Staffing and scheduling under nonstationary demand for service: A literature review

Mieke Defraeye* and Inneke Van Nieuwenhuyse
Research Center for Operations Management
Department of Decision Sciences and Information Management
KU Leuven, Belgium
mieke.defraeye@kuleuven.be
inneke.vannieuwenhuyse@kuleuven.be

April 6, 2015

Abstract

Many service systems display nonstationary demand: the number of customers fluctuates over time according to a stochastic—though to some extent predictable—pattern. To safeguard the performance of such systems, adequate personnel capacity planning (i.e., determining appropriate staffing levels and/or shift schedules) is often crucial. This article provides a state-of-the-art literature review on staffing and scheduling approaches that account for nonstationary demand. Among references published during 1991-2013, it is possible to categorize relevant contributions according to system assumptions, performance evaluation characteristics, optimization approaches and real-life application contexts. Based on their findings, the authors develop recommendations for further research.

Keywords: Nonstationary arrival process, Staffing and scheduling, Personnel planning, Performance evaluation, Capacity analysis, Optimization

1 Introduction and scope

In most service systems, staffing drives both costs and service quality. Personnel capacity planning for these systems tends to be non-trivial though, due to the many sources of variability inherent in real-life service systems (e.g., nonstationary demand, stochastic service times, different customer classes) and phenomena like customer abandonment, balking, retrials etc. The personnel capacity planning process usually gets decomposed into four steps [1, 2, 3, 4, 5, 6, 7]:

1. Forecasting demand (based on empirical data).
2. Determining staffing requirements: The staffing levels required over time are selected, in order to meet a specific performance target at minimal cost.
3. Shift scheduling: This step determines how many workers to assign to each shift type, in order to cover the staffing requirements.
4. Rostering: In this final step, employees are assigned to shifts.

*Corresponding author.

Short-term schedule updates may represent an additional step [2, 6, 8] (for an overview and analysis of available methods for online shift updating, see Hur et al. [9], Mehrotra et al. [10], and Testik et al. [11]). Because our goal with this literature review is to provide a state-of-the art overview of research on staffing and personnel scheduling in systems with nonstationary demand, we focus on steps 2 and 3, and consider steps 1 and 4 beyond the scope of this review[1].

The practical relevance of this research field can hardly be overestimated. In many real-life systems (e.g., call centers, emergency departments, toll booths), nonstationary demand is prominent, and appropriate staffing is often the only way to safeguard customer service in these systems. Despite this practical relevance, time-varying arrival rates often do not receive sufficient attention in real-life personnel capacity planning [22, 23, 24].

This research field has grown rapidly in the past two decades. We focus on the period 1991-2013, selecting 62 articles that *focus on personnel staffing and/or scheduling* and that *specifically target systems with nonstationary demand (i.e., stochastic with a time-varying rate)*. Table 1 gives an overview of the selected articles. We categorize these based on four classification criteria: system assumptions, performance evaluation characteristics, optimization approaches and real-life application context. We did not include in the categorization articles that present general staffing or scheduling algorithms for deterministic demand (as in [25, 26, 27, 28]), scheduled demand [29], and/or non-time-varying systems [30, 31, 32, 33]. We also exclude articles that focus solely on scheduling algorithms, with assumptions of exogenous staffing requirements (as in the early work of Dantzig [34] and Keith [35]; see, e.g., Van den Bergh et al. [41] for a recent, general review of scheduling algorithms), and manuscripts that centered on other types of resources (such as hospital beds; [36]).

Time range	Number of articles	References
1991 - 1995	4	[58], [167], [168], [196]
1996 - 2000	10	[42], [66], [93], [94], [103], [169], [181], [186], [197], [204]
2001 - 2005	10	[60], [74], [78], [90], [178], [184], [207], [185], [198], [201]
2006 - 2010	25	[23] [55], [61], [62], [63], [64], [68], [75], [76], [77], [79], [91], [92], [109], [110], [150], [165], [177], [179], [188], [190], [200], [202], [203], [205]
2011 - 2013	13	[53], [54], [56], [65], [84], [95], [160], [166], [176], [180], [191], [199], [206]

Table 1: Categorized articles

This overview differs in some key respects from previously published review articles in this field. For example, Gans et al. [6] and Aksin et al. [15] present surveys that specifically target call centers, discussing not only staffing problems but also various other operational problems related to this specific application area. Our review focuses solely on staffing and scheduling for nonstationary demand systems, and we discuss the relevance of different models to various application areas. Green et al. [37] and Whitt [38] offer an extensive overview of methods for staffing with nonstationary demand for service, but the methods they propose rely largely on stationary approximations (see Section 4.1) and do not include shift scheduling. Ernst et al. [39, 40] and more recently Van den Bergh et al. [41] provide comprehensive reviews of research on scheduling and rostering, but do not specifically focus on methods for nonstationary demand. We consider both staffing and scheduling, in settings with nonstationary demand for service.

[1] See [6, 12, 13, 14, 15, 16, 17], for issues related to demand forecasting. A more elaborate discussion of the rostering problem can be found in [18, 19, 20, 21], among others.

The remainder of this article is organized as follows: Section 2 describes the classification scheme used to categorize the literature. The following sections then provide an in-depth discussion of each classification criterion. Section 3 features the classification of the articles in accordance with the system assumptions, and Section 4 outlines the evaluation methods for system performance. Because performance evaluation is necessary to evaluate proposed solutions and guide the search for better solutions, it is a highly relevant subroutine in any staffing or shift scheduling approach. We offer an overview of the optimization methodologies in Section 5, then classify the articles on the basis of the suggested real-life application areas in Section 6. Finally, Section 7 contains the conclusions and identifies promising directions for further research.

2 Overview of classification criteria

Figure 1 displays a simple representation of a (single-stage) service system with nonstationary demand. Customers arrive according to a nonstationary arrival process with time-varying arrival rate λ_t (where t represents time). Typically, the arrival pattern repeats over a given cycle (e.g., day, week, month, year). The service process (with per server service rate μ) starts immediately if a server is available on arrival; otherwise, the customer joins the queue. The aggregate service rate (denoted $s_t\mu$) can be influenced by changing s_t, the number of servers available at time t. The per server service rate μ is commonly assumed to be constant, though some models allow for time-varying service rates (e.g., [42]).

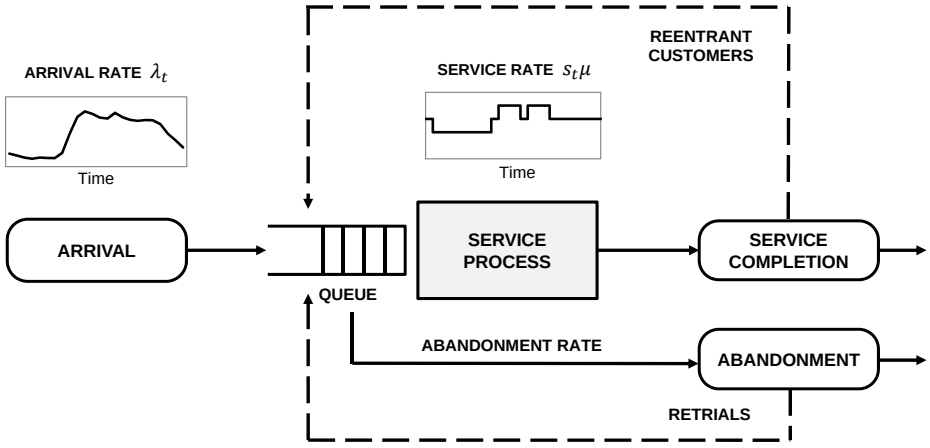

Figure 1: Schematic representation of a single-stage queueing system with nonstationary demand.

In many service systems, customers may opt to abandon by leaving the queue without being served; they are referred to as *abandonments* (or *left without being seen* [LWBS], in a healthcare context). Long waiting times are the main reason for customers to abandon (Johnson et al. [43] report that almost 77% of LWBS patients in an emergency department claim to abandon because of long waiting times). Although abandonments are undesirable from a customer service perspective, they tend to have a positive effect on system stability, especially when the system is temporarily overloaded (e.g., [37]). The abandoned customers may reenter the service system later: *retrials* refer to customers that abandoned previously either upon arrival (because the queue was too long [44, 45]), or after experiencing a positive waiting time [46]. If there are no retrials, ignoring abandonment behavior tends to cause overstaffing, implying higher labor

costs. Ignoring retrials, on the other hand, will tend to cause understaffing, in particular when retrial rates are high. Note that serviced customers may also reenter the queue if they need to be serviced several times by the same server (*reentrant customers*, see [47, 48]).

Classifier	Features	Notation
System assumptions	Kendall notation	A/B/C/K+L with A = distribution of the arrival process, B = distribution of service process, C = number of servers, K = maximum number of jobs that can be in the system, either waiting or in service, L = distribution of the abandonment process.
	Homogeneity of customers/servers	HO = homogenous and HE = heterogenous.
	Staffing interval	Y = yes; N = no.
	Queueing policy	FIFO = first-in first-out; SBR = skill-based routing; Priority = queueing based on customer priority.
	Service policy	E = exhaustive; P = preemptive; NS = not specified.
	System structure	S = single stage; N = network.
	Parameter uncertainty	Y = yes; N = no.
	Retrials	Y = yes; N = no; NS = not specified.
	Reentrant	Y = yes; N = no; NS = not specified.
Performance evaluation	Methodology	
	Performance metrics	see Table 5
Optimization approach	Methodology	
	Objective	
	Constraints	
Real-life application context	Context	
	Implementation (+results reported)	Y=yes; N=no.
	Validation by means of real-life data	Y=yes; N=no.
	Validation by means of other (fictive) examples	Y=yes; N=no.

Table 2: Overview of classifiers, features and notation

It is possible to classify previous publications by the criteria listed in Table 2: system assumptions, performance evaluation characteristics, optimization approaches and real-life application context.

For the system assumptions classifier, we rely on the commonly used Kendall notation [49] to reflect any assumptions regarding the arrival and service processes in the system. Heyman and Whitt [50] were among the first to add the notation "t" to represent the time-dependent nature of the arrival process; the notation for customer abandonments was introduced by Baccelli and Hebuterne [51]. For example, the $M_t/G/s_t/K + G$ notation represents a system with time-varying Poisson arrivals (M_t), a general service time distribution (the first G), time-varying staffing levels s_t, a limit on the maximum number of jobs that can be in the system (K), and abandonments that follow a general distribution (the last G). The parameter K can be equal to infinity; in that case, it is typically not shown. Other relevant features are the homogeneity of customers and/or servers, the presence of staffing intervals, the queueing discipline, the service

discipline, the structure of the system, and parameter uncertainty. Customers are heterogenous if the system takes different customer classes into account (e.g., due to differences in process steps, service times, or queueing discipline [52]); if only a single customer class is considered, customers are homogenous. Servers are homogenous if they all exhibit the same skills (i.e., can all handle the same types of customers at the same rate) and have the same service rate; otherwise, they are heterogenous servers.

A common assumption is that capacity changes can be made only at specific points in time; the time period during which capacity remains constant is the staffing interval. The staffing interval length can vary: e.g., Defraeye and Van Nieuwenhuyse [53] use an interval length of 15 minutes in their computational results, whereas Izady and Worthington [54] use intervals of 30 minutes or 1 hour (the methods can equally be applied to other staffing interval lengths).

The queueing policy refers to the sequence in which customers are serviced; first-in first-out (FIFO) is by far the most frequently used queueing discipline in the articles we survey, though priority-based rules are also common, particularly in the context of emergency services (e.g., priority based on the urgency of a patient's condition). The service policy reflects what happens to a customer in service when a server is scheduled to leave. Many existing models implicitly assume a preemptive service discipline, such that service is interrupted and the customer in service rejoins the queue. Under the (more realistic) exhaustive service policy, the customer service instead gets completed before the server leaves, even if this means that a server has to work beyond his or her scheduled time.

For system structure, we distinguish between systems that contain only a single service step (single-stage models) and those that contain multiple service steps (networks). Next, we check whether the model accounts for parameter uncertainty. The use of stochastic arrival rates, service rates, and abandonment rates requires an estimation of the distributional parameters, which might introduce error into the models (and cause the desired performance target to be violated). Accounting for this parameter uncertainty during the personnel capacity planning process can significantly improve the staffing solutions (though possibly at a higher staffing cost; [55, 56]). Finally, we report whether the model accounts for retrials (only relevant in case of abandonments) and/or for reentrant customers.

For the performance evaluation classifier, we categorize prior contributions according to the methodology used to evaluate the performance of a given personnel allocation, that is, given s_t values. We provide key references for each evaluation method. In addition, we list the performance metrics and discuss which metrics are most common in practice, in different application contexts.

By considering the optimization approach, we can categorize contributions according to the methodology used to optimize personnel capacity, along with the objective and the constraints. Models that vary s_t without taking into account shift requirements (e.g., shift patterns, shift durations) are staffing models (they result in *staffing requirements*); otherwise, they are shift scheduling models. We distinguish three approaches for shift schedule optimization: the two-step, the feedback-based, and the direct approach (a detailed discussion of these approaches is given in Section 5.2).

Finally, for the real-life application category, we classify articles on the basis of their application context, as suggested by the authors, as well as according to evidence of real-life implementation, validation using real-life data, or validation using other (fictive) examples.

3 Classification by system assumptions

Table 3 displays the literature classification based on the system assumptions. These assumptions are often linked with the choice of a performance evaluation method and/or capacity

Homogeneity customers / servers	Kendall notation	References	Staffing interval (Y/N)	Queueing policy	Service policy (E/P)	System structure (S/N)	Parameter uncertainty (Y/N)	Retrials	Reentrant
HO / HO	$M_t/M/s_t$	[196]	Y	FIFO*	NS	S	N	-	N
		[60]	Y	FIFO*	NS	S	N	-	N
		[63]	Y	FIFO*	P	S	N	-	N
		[62]	Y	FIFO*	P	S	N	-	N
		[177]	Y	FIFO*	NS	S	N	-	N
		[198]	Y	FIFO*	NS	S	N	-	N
		[206]	Y	FIFO	NS	S	Y	-	N
		[199]	Y	FIFO	P/E	S	Y	-	N
		[58]	Y	FIFO*	E	S	N	-	N
		[181]	Y	FIFO	NS	S	N	-	N
	$M_t/M/s_t + M$	[180]	Y	FIFO*	NS	S	N	N	N
		[191]	Y	FIFO*	NS	S	Y	N	N
		[103]	Y	FIFO*	NS	S	N	N	N
		[95]	Y	FIFO	P	S	N	N	N
		[55]	Y	FIFO*	NS	S	Y	N	N
		[90]	Y	FIFO*	NS	S	N	N	N
		[205]	Y	FIFO*	NS	S	N	N	N
	$M_t/M/s_t/K$	[185]	Y	FIFO*	NS	S	N	-	N
	$M_t/M/s_t/K + M$	[110]	N (+Y)	FIFO*	NS	S	N	N	N
	$M_t/M/s_t/K + G$	[168]	Y	FIFO*	NS	S	N	Y	N
	$M_t/M/s_t + G$	[167]	Y	FIFO*	NS	S	N	Y	N
	$M_t/G/s_t$	[178]	Y	FIFO	E	S	N	-	N
		[179]	Y	FIFO	E	S	N	-	N
		[197]	N	FIFO	NS	S	Y	-	N
	$M_t/G/s_t + G$	[53]	Y	FIFO	E	S	N	N	N
		[109]	N	FIFO*	NS	S	N	N	N
		[61]	N (+Y)	FIFO	P (+ E)	S	N	N	N
	$G_t/M/s_t$	[204]	Y	FIFO*	NS	S	N	-	N
	$G_t/G/s_t$	[176]	Y	FIFO*	NS	S	N	-	N
	$G_t/G_t/s_t$	[42]	Y	FIFO*	NS	S	N	-	N
	$G_t/G/s_t + G$	[84]	Y	FIFO*	P/E	S	N	N	N
		[160]	N	FIFO	NS	S	N	N	N
	Not specified	[68]	Y	FIFO*	NS	S	N	N	N
		[91]	Y	FIFO*	NS	S	N	N	N
		[186]	Y	FIFO*	NS	S	N	NS	NS
		[92]	Y	FIFO*	NS	S	N	NS	NS
		[93]	Y	FIFO*	NS	S	N	NS	NS
		[94]	Y	FIFO*	NS	S	N	NS	NS
		[166]	Y	FIFO*	NS	S	N	N	N
HO/HE	$M_t/M/s$	[200]	-	FIFO*	NS	S	Y	-	N
	$M_t/M/s_t$	[201]	Y	FIFO*	NS	S	Y	-	N
	$M_t/G/s_t + G$	[66]	Y	FIFO	NS	S	N	N	N
	$M_t/G/s_t/K + G$	[169]	Y	FIFO*	NS	S	N	Y	N
HE/HE	$M_t/M/s_t$	[202]	Y	SBR	NS	S	N	-	N
		[56]	N	FIFO	NS	S	Y	-	N
		[65]	Y	Priority	NS	N	N	-	N
	$M_t/M/s_t/s_t$	[79]	Y	FIFO	NS	S	N	-	N
	$M_t/M/s + M$	[75]	-	SBR	E	S	Y	N	N
		[76]	-	SBR	P	S	Y	N	N
		[74]	-	SBR	P	S	Y	N	N
	$M_t/M/s_t + M$	[78]	Y	SBR	P	S	Y	N	N
		[77]	Y	SBR	NS	S	Y	N	N
		[165]	Y	SBR	NS	S	Y	N	N
		[188]	Y	FIFO / priority	NS	S	N	N	N
		[190]	Y	SBR	NS	S	N	N	N
	$M_t/G/s$	[64]	-	Priority*	NS	N	N	-	N
	$M_t/G/s_t$	[207]	Y	FIFO*	NS	N	N	-	N
		[54]	Y	Priority	E	N	N	-	N
		[203]	Y	FIFO*	NS	S	N	-	N
	$M_t/G/s_t + G$	[150]	Y	SBR	P	S	Y	Y	Y
	Not specified	[184]	Y	SBR*	NS	N	N	NS	NS
		[23]	Y	Priority*	NS	N	N	NS	NS

* = assumed, not stated explicitly in article; - = not relevant; (·) = briefly described, such as an extension.

Table 3: Classification by system assumptions

optimization approach, as discussed further in Sections 4 and 5, respectively.

A large majority of extant studies assume that both customer types and server types are homogenous and that the system consists of a single stage. More recent work has shifted this emphasis toward models that include both customer and server heterogeneity (albeit mainly with exponential assumptions on the service and abandonment time distribution, see Table 3), as is further detailed below. The few articles that consider a service network, assume that customers and servers are heterogenous; none of these studies include abandonments.

It is worthwhile to explore in further detail the classification according to Kendall notation, irrespective of the other assumptions, as in Figure 2[2]. It shows that the large majority of contributions have focused on systems with time-varying number of servers. Among these, the $M_t/M/s_t$ model can be considered as a "base" model, which can then be extended by including abandonments, limiting system size, and/or changing exponential distribution assumptions into general distributions. The figure highlights that the inclusion of Poisson abandonments has received considerable attention, while the extension towards general distributions is somewhat less common (because performance evaluation then becomes more complex). An overwhelming majority of articles assumes a nonstationary Poisson arrival process; Kim and Whitt [57] find that this assumption is consistent with empirical arrival processes observed in call centers and emergency departments. Daily recurring demand patterns typically display one to three peaks per day [58, 59]. Authors frequently resort to sine functions to generate demand rate profiles for their computational experiments: see for example Green et al. [60], Liu and Whitt [61] (only one peak per cycle) and Ingolfsson et al. [62], Green et al. [59] (two peaks per cycle). The applicability of the staffing and scheduling models, however, does not depend on the use of the sine function. Many methods actually assume that the arrival rate is constant over the staffing interval (e.g., [60]) or over a shorter calculation interval (e.g., [62]), and therefore average the arrival rate over that interval (a more restrictive approach instead considers the maximum arrival rate over the staffing interval, [60]). This is reasonable because real-life data are often available only on an aggregate basis, e.g., per hour or half hour [54, 63, 64, 65].

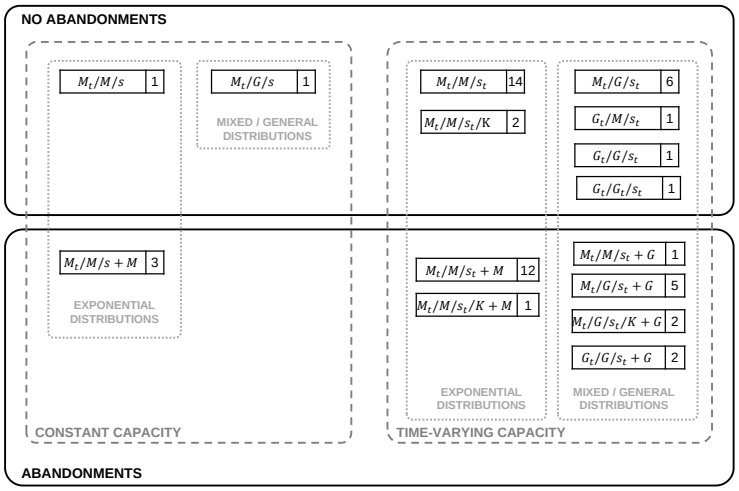

Figure 2: Classification based on Kendall notation (number of articles).

Figure 2 also reveals that a majority of published articles assume the service process is

[2]Note that the $M_t/M/s_t/s_t$ model is a special case of the $M_t/M/s_t/K$ model; the $M_t/M/s_t/s_t$ model is thus not shown separately in the figure.

exponentially distributed. Zeltyn et al. [65] and Hueter and Swart [66] largely validate this assumption using empirical data for an emergency department and restaurant setting, while non-exponential service time distributions have been reported in a call center context (e.g., Brown et al. [67] report a lognormal distribution and Castillo et al. [68] report Erlang distributed service times). Abandonments, if included at all, are also commonly assumed to follow an exponential distribution. It is known that, in systems with abandonments, the impact of the exact choice of the service and abandonment distributions depends on the system utilization. In *stationary* systems the service time distribution is more important than the patience time distribution when the systems are critically loaded [69, 70], and the patience time distribution is more important than the service time distribution when the systems are overloaded [71, 72]. Chassioti et al. (2014) have also shown that, in overloaded systems with nonstationary demand and abandonment, the time-varying expected number in system is relatively insensitive to the service time distribution. Nevertheless, as shown by Davis et al. (1995), the impact of the service time distribution on the blocking probability in a moderately loaded $M_t/PH/s/0$ queue is significant.

Table 3 shows that the queueing policy is predominantly FIFO; we find evidence of priorities or skill-based routing (SBR) only when both customers and servers are heterogenous. In practice, the use of priorities is common particularly in health care settings [54, 65], whereas call center models mostly rely on skill-based routing, which impacts the sequence in which customers receive service[3]. Accounting for customer routing adds complexity to the personnel capacity decision process, in the sense that the system's performance depends on not only staffing (or scheduling) decisions, but also on routing decisions. Harrison and Zeevi [74], Bassamboo and Zeevi [75, 76], and Bertsimas and Doan [77], among others, propose methods to solve the staffing and (dynamic) routing problems in call centers with heterogeneous servers and customers. Bassamboo and Zeevi [78] extend their previous work [75] by including admission control decisions. In the ambulance crew scheduling problem of Erdoğan et al. [79], customers and servers are heterogeneous because they differ in spatial location (the authors apply the Approximate Hypercube model of Budge et al. [80] to integrate this aspect in their model).

Many articles fail to provide details on the service policy being applied. According to Ingolfsson et al. [81], extensive literature (implicitly) assumes a preemptive service discipline, whereas in many real-life settings, the service policy is inherently exhaustive [81, 82, 83, 84]. The service policy is likely to have a large impact if the average service time is relatively long compared to the staffing interval: the amount of overtime evidently depends on the service time of the customer being in service, while shorter staffing intervals imply that capacity decreases (which potentially initiate overtime) are more frequent. The effect of the service policy will be less prominent in systems with low average utilization though, because servers are then more likely to be idle at the end of their shift.

Table 3 also shows that most articles do not account for parameter uncertainty[4] or network settings. As shown by Table 4 (that details those system assumptions for which we observed an evolution over time) these two phenomena have only appeared very recently in the literature. The inclusion of heterogenous customer and service setting (HE/HE) also gained attention in recent years. Accounting for parameter uncertainty during the personnel capacity planning process can lead to significant reductions in the total expected cost (which generally includes, besides the personnel cost, a penalty for not meeting the performance constraint; [55, 56]). As is evident from Table 3 though, staffing and scheduling models that include parameter uncertainty tend to rely on exponential assumptions for the arrival, service, and abandonment processes.

A final observation from Table 3 is that, while a considerable number of models include some

[3] An overview of problems related to staffing and routing in call centers can be found in [7].

[4] For general references on the impact and implications of parameter uncertainty, see [82, 85, 86, 87, 88, 89].

		1991 - 1995	1996 - 2000	2001 - 2005	2006 - 2010	2011 - 2013
Homogeneity customers / servers	HO/HO	4	8	5	12	10
	HO/HE	0	2	1	1	0
	HE/HE	0	0	4	12	3
Service policy	Preemptive	0	0	2	5	3
	Exhaustive	1	0	1	3	4
	Not specified	3	10	7	18	8
System structure	Single stage	4	10	8	23	11
	Network	0	0	2	2	2
Parameter uncertainty included	Yes	0	1	3	7	4
	No	4	9	7	18	9

Table 4: Trends in system assumptions (number of articles)

type of additional complexity (e.g., by considering non-exponential service and abandonment times, non-homogenous customers or servers, network settings, etc.), we found no articles that address all aspects simultaneously. Moreover, we observe that extensions toward networks of queues and exhaustive service policies are particularly underrepresented in the literature, and present challenging directions for future research.

4 Classification by performance evaluation methods and performance metrics

This section highlights the performance metrics evaluated in each article, and classifies articles according to the methodology used to evaluate system performance for given capacities.

The number of performance metrics actually used is vast, as the overview in Table 5 reveals (this table also clarifies the more concise notation we use in Tables 6, 7 and 8). We distinguish metrics based on number in system/number in queue, waiting time, abandonments/throughput, length of stay, and utilization[5]. In terms of notation, we closely adhere to that introduced in Baron and Milner [32]: We distinguish between metrics taken over the planning horizon (horizon-based, $(\cdot)_{\mathrm{HB}}$), those assessed over a smaller interval such as a staffing interval (interval-based, $(\cdot)_{\mathrm{IB}}$), and instantaneous metrics (time epoch-based, $(\cdot)_{\mathrm{TB}}$). Metrics that are based on per customer performance are represented as $(\cdot)_{\mathrm{CB}}$ (customer-based).

Table 6 contains the performance evaluation metrics and methodologies for the studied articles; it highlights that the performance metrics tend to depend on the application context. Often, specific terminology then applies. In emergency departments, waiting times and length-of-stay (LoS) metrics are most common. Abandonments are commonly referred to as *left-without-being-seen* or LWBS [63]). Call centers tend to focus either on the service level (which is then referred to as the *telephone service factor* or TSF, [55]) or the expected waiting time (*average speed of answer* or ASA, e.g., [90]). The category "other" in Table 6 includes references on personnel scheduling in restaurants [66, 91], crew scheduling for ambulances [79], personnel scheduling in retail stores [92, 93], and scheduling customs staff at airports [94]. Many metrics in these contexts relate to service levels: in ambulance scheduling, the *coverage* (which specifies the probability that the response time lies below a given time limit) is maximized [79]. In retail, on the contrary, a profit-driven approach is common. For instance, Lam et al. [93] consider profit as sales revenue minus personnel cost, and model sales revenue as a function of personnel staffing, customer arrivals, and other factors. Customer service is checked afterwards, by measuring the

[5]We do not explicitly include labor cost as a performance metric, because its calculation is usually straightforward.

Notation	Interpretation
NUMBER IN SYSTEM / QUEUE	
N_t	Number in system at time t
B_t	Number busy servers at time t
Q_t	Queue length at time t
$P_{\mathrm{TB}}(Q \geq q)$	Queue length tail probability
$E_{\mathrm{TB}}[Q]$	Expected number in queue, at time t
$E_{\mathrm{IB}}[Q]$	Expected queue length, over interval $\tilde{t}$
$E_{\mathrm{HB}}[Q]$	Expected queue length, over time horizon T
$\max_{\mathrm{HB}}\{Q\}$	Maximum queue length measured over time horizon T
$E_{\mathrm{HB}}[E_{\mathrm{IB}}[Q]]$	Expected queue length, measured over interval $\tilde{t}$ and averaged over time horizon T
$E_{\mathrm{HB}}[N]$	Expected number in system (in queue and in service) over time horizon T
WAITING TIME	
$P_{\mathrm{TB}}(W > 0)$	Probability of experiencing a positive waiting time, upon arrival at time t
$P_{\mathrm{IB}}(W > 0)$	Probability of experiencing a positive waiting time, upon arrival in interval $\tilde{t}$
$E_{\mathrm{TB}}[W]$	Expected waiting time, at time t
$E_{\mathrm{IB}}[W]$	Expected wait, measured over interval $\tilde{t}$
$E_{\mathrm{HB}}[W]$	Expected waiting time, over time horizon T
$\max_{\mathrm{HB}}\{W\}$	Maximum wait, measured over time horizon T
$E_{\mathrm{HB}}[C_{\mathrm{CB}}(W > 0)]$	Expected cost for positive wait
$E_{\mathrm{HB}}[C_{\mathrm{CB}}(W)]$	Expected cost for length of waiting time
$P_{\mathrm{TB}}(W > \tau)$	Probability of experiencing a waiting time exceeding τ, upon arrival at time t
$P_{\mathrm{IB}}(W > \tau)$	Probability of experiencing a waiting time exceeding τ, upon arrival in interval $\tilde{t}$
$P_{\mathrm{HB}}(W > \tau)$	Probability of experiencing a waiting time exceeding τ, for all arrivals over time horizon T
$E_{\mathrm{HB}}[P_{\mathrm{TB}}(W > \tau)]$	Probability of experiencing a waiting time exceeding τ, upon arrival at time t, averaged over time horizon T
$E_{\mathrm{HB}}[P_{\mathrm{IB}}(W > \tau)]$	Probability of experiencing a waiting time exceeding τ, upon arrival in interval $\tilde{t}$, averaged over time horizon T
$E_{\mathrm{TB}}[W - \tau \vert W > \tau]$	Average excess with regard to maximum allowed waiting time τ
$\min_{\mathrm{HB}}\{P_{\mathrm{TB}}(W > \tau)\}$	Minimal service level over time horizon T
$E_{\mathrm{IB}}[CGOS]$	Expected customer grade of service per interval (utility function based on waiting time)
$E_{\mathrm{HB}}[Coverage]$	Expected aggregated coverage, over time horizon T
$E_{\mathrm{IB}}[Coverage]$	Expected coverage, over interval $\tilde{t}$
ABANDONMENTS / THROUGHPUT	
Ab_t	Abandonment rate, as a function of t
$P_{\mathrm{TB}}(Ab)$	Abandonment probability, as a function of t
$E_{\mathrm{HB}}[\%Ab]$	Average percentage abandoned, over time horizon T
$E_{\mathrm{IB}}[\%Ab]$	Expected percentage abandoned, over interval $\tilde{t}$
$E_{\mathrm{HB}}[E_{\mathrm{IB}}[\%Ab]]$	Expected percentage abandoned, measured over interval $\tilde{t}$ and averaged over time horizon T
Bl_t	Blocking rate, as a function of t
$E_{\mathrm{HB}}[\%Bl]$	Expected percentage blocked, over time horizon T
$E_{\mathrm{HB}}[\%Served]$	Fraction of customers that is served, over time horizon T
$E_{\mathrm{HB}}[C_{\mathrm{CB}}(Ab)]$	Expected abandonment cost, over time horizon T
$E_{\mathrm{HB}}[C_{\mathrm{CB}}(Bl)]$	Expected blocking cost, over time horizon T
$\mathrm{throughput}_t$	Throughput, as a function of t
$E_{\mathrm{IB}}[\mathrm{throughput}]$	Expected throughput over interval $\tilde{t}$
$E_{\mathrm{HB}}[\mathrm{throughput}]$	Expected throughput over time horizon T
LENGTH OF STAY	
$E_{\mathrm{HB}}[LoS]$	Expected length of stay, over time horizon T
$P_{\mathrm{HB}}(LoS < \alpha)$	Probability of experiencing a length of stay exceeding α, over time horizon T
UTILIZATION	
U_t	Utilization, as a function of t
$E_{\mathrm{IB}}[U]$	Expected utilization over interval $\tilde{t}$
$E_{\mathrm{HB}}[U]$	Expected utilization, over time horizon T
$E_{\mathrm{HB}}[E_{\mathrm{IB}}[U]]$	Expected utilization, measured over interval $\tilde{t}$ and averaged over time horizon T
SIT_{HB}	Server idle time, over time horizon T
$E_{\mathrm{TB}}[Busy]$	Expected number of busy servers at time t
Number of hours where $U_{\mathrm{TB}} > u$	Number of hours workload exceeds a certain percentage, over time horizon T
$\max_{\mathrm{HB}}[U]$	Maximum utilization, over time horizon T

Table 5: Overview of performance metrics and compact notation

service availability for the resulting schedule (expressed by the ratio of staff number to traffic). In restaurants, Hueter and Swart [66] aim to limit the expected waiting time and percentage abandoned customers, whereas Choi et al. [91] target a constant ratio of customers to servers (*customer count per server*, or CCS).

Some authors seek to exploit the relation between performance metrics, using metrics that are easy to compute to obtain results for more complex ones. Simply-computed metrics are often sufficient to guide the search for adequate personnel schedules, e.g., Izady and Worthington [54] apply analytic results related to delay probability to determine shift schedules that meet a length-of-stay target in an emergency department. Similarly, Green et al. [63] focus on a service level (at most 20% of patients wait more than 1 hour) to realize a reduction in the percentage LWBS. Kim and Ha [95] impose an upper bound on the number of customers in the call center, which is used as a proxy metric to control the expected waiting time, the delay probability and the service level. Exploring the relationships across different performance metrics in complex nonstationary systems may open up interesting opportunities for further research, particularly for performance metrics that are difficult to compute.

We elaborate on the performance evaluation methodologies in the following sections. Section 4.1 describes how stationary models can be applied to estimate performance in systems with a nonstationary arrival process. Section 4.2 discusses discrete-event simulation and 4.3 addresses numerical methods (such as randomization and discrete-time modeling). Fluid approximations are described in Section 4.4. Section 4.5 briefly elaborates on how empirical data have been used for performance evaluation. For a good overview of models with accompanying methods, we provide a table in Appendix that integrates the main elements of Table 3 (system assumptions) and Table 6 (methodologies used). It shows that discrete-event simulation is used regardless of the system assumptions, while fluid models tend to be popular in the HE/HE category with Markovian assumptions for the service and abandonment process. Also, none of the selected articles applied numerical or empirical models in the HE/HE category.

4.1 Stationary approximations

As Table 6 shows, stationary approximations are by far the most widely adopted approach for performance evaluation in nonstationary systems. These approaches translate the nonstationary system parameters into stationary counterparts, which they feed into a (series of) stationary model(s). Various methods have been suggested; for detailed descriptions, we refer readers to Green et al. [37], Whitt [38] and Defraeye and Van Nieuwenhuyse [96]. Here, we limit ourselves to a brief discussion.

The pointwise stationary approximation (PSA; [97, 98, 99]) uses the instantaneous arrival rate λ_t at each time t in a separate stationary model. The underlying assumption here is that the steady-state is realized almost immediately, which can be the case only if the number of arrivals and service completions is sufficiently high relative to the frequency and magnitude of the arrival rate fluctuations [99]. In a stationary independent period-by-period approach (SIPP, [60]), a separate stationary model instead gets applied to each discrete time interval, with the average arrival rate as the input parameter. Green et al. [60] present extensions to the SIPP approach, such as Lag SIPP, in which the arrival rate shifts by an amount of time proportional to the expected service time [100, 101]. This approach complies with the observation that in nonstationary systems, peaks in system congestion lag behind the arrival rate peaks [58, 102], as is commonly referred to using terms such as time lag or congestion lag. A lagged variant of PSA can be applied similarly [101]. Accounting for this lag can greatly improve the accuracy of SIPP (and PSA), particularly when the average service time —and thus the time lag—is

Context	References	Stationary approximation	Discrete event simulation	Numerical methods	Fluid model	Empirical	Performance metrics related to ...				
							Number in system / queue	Waiting time	Abandonment / throughput	LoS	Utilization
General	[84]	IS						$P_{\mathrm{TB}}(W>0)$, $E_{\mathrm{TB}}[W]$, $P_{\mathrm{TB}}(W\leq\tau)$, $E_{\mathrm{TB}}[W-\tau\|W>\tau]$, $\min_{\mathrm{HB}}\{P_{\mathrm{TB}}(W>\tau)\}$, $E_{\mathrm{HB}}[P_{\mathrm{TB}}(W>\tau)]$	$P(Ab_t)$		
	[68]		x				$E_{\mathrm{HB}}[Q]$, $\max_{\mathrm{HB}}\{Q\}$	$E_{\mathrm{HB}}[W]$, $\max_{\mathrm{HB}}\{W\}$, $P_{\mathrm{HB}}(W>\tau)$	$E_{\mathrm{HB}}[\%Ab]$, $E_{\mathrm{HB}}[\%Bl]$		$E_{\mathrm{HB}}[U]$
	[109]		x					$P_{\mathrm{TB}}(W>0)$			
	[204]			x			$E_{\mathrm{HB}}[N]$				
	[60]	lag SIPP						$P_{\mathrm{IB}}(W>0)$			
	[185]			x				$P_{\mathrm{TB}}(W>0)$, $P_{\mathrm{TB}}(W>\tau)$, $E_{\mathrm{HB}}[P_{\mathrm{TB}}(W\leq\tau)]$			
	[62]			x				$P_{\mathrm{TB}}(W\leq\tau)$, $\min_{\mathrm{HB}}\{P_{\mathrm{TB}}(W>\tau)\}$			
	[42]	IS					$E_{\mathrm{TB}}[Q]$	$P_{\mathrm{TB}}(W>0)$			
	[61]	MOL/IS			x		$E_{\mathrm{TB}}[Q]$	$E_{\mathrm{TB}}[W]$, $P_{\mathrm{TB}}(W>0)$	$P_{\mathrm{TB}}(Ab)$		$E_{\mathrm{TB}}[Busy]$
	[160]				x		Q_t, N_t, B_t	$E_{\mathrm{TB}}[W]$			
	[58]	EAR						$E_{\mathrm{HB}}[P_{\mathrm{IB}}(W\leq\tau)]$			
	[181]	EAR						$P_{\mathrm{HB}}(W<\tau)$, $P_{\mathrm{IB}}(W<\tau)$			
Emergency department	[64]		x					$E_{\mathrm{HB}}[W]$	E_{HB}[throughput]		
	[207]		x							$E_{\mathrm{HB}}[LoS]$	
	[53]		x					$P_{\mathrm{TB}}(W>\tau)$, $P_{\mathrm{TB}}(W>0)$, $E_{\mathrm{IB}}[W]$			
	[63]	lag SIPP						$P_{\mathrm{IB}}(W>\tau)$	$E_{\mathrm{HB}}[\%Ab]$		
	[54]	MOL	x					$P_{\mathrm{TB}}(W>0)$, $E_{\mathrm{IB}}[W]$		$P_{\mathrm{HB}}(LoS<4h)$	U_t
	[23]		x							$E_{\mathrm{HB}}[LoS]$	U_t, Number of hours where $U_{\mathrm{TB}}>u$, $\max_{\mathrm{HB}}[U]$
	[65]	MOL	x					$P_{\mathrm{HB}}(W>\tau)$, $P_{\mathrm{IB}}(W>\tau)$, $E_{\mathrm{IB}}[W>\tau)]$		$E_{\mathrm{HB}}[LoS]$	$E_{\mathrm{IB}}[U]$
Call center	[75]	PSA			x			$E_{\mathrm{HB}}[C_{\mathrm{CB}}(W)]$	$E_{\mathrm{HB}}[C_{\mathrm{CB}}(Ab)]$		
	[196]	SIPP*						$P_{\mathrm{HB}}(W<\tau)$			
	[167]	SIPP*				x		$P_{\mathrm{HB}}(W<\tau)$	$P_{\mathrm{HB}}(\%Ab)$, $P_{\mathrm{HB}}(\%Bl)$		
	[178], [179]		x					$P_{\mathrm{IB}}(W>0)$, $P_{\mathrm{IB}}(W>\tau)$			
	[188]		x					$P_{\mathrm{IB}}(W>\tau)$, $P_{\mathrm{HB}}(W>\tau)$			
	[78]	PSA			x			$E_{\mathrm{HB}}[C_{\mathrm{CB}}(W)]$	$E_{\mathrm{HB}}[C_{\mathrm{CB}}(Bl)]$, $E_{\mathrm{HB}}[C_{\mathrm{CB}}(Ab)]$		
	[76]	PSA			x				$E_{\mathrm{HB}}[C_{\mathrm{CB}}(Ab)]$		
	[77]				x			$E_{\mathrm{HB}}[C_{\mathrm{CB}}(W)]$	$E_{\mathrm{HB}}[C_{\mathrm{CB}}(Ab)]$		
	[200]			x				$E_{\mathrm{HB}}[C_{\mathrm{CB}}(W)]$, $E_{\mathrm{HB}}[C_{\mathrm{CB}}(W>0)]$			
	[202]	SIPP*						$P_{\mathrm{IB}}(W<\tau)$			
	[190]		x					$P_{\mathrm{IB}}(W>\tau)$, $P_{\mathrm{HB}}(W>\tau)$			
	[176]		x					$P_{\mathrm{IB}}(W>0)$			
	[180]	SIPP					$E_{\mathrm{IB}}[Q]$, $E_{\mathrm{HB}}[E_{\mathrm{IB}}[Q]]$	$P_{\mathrm{IB}}(W>\tau)$, $E_{\mathrm{HB}}[P_{\mathrm{IB}}(W>\tau)]$	$E_{\mathrm{IB}}[\%Ab]$, $E_{\mathrm{HB}}[E_{\mathrm{IB}}[\%Ab]]$		$E_{\mathrm{IB}}[U]$, $E_{\mathrm{HB}}[E_{\mathrm{IB}}[U]]$
	[203]	SIPP	x					$P_{\mathrm{IB}}(W\leq\tau)$			
	[184]	SIPP*	x				$E_{\mathrm{HB}}[Q]$	$P_{\mathrm{HB}}(W>\tau)$	$E_{\mathrm{HB}}[\%Ab]$		SIT_{HB}
	[191]		x						$E_{\mathrm{IB}}[\%Ab]$, $E_{\mathrm{HB}}[\%Ab]$		
	[165]		x		x				$P_{\mathrm{HB}}(Ab_t>\alpha)$		
	[110]	MOL							$E_{\mathrm{HB}}[C_{\mathrm{CB}}(Ab)]$, $E_{\mathrm{HB}}[\%Ab]$, $E_{\mathrm{HB}}[\%Bl]$, $E_{\mathrm{HB}}[C_{\mathrm{CB}}(Bl)]$		
	[74]				x				$E_{\mathrm{HB}}[C_{\mathrm{CB}}(Ab)]$, throughput$_t$		
	[150]		x		x			$E_{\mathrm{HB}}[W]$	$E_{\mathrm{HB}}[\%Served]$		
	[186]		x					$E_{\mathrm{IB}}[CGOS]$			
	[103]	PSA						$E_{\mathrm{IB}}[CGOS]$			
	[201]	SIPP						$P_{\mathrm{IB}}(W\leq\tau)$			
	[177]		x					$P_{\mathrm{IB}}(W\leq\tau)$			
	[95]	IS					Q_t				
	[198]	SIPP						$P_{\mathrm{IB}}(W>\tau)$, $E_{\mathrm{HB}}[P_{\mathrm{IB}}(W>\tau)]$			
	[56]	SIPP						$P_{\mathrm{TB}}(W\leq\tau)$			
	[206]	SIPP						$P_{\mathrm{HB}}(W\leq\tau)$			
	[169]	SIPP	x			x			$E_{\mathrm{IB}}[\%Ab]$		
	[166]					x		$E_{\mathrm{IB}}[W]$, $E_{\mathrm{HB}}[W]$	$E_{\mathrm{IB}}[\%Ab]$, $E_{\mathrm{HB}}[\%Ab]$		
	[168]	SIPP*				x		$E_{\mathrm{HB}}[W]$, $P_{\mathrm{HB}}(W<\tau)$	$P_{\mathrm{HB}}(\%Ab)$, $P_{\mathrm{HB}}(\%Bl)$		
	[55]	SIPP*						$P_{\mathrm{HB}}(W\leq\tau)$			
	[199]		x	x				$E_{\mathrm{HB}}[P_{\mathrm{TB}}(W\leq\tau)]$, $P_{\mathrm{IB}}(W\leq\tau)$			
	[90]	SIPP	x				$E_{\mathrm{HB}}[Q]$	$P_{\mathrm{IB}}(W>0)$, $E_{\mathrm{HB}}[C_{\mathrm{CB}}(W>0)]$	$E_{\mathrm{HB}}[\%Ab]$, $E_{\mathrm{HB}}[C_{\mathrm{CB}}(Ab)]$		
	[205]	SIPP	x				$E_{\mathrm{HB}}[Q]$	$P_{\mathrm{IB}}(W>0)$, $E_{\mathrm{HB}}[C_{\mathrm{CB}}(W>0)]$	$E_{\mathrm{HB}}[\%Ab]$, $E_{\mathrm{HB}}[C_{\mathrm{CB}}(Ab)]$		
	[197]	IS						W_t			
Other	[91]								E_{IB}[throughput]		
	[79]	SIPP						$E_{\mathrm{HB}}[Coverage]$, $E_{\mathrm{IB}}[Coverage]$			
	[66]		x					$E_{\mathrm{HB}}[W]$	$E_{\mathrm{HB}}[\%Ab]$		$E_{\mathrm{HB}}[U]$
	[92]					x			E_{IB}[throughput]		
	[93]					x			E_{IB}[throughput]		
	[94]		x					$P_{\mathrm{HB}}(W\leq\tau)$			

*: assumed, not stated explicitly in article

Table 6: Classification by performance metrics

long. Henderson and Mason [103] evaluate performance by applying a smoothing algorithm to the stationary results. This method, which serves as an improvement to the PSA approach, is capable of modeling the commonly observed congestion lag. Thompson [58] puts forward an effective arrival rate approximation (EAR), that shifts the arrival rate proportional to the expected waiting time. Green and Kolesar [102] present the simple peak-hour approximation (SPHA), an approach that is popular in practice. SPHA approximates performance by a single stationary model, which takes the maximum arrival rate over the cycle as an input parameter.

The modified offered load (MOL) approximations and infinite server (IS) approximations account for the congestion lag in a different way, by relying on analytically tractable results for infinite server queues [100, 104]. In IS approximations, the time-varying number of customers N_t in the system is approximated by its infinite server counterpart N_t^∞ [42, 105] (e.g., the $M_t/G/s_t$ queue is approximated by an $M_t/G/\infty$ system). The delay probability, which can be obtained as $\Pr(N_t \geq s_t)$, is then approximated by $\Pr(N_t^\infty \geq s_t)$ [42]. In contrast, MOL entails a stationary approximation, such that at each moment in time, a stationary model gets applied, using the modified arrival rate $\lambda_t^{\text{MOL}} \equiv m_t^\infty \mu$, with m_t^∞ indicating the expected number of busy servers in an infinite-server system with the same arrival and service processes at time t. Details regarding MOL can be found in [42, 105, 106, 107, 108, 109]. Although the quantity m_t^∞ by definition disregards abandonments (as these do not occur in an infinite server system), MOL can be applied in systems with abandonments, by inserting λ_t^{MOL} in a (stationary) model with abandonments (Feldman et al. [109] report promising results for $M_t/M/s_t + M$ systems). Also, Green et al. [37] and Whitt [38] show that the number of customers in the $M_t/M/s_t + M$ model has exactly the same distribution as the infinite-server $M_t/M/\infty$ model when the abandonment rate equals the service rate. Liu and Whitt [61] suggest the delayed infinite server offered load (DIS-OL) method for staffing, an alternative offered load approach that targets overloaded systems and that is tailored to performance metrics such as abandonment probability and expected waiting time. Hampshire et al. [110] extend the MOL approach to queues with limited capacity using the so-called fluid modified offered load, which provides insights into the number of blocked and abandoned customers.

The key advantage of stationary approximations lies in their simplicity: they can be applied to any system (regardless of the assumptions on service and abandonment processes, the priority rule, the system structure), as long as the stationary counterpart is available. However, the approach also has downsides. For instance, stationary approximations cannot be obtained in (temporarily) overloaded systems without abandonments, because the stationary system then is unstable. Their applicability and accuracy is also highly linked to the validity of the underlying assumptions, such as statistical independence of delays between separate intervals and steady-state being reached quickly in each interval [60]. Moreover, the stationary model itself may already be challenging, requiring the use of approximations (for example, Whitt [71] and Iravani and Balcioğlu [111] provide approximations for the difficult $M/G/s + G$ queue). This may explain why many authors resort to the $M_t/M/s_t$ system, as closed-form results are available for the stationary $M/M/s$ queue [112]. Finally, the effect of the exhaustive service process cannot be accounted for with a stationary approximation, because the service policy is irrelevant in a stationary model.

Apart from the stationary approaches mentioned in Table 6, the literature also presents the stationary backlog-carryover (SBC) approach [113, 114]. This approach does not appear in the categorization as it has not yet been used within an optimization framework for staffing or scheduling; it has been applied successfully to analyze time-dependent delays at airport runways and check-in counters though [115, 116]. The advantage of this approach is that, unlike other stationary approaches, it can be applied in temporarily overloaded systems without abandonments. Whereas most stationary approximations assume staffing intervals to be independent,

SBC instead measures the "backlog" incurred in each period, and transfers it to the next period. As such, the link between the congestion of consecutive periods is captured.

4.2 Simulation

As is evident from Table 6, discrete-event simulation is highly popular for performance evaluation. Discrete-event simulation can model complexities that go beyond the capabilities of analytical and numerical methods (see Law and Kelton [117] for a comprehensive textbook on discrete-event simulation). Especially in healthcare contexts, simulation is a widely adopted methodology (review articles include [118, 119, 120, 121]), but it also appears in other contexts, such as call centers [122]. Although simulation models are commonly context-specific (see, e.g., [64, 123, 124, 125, 126, 127] for applications of simulation in emergency departments), several efforts have sought to develop generic simulation models [128, 129, 130, 131, 132]. However, developing and validating a simulation model is often burdensome, and the computation time tends to be high.

4.3 Numerical methods

In an $M_t/M/s_t$ system, performance can be evaluated by numerically integrating the ordinary differential equations (ODEs) that describe the system (see, e.g., Gross et al. [112] for general background; a more thorough description can be found in [81, 133, 134]). Several ODE-solvers, such as the Euler or Runge-Kutta ODE solver from the Matlab ODE Suite [135], seek to facilitate this analysis. Numerically solving ODEs offers a commonly used benchmark to assess the accuracy of stationary approximations [59, 60] or other methods.

Although Ingolfsson et al. [185] apply this approach, they also note that it requires substantial computational effort. A recent study by Ingolfsson et al. [81] compares several numerical performance evaluation methods in terms of their accuracy and speed for the $M_t/M/s_t$ system. They show that the randomization approach provides a level of accuracy similar to the ODE approach, at a substantially lower computational cost. Though randomization (or uniformization) originates in stationary queues [136, 137, 138], it can be applied successfully for personnel capacity planning in nonstationary queues too (as in Ingolfsson et al. [62]; see also Ingolfsson et al. [81], Ingolfsson [83] and Creemers et al. [139] for related work on performance evaluation in nonstationary queues using the randomization approach).

In general, both randomization and numerical solutions to ODEs rely heavily on Markovian assumptions. The majority of models use an exponential distribution for the service and/or abandonment process. Izady [140] describes how the methods can be extended to phase-type distributions, and concludes that the computational effort increases considerably (which is confirmed by the computational results in Creemers et al. [139]).

The numerical methods generally do not include abandonments or an exhaustive service policy. Ingolfsson [83] includes the exhaustive service policy in a randomization approach and outlines how abandonments can be accommodated. Creemers et al. [139] present a general randomization approach that includes abandonments, an exhaustive service policy and time-varying phase-type distributions for the service and abandonment processes.

None of the categorized articles use discrete-time modeling (DTM, [73, 140, 141, 142, 143, 144, 145]) or closure approximations [146, 147, 148, 149] with a view toward optimizing staffing or scheduling decisions; the available articles focus solely on performance evaluation. The advantage of DTM lies in its ability to accommodate general service time distributions, by approximating the service duration by a discrete process using two-moment matching (for further details we refer to [141, 142, 143, 144, 145]). Wall and Worthington [145] report distinct advantages over MOL and PSA, particularly when temporal overloading is present. However, the

computational effort of DTM may be high [140] and the existing articles all study the $M_t/G/s$ system (i.e., no time-varying number of servers). Recently, Chassioti et al. [73] put forward a DTM approach for systems with abandonments; they focus on systems with low service level targets (i.e., long customer waiting times), where congestion may be affected significantly by abandonment behavior. Closure approximations appear to be less attractive: as Ingolfsson et al. [81] show, they are cumbersome to implement and dominated by other methods (e.g., MOL) in terms of both accuracy and computation speed.

4.4 Fluid models

Deterministic fluid models are intended for systems that do not display stochasticity, but can serve as approximations to derive time-dependent performance in stochastic systems. These methods rely on so-called "fluid scaling", such that the system gets scaled up (e.g., by multiplying arrival rates and the number of servers by the same factor), and the stochastic randomness accordingly decreases in importance, relative to system dynamics (see [150] for an example). Whitt [72] points out that fluid approximations are particularly useful to assess performance in systems that are temporarily overloaded, in which contexts many traditional methods fail (e.g., stationary approximations are no longer valid, because the assumed per period stationarity will result in an infinite queue). For underloaded systems, fluid approximations often fail to capture system dynamics accurately [44, 151, 152]. Fluid approximations rely on approximating the stochastic system by its deterministic counterpart and therefore implicitly assume that queues will only start to build up if the traffic intensity exceeds 1 (hence they target overloaded systems). Fluid models regard arrival and departure processes as continuous flows rather than discrete processes, and they tend to become more accurate as the number of servers grows large [72]. For additional literature on the use of fluid approximations for systems with exponential service and abandonment processes, we refer to Mandelbaum et al. [46, 153, 154, 155, 156][6], Ridley et al. [157], and Jiménez and Koole [152]. Other systems suggest general service and/or abandonment time distributions, including $G_t/G/s + G$ models (with state-dependent arrival rates, [30]), the $G_t/G/s_t + G$ model [158, 159, 160, 161], and networks of queues [162, 163]. Aguir et al. [44] apply fluid models to gain insight into a system with retrials. Personnel capacity planning methods also can rely on fluid models; existing studies [74, 78, 75, 76, 77, 164, 165] all focus on a setting with heterogeneous customers and servers and account for uncertainty in the arrival rate.

4.5 Empirical methods

Some authors use empirical data to estimate system performance. Nah and Kim [166] apply regression to express the abandonment percentage and the mean waiting time as a function of the arrival rate per server. The resulting expressions then are inserted in a mathematical program to obtain a minimum cost shift schedule. Lam et al. [93] and Kabak et al. [92] target shift scheduling in the retail sector. They rely on empirical data to link store sales with customer arrivals, staff number, and other factors. The staffing levels are then selected to maximize the expected profit. Andrews and Parsons [167], Quinn et al. [168], Lin et al. [169] include abandonment-related performance metrics, that are derived from the service level by regression.

[6]Note that, in addition to fluid approximations, these articles also develop diffusion approximations of the studied systems.

5 Classification by optimization approach

In this section, we classify previous publications according to the approach used to optimize personnel capacity. We make a distinction between articles that solely focus on *staffing optimization*, and thus ignore shift schedule considerations (Section 5.1) and articles that take into account shift schedule requirements (Section 5.2).

5.1 Staffing optimization

Table 7 presents an overview of the different staffing methods. As is evident from this table, simple heuristics tend to be popular, such as the "smallest staffing level" (SSL) approach and the square-root staffing (SRS) rule. The SSL approach solves for the stationary model using different capacity values and selects the smallest staffing level that yields satisfactory performance. For example, the staff level s_t is selected by:

$$s_t = \arg\min\{c \in \mathbb{N} : \Pr(N_t \geq c) < \alpha\}\,, \tag{1}$$

if the performance target is to keep the delay probability $\Pr(W_t > 0) = \Pr(N_t \geq s_t)$ below a given target α, for each t. SSL requires an explicit evaluation of the performance metrics, which can be hard to obtain especially in more complex queueing systems for which closed-form results are not available (e.g., the $M_t/G/s_t + G$ queue [61]). Accordingly, Table 7 reveals that many articles that resort to SSL ignore abandonments and assume exponential service times, such that the closed-form results for the $M/M/s$ queue are applicable (this is the well-known Erlang-C formula).

The SRS rule does not explicitly evaluate the performance metrics. Instead, as a general rule-of-thumb, it sets capacity at time t equal to the offered load m_t, augmented by an amount of safety capacity that is proportional to the square root of the offered load:

$$s_t = m_t + \beta\sqrt{m_t}\,. \tag{2}$$

The safety factor β is related to the target delay probability α. Reducing the safety factor to zero results in *staffing to the offered load*; see [61]. The offered load that is inserted into the SRS formula depends on the performance approximation used: for instance, $m_t = m_t^{\infty}$ corresponds to an IS approximation, whereas $m_t = \bar{\lambda}_t/\mu$ corresponds to the SIPP method (with $\bar{\lambda}_t$ the average arrival rate over a given staffing interval). SRS can be applied as a simple heuristic to determine staffing levels in combination with either stationary approximations (e.g., SIPP, PSA, lagged SIPP or MOL [109]), as well as infinite server approximations [42], or in a network context [65]. The general background and applicability of SRS is provided in Gans et al. [6], Borst et al. [170], Whitt [171], Koole and Mandelbaum [172]. Although theoretical and empirical evidence in support of the SRS rule has grown [61, 109, 173], the main challenge in practical applications lies in determining the appropriate value for the safety factor [37, 61, 170, 174, 175].

Simulation-based heuristics use simulation as performance evaluation method in an iterative procedure, to guide the search process. They provide great flexibility in terms of system assumptions; they can be found in Feldman et al. [109], Defraeye and Van Nieuwenhuyse [53], Ahmed and Alkhamis [64], Corominas and Lusa [176], and Kim and Ha [177], among others. Feldman et al. [109] propose the promising iterative staffing algorithm (or ISA) for determining staffing requirements in $M_t/G/s_t + G$ queues, with a view toward stabilizing the delay probability. ISA repeatedly evaluates and alters the staffing function based on the distribution of the number in system at each time instant (which is estimated by simulation), until the desired performance is attained. Defraeye and Van Nieuwenhuyse [53] propose ISA(τ), addressing waiting time tail probabilities instead of delay probabilities. ISA(τ) updates the staffing vector based

Staffing method	References	Objective	Performance constraints (targets are denoted as α_1, α_2, ...)	Kendall notation
SRS	[110]	Maximize profit (revenue per customer, penalty cost for positive waiting time, penalty cost for abandonments)	$E_{\mathrm{HB}}[\%Ab] < \alpha_1$ and $E_{\mathrm{HB}}[\%Bl] < \alpha_1$	$M_t/M/s_t/K+M$
	[42]	Allocate resources to meet target performance	$P(W_t > 0) \leq \alpha$ for all t	$G_t/G_t/s_t$
	[61]	Stabilize performance	$E_{\mathrm{TB}}[W] = \alpha_1$ for all t, $P_{\mathrm{TB}}(Ab) = \alpha_2$ for all t	$M_t/G/s_t+G$
	[197]	$W_t = 0$ for all arriving customers		$M_t/G/s_t$
	[65]	Allocate resources to meet target performance	$P_{\mathrm{IB}}(W > \tau) \leq \alpha$ for all $\tilde{t}$	$M_t/M/s_t$
SSL	[84]	Provide lower bound on staffing requirements	One of the following performance constraints: $P_{\mathrm{TB}}(W > 0) \leq \alpha$ for all t, $P_{\mathrm{TB}}(W > \tau) \leq \alpha$ for all t, $E_{\mathrm{TB}}[W - \tau \mid W > \tau] \leq \alpha$ or $E_{\mathrm{HB}}[P_{\mathrm{TB}}(W > \tau)] \leq \alpha$	$G_t/G/s_t+G$
	[60]	Minimize labor cost	$P_{\mathrm{IB}}(W > 0) < \alpha$ for all intervals $\tilde{t}$	$M_t/M/s_t$
	[201]	Minimize labor cost	$P_{\mathrm{IB}}(W > \tau) \leq \alpha$ for all $\tilde{t}$	$M_t/M/s_t$
	[61]	Stabilize performance	$E_{\mathrm{TB}}[W] = \alpha_1$ for all t, $P_{\mathrm{TB}}(Ab) = \alpha_2$ for all t	$M_t/G/s_t+G$
	[160]	Stabilize performance	$E_{\mathrm{TB}}[W] = \alpha_1$ for all t	$G_t/G/s_t+G$
	[168]	Minimize cost (labor, trunk lines, monthly rental costs)	-	$M_t/M/s_t/K+G$
Dynamic program	[204]	Minimize sum of (linear function of) labor cost and expected number in system		$G_t/M/s_t$
	[103]	Minimize labor cost	$E_{\mathrm{IB}}[CGOS] > \alpha$ for all intervals $\tilde{t}$	$M_t/M/s_t+M$
	[199]	Minimize cost (staffing, penalty if service level is not met)	Constraint on number available workspaces for staff	$M_t/M/s_t$
Mathematical programming	[78]	Minimize cost (staffing, abandonment, blocking, waiting)		$M_t/M/s_t+M$
	[75]	Minimize cost (staffing, abandonment, waiting)		$M_t/M/s+M$
	[76]	Minimize cost (staffing, penalty cost per abandonment)		$M_t/M/s+M$
	[77]	Minimize cost (staffing, waiting, abandonment)		$M_t/M/s_t+M$
	[200]	Minimize cost(cost of incurring positive delay, cost of hiring servers, cost of waiting, cost of using temporary servers)	$P_{\mathrm{HB}}(W = 0) > \alpha_1$, $E_{\mathrm{HB}}[Q] < \alpha_2$	$M_t/M/s$
	[165]	Minimize cost (staffing)	$P_{\mathrm{HB}}(Ab_t > \alpha) < \delta$	$M_t/M/s_t+M$
	[74]	Minimize cost (staffing, abandonment)		$M_t/M/s+M$
	[95]	Minimize labor cost	$Q_t < \alpha$ for all times t	$M_t/M/s_t+M$
Simulation-based heuristic	[64]	Maximize E_{HB}[throughput]	$E_{\mathrm{HB}}[W] < \alpha_1$, upper bound on staffing cost	$M_t/G/s$
	[176]	Minimize labor cost	$P_{\mathrm{IB}}(W > 0) \leq \alpha$ for all $\tilde{t}$	$G_t/G/s_t$
	[53]	Minimize labor cost	$P_{\mathrm{TB}}(W > \tau) \leq \alpha$ for all time epochs t	$M_t/G/s_t+G$
	[109]	Stabilize performance	$P_{\mathrm{TB}}(W > 0) \leq \alpha$ for all time epochs t	$M_t/G/s_t+G$
	[177]	Minimize labor cost	$P_{\mathrm{IB}}(W \leq \tau) > \alpha$ for all $\tilde{t}$	$M_t/M/s_t$

Table 7: Classification by staffing method

on the observed performance, multiplying the staffing levels with a factor proportional to the deviation from the performance target. Ahmed and Alkhamis [64] present a simulation-based heuristic that does not allow the staffing level to vary over time. As such, the dimension of the solution space remains limited to the number of resources available (6 in that study). Allowing staffing changes causes a steep increase in the dimension of the solution vector; we did not find applications of this type of approach for systems with a time-varying number of servers. Finally, Kim and Ha [177] and Corominas and Lusa [176] select staffing levels chronologically, on an interval-by-interval basis; their heuristics each time take the previously selected capacity levels (i.e., in earlier staffing intervals) as given.

Most articles adopt a constraint-satisfaction approach, minimizing cost subject to one or more performance constraints that are commonly related to the quality of service (see, e.g., [53, 103]). For mathematical programming models though, the constraints are frequently included in the objective function by assigning a penalty cost (e.g., cost related to abandonments, blocking, waiting). An alternative objective is to pursue time-stable performance instead of minimizing costs [61, 109].

5.2 Shift schedule optimization

Table 8 classifies prior research according to its approach to shift schedule optimization. We distinguish three approaches:

- *Two-step approaches* construct schedules based on known staffing requirements (staffing and scheduling are two separate, consecutive steps).
- *Feedback-based approaches* address staffing and scheduling simultaneously: they use the concept of staffing requirements to fit shift requirements in an iterative manner, with a feedback loop between staffing and scheduling.
- *Direct approaches* construct shift schedules directly based on the nonstationary demand, without using the concept of staffing requirements.

According to Table 8, most articles adopt the *two-step approach*. This approach first determines the staffing levels required to meet the desired performance at low cost and then fits the minimum cost shift schedule to these requirements. Dantzig's set covering formulation [34] –though it dates back to the 50s– is still highly relevant and used frequently in the literature (see [62, 178, 179]). The staffing requirements are interpreted as strict constraints to be met in Dantzig's model. Alternatively, they can be seen as "desirable" levels that still allow for deviations, as proposed by Keith [35] (see [23, 54, 180], among others). The constraints introduced in the scheduling step are commonly related to work regulations (e.g., minimum amount of time between consecutive shifts, maximum number of working hours per week) and employee preferences (e.g., full time versus part time). As noted in general overviews of the shift scheduling literature [39, 40, 41, 181, 182, 183], most analyses rely on mathematical programming techniques to find an optimal shift schedule. Search heuristics can also be used [181, 184]. Nearly all studies that adopt the two-step approach rely on either SSL or SRS to determine the staffing requirements. The two-step approach is appealing due to the difficulty of integrating stochastic performance constraints into mathematical models; with this approach, the performance constraints are taken care of in the staffing step, such that shift scheduling becomes a deterministic problem. However, the two-step approach may lead to suboptimal shift schedules [103, 185, 186], because several equivalent staffing solutions might exist that lead to shift schedules with substantially varying costs [58, 62, 178]. Therefore, the recent literature increasingly focuses on feedback-based and direct approaches.

The *feedback-based approach* addresses staffing and scheduling simultaneously: it iteratively updates staffing requirements and fits the minimum cost shift schedule, until a satisfactory (not necessarily optimal) solution is found. Feedback-based approaches thus make use of the concept of staffing requirements to set shift requirements, and are iterative. In some cases (e.g., Atlason et al. [178, 179]), the staffing requirements may be considered as auxiliary variables (i.e., they serve to formulate the problem mathematically); nevertheless, we categorize them as feedback-based because the problem formulation explicitly uses (and iteratively adds constraints on) the staffing requirements for fitting shift requirements. A prototype of this approach can be found in Kolesar et al. [187]. The authors iterate SIPP and a mathematical model similar to Dantzig [34] to derive shift schedules for police patrol cars (though they do not provide a systematic approach for updating the staffing requirements). Table 8 reveals that the dominant solution methodology for the feedback-based approach is mathematical programming. Henderson and Mason [186], Atlason et al. [178, 179] and Avramidis et al. [188] rely on cutting plane methods [189] to determine the optimal shift schedule and conduct simulations to evaluate system performance. Atlason et al. [178, 179] extend the work of Henderson and Mason [186]. The algorithm in Atlason et al. [178] requires the service level function to be a concave function of the staffing levels; however, because the service level function tends to follow an S-shaped curve as staffing increases [62, 179], the assumption has been relaxed toward pseudo-concavity in Atlason et al. [179]. Cezik and L'Ecuyer [190] extend the method of Atlason et al. [179] toward settings with heterogeneous customers and servers. Avramidis et al. [188] use a cutting plane method for simultaneous staffing and scheduling, and apply local search techniques to further improve the solution. Ingolfsson et al. [62] present a cutting plane algorithm that relies on randomization to evaluate performance. Campello and Ingolfsson [84] derive strict lower bounds on staffing (which are not necessarily feasible with respect to the performance constraint) and use them as a starting point in the algorithm of Ingolfsson et al. [62]. The feedback-based approach avoids the type of suboptimality that may arise in the two-step approach, as it determines staffing requirements and shift schedules simultaneously. In that sense, it is superior to the two-step approach. Note, however, that the implementation of a feedback-based approach does not by definition guarantee that the obtained solution is optimal. For instance, the cutting plane algorithm in Ingolfsson et al. [62] may miss the optimum because the cuts are introduced based on *estimations* of the additional staffing that is required to meet the performance constraint. By contrast, Atlason et al. [178, 179] show that their method converges to the optimal solution as the number of replications in the simulation model grows large.

The direct approach does not rely on per-period staffing requirements; instead, it creates schedules directly from the arrival rates. Ingolfsson et al. [185] use a genetic algorithm to generate schedules directly from demand, and use numerical integration of the Chapman- Kolmogorov equations (assuming exponential arrival and service processes) to evaluate the performance of a given schedule. Gans et al. [191] adopt a stochastic programming approach that takes forecasted arrival rates as an input, using simulation to evaluate performance. Surprisingly, Castillo et al. [68] are the only ones to treat the shift scheduling problem as a multi-criteria decision problem, by using free disposable hull analysis [192] to select a set of dominant schedules with respect to several performance metrics (which are evaluated by means of simulation). Other methodologies used to evaluate performance of shift schedules in the direct approach are stationary approximations [198] and fluid models [150]; Nah and Kim [166] and Lam et al. [93] use empirical data to estimate performance (such as expected sales, expected waiting time, abandonment rate) as a function of other system parameters (system load, number of salespeople).

Each of the three approaches (two-step approach, feedback-based approach, and direct approach) has its own pros and cons. The two-step approach has the advantage of flexibility in the choice of the algorithms used in the separate staffing and scheduling steps. In spite of this flexi-

Shift optimization	References	Objective	Constraints on quality of service	Kendall notation	Staffing optimization
Two-step approach					
Mathematical programming	[167]	Minimize cost (labor, waiting, lost calls)	-	$M_t/M/s_t + G$	SSL
	[202]	Minimize labor cost	$P_{\mathrm{IB}}(W < \tau) \geq \alpha$ for all $\tilde{t}$	$M_t/M/s_t$	SSL
	[207]	Minimize labor cost	$P_{\mathrm{HB}}(LoS < \tau) \geq \alpha$	$M_t/G/s_t$	Simulation-based heuristic
	[91]	Minimize labor cost	$E_{\mathrm{IB}}[U] = \alpha$, for all intervals $\tilde{t}$	Not specified	Heuristic
	[180]	Minimize deviation between scheduled servers and staffing requirements	$P_{\mathrm{IB}}(W < \tau) \geq \alpha$ for all $\tilde{t}$	$M_t/M/s_t + M$	SSL
	[79]	Maximize expected coverage	$\max E_{\mathrm{HB}}[Coverage]$ and $maximinE_{\mathrm{IB}}[Coverage]$	$M_t/M/s_t/s_t$	Tabu search
	[203]	Minimize number of shifts	$P_{\mathrm{IB}}(W \leq \tau) > \alpha$ for all $\tilde{t}$	$M_t/G/s_t$	SSL
	[66]	Minimize labor cost	$E_{\mathrm{HB}}[W] \leq \alpha$	$M_t/G/s_t + G$	Simulation-based heuristic
	[54]	Minimize penalty of over and understaffing w.r.t. staffing requirements	$P_{\mathrm{HB}}(LoS < \tau) \geq \alpha$	$M_t/G/s_t$	SRS
	[92]	Maximize (profit - staff cost)	-	$M_t/G/s_t$	SRS
	[56]	Minimize cost (labor understaffing w.r.t. staff requirements, overtime cost)	$P(P_{\mathrm{TB}}(W \leq \tau) < \alpha) < \delta$	$M_t/M/s_t$	SSL
	[206]	Minimize labor cost	$P(P_{\mathrm{TB}}(W \leq \tau) < \alpha) < \delta$	$M_t/M/s_t$	SSL
	[169]	Minimize overtime cost and cost of uneven manpower distribution	$E_{\mathrm{IB}}[\%Ab]$	$M_t/G/s_t/K + G$	SSL
	[94]	Minimize labor cost	$P_{\mathrm{HB}}(W \leq \tau)$, for all customers of a given flight	Not specified	Simulation-based heuristic
	[55]	Minimize cost (labor cost, penalty cost for not meeting aggregate service level constraint)	$P_{\mathrm{HB}}(W \leq \tau) > \alpha$	$M_t/M/s_t + M$	SSL
	[205]	Minimize cost (labor, waiting, abandonments)	$P_{\mathrm{IB}}(W > 0) \leq \alpha$	$M_t/M/s_t + M$	SSL
	[23]	Minimize penalty of over and understaffing wrt staffing requirements; Minimize $E_{\mathrm{HB}}[LoS]$	-	Not specified	SRS/offered load
	[58]	Minimize actually worked labor hours	$E_{\mathrm{HB}}[P_{\mathrm{IB}}(W \leq \tau)] \geq \alpha$	$M_t/M/s_t$	SSL
Mathematical programming + metaheuristic	[90]	Minimize cost (labor, waiting, abandonments)	$P_{\mathrm{IB}}(W > 0) \leq \alpha$	$M_t/M/s_t + M$	SSL
(Meta)heuristics	[184]	Maximize weighted scores w.r.t. performance targets	Various customer and server-based constraints (not specified)	Not specified	SSL
	[181]	1) Minimize labor cost 2) Maximize overall service, given total staffing	1) $P_{\mathrm{IB}}\{W < \tau\} > \alpha_1$ and $P_{\mathrm{HB}}\{W < \tau\} > \alpha_2$ 2) $P_{\mathrm{IB}}\{W < \tau\} > \alpha_1$	$M_t/M/s_t$	SSL
Trial-and-error	[196]	Minimize labor cost	$P_{\mathrm{HB}}(W \leq \tau) > \alpha$	$M_t/M/s_t$	SSL
Not specified	[63]	Reduce $E_{\mathrm{HB}}[\%Ab]$, allocate resources to meet target performance	$P_{\mathrm{IB}}(W > \tau) \leq \alpha$ for all $\tilde{t}$	$M_t/M/s_t$	SSL
Feedback-based approach					
Mathematical programming	[178], [179]	Minimize labor cost	$P_{\mathrm{IB}}(W > \tau) \leq \alpha$ for all $\tilde{t}$	$M_t/G/s_t$	
	[188]	Minimize labor cost	$P_{\mathrm{HB}}(W > \tau) \leq \alpha$; $P_{\mathrm{IB}}(W > \tau) \leq \alpha$ for all $\tilde{t}$	$M_t/M/s_t + M$	
	[190]	Minimize labor cost	$P_{\mathrm{IB}}(W > \tau) \leq \alpha_1$ for all $\tilde{t}$ (per call type and overall); $P_{\mathrm{HB}}(W > \tau) \leq \alpha_2$ (per call type and per period)	$M_t/M/s_t + M$	
	[186]	Minimize labor cost	$E_{\mathrm{IB}}[CGOS] > \alpha$, for all intervals $\tilde{t}$	Not specified	
	[62]	Minimize labor cost	$P_{\mathrm{TB}}(W \leq \tau) \geq \alpha$ for all t	$M_t/M/s_t$	
Direct approach					
Free disposable hull analysis	[68]	Determine dominating schedules w.r.t. $E_{\mathrm{HB}}[W]$, $\max_{\mathrm{HB}}\{W\}$, $E_{\mathrm{HB}}[Q]$, $\max_{\mathrm{HB}}\{Q\}$, $E_{\mathrm{HB}}[\%Ab]$,$E_{\mathrm{HB}}[\%Bb]$, $P_{\mathrm{HB}}(W > \tau)$, $E_{\mathrm{HB}}[U]$	-	Not specified	-
Metaheuristic	[185]	Minimize labor cost	$P_{\mathrm{TB}}\{W = 0\} \geq \alpha_1$ (or alternatively, $P_{\mathrm{TB}}\{W \leq \tau\} \geq \alpha_2$ or $E_{\mathrm{HB}}[P_{\mathrm{TB}}(W \leq \tau)] \geq \alpha_3$)	$M_t/M/s_t/K$	-
Local search	[198]	Minimize number of shifts	$E_{\mathrm{HB}}[P_{\mathrm{IB}}(W > \tau)] \leq \alpha$	$M_t/M/s_t$	-
Mathematical programming	[191]	Minimize labor cost	$E_{\mathrm{HB}}[\%Ab] < \alpha_1$, $E_{\mathrm{IB}}[\%Ab] = \alpha_{\tilde{t}}$, for all $\tilde{t}$	$M_t/M/s_t + M$	-
	[150]	Maximize average profit	$E_{\mathrm{HB}}[W] < \alpha_1$, $E_{\mathrm{HB}}[\%Served] > \alpha_2$	$M_t/G/s_t + G$	-
	[166]	Minimize cost (labor, waiting, abandonments)	$E_{\mathrm{IB}}[\%Ab] \leq \alpha_1$ and $E_{\mathrm{HB}}[\%Ab] \leq \alpha_2$	Not specified	-
	[93]	Maximize (profit - staff cost)	-	Not specified	-

Table 8: Classification by shift scheduling method

bility, the majority of two-step approaches implement fairly basic scheduling models (e.g., similar to Dantzig's model). Although dedicated high-level scheduling algorithms that are designed to account for complex scheduling constraints in an efficient way (such as Aykin [193, 194], Rekik et al. [182], Brunner and Bard [195]) can easily be included in the two-step approach, we found no applications of such two-step algorithms in the literature. The feedback-based approach and direct approach are both appealing because they avoid the type of suboptimality that may arise with the two-step approach. However, these models are often highly complex, implying that simplifications to the system assumptions may be required to keep the models solvable. Though the feedback-based approach benefits from an intuitive problem formulation (through the use of staffing requirements), the direct approach skips the staffing step and focuses directly on the final outcome (i.e., the shift schedules). The schedule optimization then becomes more challenging, however, because the solution space is less constrained when the direct approach is applied.

6 Classification by application areas

Finally, Table 9 classifies articles on the basis of their application context. For each reference, we indicate whether the model was implemented (and the results reported), or if it was validated using real-life data or fictive examples. We only consider implementations reported in the academic literature and acknowledge that this is an incomplete indicator of practical implementation. For ease of reference, we repeat the methodology used for staffing and scheduling. As is evident from this table, emergency departments and call centers are the most popular (intended) application areas for the various types of models.

Within the set of articles we consider, Quinn et al. [168], Fukunaga et al. [184], Green et al. [63], Mason et al. [94], Hueter and Swart [66] and Choi et al. [91] are the only studies to implement a model and report the results; they all rely on the two-step approach. Quinn et al. [168], Fukunaga et al. [184] and Green et al. [63] used the (relatively unsophisticated) SSL approach to set staffing levels. Fukunaga et al. [184] complement their analysis with various search heuristics designed to select optimal shift schedules (however, they remain rather vague on the details of the proposed staffing and scheduling algorithms). Quinn et al. [168] apply a profit-driven approach, where the performance target is included in the objective such that personnel is added as long as the incremental cost does not exceed the additional revenue (a similar logic can be found in Lam et al. [93] and Kabak et al. [92], in a retail setting). Mason et al. [94] and Hueter and Swart [66] apply a simulation-based heuristic (for staffing) and mathematical programming (for scheduling). Choi et al. [91] set staffing levels based on a heuristic and further refine the schedule using mathematical programming. The implementations of staffing and scheduling models resulted in, among others, higher revenues [168], reductions in the labor cost [66, 91], less abandoned customers [168, 184, 63], better service levels and lower average waiting times [168]. However, Mason et al. [94] highlighted that the initial schedules provided by their algorithm needed adjustments, because inadequate forecasts had caused understaffing. Moreover, they reported increases in the level of sick leave, possibly caused by the higher complexity of the new schedules.

Zeltyn et al. [65], Dietz [180], and Agnihothri and Taylor [196] assert that their models were implemented, but they do not provide any results about the actual implementation. Instead, they use real-life data to validate the model. The remainder of the publications lack any real-life implementation, though the large majority provide a model validation using real-life data or other means. Henderson and Mason [186], Whitt [197], Fu et al. [204] Bassamboo and Zeevi [78], and Liu and Whitt [160] did not provide any type of implementation or validation for the proposed staffing and/or scheduling model.

Context	References	Implementation (+results reported) (Y/N)	Validation by means of real-life data (Y/N)	Validation by means of other (fictive) examples (Y/N)	Methodology staffing	Methodology scheduling
General	[84]	N	Y	Y	SSL	-
	[68]	N	Y	N	-	Free disposable hull analysis
	[109]	N	Y	Y	Simulation-based optimization	-
	[204]	N	N	N	Dynamic programming	-
	[60]	N	N	Y	SSL	-
	[185]	N	Y	N	-	Metaheuristic
	[62]	N	N	Y	Mathematical programming	id.
	[42]	N	Y	N	SRS	-
	[61]	N	N	Y	SRS;SSL	-
	[160]	N	N	N	SSL	-
	[58]	N	N	Y	SSL	Mathematical programming
	[181]	N	N	Y	SSL	Metaheuristic
Emergency department	[64]	N	Y	N	Simulation-based optimization	-
	[207]	N	Y	N	Simulation-based heuristic	Mathematical programming
	[53]	N	Y	Y	Simulation-based optimization	-
	[63]	Y (+Y)	Y	N	SSL	Not specified
	[54]	N	Y	N	SRS	Mathematical programming
	[23]	N	Y	N	SRS/offered load	Mathematical programming
	[65]	Y (+N)	Y	N	SRS	-
Call center	[196]	Y(+N)	Y	N	SSL	Trial-and-error
	[167]	N	Y	N	SSL	Mathematical programming
	[178], [179]	N	N	Y	Mathematical programming	id.
	[188]	N	Y	N	Mathematical programming	id.
	[78]	N	N	N	Mathematical programming	-
	[75]	N	N	Y	Mathematical programming	-
	[76]	N	N	Y	Mathematical programming	-
	[77]	N	Y	N	Mathematical programming	-
	[200]	N	N	Y	Mathematical programming	-
	[202]	N	Y	N	SSL	Mathematical programming
	[190]	N	Y	Y	Mathematical programming	id.
	[176]	N	N	Y	Simulation-based heuristic	-
	[180]	Y(+N)	Y	N	SSL	Mathematical programming
	[203]	N	Y	N	SSL	Mathematical programming
	[184]	Y (+Y)	Y	N	SSL	Set of heuristics
	[191]	N	Y	N	-	Mathematical programming
	[165]	N	N	Y	Mathematical programming	-
	[110]	N	N	Y	SRS	-
	[74]	N	N	Y	Mathematical programming	-
	[150]	N	N	Y	-	Mathematical programming
	[186]	N	N	N	Mathematical programming	id.
	[103]	N	Y	N	Dynamic program	-
	[201]	N	Y	N	SSL	-
	[177]	N	Y	N	Simulation-based heuristic	-
	[95]	N	Y	N	Mathematical programming	-
	[198]	N	Y	N	-	Local search
	[56]	N	Y	N	SSL	Mathematical programming
	[206]	N	Y	N	SSL	Mathematical programming
	[169]	N	Y	N	SSL	Mathematical programming
	[166]	N	Y	N	-	Mathematical programming
	[168]	Y(+Y)	Y	N	SSL	-
	[55]	N	Y	N	SSL	Mathematical programming
	[199]	N	Y	Y	Dynamic programming	-
	[90]	N	Y	N	SSL	Mathematical programming + metaheuristic
	[205]	N	Y	N	SSL	Mathematical programming
	[197]	N	N	N	SRS	-
Other	[79]	N	Y	N	Metaheuristic	Mathematical programming
	[94]	Y(+Y)	Y	N	Simulation-based heuristic	Mathematical programming
	[91]	Y(+Y)	Y	N	Heuristic	Mathematical programming
	[66]	Y(+Y)	Y	N	Simulation-based heuristic	Mathematical programming
	[92]	N	Y	N	Mathematical programming	id.
	[93]	N	Y	N	-	Mathematical programming

Table 9: Classification by real-life application

Note that not only the objectives and definition of quality of service differ between the application contexts, but also the data availability. Detailed data are often readily available in call centers – this is in general not the case in retail stores and emergency departments. Lam et al. [93] report that the models put forward in the retail literature take advantage of the data that *is* available, for instance, using sales data to construct schedules. Moreover, practical implementations may not find their way to the academic literature due to data confidentiality.

We observe a trend toward models that place a greater emphasis on practical applicability. Dietz [180] provides a spreadsheet-based scheduling approach that can easily be used by practitioners; Gans et al. [191] present an integrated approach for forecasting, staffing and scheduling under parameter uncertainty; and Sinreich and Jabali [23] and Izady and Worthington [54] both rely on a generic simulation model for staffing and scheduling in an emergency department. Nah and Kim [166] use regression analysis to link waiting times to the observed offered load in a call center; as such they avoid using (often complex) queueing models. Lin et al. [169] apply a similar approach, but resort to stationary approximations in those staffing intervals where the regression model's performance is insufficient.

7 Conclusions and implications for future research

The extensive review of extant literature we have reported leads us to draw several conclusions that may be useful for guiding further research. First, it becomes clear that this research field is growing rapidly. Researchers have become very creative in applying multiple methodologies to optimize staffing and/or scheduling in systems with nonstationary demand, and thus meet a myriad of objectives and performance constraints. Unfortunately, our analysis of the system assumptions in Section 3 reveals that, all too often, their ambitious models still rely on rather theoretical assumptions (e.g., homogeneity of customers and servers, exponential assumptions for service and abandonment processes, single-stage systems). The discussion in Section 6 shows that many models lack a real-life implementation (and the few articles that report on real-life implementation appear limited to relatively simple stationary approximations). In particular, only few contributions have tried to tackle a network setting with general service processes ([23, 54, 64], and presumably [184] too), but none of them has addressed general abandonment times in a network —despite the seemingly high relevance of this topic in many practical situations.

As we observed in Section 4, stationary approximations remain highly popular as a performance evaluation method; in recent years, fluid models have also increased in importance (in particular in the call center literature). Both methodologies often rely on rather strict assumptions. We emphasize that the attractiveness of stationary approximations lies not in their accuracy but rather in their ability to provide a simple means to obtain rough guidelines of system performance. The obtained staff levels could be improved further on the basis of, for instance, a simulation model (as in Ertogral and Bamuqabel [203], Zeltyn et al. [65], Izady and Worthington [54]). In fact, hybrid methods that combine the simplicity and insights of queueing results with the flexibility and accuracy of simulation, provide great opportunities for analyzing highly complex settings. Our analysis revealed that authors tend to stick to exponential assumptions for their model description and validation (e.g., [65]), even when using simulation-based methods that are in principle readily extendable to general assumptions.

Our analysis in Section 4 also reveals the wide range of performance metrics being used in current research. It is intuitively clear that logical links exist among the different metrics (e.g., waiting time-related performance metrics relate to abandonment metrics and length-of-stay metrics). Surprisingly though, we were unable to find a single publication that explicitly aimed to uncover these links in complex settings (e.g., network settings with time-varying

arrival rates, general service and abandonment times). Further examination of the relationships across different performance metrics in complex nonstationary systems may open up interesting opportunities for continued research, especially in relation to performance metrics that are difficult to compute. The discovery of an easy-to-compute proxy metric can then substantially simplify the performance evaluation phase and may often be sufficient to guide the search for adequate personnel schedules (as in Izady and Worthington [54] and Green et al. [63]).

We found that some promising performance evaluation methods (e.g., the SBC or DTM approaches) have not yet found their way to staffing and scheduling algorithms; instead, the algorithms tend to resort to those methods that are the most common or straightforward (e.g., stationary approximations such as SIPP or PSA). A challenging direction for future research consists in achieving a better connection between the research fields on performance evaluation on the one hand, and staffing and scheduling on the other hand. However, performance evaluation should be well-aligned with the optimization methodology, especially in terms of computational requirements (computationally expensive evaluation methods ideally require optimization algorithms that quickly find a good solution). Another challenge is posed by the opportunity to model human behavior, as handling speed may be influenced by workload and other factors (see, e.g., [52, 208, 209, 210]).

The applicability of the models extends beyond the typical contexts presented in academic research (i.e., call centers and healthcare systems) to other settings, such as queues in retail stores, restaurants and banking. Care should be taken, though, because these systems tend to have a much smaller scale. Consequently, adding or removing a single server can cause drastic changes in performance. Current literature does not devote much attention to this inherent discreteness of capacity or its implications for model performance (e.g., Feldman et al. [109] report weak performance of the ISA algorithm in case arrival rates are extremely low). The further development of models and algorithms that specifically target small-scale systems provides a promising avenue for further research.

Moreover, the use of the models need not necessarily be restricted to personnel planning; they are in fact relevant to a broad range of problem settings. Zhang et al. [36] provide an interesting illustration: the authors determine the year-by-year capacity of beds in a hospital by a simulation-optimization algorithm that is similar to the method of Defraeye and Van Nieuwenhuyse [53], and compare the results with MOL and SIPP.

Finally, to limit the scope of this article, we did not elaborate on demand forecasting or rostering. Inaccurate forecasts may cause inadequate staffing and scheduling (see for instance the implementation results in Mason et al. [94]). This can be accommodated by including parameter uncertainty in the model (see also Section 3). The integrated approach for forecasting, staffing, and scheduling with parameter uncertainty of Gans et al. [191] represents an important step towards achieving a closer integration of the different steps in the capacity planning process. Our research reveals that staffing and scheduling algorithms for systems with nonstationary demand currently do not tend to integrate the rostering step. Though it can be expected that complexity will increase severely, including the rostering step in an integrated approach is likely to be valuable to avoid suboptimality.

Acknowledgments

This research was supported by the Research Foundation-Flanders (FWO) (grant no G.0547.09). We also thank the three anonymous referees for their remarks, which have considerably improved this article.

Appendix

Homogeneity customers / servers	Kendall notation	References	Staffing interval (Y/N)	Queueing policy	Service policy (E/P)	System structure (S/N)	Parameter uncertainty (Y/N)	Retrials	Reentrant	Stationary approximation	Discrete event simulation	Numerical methods	Fluid model	Empirical
HO / HO	$M_t/M/s_t$	[196]	Y	FIFO*	NS	S	N	-	N	SIPP*				x
		[60]	Y	FIFO*	NS	S	N	-	N	lag SIPP				
		[63]	Y	FIFO*	P	S	N	-	N	lag SIPP				
		[62]	Y	FIFO*	P	S	N	-	N			x		
		[177]	Y	FIFO*	NS	S	N	-	N		x			
		[198]	Y	FIFO*	NS	S	N	-	N	SIPP				
		[206]	Y	FIFO	NS	S	Y	-	N	SIPP				
		[199]	Y	FIFO	P/E	S	Y	-	N		x	x		
		[58]	Y	FIFO*	E	S	N	-	N	EAR				
		[181]	Y	FIFO	NS	S	N	-	N	EAR				
		[180]	Y	FIFO*	NS	S	N	-	N	SIPP				
$M_t/M/s_t + M$		[191]	Y	FIFO*	NS	S	Y	N	N		x			
		[103]	Y	FIFO*	NS	S	N	N	N	PSA				
		[95]	Y	FIFO	P	S	N	N	N	IS				
		[55]	Y	FIFO*	NS	S	Y	N	N	SIPP*		x		
		[90]	Y	FIFO*	NS	S	N	N	N	SIPP	x			
		[205]	Y	FIFO*	NS	S	N	N	N	SIPP	x			
		[185]	Y	FIFO*	NS	S	N	-	N			x		
		[110]	N (+Y)	FIFO*	NS	S	N	N	N	MOL				x
$M_t/M/s_t/K$		[168]	Y	FIFO*	NS	S	N	Y	N	SIPP*				x
$M_t/M/s_t/K + M$		[167]	Y	FIFO*	NS	S	N	Y	N	SIPP*				
$M_t/M/s_t/K + G$		[178]	Y	FIFO	E	S	N	-	N		x			
$M_t/M/s_t + G$		[179]	Y	FIFO	E	S	N	-	N		x			
$M_t/G/s_t$		[197]	N	FIFO	NS	S	Y	-	N	IS				
		[53]	Y	FIFO	E	S	N	N	N		x			
		[109]	Y	FIFO*	NS	S	N	N	N		x			
		[61]	N (+Y)	FIFO	P (+ E)	S	N	N	N	MOL/IS				
$M_t/G/s_t + G$		[204]	Y	FIFO*	NS	S	N	-	N			x		
		[176]	Y	FIFO*	NS	S	N	-	N				x	
$G_t/M/s_t$		[42]	Y	FIFO*	NS	S	N	-	N	IS	x			
$G_t/G/s_t$		[84]	Y	FIFO*	P/E	S	N	-	N	IS				
$G_t/Gt/s_t$		[160]	N	FIFO	NS	S	N	N	N				x	
$G_t/G/s_t + G$		[68]	Y	FIFO*	NS	S	N	N	N		x	x		
Not specified		[91]	Y	FIFO*	NS	S	N	N	N					
		[186]	Y	FIFO*	NS	S	N	N	NS		x			
		[92]	Y	FIFO*	NS	S	N	N	NS					x
		[93]	Y	FIFO*	NS	S	N	N	NS					x
		[94]	Y	FIFO*	NS	S	N	N	NS		x			
		[166]	Y	FIFO*	NS	S	N	N	N					x
HO/HE	$M_t/M/s$	[200]	-	FIFO*	NS	S	Y	-	N			x		
	$M_t/M/s_t$	[201]	Y	FIFO*	NS	S	Y	-	N	SIPP				

(continued on next page)

(continued from previous page)

Homogeneity customers / servers	Kendall notation	References	Staffing interval (Y/N)	Queueing policy	Service policy (E/P)	System structure (S/N)	Parameter uncertainty (Y/N)	Retrials	Reentrant	Stationary approximation	Discrete event simulation	Numerical methods	Fluid model	Empirical
	$M_t/G/s_t + G$	[66]	Y	FIFO	NS	S	N	N	N		x			
	$M_t/G/s_t/K + G$	[169]	Y	FIFO*	NS	S	N	Y	N	SIPP	x			x
HE/HE	$M_t/M/s_t$	[202]	Y	SBR	NS	S	N	-	N	SIPP*				
		[56]	N	FIFO	NS	S	Y	-	N	SIPP				
		[65]	Y	Priority	NS	N	N	-	N	MOL	x			
	$M_t/M/s_t/s_t$	[79]	Y	FIFO	NS	S	N	-	N	SIPP				
	$M_t/M/s + M$	[75]	-	SBR	E	S	Y	N	N	PSA			x	
		[76]	-	SBR	P	S	Y	N	N	PSA			x	
		[74]	-	SBR	P	S	Y	N	N				x	
	$M_t/M/s_t + M$	[78]	Y	SBR	P	S	Y	N	N	PSA			x	
		[77]	Y	SBR	NS	S	Y	N	N				x	
		[165]	Y	SBR	NS	S	Y	N	N		x		x	
		[188]	Y	FIFO / priority	NS	S	N	N	N		x			
		[190]	Y	SBR	NS	S	N	N	N		x			
	$M_t/G/s$	[64]	-	Priority*	NS	N	N	-	N		x			
	$M_t/G/s_t$	[207]	Y	FIFO*	NS	N	N	-	N		x			
		[54]	Y	Priority	E	N	N	-	N	MOL	x			
		[203]	Y	FIFO*	NS	S	N	-	N	SIPP	x			
	$M_t/G/s_t + G$	[150]	Y	SBR	P	S	Y	Y	Y		x		x	
	Not specified	[184]	Y	SBR*	NS	N	N	NS	NS	SIPP*	x			
		[23]	Y	Priority*	NS	N	N	NS	NS		x			

* = assumed, not stated explicitly in article; - = not relevant; (·) = briefly described, such as an extension.

Table 10: Classification by system assumptions and methodology

References

[1] Buffa, E. S., M. J. Cosgrove, B. J. Luce. 1976. An integrated work shift scheduling system. Decision Sciences 7(4) 620-630.

[2] Thompson, G.M. 1995. Labor scheduling using NPV estimates of the marginal benefit of additional labor capacity. Journal of Operations Management 13(1) 67-86.

[3] Thompson, G. M. 1998a. Labor scheduling, Part 1: Forecasting demand. The Cornell Hotel and Restaurant Administration Quarterly 39(5) 22-31.

[4] Thompson, G. M. 1998b. Labor scheduling, Part 2. The Cornell Hotel and Restaurant Administration Quarterly 39(6) 26-37.

[5] Thompson, G. M. 1999a. Labor Scheduling, Part 3: Developing a workforce schedule. Cornell Hotel and Restaurant Administration Quarterly 40(1) 86-94.

[6] Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospect. Manufacturing & service operations management 5(2) 79-141.

[7] Koole, G., A. Pot. 2006. An overview of routing and staffing algorithms in multi-skill customer contact centers. Working paper, VU University Amsterdam. Available online at http://www.math.vu.nl/

[8] Thompson, G. M. 1999b. Labor Scheduling, Part 4 Controlling workforce schedules in real time. Cornell Hotel and Restaurant Administration Quarterly 40(3) 85-96.

[9] Hur, D., V. Mabert, K. Bretthauer. 2004. Real-time work schedule adjustment decisions: an investigation and evaluation. Produstion and Operations Management 13 322-339.

[10] Mehrotra, V., O. Ozluk, R. Saltzman. 2010. Intelligent procedures for intra-day updating of call center agent schedules. Production and Operations Management 19(3) 353-367.

[11] Testik, M. C., J. K. Cochran, G. C. Runger. 2004. Adaptive server staffing in the presence of time-varying arrivals: a feed-forward control approach The Journal of the Operational Research Society 55(3) 233-239.

[12] Saccani, N. 2012. Forecasting for capacity management in call centres: combining methods, organization, people and technology. IMA Journal of Management Mathematics Advance Access published January 31, 2012.

[13] Matteson, D. S., M. W. McLean, D. B. Woodard, S.G. Henderson. 2011. Forecasting emergency medical service call arrival rates. Annals of Applied Statistics 5(2B) 1379-1406.

[14] Aldor-Noiman, S., P. D. Feigin, A. Mandelbaum. 2009. Workload forcasting for a call center: methodology and a case study. Annals of Applied Statistics 3(4) 1403-1447.

[15] Aksin, Z., M. Armony, V. Mehrotra. 2007. The Modern Call Center: A multi-disciplinary perspective on operations management research. Production and Operations Management 16(6) 665-688.

[16] Vile, J. 2013. Time-dependent stochastic modelling for predicting demand and scheduling of emergency medical services, PhD dissertation, Cardiff University.

[17] Taylor, J. W., R. D. Snyder. 2012. Forecasting intraday time series with multiple seasonal cycles using parsimonious seasonal exponential smoothing. Omega 40(6) 748-757.

[18] De Causmaecker, P., G. Vanden Berghe. 2011. A categorisation of nurse rostering problems. Journal of Scheduling 14(1) 3-16.

[19] Burke, E. K., P. De Causmaecker, G. Vanden Berghe, H. Van Landeghem. 2004. The state of the art of nurse rostering. Journal of Scheduling 7(1) 441-499.

[20] Örmeci, E. L., F. S. Salman, E. Yücel. 2014. Staff rostering in call centers providing employee transportation. Omega 43 41-53.

[21] Li, J., E. K. Burke, T. Curtois, S. Petrovic, R. Qu. 2012. The falling tide algorithm: A new multi-objective approach for complex workforce scheduling. Omega 40(3) 283-293.

[22] Bard, J. F., H. W. Purnomo. 2006. Incremental changes in the workforce to accommodate changes in demand. Health Care Management Science 9(1) 71-85.

[23] Sinreich, D., O. Jabali. 2007. Staggered work shifts: a way to downsize and restructure an emergency department workforce yet maintain current operational performance. Health Care Managent Science 10 293-308.

[24] Green, L. V. 2005. Capacity planning and management in hospitals. In: Operations research and Health Care. M.L. Brandeau, F. Sainfort, and W.P. Pierskalla (eds.). International Series in Operations Research & Management Science 70(2) 15-41.

[25] Beliën, J., E. Demeulemeester, P. De Bruecker, J. Van den Bergh, B. Cardoen. 2013. Integrated staffing and scheduling for an aircraft line maintenance problem. *Computers & Operations Research* **40**(4) 1023–1033.

[26] Komarudin, M.-A. Guerry, T. De Feyter, G. Vanden Berghe. 2013. The roster quality staffing problem – A methodology for improving the roster quality by modifying the personnel structure.*European Journal of Operational Research*, To appear.

[27] Venkataraman, R., M. J. Brusco. 1996. An integrated analysis of nurse staffing and scheduling policies. Omega 24(1) 57-71.

[28] Maenhout, B., M. Vanhoucke. 2013. An integrated nurse staffing and scheduling analysis for longer-term nursing staff allocation problems. Omega 41(2) 485-499.

[29] Kortbeek, N., Zonderland, M. E., Boucherie, R. J., Litvak, N., Hans, E. W. 2011. Designing cyclic appointment schedules for outpatient clinics with scheduled and unscheduled patient arrivals. Available online at http://doc.utwente.nl

[30] Whitt, W. 2006. Staffing a call center with uncertain arrival rate and absenteeism. Production and Operations Management 15(1) 88-102.

[31] Pot, A., S. Bhulai, G. Koole. 2008. A simple staffing method for multiskill call centers. Manufacturing & Service Operations Management 10(3) 421-428.

[32] Baron, O., J. Milner. 2009. Staffing to maximize profit for call centers with alternate service-level agreements. Operations Research 57(3) 685-700.

[33] Avramidis, A. N., W. Chan, P. L'Ecuyer. 2009. Staffing multi-skill call centers via search methods and a performance approximation. IIE Transactions 41(6) 483-497.

[34] Dantzig, G. 1954. A comment on Edies traffic delay at toll booths. Operations Research 2 339-341.

[35] Keith, E.G. 1979. Operator scheduling. AIII Trans 11 37-41.

[36] Zhang, Y., M. L. Puterman, M. Nelson, D. Atkins. 2012. A simulation optimization approach to long-term care capacity planning. Operations Research 60(2) 249-261.

[37] Green, L. V., P. J. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. Production and Operations Management 16(1) 13-39.

[38] Whitt, W. 2007. What you should know about queueing models to set staffing requirements in service systems. Naval Research Logistics 54(5) 476-484.

[39] Ernst, A. T., H. Jiang, M. Krishnamoorthy, D. Sier. 2004. Staff scheduling and rostering: A review of applications, methods and models. European Journal of Operational Research 153(1) 3-27.

[40] Ernst, A. T., H. Jiang, M. Krishnamoorthy, B. Owens, D. Sier. 2004. An Annotated Bibliography of Personnel Scheduling and Rostering. Annals of Operations Research 127 21-144.

[41] Van den Bergh, J., J. Beliën, P. De Bruecker, E. Demeulemeester, L. De Boeck. 2013. Personnel scheduling: A literature review. European Journal of Operational Research 226(3) 367-385.

[42] Jennings, O. B., A. Mandelbaum, W.A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. Management Science 42(10) 1383-1394.

[43] Johnson, M., S. Myers, J. Wineholt, M. Pollack, A.vL. Kusmiesz. 2009. Patients who leave the emergency department without being seen. Journal of Emergency Nursing 35(2) 105-108.

[44] Aguir, S., F. Karaesmen, O.Z. Akskin, F. Chauvet. 2004. The impact of retrials on call center performance. OR Spectrum 26(3) 353-376.

[45] Aguir, S., O. Z. Aksin, F. Karaesmen, Y. Dallery. 2008. On the interaction between retrials and sizing of call centers. European Journal of Operational Research 191(2) 398-408.

[46] Mandelbaum, A., W. A. Massey. M.I. Reiman, A. Stolyar, B. Rider. 2002. Queue lengths and waiting times for multiserver queues with abandonment and retrials. Telecommunication Systems 21(2-4) 149-171.

[47] Yom-Tov, G. 2010. Queues in Hospitals: Queueing networks with reentering customers in the QED regime. Ph.D. Thesis, Technion - Israel Institute of Technology, Haifa, Israel. Available online at http://ie.technion.ac.il/serveng/References/references.html

[48] Koole, G., R. Righter. 1998. Optimal control of tandem reentrant queues. Queueing Systems - Theory and Applications 28(4) 337-347.

[49] Kendall, D. G. 1953. Stochastic Processes occurring in the theory of queues and their analysis by the method of the imbedded markov chain. The Annals of Mathematical Statistics 24(3) 338-354.

[50] Heyman, D. P., W. Whitt. 1984. The asymptotic behavior of queues with time-varying arrival rates. Journal of Applied Probability 21(1) 143-156.

[51] Baccelli, F., G. Hebuterne. 1981. On queues with impatient customers. Performance 81. North-Holland, Amsterdam, The Netherlands, 159-179.

[52] Gans, N., N. Liu, A. Mandelbaum, H. Shen, H. Ye. 2010. Service times in call centers: Agent heterogeneity and learning with some operational consequences. Festschrift for Lawrence D. Brown(6), IMS Collections, Beachwood, 99-123.

[53] Defraeye, M., I. Van Nieuwenhuyse. 2013. Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm. Decision Support Systems, Decision Support Systems 54(4) 1558-1567.

[54] Izady, N., D. J. Worthington. 2012. Setting staffing requirements for time-dependent queueing networks: The case of accident and emergency departments. European Journal of Operational Research 219 531-540.

[55] Robbins, T. R., T.P. Harrison. 2010. A stochastic programming model for scheduling call centers with global service level agreements. European Journal of Operational Research 207 1608-1617.

[56] Liao, S., G. Koole, C. van Delft, O. Jouini. 2012. Staffing a call center with uncertain non-stationary arrival rate and flexibility. OR Spectrum 34 691-721.

[57] Kim, S. H., W. Whitt. 2013. Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? Working paper, Columbia University, New York. Available online at: http://www.columbia.edu/ ww2040/allpapers.html

[58] Thompson, G. M. 1993. Accounting for the multi-period impact of service when determining employee requirements for labor scheduling. Journal of Operations Management 11(3) 269-287.

[59] Green, L. V., P.J. Kolesar, J. Soares. 2003. An improved heuristic for staffing telephone call centers with limited operating hours. Production and Operations Management 12(1) 46-61.

[60] Green, L. V., P. J. Kolesar, J. Soares. 2001. Improving the SIPP Approach for Staffing Service Systems That Have Cyclic Demands. Operations Research 49(4) 549-564.

[61] Liu, Y., W. Whitt. 2012. Stabilizing customer abandonment in many-server queues with time-varying arrivals. Operations Research, 60(6) 1551-1564.

[62] Ingolfsson, A., F. Campello, X. Wu, E. Cabral. 2010. Combining integer programming and the randomization method to schedule employees. European Journal of Operational Research 202(1) 153-163.

[63] Green, L. V., J. Soares, J. F. Giglio, R. A. Green. 2006. Using queueing theory to increase the effectiveness of emergency department provider staffing. Academic Emergency Medicine 13(1) 61-68.

[64] Ahmed, M. A., T. M. Alkhamis. 2009. Simulation optimization for an emergency department healthcare unit in Kuwait. European Journal of Operational Research 198(3) 936-942.

[65] Zeltyn, S., Y.N. Marmor, A. Mandelbaum, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, A. Shtub, T. Lauterman, D. Schwartz, K. Moskovitch, S. Tzafrir, F. Basis. 2011. Simulation-based models of emergency departments: Operational, tactical, and strategic staffing. ACM Transactions on Modeling and Computer Simulation 21(4).

[66] Hueter, J., W. Swart. 1998. An integrated labor-management system for Taco Bell. Interfaces 28(1) 75-91.

[67] Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing perspective. Journal of the American Statistical Association 100(469) 36-50.

[68] Castillo, I., T. Joro, Y.Y. Li. 2009. Workforce scheduling with multiple objectives. European Journal of Operational Research 196(1) 162-170.

[69] Dai, J. G., He, S. 2010. Customer abandonment in many-server queues. Mathematics of Operations Research 35(2) 347-362.

[70] Mandelbaum, A., P. Momğilović. 2012. Queues with many servers and impatient customers. Mathematics of Operations Research 37(1) 41-65.

[71] Whitt, W. 2005. Engineering solution of a basic call-center model. Management Science 51(2) 221-235.

[72] Whitt, W. 2006. Fluid models for multiserver queues with abandonments. Operations Research 54(1) 37-54.

[73] Chassioti, E., D. Worthington, K. Glazebrook. 2013. Effects of state-dependent balking on multi-server non-stationary queueing systems. Journal of the Operational Research Society 1-13.

[74] Harrison, J. M., A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. Manufacturing & Service Operations Management 7(1) 20-36.

[75] Bassamboo, A., J. M. Harrison, A. Zeevi. 2006. Design and control of a large call center: asymptotic analysis of an LP-based method. Operations Research 54(3) 419-435.

[76] Bassamboo, A., A. Zeevi. 2009. On a data-driven method for staffing large call centers. Operations Research 57(3) 714-726.

[77] Bertsimas D., X. V. Doan. 2010. Robust and data-driven approaches to call centers. European Journal of Operational Research 207(2) 1072- 1085.

[78] Bassamboo, A., A. Zeevi. 2005. Dynamic routing and admission control in high-volume service systems: asymptotic analysis via multi-scale fluid limits. Queueing Systems 51(3-4) 249-285.

[79] Erdoğan, G., E. Erkut, A. Ingolfsson, G. Laporte. 2010. Scheduling ambulance crews for maximum coverage. Journal of the Operational Research Society 61(4) 543-550.

[80] Budge, S., A. Ingolfsson, E. Erkut. 2009. Technical Note-Approximating Vehicle Dispatch Probabilities for Emergency Service Systems with Location-Specific Service Times and Multiple Units per Location. Operations Research 57(1) 251-255.

[81] Ingolfsson, A., E. Akhmetshina, S. Budge, Y. Li. 2007. A survey and experimental comparison of service level approximation methods for non-stationary $M_t/M/s_t$ queueing systems with exhaustive discipline. INFORMS Journal on Computing 19(2) 201-214.

[82] Chen, B. K. P, S. G. Henderson. 2001. Two issues in setting call centre staffing levels. Annals of Operations Research 108(1-4) 175-192.

[83] Ingolfsson, A., 2005. Modeling the $M_t/M/s_t$ queue with an exhaustive discipline. Working paper, University of Alberta, Canada. Available online on http://www.business.ualberta.ca/aingolfsson/publications.htm

[84] Campello, F., A. Ingolfsson. 2011. Exact Necessary Staffing Requirements based on Stochastic Comparisons with Infinite-Server Models. Working paper, University of Alberta.

[85] Maman S. 2009. Uncertainty in the demand for service: the case of call centers and emergency departments. M.Sc. Thesis, Technion, Israel Institute of Technology.

[86] Robbins, T. R., D. J. Medeiros, P. Dum. 2006. Evaluating arrival rate uncertainty in call centers. In L.F. Perrone, F.P.Wieland, J. Liu, B.G. Lawson, D.M. Nicol, and R.M. Fujimoto (eds.), Proceedings of the 2006 Winter Simulation Conference 2180-2187.

[87] Steckley, S. G., S. G. Henderson, V. Mehrotra. 2004. Service system planning in the presence of random arrival rate. working paper, Cornell University.

[88] Steckley, S. G., S. G. Henderson, and V. Mehrotra. 2009. Forecast errors in service systems. Probability in the Engineering and Informational Sciences 23(2) 305-332.

[89] Robbins, T. R. 2007. Managing service capacity under uncertainty. Ph.D. thesis, Penn State University.

[90] Saltzman, R. M. 2005. A hybrid approach to minimize the cost of staffing a call center. International Journal of Operations and Quantitative Management 11(1) 1-14.

[91] Choi, K., J. Hwang, M. Park. 2009. Scheduling restaurant workers to minimize labor cost and meet service standards. Cornell Hospitality Quarterly 50(2) 155-167.

[92] Kabak, Ö., F. Ülengin, E. Aktaş, Ş. Önsel, Y. I. Topcu. 2008. Efficient shift scheduling in the retail sector through two-stage optimization. European Journal of Operational Research 184(1) 76-90.

[93] Lam, S., M. Vandenbosch, M. Pearce. 1998. Retail sales force scheduling based on store traffic forecasting. Journal of Retailing 74(1) 61-88.

[94] Mason, A. J., D. M. Ryan, D. M. Panton. 1998. Integrated simulation, heuristic and optimisation approaches to staff scheduling.Operations research46(2) 161-175.

[95] Kim, J.W., S.H. Ha. 2012. Advanced workforce management for effective customer services. Quality & Quantity 46(6) 1715-1726.

[96] Defraeye, M., I. Van Nieuwenhuyse. 2011. Setting staffing levels in an emergency department: opportunities and limitations of stationary queuing models. Review of Business and Economics 56 (1) 73-100.

[97] Green, L. V., P. J. Kolesar, A. Svoronos. 1991. Some effects of nonstationarity on multiserver markovian queueing systems. Operations Research 39(3) 502-511.

[98] Green, L. V., P. J. Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. Management Science 37(1) 84-97.

[99] Whitt, W. 1991. The pointwise stationary approximation for $M_t/M_t/s$. Management Science 37(3) 307-314.

[100] Eick, S. G., W.A. Massey, W. Whitt. 1993a. The physics of the $M_t/G/\infty$ queue. Operations Research 41(4) 731-742.

[101] Green, L. V., P. J. Kolesar. 1997. The Lagged PSA for Estimating Peak Congestion in Multiserver Markovian Queues with Periodic Arrival Rates. Management Science 43(1) 80-87.

[102] Green, L. V., P. J. Kolesar. 1995. On the accuracy of the simple peak hour approximation for markovian queues. Management Science 41(8) 1353-1370.

[103] Henderson, S. G., A.J. Mason, I. Ziedins, R. Thomson. 1999. A Heuristic for determining efficient staffing requirements for call centres. Working paper, University of Auckland, New Zealand.

[104] Eick, S. G., W. A. Massey, W. Whitt. 1993b. $M_t/G/\infty$ queues with sinusoidal arrival rates. Management Science 39(2) 241-252.

[105] Massey, W. A., W. Whitt. 1997. Peak congestion in multi-server service systems with slowly varying arrival rates. Queueing Systems 25(1) 157-172.

[106] Jagerman, D.L.. 1975. Nonstationary blocking in telephone traffic. Bell Syst. Tech. 54 625-661.

[107] Massey, W. A., W. Whitt. 1994. An analysis of the modified offered-load approximation for the nonstationary erlang loss model. The Annals of Applied Probability 4(4) 1145-1160.

[108] Davis, J. L., W. A. Massey, W. Whitt. 1995. Sensitivity to the service-time distribution in the nonstationary erlang loss model. Management Science 41(6) 1107-1116.

[109] Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. Management Science 54(2) 324-338.

[110] Hampshire, R. C., O.B. Jennings, W. A. Massey. 2009. A timevarying call center design via Lagrangian mechanics. Probability in the Engineering and Informational Sciences 23 231-259.

[111] Iravani, F., B. Balcioğlu. 2008. Approximations for the M/GI/N +GI type call center. Queueing Systems 58(2) 137-153.

[112] Gross, D., J.F. Shortle, J. M. Thompson, C.M. Harris. 2008. Fundamentals of queueing theory (4th Edition). Wiley Series in Probability and Statistics, Wiley-Blackwell.

[113] Stolletz, R. 2008a. Approximation of the non-stationary $M_t/M_t/c_t$-queue using stationary queueing models: The stationary backlog-carryover approach. European Journal of Operational Research 190 478-493.

[114] Stolletz, R., S. Lagershausen. 2013. Time-dependent performance evaluation for loss-waiting queues with arbitrary distributions. International Journal of Production Research 51(5) 1366-1378.

[115] Stolletz, R. 2008b. Non-stationary delay analysis of runway systems. OR Spectrum 30(1) 191213.

[116] Stolletz, R. 2011. Analysis of passenger queues at airport terminals. Research in Transportation Business & Management 1(1) 144-149.

[117] Law, A. M., W.D. Kelton. 2000. Simulation modeling and analysis. McGraw-Hill series in industrial engineering and management science, McGraw-Hill (Boston).

[118] Gunal, M. M., M. Pidd. 2010. Discrete event simulation for performance modelling in health care: a review of the literature. Journal of Simulation 4(1) 42-51.

[119] Jacobson, S. H., R. W. Hall, S. N. Hall. 2006. Discrete-event simulation of health care systems. Chapter in: Patient Flow: Reducing Delay in Healthcare Delivery. International Series in Operations Research & Management Science 91 211-252.

[120] Jun, J. B., S. H. Jacobson, J. R. Swisher. 1999. Application of discrete-event simulation in health care clinics: A survey. Journal of the Operational Research Society 50(2) 109-123.

[121] White P. K. Jr. 2005. A survey of data resources for simulating patient flows in healthcare delivery systems. Proceedings of the 2005 Winter Simulation Conference. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines (Eds.), 926-935.

[122] Mehrotra, V., J. Fama. 2003. Call center simulation modeling: methods, challenges and opportunities. Proceedings of the 2003 Winter Simulation Conference, 135-143.

[123] Evans, G.W., T.B. Gor, E. Unger. 1996. A simulation model for evaluating personnel schedules in a hospital emergency department. In Proceedings of the 28th conference on Winter simulation (WSC '96), J.M. Charnes, D.J. Morrice, D.T. Brunner, J.J. Swain (Eds.). IEEE Computer Society, Washington, DC, USA, 1205-1209.

[124] Garcia, M. L, M. A. Centeno, C. Rivera, N. DeCario. 1995. Reducing time in an emergency room via a fast-track. In Proceedings of the 27th conference on Winter simulation (WSC '95), C. Alexopoulos, K. Kang (Eds.). IEEE Computer Society, Washington, DC, USA, 1048-1053.

[125] Hung, G. R., S. R. Whitehouse, C. B. O'Neill, A. P. Gray, N. Kissoon. 2007. Computer modeling of patient flow in a pediatric emergency department using discrete event simulation. Pediatric Emergency Care 23(1) 5-10.

[126] McGuire, F. 1994. Using simulation to reduce length of stay in emergency departments. In Proceedings of the 26th conference on Winter simulation (WSC '94), M.S. Manivannan, J.D. Tew (Eds.). Society for Computer Simulation International, San Diego, CA, USA, 861-867.

[127] Takakuwa, S., H. Shiozaki. 2004. Functional analysis for operating emergency department of a general hospital. In Proceedings of the 36th conference on Winter simulation(WSC '04). Winter Simulation Conference 2003-2011.

[128] Fletcher, A., D. Halsall, S. Huxham, D. Worthington. 2007. The DH accident and emergency department model: a national generic model used locally. Journal of the Operational Research Society 58 1554-1562.

[129] Fletcher, A., D. J. Worthington. 2007. What is a "generic" hospital model? Working Paper, Department of Management Science, Lancaster University.

[130] Pitt, M. 1997. A generalised simulation system to support strategic resource planning in healthcare. In Proceedings of the 29th conference on Winter simulation (WSC '97), S. Andradottir, K.J. Healy, D.H. Withers, B.L. Nelson (Eds.). IEEE Computer Society, Washington, DC, USA, 1155-1162.

[131] Sinreich, D., Y. N. Marmor. 2004. A simple and intuitive simulation tool for analyzing emergency department operations. In Proceedings of the 36th conference on Winter simulation(WSC '04). Winter Simulation Conference 1994-2002.

[132] Gunal, M. M., M. Pidd. 2009. Understanding target-driven action in emergency department performance using simulation. Emergency medicine journal 26(10) 724-727.

[133] Koopman, B. O. 1972. Air-terminal queues under time-dependent conditions. Operations Research 20(6) 1089-1114.

[134] Green, L. V., J. Soares. 2007. Computing time-dependent waiting time probabilities in $M(t)/M/s(t)$ queueing systems. Manufacturing & service operations management 9 (1) 54-61.

[135] Shampine, L. F., M.W. Reichelt. 1997. The MATLAB ODE Suite. SIAM Journal on Scientific Computing 18(1) 1-22.

[136] Jensen, A. 1953. Markov Chains as an Aid in the Study of Markov Processes. Skand. Aktuarietidskrift 3 87-91.

[137] Grassmann, W. K. 1977. Transient solutions in markovian queueing systems. Computers & Operations Research 4(1) 47-53.

[138] Gross, D., D. R. Miller. 1984. The randomization technique as a modeling tool and solution procedure for transient Markov processes. Operations Research 32(2) 343-361.

[139] Creemers, S., Defraeye, M., Van Nieuwenhuyse, I. (2014). G-RAND: a phase-type approximation for the nonstationary G(t)/G(t)/s(t)+G(t) queue. Performance Evaluation, forthcoming.

[140] Izady, N. 2010. On Queues with Time-varying Demand. PhD Dissertation, University of Lancaster, Lancaster, UK.

[141] Chassioti, E., D. J. Worthington. 2004. A new model for call centre queue management. The Journal of the Operational Research Society 55(12) 1352-1357.

[142] Brahimi, M. 1990. Approximating multi-server queues with inhomogeneous arrival rates and continuous service time distributions. PhD Dissertation, University of Lancaster, Lancaster, UK.

[143] Brahimi, M., D. J. Worthington. 1991. The finite capacity multi-server queue with inhomogeneous arrival rate and discrete service time distribution and its application to continuous service time problems. European Journal of Operational Research 50(3) 310-324.

[144] Wall, A. D., D. J. Worthington. 1994. Using Discrete Distributions to Approximate General Service Time Distributions in Queueing Models. The Journal of the Operational Research Society 45(12) 1398-1404.

[145] Wall, A. D., D. J. Worthington. 2007. Time-dependent analysis of virtual waiting time behaviour in discrete time queues. European Journal of Operational Research 178(2) 482-499.

[146] Rothkopf, M. H., S. S. Oren. 1979. A closure approximation for the nonstationary $M/M/s$ Queue. Management Science 25(6) 522-534.

[147] Clark, G.M. 1981. Use of Polya distributions in approximate solutions to nonstationary $M/M/s$ queues. Commun. ACM 24(4) 206-217.

[148] Taaffe, M., K. Ong. 1987. Approximating nonstationary Ph(t)/Ph(t)/l/c queueing systems. Annals of Operations Research 8(1) 103-116.

[149] Tsai, W. K., P. E. Cantrell. 1989. A simple derivation of transient queue statistics and applications. Performance evaluation 10(2) 103-114.

[150] Helber, S., K. Henken. 2010. Profit-oriented shift scheduling of inbound contact centers with skills-based routing, impatient customers, and retrials. OR Spectrum 32(1/4) 109-134.

[151] Altman, E., T. Jiménez, G. Koole. 2001. On the comparison of queueing systems with their fluid limits. Probability in the Engineering and Informational Sciences 15 165-178.

[152] Jiménez, T., G. Koole. 2004. Scaling and comparison of fluid limits of queues applied to call centers with time varying parameters. OR Spectrum 26 (3) 413-422.

[153] Mandelbaum, A., W. A. Massey. 1995. Strong approximations for time-dependent queues. Mathematics of Operations Research 20(1), 33-64.

[154] Mandelbaum, A., W. Massey, M. Reiman. 1998. Strong approximations for Markovian service networks. Queueing Systems 30(1) 149-201.

[155] Mandelbaum, A., W. A. Massey, M. I. Reiman, R. Rider. 1999. Time varying multiserver queues with abandonments and retrials. Proceedings of the 16th International Teletraffic Conference 3 355-364.

[156] Mandelbaum, A., W. A. Massey, M. I. Reiman, A. Stolyar. 1999. Waiting time asymptotics for time varying multiserver queues with abandonment and retrials. Proc. 37th Allerton Conf. Monticello, IL, 1095-1104.

[157] Ridley, A. D., M. C. Fu, W. A. Massey. 2003. Customer relations management: call center operations: Fluid approximations for a priority call center with time-varying arrivals. Proceedings of the 35th Conference on Winter Simulation, New Orleans, LA, 2, 1817-1823.

[158] Liu, Y., W. Whitt. 2010. A Fluid Approximation for the $GI_t/GI/s_t+GI$ Queue. Working paper, Columbia University, New York. Available online at: http://www.columbia.edu/ ww2040/allpapers.html

[159] Liu, Y., W. Whitt. 2011. Large-Time Asymptotics for the $G_t/M_t/s_t + GI_t$ Many-Server Fluid Queue with Abandonment. Queueing systems 67 (2) 145-182.

[160] Liu, Y., W. Whitt. 2012. The $G_t/GI/s_t + GI$ many-server fluid queue. Queueing Systems 71(4) 405-444.

[161] Liu, Y., W. Whitt. 2012. A many-server fluid limit for the $G_t/GI/s_t + GI$ queueing model experiencing periods of overloading. OR Letters 40 307-312.

[162] Liu, Y., W. Whitt. 2011. A network of time-varying many-server fluid queues with customer abandonment. Operations Research 59(4) 835-846.

[163] Liu, Y., W. Whitt. 2013. Algorithms for time-varying networks of many-server fluid queues. INFORMS Journal on Computing. To appear.

[164] Choudhury, G. L., A. Mandelbaum, M. I. Reiman, W. Whitt. 1997. Fluid and diffusion limits for queues in slowly changing random environments. Stochastic Models 13(1) 121-146.

[165] Gurvich, I., J. Luedtke, T. Tezcan. 2010. Staffng call centers with uncertain demand forecasts: a chance-constrained optimization approach. Management Science 56(7) 1093-1115.

[166] Nah, J. E., S. Kim. 2013. Workforce planning and deployment for a hospital reservation call center with abandonment cost and multiple tasks. Computers & Industrial Engineering 65(2) 297-309.

[167] Andrews, B., H. Parsons. 1993. Establishing telephone-agent staffing levels through economic optimization. Interfaces 23(2) 14-20.

[168] Quinn, P., B. Andrews, H. Parsons. 1991. Allocating telecommunications resources at L. L. Bean, Inc. Interfaces 21(1) 75-91.

[169] Lin, C. K. Y., K. F. Lai, S. L. Hung. 2000. Development of a workforce management system for a customer hotline service.Computers & Operations Research 27(10) 987-1004.

[170] Borst, S., A. Mandelbaum, M. I. Reiman. 2004. Dimensioning large call centers. Operations Research 52(1) 17-34.

[171] Whitt, W. 1992. Understanding the efficiency of multi-server service systems. Management Science 38(5) 708-723.

[172] Koole, G., A. Mandelbaum. 2002. Queueing models of call centers: an introduction. Annals of Operations Research 113(1) 41-59.

[173] Whitt, W. 2013. OM Forum – Offered load analysis for staffing. Manufacturing & Service Operations Management 15(2) 166-169.

[174] Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. Manufacturing & Service Operations Management 4(3) 208-227.

[175] Halfin, S., W. Whitt. 1981. Heavy-traffic limits for queues with many exponential servers. Operations Research 29(3) 567-588.

[176] Corominas, A., A. Lusa. 2012. LETRIS: Staffing service systems by means of simulation. Journal of Industrial Engineering and Management 5(2) 285-296.

[177] Kim, J. W., S. H. Ha. 2010. Consecutive staffing solution using simulation in the contact center.Industrial Management & Data Systems 110(5) 718-730.

[178] Atlason, J., M.A. Epelman, S. G. Henderson. 2004. Call center staffing with simulation and cutting plane methods. Annals of Operations Research 127(1) 333-358.

[179] Atlason, J., M. A. Epelman, S.G. Henderson. 2008. Optimizing Call Center Staffing Using Simulation and Analytic Center Cutting-Plane Methods. Management Science 54(2) 295-309.

[180] Dietz, D. C. 2011. Practical scheduling for call center operations. Omega 39 550-557.

[181] Thompson, G. M. 1997. Labor staffing and scheduling models for controlling service levels. Naval Research Logistics 44(8) 719-740.

[182] Rekik, M., J.-F. Cordeau, F. Soumis. 2010. Implicit shift scheduling with multiple breaks and work stretch duration restrictions. Journal of Scheduling 13(1) 49-75.

[183] Brunner, J. O., J. F. Bard, R. Kolisch. 2010. Midterm scheduling of physicians with flexible shifts using branch and price. IIE Transactions 43(2) 84-109.

[184] Fukunaga, A., E. Hamilton, J. Fama, D. Andre, O. Matan, I. Nourbakhsh. 2002. Staff scheduling for inbound call centers and customer contact centers. AI Magazine Volume 23(4) 30-40.

[185] Ingolfsson, A., A. Haque, A. Umnikov. 2002. Accounting for time-varying queueing effects in workforce scheduling. European Journal of Operational Research 139(3) 585-597.

[186] Henderson, S. G., A. J. Mason. 1998. Rostering by iterating integer programming and simulation. Winter Simulation Conference Proceedings (WSC'98) 1 677-683.

[187] Kolesar, P.J., K.L. Rider, T.B. Crabill, W.E. Walker. 1975. A queuing-linear programming approach to scheduling police patrol cars. Operations Research 23(6) 1045-1062.

[188] Avramidis, A. N., W. Chan, M. Gendreau, P. LEcuyer, O. Pisacane. 2010. Optimizing daily agent scheduling in a multiskill call center. European Journal of Operational Research 200 822-832.

[189] Kelley, J. E. Jr. 1960. The Cutting-Plane Method for Solving Convex Programs. J. Soc. Indust. and Appl. Math. 8(4) 703-712.

[190] Cezik, M. T., P. L'Ecuyer. 2008. Staffing multiskill call centers via linear programming and simulation. Manag Science 54(2) 310-323.

[191] Gans, N., H. Sheng, Y.-P. Zhou, N. Korolev, A. McCord, H. Ristock. 2012. Parametric Stochastic Programming Models for Call-Center Workforce Scheduling. Working paper, University of Washington. Available online at http://faculty.washington.edu/

[192] Tulkens, H. 1993. On FDH analysis: some methodological issues and applications to retail banking, courts and urban transit. Journal of Productivity Analysis 4 183-210.

[193] Aykin, T. 1996. Optimal Shift Scheduling with Multiple Break Windows. Management Science 42(4) 591-602.

[194] Aykin, T. 1998. A Composite Branch and Cut Algorithm for Optimal Shift Scheduling with Multiple Breaks and Break Windows. The Journal of the Operational Research Society 49(6) 603-615.

[195] Brunner, J. O., J. F. Bard. 2013. Flexible weekly tour scheduling for postal service workers using a branch and price. Journal of Scheduling 16(1) 129-149.

[196] Agnihothri, S. R., P. F. Taylor. 1991. Staffing a Centralized Appointment Scheduling Department in Lourdes Hospital. Interfaces 21(5) 1-11.

[197] Whitt, W. 1999. Dynamic staffng in a telephone call center aiming to immediately answer all calls. Operations Research Letters 24 205-212.

[198] Koole, G., E. van der Sluis. 2003. Optimal shift scheduling with a global service level constraint. IIE Transactions 35(11) 1049-1055.

[199] Roubos, A., S. Bhulai, G. Koole. 2011. Flexible staffing for call centers with non-stationary arrival rates. Working paper, VU University Amsterdam. Available online at http://www.cs.vu.nl/

[200] Bhandari, A., A. Scheller-Wolf, M. Harchol-Balter. 2008. An exact and efficient algorithm for the constrained dynamic operator staffing problem for call centers. Manag Science 54(2) 339-353.

[201] Jongbloed, G., G. Koole. 2001. Managing uncertainty in call centres using Poisson mixtures. Applied Stochastic Models in Business and Industry 17(4) 307-318.

[202] Bhulai, S., G. Koole, A. Pot. 2008. Simple methods for shift scheduling in multiskill call centers. Manufacturing & Service Operations Management 10(3) 411-420.

[203] Ertogral, K., B. Bamuqabel. 2008. Developing staff schedules for a bilingual telecommunication call center with flexible workers.Computers & Industrial Engineering 54(1) 118-127.

[204] Fu, M. C., S. I. Marcus, I. J. Wang. 2000. Monotone optimal policies for a transient queueing staffing problem. Operations Research 48(2) 327-331.

[205] Saltzman, R., V. Mehrotra. 2007. Managing trade-offs in call center agent scheduling: methodology and case study. In: Proceedings of the 2007 summer computer simulation conference. Society for Computer Simulation International, 2007, 643-651.

[206] Liao, S., C. van Delft, J.P. Vial. 2013. Distributionally robust workforce scheduling in call centres with uncertain arrival rates. Optimization Methods and Software 28(3) 501-522.

[207] Centeno, M. A., R. Giachetti, R. Linn, A. M. Ismail. 2003. Emergency departments II: a simulation-ILP based tool for scheduling ER staff. In: Proceedings of the 35th conference on Winter simulation: driving innovation. Winter Simulation Conference 1930-1938.

[208] Kc, D.S., C. Terwiesch. 2009. Impact of workload on service time and patient safety: an econometric analysis of hospital operations. Management Science 55(9) 1486-1498.

[209] Chan, C.W., G. Yom-Tov, G. Escobar. 2014. When to use speedup: an examination of service systems with returns. Operations Research 62(2) 462-482.

[210] Dong, J., P. Feldman and G. Yom-Tov. 2013. Slowdown services: Staffing service systems with load-dependent service rate. Available at SSRN: http://ssrn.com/abstract=2317410 or http://dx.doi.org/10.2139/ssrn.2317410.