customers are grouped into two classes of "high priority" and "low priority," and the services are typically evaluated according to the proportion of customers who are responded to within targeted response times. To date, only approximation methods have been explored to generate staffing requirements for time-dependent dual-class services, but we propose a tractable numerical approach to evaluate system behavior and generate safe minimum staffing levels using mixed discrete-continuous time Markov chains (MDCTMCs). Our approach is delicate in that it accounts for the behavior of the system under a number of different rules that may be imposed on staff if they are busy when due to leave and involves explicitly calculating delay distributions for two customer classes. Ultimately, we embed our methodology in a proposed extension of the Euler method, coined Euler Pri, that can cope with two customer classes, and use it to recommend staffing levels for the Welsh Ambulance Service Trust (WAST). [Submitted: January 29, 2015. Revised: May 10, 2016. Accepted: May 16, 2016.]

Subject Areas: Exhaustive Service Discipline, Markov Chain, Numerical Methods, Priority Queues, Staffing, Time-Dependent Queues, and Waiting Time Distribution.

*This research was supported by GS501100000266 EPSRC http://dx.doi.org/10.13039/501100000266 grant GS501100000266 EP/F033338/1 as part of the LANCS initiative. This research was also funded by EPSRC grant EP/F033338/1 (part of the LANCS initiative) and the data underpinning the project was provided by the Welsh Ambulance Services NHS Trust (WAST). The authors would particularly like to thank WAST Clinical R&D Manager Richard Whitfield for his useful feedback at various stages of this research and the anonymous referees for their helpful comments and insights on this paper

†Corresponding author.

INTRODUCTION

With high public expectations and ever increasing competition levels in today's society, service systems are facing escalating pressures to uphold minimum service quality standards in response to time-varying demands from customers requiring assistance with varying levels of urgency. For example, targets specifying that certain proportions of customers must be served within predefined time limits are imposed upon ambulance services, health clinics, call centers, and roadside assistance organizations, to name just a few. All too often, managers employ staff on cyclical bases and only increase staffing levels when systems begin to escalate into critical situations; often too late to avoid surges in response times (Mohan, Alam, Fowler, Gopalakrishnan, & Printezis, 2014).

Whitt (2008) reviewed several techniques that have been developed to cope with time-varying demand for services that serve a homogeneous set of customers; but the application of such techniques remains a fruitful area of research for time-dependent systems that additionally prioritize customers, because many of the established methods become unmanageable with the extra level of complexity. Thus, despite the widespread prevalence of such queueing models in society, most previous research concerned with evaluating system behavior has focussed upon the development of approximation schemes. We show that despite the intricacies involved in multiserver time-dependent systems, it is possible to extend the methodology presented by Ingolfsson et al. (2007) to track performance in systems, which serve a homogeneous set of customers, to priority service systems. As such, this article develops a tractable numerical approach to evaluate the behavior of time-dependent service systems that serve both high priority (HP) and low priority (LP) customers, using mixed discrete-continuous time Markov chains (MDCTMCs).

Our analysis derives formulas to accurately track marginal delay distributions for two customer classes and defines instantaneous transitions to account for workforce changes over time. We ultimately embed the theory in an extension of the numerical Euler method (Izady & Worthington, 2012) and demonstrate how it can be used to generate a reliable low-cost staffing profile that enables the service to meet national quality monitoring standards.

Our article is motivated by the following example, described in more depth toward the end of this article. We have worked extensively with the Welsh Ambulance Service Trust (WAST) who allocate emergency patients to ambulance crews according to a nonpreemptive priory queue. Patients are labeled as HP or LP and there are government targets for the percentage of patients responded to by an emergency ambulance in an allotted time. Demand is time-dependent, there are two customer classes, and staff rosters are required so that demand and government targets on response times are met. There is an added complication in that staff cannot finish their shift until the patient has been dealt with. Often this is when the patient has been passed onto the hospital. Consequently, the main contributions of this article are as follows:

- We consider two methods to set staffing levels in time-dependent dual-class systems: Euler Pri is an exact numerical method and extends the Euler method previously derived for a single customer class to a priority queue with two customer classes. SIPP Pri is an extension of the Stationary Independent Period

by Period (SIPP) method to a priority queue with two customer classes. We benchmark the accuracy of the SIPP Pri approximations against the Euler Pri requirements to demonstrate the increased precision offered by our proposed numerical methodology.
- In extending the numerical methodology beyond its standard setup with a single customer class, we:
  – Present a tractable approach to accurately model the behavior of two customer classes in systems with time-varying arrival rates;
  – Define different shift boundary types and instantaneous transitions necessary to apply to the probability vectors tracking the composition of customers in the system to accurately track behavior at instants where minor modifications may be made to the workforce in line with peaks and troughs in demand (using "partial" shift boundaries) and at instants where the entire workforce turns over (using "full" shift boundaries);
  – Analytically determine the probability of excessive waits for two customer classes over time, with adjustments to correct for workforce changes throughout the day.
- The methodology we propose facilitates the generation of reliable minimum staffing recommendations that are relevant to a wide range of services called to deliver consistent quality despite varying demands, such as call centers and emergency departments, to name but a few.

The remainder of this article is organized as follows. Section "Related Literature" reviews the literature and section "$M(t)/M/s(t)/NPRP$ Systems" overviews $M(t)/M/s(t)$ queueing systems with two customer classes and no preemption. In particular, we present the equilibrium equations that track the movement of customers through such facilities over time and define mappings that account for workforce adjustments when the overall number of servers (i) remains the same, (ii) increases, or (iii) decreases. Then, in section "Evaluation of the Excessive Wait Probability," we formally define the virtual waiting time distribution and outline how the probability of an excessive wait for two customer classes may be accurately tracked over time, taking account of various staffing changes as described earlier. Section "Numerical Approach (Euler Pri)" demonstrates how our newly proposed numerical theory may be ultimately embedded in Euler methodology in order to generate suitable low-cost staffing profiles, and section "Approximate Approach (SIPP Pri)" formally defines SIPP Pri—an approximate approach that extends the SIPP method to a priority queue with two customer classes. Data from WAST is used to benchmark the performance of SIPP Pri against Euler Pri in section "Illustrative Case Study," and it is found that SIPP Pri produces staffing levels that are often close to the Euler Pri recommendations, but not exact.

RELATED LITERATURE

The task to deliver both low operating costs and high service quality is a fundamental challenge faced by managers of tertiary organizations, which is becoming ever more volatile as public expectations and competition between businesses are continually rising. Meeting these potentially conflicting objectives and

ensuring that the right number of staff are scheduled to meet an uncertain, time-varying demand for service involves decisions about forecasting demand, acquiring capacity, and deploying resources (Askin, Armony, & Mehrota, 2007). The most current practice to optimize personnel scheduling follows the general approach originally provided by Buffa, Cosgrove, and Luce (1976), which recommends that the following steps be taken to roster employees: (i) forecast demand; (ii) convert demand forecasts into staffing requirements; (iii) schedule shifts optimally; and (iv) assign employees to shifts. This article considers various ways to perform step (ii), although we note that forecasting methods to achieve step (i) have previously been investigated for WAST in Vile, Gillard, Harper, & Knight (2012), and accordingly used to provide estimated demand levels to be input to the staffing methods considered within this article. Steps (iii) and (iv) are important, but outside the scope of this article.

We note that most prior approaches concerned with evaluating service quality within time-dependent priority service systems relate to delay probabilities or average waiting times, but for many services, the probability of a user waiting longer than a prespecified time interval before being served is a more common performance measure of interest (Ren & Zhou, 2008). Several organizations have devised targets that specify that this wait should be kept below a given threshold for a given proportion of customers. For example, WAST is expected to reach 60% of life-threatening incidents within 8 minutes and 95% of all other emergency calls within 14, 18, or 21 minutes, depending on the location of the incident (Welsh Government, 2012). Green and Soares (2007) have previously discussed how the virtual waiting time can be computed in time-dependent systems that serve a single class of customer, but despite the widespread prevalence of services seeking to manage both time-dependent and prioritized demands, equivalent formulas have not yet been developed for systems with two customer classes. This research extends the methodology presented in Green and Soares (2007) by demonstrating how the virtual waiting time for two customer classes may specifically be tracked over time, and calculated as a function of the state probabilities. It later proceeds to use the results to derive low-cost staffing profiles that comply with response time targets.

The SIPP approach has been considered to estimate the time-dependent behavior of $M(t)/M/s(t)/FIFO$ queueing systems by several researchers (see Green, Kolesar, & Soares, 2001; Green, Kolesar, & Whitt, 2007; Dietz, 2011, and references within). The method approximately tracks performance levels over time by segmenting an entire period into a number of smaller consecutive independent periods and using the average arrival rate for each one as input to a series of stationary analyses. The task to extend SIPP to dual-class systems concerned with controlling excessive waiting times is however unfortunately complicated by the fact that closed-form expressions for the steady-state customer waiting time distribution are not available for two customer classes. It has nevertheless previously been considered to set staffing levels in a priority service system by Chen and Henderson (2001) (although the authors did not formally acknowledge their staffing approach as following SIPP methodology). In their investigation, Chen and Henderson (2001) incorporated an inversion of the Laplace-Stieltjes Transform (LST) to compute the probability of an excessive wait for HP customers originally proposed by Wagner (1997) along with an inequality they proposed to

provide a lower bound on the waiting tail probabilities for LP customers. The approximation method that we consider to set staffing levels in the closing sections of this article furthers the work of Chen and Henderson (2001) by formally describing the extension of the SIPP methodology to a format labeled as "SIPP Pri" and incorporating formulas that computes the probability that an LP customer experiences an excessive wait for service to be evaluated to a greater degree of accuracy.

In situations where the approximation approaches do not work well, numerical methods can offer more accurate insights of system behavior. Ingolfsson (2005) has previously shown that when the assumption of a homogeneous arrival rate is relaxed in a $M/M/s/FIFO$ system and replaced by a piecewise function that may be modified (along with the number of servers), the system can be modeled as a MDCTMC. In particular, Ingolfsson (2005) investigated the influence of departing servers actions dealing with a single class of customer in MDCTMCs, further covering scenarios where servers may stop accepting customers Δt units before their shift is due to end. The study concluded that it is possible to model the evolution of such systems over time using a modified set of equilibrium equations, provided that instantaneous transitions are applied to the state probability vector denoting the probabilities of various numbers of customers present when workers start and end their shifts at predefined times, referred to as "shift boundaries" herein.

Prior research works have also stressed the importance of considering the effect of a departing server if they are providing service when they are scheduled to leave, because two main outcomes are possible in this situation: the server may either follow exhaustive discipline guidelines to first complete the service currently in operation and leave when it is accomplished, or he may leave instantaneously (operating under the nonexhaustive discipline) so the customer in service is rerouted to join the queue. Much early research assumed a nonexhaustive discipline (Bondi & Buzen, 1984; Ngo & Lee, 1990; Gail, Hantler, & Taylor, 1992; Feng, Kowada, & Adachi, 2001) but Ingolfsson (2005) revealed that the performance predictions resulting from the incorporation of this discipline can widely differ from that associated with exhaustive guidelines. Because the nonexhaustive discipline is rarely realistic when the customers are humans, we focus on the exhaustive service discipline in this article. Similarly to Ingolfsson (2005), we also assume a nonpreemptive priority (NPRP) queueing discipline: that is, a HP customer can move ahead of all LP customers waiting in the queue, but cannot preempt a nonpriority in service; because this policy is enforced by most call centers, Emergency Departments (EDs), and ambulance service providers.

The extent to which the research works described above can be applied to priority service systems is however limited, due to the extra level of complexity associated with prioritization rules. For this reason, the majority of research papers analyzing time-dependent priority queues have tended to focus on the long-run steady-state performance of the system (Gail, Hantler, & Taylor, 1988; Kao and Wilson, 1999). Although a novel matrix-analytic method to analyze the expected waiting time of two customer classes in multiple priority dual queues has recently been proposed by Zeephongsekul and Bedford (2006), their analysis is again restricted to scenarios where there is a single server and a consistent arrival rate.

The research contained within this article extends the methodology of Gail et al. (1988) to cover priority queueing systems and fuses it with the MDCTMC

approach introduced by Ingolfsson (2005) for $M(t)/M/s(t)$ queues, in order to provide a tractable approach to track behavior in time-dependent priority queues. We importantly further define the corresponding instantaneous transitions necessary to correct for situations where (i) the entire workforce turns over at truly exhaustive "*full*" shift boundaries, and (ii) only minor changes are made to the workforce, at instants we coin "*partial*" shift boundaries (i.e., instants where small adjustments are made to the staffing function to more closely align capacity with peaks and troughs in demand). Both sets of shift boundaries are formally defined in section "$M(t)/M/s(t)/NPRP$ Systems."

The principal advantages of our numerical approach to track system performance over simulation and approximation methods is that in addition to providing rigorous results, it is easily generalizable and can be used to accurately model any priority service system serving two categories of customers subject to demand that cannot be backlogged, is heavily time-dependent, and highly variable.

$M(t)/M/s(t)/NPRP$ SYSTEMS

We first describe how we model the $M(t)/M/s(t)/NPRP$ system with two customer classes and an exhaustive service discipline as a MDCTMC. Our model assumes that HP customers arrive according to an inhomogeneous Poisson process with mean rate $\lambda_H(t)$ at time t and LP customers arrive with rate $\lambda_L(t)$, so the rate of customers arriving for service at time t is $\lambda(t) = \lambda_H(t) + \lambda_L(t)$. Both sets of customers are served by one of $s(t)$ servers and processed according to NPRP discipline guidelines, meaning that servers may only attend LP customers if there are no HP customers in the queue. However, once they begin serving a LP customer, they cannot be reassigned to a HP customer until they complete their current service. Service times are independently and identically distributed (not class-dependent) with mean time $\frac{1}{\mu}$ and all servers have identical capabilities. They operate under the exhaustive service discipline, and if multiple servers are available to process a job, each available server has an equal probability of taking on this job.

For the purpose of this research, we assume that $\mu(t) = \mu$, for all t. The assumption is fairly realistic because the service rate generally varies more slowly than the arrival rate (Ingolfsson, Akhmetshina, Budge, Li, & Wu, 2007) and commonly used in the literature for tractability.

In the formulas presented to analyze the system below, i and j represent the number of HP and LP customers in service, respectively; and h and l are used to denote the number of HP and LP customers in the queue. Hence, the inequalities $i, j, h, l \geq 0$ and $i + j \leq s(t)$ must hold at all times. In cases where it is relevant to track the total number of customers in the system, n is used to represent the cumulative total of HP and LP customers.

In order to accurately track the movement of all customers through the system, it is necessary to compute the number of customers of types i, j, h, and l in the system over time, represented by the quadruple $S = (i, j, h, l)$. Following the methodology presented by Gail et al. (1988), it is easily shown that the description of this state space quadruple $S = (i, j, h, l)$ may be reduced to

- $S = (i, j)$ if at least one server is idle (as both h and l must both be null, because there will be no customers in the queue)

- $S = (i, h, l)$ if all servers are busy (because j may be derived from the description of the other parameter values)

As such, this convenient notation simplifies the state space description and increases the computational efficiency of a numerical solver to track system behavior over time.

The equilibrium equations that define the evolution of the system are well known for $M/M/s/FIFO$ queues (Gross & Harris, 1998). When the assumption of a homogeneous arrival rate is relaxed and replaced by a piecewise function, the equilibrium equations may be modified (replacing λ with $\lambda(t)$) and solved numerically to model the progression of the system over time (Izady, 2010). In our representation of the system as a MDCTMC, we consider $t_z,\ z = 1, 2, ...$ to be the set of predefined shift boundaries in the service system and assume that during each interval $(0, t_1), (t_1, t_2), ...$ the system operates in a steady-state condition (so continuous-time Markov chains may be used to model the equilibrium equations). We revise the demand rate in a stepwise fashion and allow the workforce to change at the shift start and end points, t_z, in response to the time-varying arrival rate. Thus, at time points t_z the system behaves as a discrete time Markov chain, and like a continuous time Markov chain between these instants.

Using the notation:

$$\begin{aligned} P(i, h, l)(t) &= \text{Prob } \{i \text{ HP in service, } h \text{ HP in queue, } l \text{ LP in queue and all servers busy at time } t\}, \\ P(i, j)(t) &= \text{Prob } \{i \text{ HP in service and } j \text{ LP in service, no customers in the queue and at least one server idle at time } t\}, \end{aligned}$$

we extend the methodology outlined by Gail et al. (1988) to define equilibrium equations for priority service systems, where time-variable and stochastic demand is served by a time-varying number of servers, $s(t)$, in order to maintain a consistent level of service throughout the period of service operation.

The equilibrium equations for the case where at least one server is idle (i.e., $i + j < s(t)$) are

$$\begin{aligned} (\lambda(t) + (i + j)\mu)P(i, j)(t) = {} & \lambda_H(t)P(i - 1, j)(t) & & \\ & + \lambda_L(t)P(i, j - 1)(t) & & \\ & + (i + 1)\mu P(i + 1, j)(t) & \text{for } & 0 < i, 0 < j \\ & + (j + 1)\mu P(i, j + 1)(t), & \text{and } & i + j < s(t) \\ (\lambda(t) + i\mu)P(i, 0)(t) = {} & \lambda_H(t)P(i - 1, 0)(t) & & \\ & + (i + 1)\mu P(i + 1, 0)(t) & & \\ & + \mu P(i, 1)(t), & \text{for } & 0 < i < s(t) \\ (\lambda(t) + j\mu)P(0, j)t = {} & \lambda_L(t)P(0, j - 1)(t) & & \\ & + (j + 1)\mu P(0, j + 1)(t) & & \\ & + \mu P(1, j)(t), & \text{for } & 0 < j < s(t) \\ \lambda(t)P(0, 0)(t) = {} & \mu P(1, 0)(t) + \mu P(0, 1)(t), & \text{otherwise.} & \end{aligned} \tag{1}$$

For states in which all servers are busy, and only LP customers are in service (i.e., $i = 0$), the equations presented in (1) become:

$$\begin{aligned}
(\lambda(t) + s(t)\mu)P(0,h,l)(t) &= \lambda_H(t)P(0,h-1,l)(t) \\
&\quad + \lambda_L(t)P(0,h,l-1)(t), \quad \text{for } 0 < h, 0 < l \\
(\lambda(t) + s(t)\mu)P(0,h,0)(t) &= \lambda_H(t)P(0,h-1,0)(t), \quad \text{for } 0 < h \\
(\lambda(t) + s(t)\mu)P(0,0,l)(t) &= \lambda_L(t)P(0,0,l-1)(t) \\
&\quad + s(t)\mu P(0,0,l+1)(t) \\
&\quad + \mu P(1,0,l+1)(t), \quad \text{for } 0 < l \\
(\lambda(t) + s(t)\mu)P(0,0,0)(t) &= s(t)\mu P(0,0,1)(t) \\
&\quad + \mu P(1,0,1)(t) \\
&\quad + \lambda_L(t)P(0,s(t)-1)(t), \quad \text{otherwise.}
\end{aligned} \tag{2}$$

If all servers are busy and at least one HP and LP customer is in service (i.e., $0 < i < s(t)$):

$$\begin{aligned}
(\lambda(t) + s(t)\mu)P(i,h,l)(t) &= \lambda_H(t)P(i,h-1,l)(t) \\
&\quad + \lambda_L(t)P(i,h,l-1)(t) \\
&\quad + i\mu P(i,h+1,l)(t) \\
&\quad + (s(t) - i + 1) \\
&\quad \times \mu P(i-1,h+1,l)(t), \quad \text{for } 0 < h, 0 < l \\
(\lambda(t) + s(t)\mu)P(i,h,0)(t) &= \lambda_H(t)P(i,h-1,0)(t) \\
&\quad + i\mu P(i,h+1,0)(t) \\
&\quad + (s(t) - i + 1) \\
&\quad \times \mu P(i-1,h+1,0)(t), \quad \text{for } 0 < h \\
(\lambda(t) + s(t)\mu)P(i,0,l)(t) &= \lambda_L(t)P(i,0,l-1)(t) \\
&\quad + i\mu P(i,1,l)(t) \\
&\quad + (s(t) - i + 1)\mu P(i-1,1,l)(t) \\
&\quad + (s(t) - i)\mu P(i,0,l+1)(t) \\
&\quad + (i+1)\mu P(i+1,0,l+1)(t), \quad \text{for } 0 < l \\
(\lambda(t) + s(t)\mu)P(i,0,0)(t) &= i\mu P(i,1,0)(t) \\
&\quad + (s(t) - i + 1)\mu P(i-1,1,0)(t) \\
&\quad + (s(t) - i)\mu P(i,0,1)(t) \\
&\quad + (i+1)\mu P(i+1,0,1)(t) \\
&\quad + \lambda_H(t)P(i-1,s(t)-i)(t) \\
&\quad + \lambda_L(t)P(i,s(t)-i-1)(t), \quad \text{otherwise.}
\end{aligned} \tag{3}$$

Finally, if all servers are busy, and only HP customers are in service (i.e., $i = s(t)$), the equilibrium equations are

$$\begin{aligned}
(\lambda(t) + s(t)\mu)P(s(t), h, l)(t) = {} & \lambda_H(t)P(s(t), h-1, l)(t) \\
& + \lambda_L(t)P(s(t), h, l-1)(t) \\
& + s(t)\mu P(s(t), h+1, l)(t) \quad \text{for } 0 < h \\
& + \mu P(s(t)-1, h+1, l)(t), \quad \text{and } 0 < l \\
(\lambda(t) + s(t)\mu)P(s(t), h, 0)(t) = {} & \lambda_H(t)P(s(t), h-1, 0)(t) \\
& + s(t)\mu P(s(t), h+1, 0)(t) \\
& + \mu P(s(t)-1, h+1, 0)(t), \quad \text{for } 0 < h \\
(\lambda(t) + s(t)\mu)P(s(t), 0, l)(t) = {} & \lambda_L(t)P(s(t), 0, l-1)(t) \\
& + s\mu P(s(t), 1, l)(t) \\
& + \mu P(s(t)-1, 1, l)(t), \quad \text{for } 0 < l \\
(\lambda(t) + s(t)\mu)P(s(t), 0, 0)(t) = {} & s(t)\mu P(s(t), 1, 0)(t) \\
& + \mu P(s(t)-1, 1, 0)(t) \\
& + \lambda_H P(s(t)-1, 0)(t), \quad \text{otherwise.}
\end{aligned} \tag{4}$$

The probabilities of the various combinations of HP and LP customers in the system fluctuate during each interval $(0, t_1), (t_1, t_2), \ldots$ according to the above equations, but the complicating factor is how these probabilities evolve at shift boundaries, where the arrival rates and number of servers on duty are permitted to change. For example, if $(P(i, h, l)(t))$ represents the vector $(P(0, 0, 0)(t), P(1, 0, 0)(t), P(1, 1, 0)(t), \ldots)$, then at shift boundaries (i.e., time points where $t = t_z$), the vector is subject to an instantaneous transition $(P(i, h, l)(t)) = (P(i, h, l)(t))^- B(t)$, where $(P(i, h, l)(t))^- = \lim_{r \to t_z}(P(i, h, l)(r))$ (i.e., the probability vector immediately before the shift boundary) and $B(t)$ is a probability matrix.

We extend the approach followed by Ingolfsson (2005) to track behavior within priority service systems and present adjustments to account for various workforce changes at different types of shift boundaries, namely:

- A full shift boundary: These boundaries mark the set times at which predefined shifts (e.g., morning, afternoon, night) finish. It is assumed that "all" staff leave at the end of this shift and are replaced by an entirely new set of staff in the following shift.
- A partial shift boundary: This type of boundary occurs at instants at which the number of servers is permitted to change, say at the end of every hourly period. At a partial shift boundary, it is assumed that the same servers are employed if the equivalent number (or more) are required for the following period. If more staff are required, these work alongside the existing servers for as many hourly periods as needed. However, if fewer staff are needed, the probability that a particular server is selected to leave is independent of whether they are busy or idle at that time point, because customers are assigned to servers at random. A further benefit of the partial boundary is that it may be applied to determine

minimum staffing levels for short periods to maintain an acceptable service quality throughout the day, assuming that shifts are permitted to overlap and be carefully selected such that the number of staff employed may be flexed up and down to exactly match the minimum number required in each short period.

Under exhaustive guidelines, all servers that are busy when scheduled to leave, continue serving the customers they are currently dealing with until they complete their service—thus all customers being attended to by such servers at the shift boundary are consequently "ejected" from the system (i.e., no longer mathematically recognized as being present) as they no longer require the assistance of a resource scheduled to work, although in reality they continue to receive assistance from servers working beyond their scheduled departure times. Concurrently, the servers scheduled to work in the next period begin attending to customers waiting in the queue. Because the staff working beyond their scheduled departure time are assumed to continue serving customers concurrently with the new servers, this research implicitly assumes that resource shortages never surface, that is, sufficient equipment is available for both staff sets to operate simultaneously.

In order to capture the effect of workforce changes, it is necessary to apply instantaneous transitions to the state probabilities $P(i, j)(t)$ and $P(i, h, l)(t)\ \forall i, j, h, l \geq 0$ and $i + j \leq s(t)$ over each type of shift boundary. In the case of a partial boundary, the transition further depends on if servers are busy when scheduled to leave. In the mappings and probability matrices we define for this purpose below, $s(t)^-$ and $s(t)^+$ to denote the number of servers on duty for the shifts preceding and following the boundary, respectively.

Mappings of State Probability Vectors across Full Shift Boundaries

At the end of a planning period bordered by a full shift boundary, all customers in service are ejected from the system under exhaustive discipline rules; thus the probability vector mappings are identical for all adjustments made to the number of servers on servers on duty (i.e., independent of whether this number increases, decreases or remains the same). The new servers begin serving the customers in the queue at the immediate commencement of their shift, so all customers at the front of the queue move into service.

In our description of the transitions, we impose a limit G on the number of customers considered in the system to aid the application of numerical methods to solve the equations by reducing the infinite set of equilibrium equations to a finite set (Kao & Narayanan, 1990). Recalling that LP customers are only served when a server becomes free if there are no HP customers in the queue, then:

- The new number of HP customers in service, i, after the boundary might arise from there being any number between 0 (if all servers were previously busy serving LP customers) and $s(t)^-$ (if all servers were previously busy serving HP customers) HP customers present before the boundary;
- The new number of HP customers in the queue, h, equals the number of HP customers in the queue before the boundary plus any who might move into service; and

- The new number LP customers in the queue, l, equals the number of LP customers in the queue before plus any who might move into service.

Hence, the mappings that define the instantaneous transitions of the probability vectors may be expressed as follows:

For $0 \leq h + l + s(t)^+ \leq G, i <= s(t)^+$(if $i = s(t)^+$ then $P(i, h, l)(t)$ is only defined for $h = 0$):

$$P(i, h, l)(t) = \sum_{u=0}^{s(t)^-} P(u, h + i, l + (s(t)^+ - i))(t)^-, \tag{5}$$

and the transitions for the dual state vectors, defined for the case where $i + j < s(t)^+$ are

$$\begin{aligned} &\text{for } i + j = 0: \\ &\qquad P(i, j)(t) = \sum_{u=0}^{s(t)^-} P(u, 0, 0)(t)^- + \sum_{u=0}^{s(t)^- - 1} \sum_{q=0}^{s(t)^- - 1 - q} P(u, q)(t)^-; \\ &\text{for } 0 < i + j < s(t)^+: \\ &\qquad P(i, j)(t) = \sum_{u=0}^{s(t)^-} P(u, i, j)(t)^-. \end{aligned} \tag{6}$$

Due to the way in which the artificial limit G is placed on the number of customers considered within the system to allow computation of the solution in reasonable time, there will be some cases where the revised probability vectors will be assigned zero values (e.g., if $h + l + s(t)^- > G$ then $P(i, h, l)(t)^-$ will not be defined).

Mappings of State Probability Vector across Partial Shift Boundaries

The transitions are more complex to define in the case of a partial boundary, as it becomes necessary to account for the actions of departing servers. In the case where the number of servers remains the same or is increased, the probability vectors require little or no modification, because the same set of staff are assumed to work both shifts (thus each server may continue working without disruption). The only potential modification that needs to be accounted for by a mapping is the movement of customers from the head of the queue into service that receive service from any additional employees who join the team at the shift boundary. However, if the number of servers is reduced over the shift boundary, the behavior of the system is additionally dependent on the current occupation of the servers who are selected to leave. Thus, the precise mappings necessary for each of the three scenarios are separately defined in cases (A)–(C) below.

Case (A): Number of servers remains the same

If the number of servers on duty over two consecutive shifts remains consistent, then the Markov process evolves as a continuous time Markov chain, as each server is available to work at all times across the shifts. Thus, all probability

vectors remain identical across the shift boundary, and may be defined as follows:

$$\begin{aligned} P(i,h,l)(t) &= P(i,h,l)(t)^- \quad \text{for } 0 \le i \le s(t)^+, 0 \le h+l+s(t)^+ \le G, \\ P(i,j)(t) &= P(i,j)(t)^- \quad \text{for } 0 \le i+j < s(t)^+. \end{aligned} \tag{7}$$

Case (B): Number of servers is increased

For the case where more servers are supplied in the period following a partial shift boundary, vector mappings are required to account for the fact that customers at the front of the queue move into service to be attended to by the additional servers who commence their duty at time t. Using $s_c = (s(t)^+ - s(t)^-)$ to represent the change in the number of servers, which in case B will always be positive; the instantaneous transitions may be defined for the triple state probability vector as:

$$\begin{aligned} &\text{For } i = s(t)^+, 0 \le h+l+s(t)^+ \le G: \\ &\qquad P(i,h,l)(t) = P(i-s_c, h+s_c, l)(t)^-. \\ &\text{For } i < s(t)^+, 0 \le l+s(t)^+ \le G: \\ &\qquad P(i,0,l)(t) = \sum_{u=\max(0,i-s(t)^-)}^{\min(i,s_c)} P(i-u, u, l+s_c-u))(t)^-. \\ &\text{For } s_c \le i < s(t)^+, 0 < h, 0 \le h+l+s(t)^+ \le G: \\ &\qquad P(i,h,l)(t) = P(i-s_c, h+s_c, l)(t)^-. \end{aligned} \tag{8}$$

Concurrently the dual state space probability vectors remain identical, except for extra states which may arise if the number of customers in the queue is less than the quantity of additional servers joining at the boundary. Thus, for $0 < i + j < s(t)^+$:

$$\begin{aligned} &\text{For } i+j < s(t)^-: \\ &\qquad P(i,j)(t) = P(i,j)(t)^-. \\ &\text{For } i+j = s(t)^-: \\ &\qquad P(i,j)(t) = P(i,0,0)(t)^-. \\ &\text{For } s(t)^- < i+j < s(t)^+: \\ &\qquad P(i,j)(t) = \sum_{u=max(0,i-s(t)^-)}^{min(i,s_c,i+j-s(t)^-)} P(i-u, u, i+j-s(t)^- - u)(t)^-. \end{aligned} \tag{9}$$

Case (C): Number of servers is reduced

In our analysis, we assume that customers are randomly assigned to servers and the probability that each server is selected to leave at the boundary of an interval is independent of whether they are busy or idle; thus, following a similar argument to that presented in Ingolfsson (2005), it is clear that that the number of customers ejected follows a hypergeometric distribution (Johnson, Kotz, & Kemp, 1993). However, in order to approximate the likelihood that a departing server is serving

an HP or LP customer, then after initially calculating the probabilities of various numbers of busy servers departing (equivalent to the total number of customers ejected) using a specific hypergeometric distribution, we perform additional calculations to compute the various compositions of HP and LP customers that could comprise this total quantity.

Recalling that $s(t)^-, i$ and j denote the number of servers on duty, HP customers in service and LP customers in service *before* the shift boundary, respectively; and letting δn represent the total number of customers ejected from system and δs represent the total number of servers leaving at shift boundary; then the probability that δi HP customers are ejected from the system is defined for

$$\begin{aligned}&\max(0, i + j \quad - s(t)^+) \leq \delta n \leq \min(\delta s, i + j)\\ &\text{and} \quad \max(0, \delta\, n - j) \leq \delta i \leq \min(\delta n, i),\end{aligned}$$

and given by

$$\varphi(\delta n; \delta i; s(t)^-, i + j, i) = \frac{\binom{i+j}{\delta n}\binom{s(t)^- - i - j}{\delta s - \delta n}}{\binom{s(t)^-}{\delta s}} \times \frac{\binom{i}{\delta i}\binom{j}{\delta n - \delta i}}{\binom{i+j}{\delta n}}. \tag{10}$$

Equation (10) can be used to compute the dual state probability state vectors. Considering the different ways in which each of these states may arise, it directly follows that the transitions are given by

$$\text{For } i + j < s(t)^+ :$$
$$P(i, j) = \sum_{\delta n=0}^{\delta s} \sum_{\delta i=0}^{\delta n} \varphi(\delta n; \delta i; s(t)^-, i + j, i + \delta i) P(i + \delta i, j + (\delta n - \delta i))^-. \tag{11}$$

For all cases where $i + j \geq s(t)^+$, probabilities are derived by considering the triple state vectors, as defined below.

The triple state probability vectors $P(i, h, l)$ are defined at the partial shift boundary for cases where all servers are busy. The probability that δi HP customers are ejected from the system is somewhat simpler to define for this scenario, because it is certain that all departing servers will each eject a customer from the system, so it is only necessary to take into account the probability that those ejected are HP or LP customers. If there are no idle servers, it can be easily shown that the probability that δi HP customers are ejected from the system follows a hypergeometric distribution. Following the notation that has been defined above, this is given by

$$\theta(\delta i; \delta s, s(t)^-, i + j) = \frac{\binom{i+j}{\delta i}\binom{s(t)^- - i - j}{\delta s - \delta i}}{\binom{s(t)^-}{\delta s}}. \tag{12}$$

One may observe that the number of HP and LP customers in the queue remain identical over the partial boundary: because staff numbers decrease, there are no additional servers available to accept new customers at the commencement of the new shift. Thus the only parameter value to experience a transition in the triple state vector over the shift boundary is i (representative of the number of HP customers in service). Ingolfsson (2005) has demonstrated that if servers depart at the shift

boundary, the nonpriority probability vector systems serving a single customer class experiences an instantaneous transition according to $P(t) = P(t)^{-}B(t)$, where $B(t)$ is a probability matrix. The same methodology is applied here to model the instantaneous transitions for $(P(i,h,l)(t)) = (P(i,h,l)(t))^{-}B(t)$, giving

$$\begin{aligned} &\text{for } 0 < h + l, 0 \le s(t)^{+} + h + l \le G: \\ &P(i,h,l)(t) = \sum_{u=0}^{s(t)^{+}} P(u,h,l)(t)^{-}B(t), \end{aligned} \tag{13}$$

where transition matrix $B(t)$ has the following nonzero entries:

$$b_{n,n-\delta i} = \theta(\delta i; \delta s, s(t)^{-}, n) \quad \begin{cases} \text{for } n = 0, 1, \ldots s(t)^{-} - 1 \text{ and} \\ \max(0, n - s(t)^{+}) \le \delta i \le \min(\delta s, n). \end{cases} \tag{14}$$

Equation (11) is valid for $i + j < s(t)^{+}$. Yet it also gives the probability for the boundary state for the case when all idle servers leave the system, leaving no customers in the queue, but all remaining servers busy. Thus, the triple state probability vector defining the case where there are no customers in the queue additionally needs to take into account this event, so the transition may be defined by

$$\begin{aligned} &\text{for } i \le s(t)^{+}: \\ &\quad P(i,0,0) = \sum_{\delta n=0}^{\delta s} \sum_{\delta i=0}^{\delta n} \varphi(\delta n; \delta i; s(t)^{-}, i + j, i + \delta i) P(i + \delta i, j \\ &\qquad + (\delta n - \delta i))^{-} + \sum_{u=0}^{s(t)^{+}} P(u,0,0)(t)^{-}B(t). \end{aligned} \tag{15}$$

EVALUATION OF THE EXCESSIVE WAIT PROBABILITY

The virtual waiting time is defined to be the time that a customer arriving at the system at time t waits before commencing service. We are interested in ensuring that this wait remains below a given threshold, x, for a target proportion of the population. For a nonpriority service system, this can be expressed as

$$P(W_q(t) > x) \le \alpha, \tag{16}$$

where $W_q(t)$ represents the virtual waiting time of a customer arriving at time t, x indicates the maximum acceptable waiting time, and α denotes the targeted (i.e., the maximum allowed) excess wait probability.

If $p_n(t)$ indicates the probability there are n customers in the system at time t and $W_q^n(t)$ represents the virtual waiting time of a customer who arrives to find n people ahead in the system then we may write

$$P(W_q(t) > x) = \sum_{n=s}^{\infty} P\left(W_q^n(t) > x\right) p_n(t). \tag{17}$$

Observing that the wait in the queue will be greater than time x if less than n services are completed in the time x, for stationary $M/M/s$ systems with

a constant arrival rate λ and service rate μ, we have $p_n(t) = p_n \; \forall t$ and Equation (17) may be reduced to the following closed-form formula (Gross & Harris, 1998):

$$P(W_q > x) = \left(\frac{\left(\frac{\lambda}{\mu}\right)^s p_0}{s!\left(1 - \frac{\lambda}{s\mu}\right)} \right) \left(e^{-(s\mu-\lambda)t}\right). \tag{18}$$

In the analysis that follows, we present exact expressions for $P(W_q^n(t) > x)$ in priority service systems, together with appropriate adjustments to account for the effect of full and partial shift boundaries on the performance measure. The expressions presented are applicable to services in which the number of servers is permitted to change at most once during the maximal allowed waiting time.

In systems staffed by a time-varying number of servers, the evaluation of Equation (17) is complicated by the fact that $P(W_q^n(t) > x)$ depends not only on the number of staff present at time t, $s(t)$, but also on the number of servers present over the time interval $(t, t + x]$. We proceed to define exact expressions for $P(W_q^n(t) > x)$ for cases where the number of servers changes at most once in the interval $[t, t + x]$, assuming that the infinite dimensional vector $(p_n(t))$ is known. Letting $\tilde{n}$ represent the cumulative number of LP customers in service and HP customers in the system (i.e., all customers in the system excluding LP customers in the queue), $p_{\tilde{n}}(t)$ denote the probability that the system is in each state, and $W_{qH}^{\tilde{n}}$ denote the waiting time for HP customers that arrive to find $\tilde{n}$ people in the system ahead with s servers on duty; the probability that an HP customer waits longer than a predefined acceptable time x_H in the queue is given by

$$P(W_{qH}(t) > x_H) = \sum_{\tilde{n}=s}^{\infty} P\left(W_{qH}^{\tilde{n}} > x_H\right) p_{\tilde{n}}(t). \tag{19}$$

Because the system evolves as a continuous time Markov chain across this interval, one may compute the probability of an excessive wait using the fact that the departure process behaves as a nonhomogeneous Poisson process with rate μs ($s(t) = s \; \forall t$), provided that all servers are busy over the interval, so the mean number of departures over $[t, t + x_H]$ is given as below (for further details, see Green et al., 2007):

$$a = \mu s x_H. \tag{20}$$

Thus, the probability that an HP customer will wait greater than the acceptable waiting time threshold x_H is equivalent to $P(\text{“}\tilde{n} - s$ or fewer departures over $[t, t + x_H]\text{”})$, which may be computed as

$$\sum_{b=0}^{\tilde{n}-s} \frac{a^b e^{-a}}{b!}. \tag{21}$$

$P(W_{qH}^{\tilde{n}}(t) > x_H)$ can thus be evaluated for each $\tilde{n}$ using

$$P\left(W_{qH}^{\tilde{n}}(t) > x_H\right) = \begin{cases} \sum_{b=0}^{\tilde{n}-s} \frac{a^b e^{-a}}{b!} & \text{if } \tilde{n} \geq s, \\ 0 & \text{if } \tilde{n} < s. \end{cases} \tag{22}$$

Combining these results yields

$$P(W_{qH}(t) > x_H) = \begin{cases} \sum_{\tilde{n}=s}^{\infty} \sum_{b=0}^{\tilde{n}-s} \frac{a^b e^{-a}}{b!} p_{\tilde{n}}(t) & \text{if } \tilde{n} \geq s, \\ 0 & \text{if } \tilde{n} < s. \end{cases} \tag{23}$$

A similar approach can be followed to compute the waiting tail probability for LP customers. Letting W_{qL}^n denote the waiting time for LP customers that arrive to find n people in the system ahead with s servers on duty and x_L denote maximum acceptable waiting time for LP customers, $P(W_{qL} > x_L)$ may be computed as

$$P(W_{qL}(t) > x_L) = \sum_{n=s}^{\infty} P\left(W_{qL}^n > x_L\right) p_n(t), \tag{24}$$

where

$$P\left(W_{qL}^n(t) > x_L\right) = \begin{cases} \sum_{f=0}^{\infty} P(f \text{ HPs arrive in } x_L) \times \sum_{b=0}^{n-s+f} \frac{a^b e^{-a}}{b!} & \text{if } n \geq s, \\ 0 & \text{if } n < s. \end{cases} \tag{25}$$

Here, $a = \mu s x_L$ and because HP customers are assumed to arrive in a Poisson fashion, the probability of f HP arrivals in time x_L may be calculated as

$$P(f \text{ HPs arrive in } x_L) = \frac{(\lambda_H x_L)^f e^{-(\lambda_H x_L)}}{f!}. \tag{26}$$

Thus, combining these results yields

$$P(W_{qL}(t) > x_L) = \begin{cases} \sum_{n=s}^{\infty} \left(\sum_{f=0}^{\infty} \frac{(\lambda_H x_L)^f e^{-(\lambda_H x_L)}}{f!} \sum_{b=0}^{n-s+f} \frac{a^b e^{-a}}{b!} \right) p_n(t) & \text{if } n \geq s, \\ 0 & \text{if } n < s. \end{cases} \tag{27}$$

If the number of staff however changes within $[t, t + x_H]$ or $[t, t + x_L]$, we observe that the waiting time formulas cannot be simply extended by replacing s with $s(t)$, because if the number of servers is increased exactly once over the interval for example, say at time $t + \Delta t$, where $\Delta t < x_H$, then fewer than $n - s(t)$ departures may result in an arriving HP customer waiting less than time x_H before being served (Green et al., 2007). This is because the additional staff starting at time Δt will each acquire a customer as soon as their shift begins, so fewer departures than service commencements need to occur across the interval, to meet the waiting time target.

Green et al. (2007) and Ingolfsson (2005) have previously considered this issue for services that serve a single class of customer, and shown that for a

maximal allowed waiting time, x, if the number of servers changes exactly once in $[t + \Delta t, t + x]$, there exists some $\Delta \leq x$ such that:

$$s(u) = \begin{cases} s(t)^- & \text{if } u \in [t, t + \Delta t], \\ s(t)^+ & \text{if } u \in [t + \Delta t, t + x]. \end{cases} \tag{28}$$

Thus, if a staffing change occurs between the time that a customer arrives and the maximal allowable waiting time for that customer comes to pass, a must be redefined as

$$a = \mu \int_t^{t+x} s(u)\,\mathrm{d}u, u \in [t, t + x] = \mu s(t)^- \Delta t + \mu s(t)^+ (x - \Delta t) \tag{29}$$

to reflect the mean number of departures expected over an interval covered by two different staffing teams. Note that when $n < \max(s(t)^-, s(t)^+)$, it will always be true that $P(W_q^n(t) > x) = 0$ because the $(n + 1)$th customer will begin either begin service immediately (if $s(t)^- > s(t)^+$) or at time $t + \Delta t$ if $s(t)^- < s(t)^+$. When $n \geq \max(s(t)^-, s(t)^+)$, then $P(W_q^n(t) > x)$ will be dependent on the number of servers in time period $[t, t + x]$.

In sections "Adjustment at full shift boundaries" and "Adjustment at partial shift boundaries," we proceed to extend this analysis to $M(t)/M/s(t)/NPRP$ systems, and present adjustments that can be applied to Equations (23) and (27) to account for staffing changes at full and partial shift boundaries. As defined in Equation (29), $a = \mu s(t)^- \Delta t + \mu s(t)^+ (x - \Delta t)$ will be used to represent the mean departure rate over $[t, t + x]$ (where x will be adjusted to equal x_H or x_L, as required).

Adjustment at Full Shift Boundaries

At a full shift boundary where servers operate under the exhaustive discipline, all customers in service are ejected from the system. Because all servers leave the system and are replaced by an entirely new set, one may observe that only one standard adjustment is needed to account for all possible changes in staffing levels (i.e., an increase, decrease or equal levels) giving,

$$\begin{aligned} P(W_{qH}(t) > x_H) &= \sum_{\tilde{n}=s(t)^-+s(t)^+}^{\infty} P\left(W_{qH}^{\tilde{n}}(t) > x_H\right) p_{\tilde{n}}(t), \quad \text{where} \\ P\left(W_{qH}^{\tilde{n}}(t) > x_H\right) &= \sum_{b=0}^{\tilde{n}-s(t)^--s(t)^+} \frac{a^b e^{-a}}{b!} \quad \text{if } \tilde{n} \geq s(t)^- + s(t)^+ \end{aligned} \tag{30}$$

and

$$\begin{aligned} P(W_q L(t) > x_L) &= \sum_{n=s(t)^-+s(t)^+}^{\infty} P\left(W_q^n L > x_L\right)\ p_n(t)\,, \text{where} \\ P\left(W_q^n L > x_L\right) &= \sum_{f=0}^{\infty} P(f \text{ HPs arrive in } x_L) \\ &\quad \times \sum_{b=0}^{n-s(t)^--s(t)^++f} \frac{a^b e^{-a}}{b!} \quad \text{if } n \geq s(t)^- + s(t)^+. \end{aligned} \tag{31}$$

Adjustment at Partial Shift Boundaries

Case 3.2(A): Number of servers remains the same

If the number of servers remains unchanged over the shift boundary, the same formula may be used to calculate the probability of an excessive wait as if no boundary was imposed, because the servers are unaffected by the occurrence of the shift boundary and continue working as normal. Note that as $s(t)^- = s(t)^+ = s$, they may be used interchangeably within the expression:

$$\begin{aligned} P(W_{qH}(t) > x_H) &= \sum_{\tilde{n}=s(t)^+}^{\infty} P\left(W_{qH}^{\tilde{n}}(t) > x_H\right) p_{\tilde{n}}(t), \quad \text{where} \\ P\left(W_{qH}^{\tilde{n}}(t) > x_H\right) &= \sum_{b=0}^{\tilde{n}-s(t)^+} \frac{a^b e^{-a}}{b!} \quad \text{if } \tilde{n} \geq s(t)^+ \end{aligned} \tag{32}$$

and

$$\begin{aligned} P(W_q L(t) > x_L) &= \sum_{n=s(t)^+}^{\infty} P\left(W_q^n L > x_L\right) p_n(t), \text{ where} \\ P\left(W_q^n L > x_L\right) &= \sum_{f=0}^{\infty} P(f \text{ HPs arrive in } x_L) \sum_{b=0}^{n-s(t)^+ + f} \frac{a^b e^{-a}}{b!} \quad \text{if } n \geq s(t)^+. \end{aligned} \tag{33}$$

Case 3.2(B): Number of servers is increased

If the number of servers increases at time $t + \Delta t$, so $s(t)^+ > s(t)^-$ the waiting tail probabilities $P(W_q^n H(t) > x_H)$ and $P(W_q^n L(t) > x_L)$ may be calculated as

$$\begin{aligned} P(W_{qH}(t) > x_H) &= \sum_{\tilde{n}=s(t)^+}^{\infty} P\left(W_{qH}^{\tilde{n}}(t) > x_H\right) p_{\tilde{n}}(t), \quad \text{where} \\ P\left(W_{qH}^{\tilde{n}}(t) > x_H\right) &= \sum_{b=0}^{\tilde{n}-s(t)^+} \frac{a^b e^{-a}}{b!} \quad \text{if } \tilde{n} \geq s(t)^+ \end{aligned} \tag{34}$$

and

$$\begin{aligned} P(W_q L(t) > x_L) &= \sum_{n=s(t)^+}^{\infty} P\left(W_q^n L > x_L\right) p_n(t), \quad \text{where} \\ P\left(W_q^n L > x_L\right) &= \sum_{f=0}^{\infty} P(f \text{ HPs arrive in } x_L) \sum_{b=0}^{n-s(t)^+ + f} \frac{a^b e^{-a}}{b!} \quad \text{if } n \geq s(t)^+. \end{aligned} \tag{35}$$

Case 3.2(C): Number of servers is reduced

Finally considering the case where the number of servers decreases at time $t + \Delta t$, that is, $s(t)^+ < s(t)^-$, the probability of an excessive wait may be computed as

$$\begin{aligned} P(W_{qH}(t) > x_H) &= \sum_{\tilde{n}=s(t)^-}^{\infty} P\left(W_{qH}^{\tilde{n}}(t) > x_H\right) p_{\tilde{n}}(t), \quad \text{where} \\ P\left(W_{qH}^{\tilde{n}}(t) > x_H\right) &= \sum_{b=0}^{\tilde{n}-s(t)^-} \frac{a^b e^{-a}}{b!} \quad \text{if } \tilde{n} \geq s(t)^- \end{aligned} \tag{36}$$

and

$$P(W_q L(t) > x_L) = \sum_{n=s(t)^-}^{\infty} P\left(W_q^n L > x_L\right)\ p_n(t), \quad \text{where}$$
$$P\left(W_q^n L > x_L\right) = \sum_{f=0}^{\infty} P(f \text{ HPs arrive in } x_L) \sum_{b=0}^{n-s(t)^-+f} \frac{a^b e^{-a}}{b!} \quad \text{if } n \geq s(t)^-. \tag{37}$$

NUMERICAL APPROACH (EULER Pri)

Following the analysis contained in Ingolfsson et al. (2007) and Izady and Worthington (2012), which present the Euler method as a comprehensive numerical technique, this section considers the potential of the Euler approach to be extended, through incorporating the methodology, which we have described above, to accurately evaluate system performance in priority service systems in a method we call "Euler Pri." The Euler method is a general approach for solving ordinary differential equations with the advantage that it may be implemented to provide solutions at a quicker rate and does not require an ordinary differential equation solver (Izady, 2010). The Euler method has been investigated in several papers dealing with $M/M/s/FIFO$ systems (Gail et al., 1988; Wagner, 1997); and by embedding the theory presented in sections "$M(t)/M/s(t)/NPRP$ Systems" and "Evaluation of the Excessive Wait Probability," we enable its application within time-dependent dual-class $M(t)/M/s(t)/NPRP$ systems, where staff operate under the exhaustive service discipline.

The Euler method determines the solution by evaluating the equations at a starting value, and then at steps separated by small time intervals (between which the solution is not expected to have changed greatly). Smaller step sizes generate solutions with higher accuracies, but this requires greater computation time. When evaluating performance methods as a function of time over a period of service operation $(0, T]$, it is commonly assumed that the period $(0, T]$ is divided into planning periods of length δ_{pp} and the performance measure is evaluated at calculation periods separated by an interval of length δ_c, which is a divisor of δ_{pp} small enough to guarantee convergence to the actual solution. As δ_c is decreased, both the accuracy and computation time increase (Izady, 2010).

For computational efficiency, Kao and Wilson (1999) comment that it is impractical to avoid truncation (in terms of the number of equations considered in the set of balance equations) in a numerical solution of the problem, because the equations must be reduced to a finite set to be solved numerically. Thus, a limit G is imposed on the number of customers considered in the system, which must be large enough to allow accurate analysis while ensuring that the dimension of the Markov chain is finite. The same approximation is clearly required for numerical analysis of $M(t)/M/s(t)/NPRP$ systems, and following similar reasoning to the case presented by Izady (2010) for nonpriority systems, this research recommends this upper limit be chosen such that $P_G(t) \leq 10^{-6}\ \forall t$, where $P_G(t)$ denotes the probability that there are G customers present in the system (i.e., the cumulative total of all HP and LP customers in the queue and in service) at time t.

APPROXIMATE APPROACH (SIPP Pri)

Acknowledging that SIPP is a widely used method to approximate the time-dependent behavior of $M(t)/M/s(t)/FIFO$ systems, this section formally defines how the methodology can be extended to approximate the time-varying behavior of a dual class priority queueing system. Our proposed extension, which we call SIPP Pri, can be used to generate suitable staffing profiles by computing stationary measures in a set of stationary systems, which are subsequently adjoined by the technique, as follows:

(I) Segment the scheduling horizon into a number of smaller distinct intervals;

(II) Find the average arrival rate of HP and LP customers within each interval;

(III) Assume the system reaches steady-state within each interval, so each interval may be modeled as a $M/M/s/NPRP$ system;

(IV) Use mathematical expressions to evaluate performance measures in each interval and use these to set staffing levels based on system quality.

Hence, by assuming that the behavior of the system in consecutive intervals is statistically independent and that the system reaches steady state within each one, stationary measures may be used to approximate the system behavior and recommend minimum staffing levels that ensure the required performance metrics are attained at all times.

In their approach, Chen and Henderson (2001) calculated the probability that HP customers waited longer than the acceptable waiting time, x_H, before commencing service using the inversion of the LST:

$$P(W_{qH} > x_H) = P(\text{All servers busy})\, e^{-(s\mu-\lambda_H)x_H}. \tag{38}$$

Yet, because the equivalent inversion is analytically intractable for LP customers, they proposed a lower bound to calculate the waiting tail probabilities for LP customers, as follows:

$$P(W_{qL} > x_L) \leq \min\left(\frac{\overline{W_{qL}}}{x_L}, \frac{\overline{W_{qL}}^2}{x_L^2}\right), \tag{39}$$

where $\overline{W_{qL}}$ is the average expected waiting time of LP customers in the queue.

This bound provides a conservative estimate of waiting time; thus, if implemented as part of a wider SIPP model to evaluate performance measures and recommend minimum staffing levels, it will provide staffing levels that will ensure the required performance target will be certainly met in $M/M/s/NPRP$ service systems, given the assumptions of SIPP are met. However, while the staffing levels it recommends will always be sufficient, they may be higher than necessary. In order to overcome the risk of setting staffing schedules with unnecessarily high staff quantities, we further the work of Chen and Henderson (2001) by replacing the lower bound presented in Equation (39) with the exact expression presented in Equation (27)—in which, the time-dependent probability $p_n(t)$ is replaced with the steady state probability vector, p_n for each interval (assumed independent).

Thus, for $a = \mu s x_L$, the probability that a LP customer experiences an excessive wait is given by

$$P(W_{qL} > x_L) = \begin{cases} \sum_{n=s}^{\infty} \left(\sum_{f=0}^{\infty} P(f \text{ HPs arrive in } x_L) \times \sum_{b=0}^{n-s+f} \frac{a^b e^{-a}}{b!} \right) p_n & \text{if } n \geq s, \\ 0 & \text{if } n < s. \end{cases} \quad (40)$$

While SIPP Pri can be used to provide reasonable approximations at speed, its approximate nature means it will almost always be subject to a certain degree of error, by definition. In order to provide reliable approximations, it is clear that the same conditions are required as its nonpriority counterpart (i.e., the behavior of the system in consecutive intervals is statistically independent and that the system reaches steady state within each one), and the extent to which these assumptions are violated determines whether it is reasonable to use it.

ILLUSTRATIVE CASE STUDY

We present here a case study demonstrating the different number of crews the approximate and numerical methodologies recommend that WAST should deploy in the Cardiff region for two given demand profiles—one typical for a 28-day period in July and another for a 28-day period in December. Call handlers at WAST allocate patients to crews according to the rules specified for a nonpreemptive resume priority queue.

Our test model is an extension of the models considered in Green et al. (2001) and Green, Kolesar, and Soares (2003), which are call centers that can be modeled as $M(t)/M/s(t)$ queueing systems, because WAST have the additional complication that they are required to prioritize certain patient requests. We consider the challenge to determine hourly crew requirements for WAST, given the government target that 95% of HP and LP calls should be responded to by an emergency ambulance within 14 minutes. Similarly to Ganguly, Lawrence, and Prather (2014), we choose to determine staffing levels for hourly periods because Green et al. (2001) have previously shown that SIPP performs better when applied to short planning periods but hourly staffing changes are the most frequent WAST could realistically impose. In addition to coinciding with shift boundaries, hourly intervals are small enough to capture granularity of deviation without requiring excessive computation time. The minimum hourly staffing function output is ultimately intended to inform the development of a shift schedule (outside the scope of this article).

In order to impose a consistent response time target in the analysis, the data employed in this study are confined to demand arising within a single region of Wales only, namely Cardiff, because the response target in reality varies throughout Wales. The specific guidelines issued by the Welsh Government (2012) for fully equipped emergency ambulances (EAs) in the Cardiff region are:

- To respond to 95% of life-threatening (HP) calls within 14 minutes (as a follow-up vehicle to a first responder vehicle which should have arrived within 8 minutes)

- To respond to 95% of all other emergency (LP) calls within 14 minutes (as the first responder)

We note that although WAST are additionally expected to send a first responder to arrive at the scene of life-threatening incidents within 8 minutes, separate ambulance officers and vehicles are often used for this purpose; thus, we restrict this case study to solely concern the deployment of EAs.

The expected number of HP and LP emergencies requiring EA assistance for each period of each day in the scheduling horizons are obtained from Singular Spectrum Analysis (SSA) forecasts. Briefly, SSA is a nonparametric method of time series analysis suitable for forecasting data with clear seasonal structure—see Vile et al. (2012) for further details. Analysis of WAST data finds an average service time of $\mu = 54.55$ minutes for HP and LP calls. However, the average travel times to reach HP and LP emergencies (assumed out of the crew's control) are found to be significantly lower for LP calls (4.79 minutes) than HP calls (5.73 minutes) because control room workers have scope to delay responses to less serious incidents by a few minutes if all conveniently located ambulances are busy. We subtract travel times from the 14-minute targeted response time and seek to find a desirable staffing function $s(t)$, which defines the minimum number of EA crews that must be deployed in each hourly interval, to limit the proportion of HP and LP patients waiting longer than targeted 14-minute response target time to a maximum of 5% at all times, in line with the government expectation. This may be expressed via two equations, which must both be satisfied at all time points:

$$P(W_{qH}(t) > 8.27) \leq 0.05 \tag{41}$$

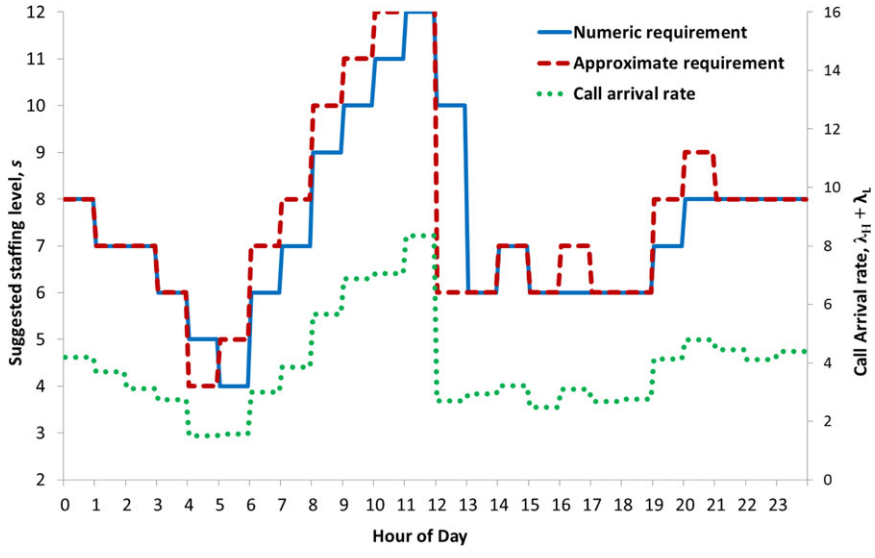
and

$$P(W_{qL}(t) > 9.21) \leq 0.05. \tag{42}$$

Minimum EA crew requirements are sought for 1-hour planning periods throughout the scheduling horizons. Within each hour, requests for assistance are assumed to arrive according to a homogeneous Poisson process with the forecasted mean rate for that hour, and all servers are assumed to have independent exponentially distributed service times, with the same mean length. The numerical analysis assumes that the minimum service level must hold for every time point in the scheduling horizon (rather than being considered as an aggregate service level that is to be achieved on average across all hourly periods in the scheduling horizon), to ensure a consistent quality of service is provided.

In our implementation of the described Euler Pri approach, we incorporate the formulas that have been developed in sections "$M(t)/M/s(t)/NPRP$ Systems," "Evaluation of the Excessive Wait Probability," and "Numerical Approach (Euler Pri)" to compute the time-dependent waiting time distribution in priority $M(t)/M/s(t)/NPRP$ service systems. We use the equilibrium equations (1)–(4) to track system behavior between shift boundaries, and the mappings of the state probability vectors across partial shift boundaries defined in section "Mappings of state probability vector sacross partial shift boundaries" to track the evolution of the system every over time. (A partial boundary is with exhaustive discipline is applied at every hourly interval, because the ultimate aim of the analysis is to construct

Figure 1: Exemplar approximate and numerical crew requirements for the Cardiff region.



hourly coverage requirements to inform the development of a staff/ambulance rota in which crews are permitted to join and leave the workforce, as appropriate, at hourly boundaries.) Finally, we generate a suitable staffing profile by iteratively computing the expected probabilities of excessive waits for different crew sizes using the formulas presented in section "Evaluation of the Excessive Wait Probability" and selecting the minimum number that satisfy the inequality presented in Equations (41) and (42) for each period.

Before performing meaningful analysis of the system, we also incorporate a warm-up period of 1 day to ensure that dynamic steady state is reached (see Heyman and Whitt, 1984, for a definition and necessary conditions to achieve this state). We use calculation periods of $\delta_c = 0.04$ hours and place a cap of $G = 40$ imposed on the number of patients considered in the system at any specific time instance for computational efficiency.

SIPP Pri is implemented by assuming that the behavior of the system in consecutive intervals is statistically independent and that the system reaches steady state within each one. As with Euler Pri, we use SIPP methodology to generate a suitable staffing profile by iteratively computing the expected probabilities of excessive waits for different crew sizes using the steady state formulas given in Equations (38) and (40), and selecting the minimum number that satisfy the response time targets for each hourly period.

Prior to discussing the results generated by all methods for longer 28-day scheduling horizons (i.e., 672 hourly periods), a graphical illustration of the results generated for the first day in the July scheduling horizon is provided so as to aid with the interpretation of the results. Figure 1 illustrates the call arrival rate on a typical

Monday in the Cardiff region, together with the minimum staffing levels recommended by the numerical Euler Pri and approximate SIPP Pri techniques. The pattern of increasing call rates throughout the morning that peak at noon is consistent with most ambulance services (e.g., Matteson, McLean, Woodard, & Henderson, 2011). However, while most ambulance services also observe lower demand in the afternoon, the call rates tend to drop at a gentler rate. The reason for the sudden drop at 12 p.m. might indeed be related to the topography of Cardiff—because it is a built-up urban area, less city workers tend to drive during their lunch break.

Figure 1 highlights that the approximate methodology is capable of recommending staffing levels that are close to the exact WAST requirements generated by the numerical methodology, but often not identical as it overstaffs several periods by one crew. The main reason for the disparity is that SIPP Pri does not account for the effect of the service level in the previous period upon the current period and assumes the system operates in a steady-state fashion throughout each 1-hour period. As such, the methodology fails to recognize that in periods where demand is strictly increasing (e.g., between 6:00 and 12:00), it takes time for the queue to build up to a level great enough to justify the employment of additional staff. Also, between 12:00 and 14:00 SIPP recommends that a constant level of six crews should be employed; but the exact method recognizes that more crews are needed between 12:00 and 13:00 to deal with the backlog of patients who requested assistance during the previous hour that are still awaiting treatment. By assuming the requirements can be generated independently for each period, SIPP estimates that far fewer crews are required for this hour. Furthermore SIPP Pri can be very sensitive to small changes in the arrival rate because the relationship between the arrival rate and number of crews required is effectively a step function; hence a small change in the average hourly arrival rate can make the difference between SIPP Pri recommending the deployment of x or $x + 1$ crews. For example, SIPP Pri recommends that six, seven, and six crews should be deployed between 15:00 and 16:00, 16:00 and 17:00, and 17:00 and 18:00, respectively, whereas the exact method recognizes that during 15:00–16:00 considerably more than 95% of patients are responded to within the target response time, so less crews are required for the following period as there is less congestion in the system at its commencement.

Table 1 displays the average root mean square error (RMSE) associated with the staffing requirements generated for each hour of each day for 28-day forecasting horizons by SIPP Pri when compared with the Euler Pri requirements, where periods with an RMSE greater than or equal to 1 are highlighted with bold text. The SIPP Pri approach provides identical results to Euler Pri in $\frac{404}{672} = 60\%$ of cases for July and $\frac{397}{672} = 59\%$ for December. The main problem with SIPP Pri is that it overstaffs a number of periods in this case study. For example, it overestimates $\frac{227}{672} = 34\%$ of the hourly periods for July (although never by more than a single crew), and underestimates $\frac{41}{672} = 6\%$ of the hourly periods (18 of which are underestimated by a single crew, 11 by two crews, 10 by three crews, and 2 by four crews). Table 1 reveals that SIPP Pri predominantly fails to produce reliable requirements for the 08:00–09:00 and 12:00–13:00 periods, for the reasons detailed above.

Table 1: SIPP Pri average accuracy (July/December)

Hour	0	1	2	3	4	5	6	7	8	9	10	11
$\lambda_H + \lambda_L$ (July)	4.8	4.5	3.8	3.1	2.2	1.9	3.2	3.8	5.4	6.9	7.7	7.5
RMSE (July)	0.37	0.42	0.33	0.46	0.53	0.53	0.76	0.85	**1.00**	0.98	0.93	0.63
$\lambda_H + \lambda_L$ (December)	5.1	4.8	4.0	3.3	2.3	2.0	3.5	4.2	5.9	7.4	8.3	8.1
RMSE (December)	0.62	0.38	0.33	0.65	**1.05**	0.60	0.94	0.73	**1.10**	**1.05**	0.82	0.65
Hour	12	13	14	15	16	17	18	19	20	21	22	23
$\lambda_H + \lambda_L$ (July)	2.9	3.1	3.1	3.0	3.1	3.0	3.2	4.9	4.9	5.3	5.1	5.0
RMSE (July)	**2.43**	0.65	0.00	0.38	0.46	0.60	0.00	**1.00**	0.53	0.57	0.50	0.19
$\lambda_H + \lambda_L$ (December)	3.2	3.5	3.5	3.4	3.4	3.3	3.5	5.3	5.3	5.7	5.4	5.3
RMSE (December)	**2.34**	0.38	0.50	0.42	0.42	0.19	0.00	0.87	0.65	0.42	0.38	0.33

Notes: SIPP Pri = Stationary Independent Period by Period (to a priority queue); RMSE = root mean square error.

Table 2: Run times required to execute SIPP Pri and Euler Pri for various forecasting horizons

Forecasting Horizon	SIPP Pri (Minutes)	Euler Pri (Minutes)
7 days	0.3	10
28 days	4	40
3 months	10	120

A final consideration that must be taken into account when comparing the staffing algorithms is the computational cost involved in executing each method. When executed on a 3 GHz machine with 2.96 GHz of RAM, the staffing functions output by SIPP Pri and Euler Pri took around 4 and 40 minutes to be generated, respectively, for this case study. Table 2 illustrates that the performance of SIPP is not overly sensitive to the selected time length—it typically takes around four times longer than Euler Pri to generate a set of minimum staffing requirements, but has the advantage that the results output are accurate. Hence, when deciding which method to execute, analysts should consider the importance of obtaining an accurate forecast as opposed to a quick approximation. For example, WAST planners commended both methods as being significantly quicker than the several days it would typically take an analyst to compute the requirements using spreadsheet models. They viewed the extra effort involved in computing the accurate Euler Pri requirements to be small compared to the importance of obtaining accurate staffing requirements, because a prompt ambulance response can make the difference between life and death.

In situations where accuracy is of upmost importance, we recommend that Euler Pri should always be selected, because the approximate methods will always

be susceptible to a certain degree of error. However, because Euler Pri takes around four times longer to run on a standard office computer; it would be worth investigating the potential of developing a hybrid approach to generate staffing requirements in future work, where the SIPP Pri outputs could be used as initial staffing quantities to be optimized using a Euler solver.

CONCLUSIONS

Using a modification MDCTMC model, this article has presented a tractable approach to model and analyze complex time-dependent dual-class queueing systems. While Green and Soares (2007) have previously described how the virtual waiting time distribution may be computed in $M(t)/M/s(t)/FIFO$ queues and Chen and Henderson (2001) have provided an approximate solutions for $M(t)/M/s(t)/NPRP$ service systems, this research represents the first time that numerical methodology has been developed to accurately track the probability of an excessive wait for two customer classes over time, which allows reliable minimum staffing requirements to be generated for situations in which the SIPP Pri approximations are poor. Moreover, we have outlined the instantaneous transitions necessary to apply to the state probability vector and waiting time formulas to incorporate the effect of both full workforce turnovers and small staffing changes. In presenting methods that accurately evaluate the probability of an excessive wait for two customer classes and extending the analysis to cover the commonly occurring case in which not all servers leave at the same time, the analysis greatly contributes to the analysis of realistic service systems.

The benefit of the theory that has been derived within this article is that it may be embedded in approximate and numerical techniques, to set staffing levels throughout the day in multiserver priority systems subject to time-dependent demand as illustrated with the case study in section "Illustrative Case Study." While the methodology could be applied to a range of services, it has to date been presented to the clinical research and development (R&D) manager at WAST, who commented that "The work is an extremely relevant contribution to implementing policy and procedural changes at WAST."

References

Askin, Z., Armony, M., & Mehrota, V. (2007). The modern call center: A multidisciplinary perspective on operations management research. *Production and Operations Management*, *16*(6), 665–688.

Bondi, A., & Buzen, J. (1984). The response times of priority classes under preemptive resume in $M/G/m$ queues. *Sigmetrics* (August), *12*(3), 195–1201.

Buffa, E., Cosgrove, M., & Luce, B. (1976). An integrated work shift scheduling system. *Decision Sciences*, *7*, 620–630.

Chen, B., & Henderson, S. (2001). Two issues in setting call centre staffing levels. *Annals of Operations Research*, *108*, 175–192.

Dietz, D. (2011). Practical scheduling for call center operations. *Omega*, *39*, 550–557.

Feng, W., Kowada, M., & Adachi, K. (2001). Analysis of a multiserver queue with two priority classes and (M, N)-threshold service schedule II: Preemptive priority. *Asia-Pacific Journal of Operational Research*, *18*, 23–34.

Gail, H., Hantler, S., & Taylor, B. (1988). Analysis of a non-preemptive priority multiserver queue. *Advances in Applied Probability*, *20*, 852–879.

Gail, H., Hantler, S., & Taylor, B. (1992). On a preemptive Markovian queue with multiple servers and two priority classes. *Mathematics of Operations Research*, *17*(2), 365–391.

Ganguly, S., Lawrence, S., & Prather, M. (2014). Emergency department staff planning to improve patient care and reduce costs. *Decision Sciences*, *45*(1), 115–145.

Green, L., Kolesar, P., & Soares, J. (2001). Improving the SIPP approach for staffing service systems that have cyclic demands. *Operations Research*, *49*, 549–564.

Green, L., Kolesar, P., & Soares, J. (2003). An improved heuristic for staffing telephone call centers with limited operating hours. *Production and Operations Management*, *12*(1), 46–61.

Green, L., Kolesar, P., & Whitt, W. (2007). Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, *16*, 13–39.

Green, L., & Soares, J. (2007). Computing time-dependent probabilities in $M(t)/M/s(t)$ queuing systems. *Manufacturing & Service Operations Management*, *9*, 54–61.

Gross, D., & Harris, C. (1998) . *Fundamentals of queueing theory* (3rd ed.). New York: Wiley.

Heyman, D. P. & Whitt, W. (1984). The asymptotic behaviour of queues with time-varying arrival rates. *Journal of Applied Probability*, *21*, 143–156.

Ingolfsson, A. (2005). Modelling the $M(t)/M/s(t)$ queue with an exhaustive discipline. *Thinking Beyond the Old 80/20 Rule. Call Center Magazine*, *15*, 54–56.

Ingolfsson, A., Akhmetshina, E., Budge, S., Li, Y., & Wu, X. (2007). A survey and experimental comparison of service-level-approximation methods for nonstationary $M(t)/M/s(t)$ queueing systems with exhaustive discipline. *INFORMS Journal on Computing*, *19*(2), 201–214.

Izady, N. (2010). *On queues with time-varying demand*, PhD thesis, Lancaster University Management School.

Izady, N., & Worthington, D. (2012). Setting staffing requirements for time dependent queueing networks: The case of Accident and Emergency departments. *European Journal of Operational Research*, *219*(3), 531–540.

Johnson, N., Kotz, S., & Kemp, A. (1993). *Univariate discrete distributions*. New York, NY: John Wiley & Sons.

Kao, E., & Narayanan, K. (1990). Computing steady-state probabilities of a non-preemptive multiserver queue. *Journal on Computing*, *2*(3), 211–218.

Kao, E., & Wilson, S. (1999). Analysis of nonpreemptive priority queues with multiple servers and two priority classes. *European Journal of Operational Research*, *118*, 181–193.

Matteson, D., McLean, M., Woodard, D., & Henderson, S. (2011). Forecasting emergency medical service call arrival rates. *The Annals of Applied Statistics*, *5*(2B), 1379–1406.

Mohan, S., Alam, F., Fowler, J., Gopalakrishnan, M., & Printezis, A. (2014). Capacity planning and allocation for web-based applications. *Decision Sciences*, *45*(3), 535–567.

Ngo, B., & Lee, H. (1990). Analysis of a preemptive priority $M/M/c$ model with two types of customers and restriction. *Electronic Letters*, *26*, 1190–1192.

Ren, Z. J., & Zhou, Y.-P. (2008). Call center outsourcing: Coordinating staffing level and service quality. *Managment Scence*, *54*(2), 369–383.

Vile, J., Gillard, J., Harper, P., & Knight, V. (2012). Predicting ambulance demand using singular spectrum analysis. *Journal of the Operational Research Society*, *63*(11), 1556–1565.

Wagner, D. (1997). Waiting time of a finite-capacity multi-server model with non-preemptive priorities. *European Journal of Operational Research*, *102*, 227–241.

Welsh Government. (2012). Ambulance services in Wales: September 2012. Technical Report. SDR 187/2012.

Whitt, W. (2008). What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics*, *54*, 476–484.

Zeephongsekul, P., & Bedford, A. (2006). Waiting time analysis of the multiple priority dual queue with a preemptive priority service discipline. *European Journal of Operational Research*, *17*, 886–908.

Julie Vile is a performance analysis manager at the NHS Wales Delivery Unit. Having previously worked as a lecturer at Cardiff School of Mathematics and an embedded mathematical modeler at Aneurin Bevan University Health Board, she is passionate about bridging the gap between academic theory and practice. She currently manages a portfolio of modeling and data analysis projects for the NHS, is chair of the South Wales Operational Research Society, and provides specialist advice/training on complex statistical information to senior health care professionals.

Jonathan Gillard is a senior lecturer in statistics at Cardiff University. After completing a PhD in the area measurement error models, he broadened his research interests into the area of low rank approximation. He maintains an interest in developing novel mathematics for applied and interdisciplinary mathematics.

Paul Harper is professor of operational research and deputy head of the School of Mathematics, Cardiff University. He is also director of the Health Modelling

Centre Cymru (hmc2), a pan-Wales center for modeling in healthcare, and director of engagement for mathematics. His research interests are primarily in OR modeling and stochastic methods applied to health care systems, and he has been an investigator on in excess of £6 million of funding from various research councils and direct from the health service. He is an editor for the journal *Health Systems* (Palgrave Macmillan), author of more than 70 peer-reviewed papers and book chapters, and a fellow of the Learned Society of Wales. In 2015, he was awarded the UK Times Higher Education prize for outstanding Contribution to Innovation and Technology. www.profpaulharper.com

Vincent Knight is a lecturer of operational research at Cardiff University. His research interests are in stochastic modeling and game theory. His particular interests include the modeling of strategic queuing behavior as well as health care systems.