Setting staffing levels in an emergency department: opportunities and limitations of stationary queueing models

Mieke Defraeye and Inneke Van Nieuwenhuyse

Research Center for Operations Management, Department of Decision Sciences and Information Management, K.U.Leuven
Naamsestraat 69
3000 Leuven
Mieke.defraeye@econ.kuleuven.be, inneke.vannieuwenhuyse@econ.kuleuven.be

Abstract. Within the operations management literature, the field of healthcare operations management provides one of the most intensively researched areas of the past two decades. Healthcare systems (like many other service systems) tend to exhibit characteristics which complicate the direct application of traditional models. One of these typical characteristics is the time-varying demand for service: in an emergency department (ED) for instance, patient arrival rates tend to vary throughout the day. This severely complicates the process of determining appropriate staffing levels in order to ensure timely service to the patients. Indeed, an important goal of an ED is to strive for patient waiting times that are sufficiently low for all patients, independent of the arrival time.

In this paper, we highlight the different approaches which have been put forward in the literature in order to (1) approximate the behavior of time-varying systems by traditional (stationary) queueing models, and (2) determine appropriate staffing levels for these systems, in order to meet given waiting time related targets. These models provide a means to evaluate waiting time related performance measures, and can be used as a basis to determine appropriate staffing levels in view of reaching specific performance targets. The applicability and appropriateness of each method in the specific context of an emergency department is discussed, along with the main advantages and drawbacks.

I. INTRODUCTION

In many real life service systems, the demand for service is not constant over time, i.e. fluctuations are usually present on a daily, weekly, monthly or yearly basis. The question that is addressed here, is how personnel capacity planning can be used to deal with this time-varying nature of demand. Moreover, we focus on an emergency department setting (ED) and consequently, some characteristics specific to this context should be taken into account.

In general, variability in demand is to some extent predictable and can be dealt with in several ways, for instance by making appropriate capacity decisions, choosing a suitable appointment system or using waiting lists (Hall et al., 2006). In an ED however, the last two options are inappropriate: consequently,
waiting times are mainly controlled through adequate capacity decisions. In general, an emergency department can be seen as a network, in which patients arrive according to some time-varying arrival process and subsequently move through several process steps in order to receive treatment (an illustration of a simplified ED system is provided in Figure 1).

![Figure 1 Example of a simplified ED system](image)

An ED setting exhibits a range of characteristics that tend to complicate the analysis of waiting time related performance measures. For instance, the *routing of patients is typically not fixed*: i.e., the number and type of process steps tends to be quite different from patient to patient (Williams and Crouch (2006); in Figure 1, an example path followed by two types of patients is shown.). Moreover, *patients are typically treated based on urgency*: priorities are assigned to each patient by means of an urgency code that determines the sequence in which patients are treated (Moll (2010), George et al. (1992)). The more severe the patient’s illness or injuries are, the sooner he should be treated.

One of the main challenges, however, is that the *demand for service is time-varying*, meaning that the number of patients arriving at the ED is not constant over time (Flottemesch et al. (2007), McCarthy et al. (2008), Green and Kolesar (1995), Green et al. (2006)). Fluctuations can occur according to a daily, weekly, monthly or yearly pattern, though the fluctuations on a daily basis tend to be most outspoken (Hall et al., 2006). The number of patients that is expected to arrive per hour can usually be predicted to some extent. As an illustration, Figure 2a shows the hourly arrival rates (for weekdays and weekends) at the ED of the St.-Elisabeth Hospital in Turnhout (Belgium), while Figure 2b displays the arrival rates observed in a NY city emergency department (Green et al., 2006).

![Figure 2 Examples of time-varying arrival rates](image)

- a. Hourly arrival rates St.-Elisabeth Hospital, Turnhout
- b. Hourly arrival rates NY Emergency department (Green et al., 2006)
Clearly, both graphs display similar patterns, e.g. a distinct peak can be distinguished around noon and nightly arrival rates are rather low. These figures, however, only show the expected arrival rates: random (and, hence, unpredictable) fluctuations around these hourly averages will be present, and will influence the system’s performance.

The presence of time-varying demand may lead to temporary overloading of the resources: at some points in time, the arrival rate will exceed the overall service capacity, causing a buildup of patients waiting in queue. Green et al. (2007) state that this feature has a major impact on performance, and hence should not be neglected. The buildup of queues also typically influences the patients’ behavior: if a patient’s waiting time becomes too long, he might choose to leave the ED without being seen by a doctor, resulting in so-called abandonments, or left without being seen (Johnson et al. (2009), Pham et al. (2009)). As mentioned in Green et al. (2007), the presence of abandonments often has a positive effect on the stability of the system, especially when temporary overloading is present.

The remainder of this paper presents a review of the different approaches which have been put forward in the literature in order to (1) approximate the behavior of time-varying systems by traditional (stationary) queuing models, and (2) determine appropriate staffing levels for these systems, in order to meet given waiting time related targets. Traditionally, the literature has focused on so-called single-stage systems (i.e., the network structure is neglected). In section II, we briefly introduce the notation and terminology needed to describe single-stage queueing systems with time-varying behavior, along with the main measures to estimate performance. Section III presents the actual literature review: section III.A discusses the different stationary approximations, while section III.B focuses on the main methodologies to set staffing levels. Finally, section IV presents the conclusions.

II. NOTATION, TERMINOLOGY AND PERFORMANCE MEASURES

Figure 3 gives a schematic representation of a single stage queueing system with time-varying demand. Patients enter the system according to a time-varying pattern with arrival rate $\lambda(t)$. They may leave the system prematurely: the abandonment rate is denoted by $\theta$. Let $\mu$ denote the service rate of the (single-stage) service process, i.e. the rate at which one server (or doctor) can treat customers (or patients). Note that, in general, the service and abandonment rates are assumed to be time-invariant.
The main goal is then to determine an appropriate staffing requirement function \( s(t) \), which defines the number of servers to be assigned at each time instant \( t \) (with \( 0 \leq t \leq T \)). In doing so, the desired system performance (e.g. targets w.r.t. the length of waiting times) must be taken into consideration. Moreover, the number of servers is usually only allowed to change at fixed points in time. The planning period over which the number of servers is assumed to remain unchanged will be referred to as the staffing interval.

When deciding on staffing levels, one first needs to specify what defines a “good” result: an appropriate performance measure must be chosen and a target must be set w.r.t. that performance measure. The performance target should be interpreted as a goal rather than a strict constraint (Green et al., 2001). There are numerous possibilities concerning the choice of appropriate performance measures (an overview of some ED-specific performance measures is provided in Welch et al. (2006)), but here we limit ourselves to a brief description of performance measures that are either interesting in an ED context, or used frequently in the staffing models available in the literature (see for instance Green et al. (2007)):

- **Delay probability**: This is the probability that upon arrival, a customer (or patient) cannot be helped immediately and therefore has to wait until a server (or physician) is idle. This measure is often used in the literature on time-varying demand for service (e.g. Green et al., 2003), Feldman et al. (2008)). Advantages include insensitivity to model details, an interpretation independent of scale (Halfin and Whitt, 1981) and ease of computation (Green et al., 2007). Moreover, choosing delay probability targets can be convenient for some staffing methods, because the link between the desired performance and the number of servers needed to obtain these results, can be established easily if the delay probability is chosen as performance measure (Whitt, 2007). A performance target related to the delay probability could then be defined as in Feldman et al. (2008), namely: choose the staffing level in each interval as small as possible, while meeting the condition that delay probability must lie below some target in all staffing intervals.

- **Expected delay**: In this case, the aim is to obtain waiting times that – on average – lie below some target value (e.g. see Green and Kolesar (1995), Liu and Whitt (2009)). Usually virtual waiting times are used, i.e. the time that a customer would have to spend in queue if he were to arrive at a specific point in time (Gross et al. (2008)). An important advantage of this performance measure is that only little information is needed on the distribution of the waiting time. This advantage is directly linked to the main drawback: when using expected delays, the remainder of the waiting time distribution is neglected. While the length of the wait might be acceptable on average, waiting times might still be intolerably high for some customers, which is of course to be avoided.

- **Service levels**: These can be specified in terms of queue length or waiting times. The former implies controlling for the probability that the queue length is too high, whereas the latter seems more suited to an ED context, as it rather focuses on avoiding large waiting times. A service level target could be to aim at a sufficiently large probability that a patient (who has not abandoned) has a waiting time shorter than some chosen maximal waiting time (Green et al., 2007), i.e. for each patient, there should be only a small chance that the waiting time is “too long”. An example of such a service level constraint can be expressed as

\[
P(W(t) \leq \tau \text{ and served}) \geq x \quad \text{for all } t
\]

where \( W(t) \) represents the virtual waiting time of a customer arriving at time \( t \) and \( \tau \) denotes the maximal allowed waiting time. In words, this means that the probability of an excessive wait should be sufficiently small for all patients, independent of the time of arrival. The importance of meeting a performance goal performance stably during the day is emphasized by Feldman et al. (2008) and Ingolfsson et al. (2010).

---

1 For a number of other performance measures based on the delay probability, we refer to Green et al. (2001).
2 Expected waiting time can be conditioned on whether the customer enters service or not (see for instance the performance measures used in Whitt (2005) and Feldman et al. (2008) or on waiting times being larger than zero (e.g. in Feldman et al. (2008)).
• **Abandonment probability**: If the system is characterized by abandonments, it might be desirable from a managerial perspective to keep the abandonment probability low, rather than to control for the often used delay probability. Intuitively, this performance measure represents customers who leave the system because the waiting time was too long and is therefore related to the customer’s perception of service.

• **Cost**: Keeping labor costs as small as possible, while providing sufficiently good service is an obvious managerial objective. In fact, a tradeoff can be made between labor costs on the one hand, and costs related to the *quality of service (QoS)* on the other hand, as pointed out by Borst et al. (2004). Quality of service is related to customer satisfaction, and a low QoS can be reflected through a penalty for long waiting times or abandonments. The emphasis to be placed on either labor costs or QoS costs depends on the specific problem setting. Based on this observation, three different operational regimes have been suggested by Garnett et al. (2002). In the *efficiency-driven regime*, labor costs are substantially higher than the cost of a low QoS and therefore emphasis is placed on minimizing the staffing levels, which generally results in higher delay probabilities and severely loaded systems. On the other end of the spectrum lies the *quality-driven regime*, where a high QoS is considered more important than keeping labor costs low, which rather leads to overstaffing. The *rationalized or quality-and-efficiency driven (QED)* regime can be situated between these two extremes; here the aim is to provide good service, whilst keeping staffing costs acceptably low. For details on these operational regimes, we refer to Garnett et al. (2002), Koole and Mandelbaum (2002), Zeltyn and Mandelbaum (2005).

### III. LITERATURE REVIEW

An excellent overview of past research can be found in Green et al. (2007) and Whitt (2007). Several methods that are applicable for dealing with time-varying demand when determining staffing requirements in an $M(t)/GI/s(t)+GI$ environment, are touched upon. The emphasis mainly lies on telephone call center applications, although other examples are also provided (e.g. an application in an emergency department). Our aim is to extend this work by including some recent research that has been done in the domain of service systems with time-varying demand patterns. However, all methods are discussed with their specific applicability within the context of an emergency department in mind.

When choosing appropriate staffing levels if demand for service varies over time, two aspects are important (Ingolfsson et al., 2007). On the one hand, there is the problem of finding staffing requirements that satisfy the chosen performance target. An example in an ED context might be to strive for a sufficiently small probability that a patients waiting time exceeds some predetermined maximum wait. On the other hand, the system performance that results from a given staffing level function, must be assessed. Therefore, one needs a method to evaluate this probability (ideally with high accuracy and limited computational effort).

For ease of reference, Table 1 gives an overview of the different performance evaluation methods that are discussed in the remainder of this section, along with some key references and information on intended use, advantages and drawbacks of each method. Table 2 gives an overview of the staffing methods that will be covered.

The remainder of the section is organized as follows: in Section III.A, we discuss the performance evaluation methods that rely on stationary results in order to approximate the nonstationary system. Section III.B discusses the staffing methods.

---

3 The abandonment probability is sometimes conditioned on waiting times being larger than a target value (e.g. Mandelbaum and Zeltyn, 2005). An additional refinement that is applied occasionally, is to exclude customers with very small waits from the abandonment probability (Mandelbaum and Zeltyn, 2004).
### PERFORMANCE EVALUATION METHODS

<table>
<thead>
<tr>
<th>Method</th>
<th>Relevant literature</th>
<th>Intended use &amp; advantages</th>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSA</td>
<td>Green et al. (1991)</td>
<td>Appropriate for small systems, long service times, limited nonstationarity, Fast computation, easy to implement</td>
<td>Only 1 model for the entire time horizon, Nonstationarity is neglected, Fails if system is overloaded</td>
</tr>
<tr>
<td>SPHA</td>
<td>Bear (1980)</td>
<td>Appropriate for long service times (compared to change in arrival rate), Fast computation, easy to implement</td>
<td>Only 1 model for the entire time horizon, Overstaffing, high labor costs, Fails if system is overloaded</td>
</tr>
<tr>
<td>PSA</td>
<td>Green and Kolesar (1991), Jennings et al. (1996), Whitt (1991)</td>
<td>Appropriate for large systems, limited nonstationarity, short service times, Fast computation, easy to implement</td>
<td>Staffing intervals not considered, Poor results for longer service times, Fails if system is overloaded</td>
</tr>
<tr>
<td>SIPP &amp; Segmented PSA</td>
<td>Green et al. (2001), Whitt (1991)</td>
<td>Appropriate for small systems, short service times, short staffing intervals, Fast computation, easy to implement</td>
<td>Length of staffing interval influences performance, poor results for longer staffing intervals, due to averaging, Fails if system is overloaded</td>
</tr>
<tr>
<td>Models with time lag</td>
<td>Green and Kolesar (1995), Green et al. (2001), Green et al. (2003)</td>
<td>Appropriate if service times are somewhat longer, Fast computation, easy to implement</td>
<td>Fails if system is overloaded</td>
</tr>
<tr>
<td>MOL &amp; IS</td>
<td>Jagernem (1975) Massey and Whitt (1994), Massey and Whitt (1997), Liu and Whitt (2009), Jennings et al. (1996)</td>
<td>Somewhat more appropriate if target delay probability is small and system size is large, but good results in various settings, Fast computation, easy to implement</td>
<td>Fails if system is overloaded</td>
</tr>
<tr>
<td>Effective arrival rate approximation</td>
<td>Thompson (1993)</td>
<td>Usually similar accuracy as MOL, Fast computation, easy to implement</td>
<td>Fails if system is overloaded</td>
</tr>
</tbody>
</table>

Table 1 Summary of performance evaluation methods covered in Section III.A

### STAFFING ALGORITHMS

<table>
<thead>
<tr>
<th>Method</th>
<th>Relevant literature</th>
<th>Intended use &amp; advantages</th>
<th>Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Choose smallest staffing level satisfying constraint</td>
<td>E.g. Green et al. (2001)</td>
<td>Intuitive method</td>
<td>Inefficient, Computationally expensive (especially in large systems)</td>
</tr>
<tr>
<td>SRS-rule</td>
<td>Whitt (1992), Koole and Mandelbaum (2002)</td>
<td>Simple rule of thumb, Robust in many situations (empirically and theoretically validated)</td>
<td>Choice of appropriate parameter not always easy (depends on target performance measure), Guideline, so better solutions</td>
</tr>
</tbody>
</table>
A. Stationary Approximations For Performance Evaluation

A first way to deal with time-varying nature of the demand for service, is to try to approximate the nonstationary, time-varying system by one (or more) related stationary model(s). To obtain staffing levels for a nonstationary system using a stationary model, one needs a way to translate the nonstationary system to a similar stationary system. That is, the system characteristics of the nonstationary model have to be translated into appropriate input parameters for the stationary model(s). Often multiple stationary models are applied on smaller intervals, i.e. by dividing the time horizon into subintervals and subsequently applying a stationary model in each separate interval.

Using several stationary models in a nonstationary environment, implies that some assumptions are made (Green et al., 2001), a.o. that (1) delays between separate intervals are statistically independent; (2) steady state is reached in each interval, and (3) the arrival rate remains constant over the staffing interval.

Of course, the validity of these assumptions influences the performance of the resulting staffing levels. Whereas appropriateness of the first two assumptions tends to increase as staffing intervals become longer and service times smaller (cfr. Green et al., 2001) the last assumption is more likely to be approximately valid for shorter staffing intervals (Green et al., 2007). An important additional implication of using stationary models is that no overloading should be present within any given interval, as this would cause instability in the stationary model.

In the remainder of this section, the different performance evaluation methods listed in Table 1 are discussed in further detail.

1. Simple Stationary Approximation (SSA)

The first and most straightforward way to approximate a system with time-varying arrival rates by a stationary system, is to apply a so-called Simple Stationary Approximation (SSA). The idea is to ignore the nonstationarity in the arrival process and simply use the average arrival rate over the entire time period as an input parameter for one single stationary model (Green et al., 1991). Surprisingly, this approach is used frequently in practice, although (not surprisingly) it often leads to poor results because the time-varying nature of demand is neglected.

SSA : apply one M/G/s+G model with arrival rate $\lambda_{\text{SSA}} \equiv \bar{\lambda} = \frac{1}{T} \int_0^T \lambda(t)dt$

Some results on the accuracy of this method are provided in Green et al. (1991). A comparison is made between the actual performance of the nonstationary system on the one hand, which is determined by numerically solving the differential equations (ODE’s) for the M(t)/M/s model, and the performance for the stationary model with an arrival rate equal to the daily average arrival rate on the other hand. Arrival rates are assumed to vary according to a daily sinusoid pattern (Poisson distributed), with a cycle length of one day. The performance measures of interest are the expected delay, queue length and the probability of delay. To allow for a comparison between the performance measures obtained by numerically solving the system’s ODE’s (available \textit{instantaneously}) and SSA performance (where only
one value is available for the entire time horizon), the actual performance values will be averaged over the time horizon. Relative errors, defined as \((\text{actual value} - \text{stationary value})/\text{actual value}\), are calculated. Note that all performance measures concern daily (customer) averages, so no information is provided about the variability in performance throughout the cycle. Experiments were performed to check under what conditions (amplitude, frequency of events, traffic intensity, system size) simply ignoring the nonstationarity of the system leads to sufficiently accurate results. In short, Green et al. (1991) conclude that the SSA systematically underestimates performance and therefore provides a lower bound on all considered performance measures. The SSA performs poorly even for systems with a low degree of nonstationarity and its accuracy deteriorates as the relative amplitude \((\text{RA} = \frac{A}{\lambda})\), where \(A\) represents the amplitude of the sine), system size (measured by the number of servers) or event frequency (i.e. the magnitude of arrival and service rates, for fixed traffic intensity) increase. From a practical point of view, ignoring nonstationarity seems only acceptable for small systems (namely 1 or 2 servers) with very low relative amplitudes (<10%) and event frequencies (low arrival and service rates). According to Jennings et al. (1996), the SSA can be used successfully when service times are long, compared to the change in arrival rate.

2. Simple Peak Hour Approximation

A second approach that approximates the nonstationary system with only one stationary model, relies on the idea that it is sufficient to choose staffing levels such that good performance is guaranteed during the peak period. This is referred to as Simple Peak Hour Approximation (SPHA) and is described in Green and Kolesar (1995, 1997), although a similar principle was proposed in Bear (1980), named busy-hour engineering. The system’s performance is approximated by a stationary model, with an arrival rate equal to the average arrival rate over the peak hour (i.e. the hour surrounding the time instant at which the arrival rate peaks) or peak epoch\(^4\) (i.e. the time instant at which the arrival rate peaks). SPHA can lead to good results for high service rates (e.g. 100 customers per hour or more), again because steady state is then reached faster (Green and Kolesar, 1995). When staffing intervals are long (consider for instance 8-hour shifts), SPHA might be a good choice as well, provided that the targeted quality of service is high (because staffing for peak demand obviously results in overcapacity during calmer periods).

\[
\text{SPHA} : \text{apply one M/G/s+G model, with arrival rate } \lambda_{\text{SPHA}} = \frac{1}{x_2-x_1} \int_{x_1}^{x_2} \lambda(t) \, dt \\
\text{where } [x_1, x_2] \text{ represents the peak period}
\]

3. Pointwise Stationary approximation (PSA)

In the Pointwise Stationary Approximation (PSA) implies that steady state performance measures are calculated for each moment in time, using the arrival rate at each moment in time instant in a stationary model (Green and Kolesar (1991), Jennings et al. (1996), Whitt (1991)). Consequently, PSA does not account for the presence of staffing intervals (during which the capacity remains unchanged).

\[
\text{PSA} : \text{apply a M/G/s+G model on each time instant } t, \text{ with arrival rate } \lambda_{\text{PSA}} = \lambda(t)
\]

The effectiveness of this approach is discussed in Green and Kolesar (1991). Analogous to Green et al. (1991), daily average expected queue length, daily average expected waiting time, daily average

\(^4\) Using a peak period (e.g. an hour) or a peak epoch only leads to minor changes in results.
probability of delay and daily average probability of all servers busy are chosen as performance
measures (again, these are all averages over the time horizon of one day). Relative errors are calculated
for each measure, this time comparing the actual performance to the PSA performance. The effect of
changes in system parameters (in an M(t)/M/s system) are explored for sinusoidal Poisson arrival
patterns. The parameters that are varied include the event frequency, service rate and maximal traffic
intensity (equal to the maximum of the time-dependent traffic intensity \( \rho(t) = \frac{\lambda(t)}{s(t)} \mu \), which is
influenced by the amplitude of the arrival rate, the average arrival rate and the number of servers).

PSA becomes more accurate as event frequency increases, i.e. as the average arrival rate \( \bar{\lambda} \) and
service rate \( \mu \) increase with the same amount, while keeping the offered load constant (or equivalently,
as the time horizon becomes longer); a result that is consistent with Green et al. (1991), where
empirical evidence implied that the expected queue length and delay probability asymptotically
approach the PSA as the rates increase (provided that the PSA exists) and. A formal proof of this is
provided in Whitt (1991), who shows for M(t)/M/s queues that PSA is asymptotically correct as event
frequency increases. Alnowibet (2004) point out that this effect is due to the steady state being reached
more quickly for higher arrival and service rates, leading to a better correspondence to steady state
behavior at each moment in time (which is implicitly assumed). Increasing the arrival or service
separately lead to the following conclusions: effects concerning an increase in the arrival rate are
unclear, but performance can be estimated more accurately for higher service rates. Accuracy seems
acceptable for sufficiently high service rates, e.g. of at least 2 customers per hour (Green and Kolesar,
1991). Results on the effect of increasing maximal traffic intensity indicate that the accuracy of PSA
decreases as maximal traffic intensity approaches one. For a maximal traffic intensity that exceeds 1,
PSA does not produce finite values anymore, because of instability in the stationary models. For all
components influencing the maximal arrival rate (amplitude, average arrival rate and number of
servers), the effect is as could be expected: larger amplitudes and average arrival rates result in larger
errors, whereas an increase in the number of servers leads to lower errors. Also, experiments with
varying amplitudes indicate that PSA is most appropriate when arrival rate changes relatively slowly
compared to the service rate. This is intuitive: over an interval in time, there is a higher resemblance
between a slowly fluctuating arrival rate function and a constant arrival rate (i.e. the starting point for a
stationary model), compared to an arrival rate function that changes more rapidly. Because PSA
implies that a stationary model is applied on each moment in time, this approach will evidently perform
better as the resemblance to a stationary system increases. In Jennings et al. (1996), this intuition is
confirmed for infinite server models with a sufficiently smooth arrival rate, by means of the quadratic
approximation for the offered load that was provided in Eick et al. (1993a) (which will be discussed
more in detail in a next subsection). There, it was shown that PSA is approximately correct (and thus
more effective) if the arrival rate changes slowly and if service times are small. It has been proven that
the PSA approximation provides an upper bound on the expected queue size in an M(t)/M/s system
under the condition that stationary models are applicable, i.e. traffic intensity never exceeds one (cfr.
Grassmann (1983), who first used this approach, and Green et al. (1991)).

The PSA approach does not impose capacity to remain constant throughout a staffing interval. To
account for the fact that capacity can usually only be altered at specific points in time, two refinements
to the PSA approach can be used: SIPP and Segmented PSA.

4. PSA-based approaches accounting for staffing intervals

Intuitively, SSA and PSA can be considered as two opposites. In SSA, arrival rate is averaged over the
entire time horizon and used in one single stationary model, whereas the PSA approach uses no
averaging at all and applies a stationary model on each moment in time. According to Whitt (1991), an
approach that lies between these two extremes (for instance by averaging arrival rates over an interval)
could be useful, although it is not clear how the length of this interval should be determined. A
common choice is 30 minutes (cfr. Green et al., 2001) but an interval length proportional to the
expected service rate (Whitt, 1991) can be used as well. As stated in Massey and Whitt (1996), this
interval should be small enough to obtain a sufficiently precise performance measure, yet large enough to capture the “relevant past” within the interval (particularly if independence between intervals is assumed).

**Stationary independent period-by-period method (SIPP)**

One approach to account for the length of staffing intervals is to use a stationary model in each interval, with the arrival rate averaged over that interval as input parameter. This is commonly referred to as the *Stationary Independent Period-by-Period (SIPP)* approach (Green et al., 2001), although in Whitt (1991) a similar approach is proposed, named the *Average Stationary Approximation*. From a practical point of view, note that SIPP is advantageous in the sense that it tends to avoid arrival rate estimation and smoothing issues, because call centres often only provide piecewise constant arrival rate functions. An application of SIPP to set staffing levels in an ED, is provided in Green et al. (2006).

**SIPP** : apply one M/G/s+G model per interval \( i \) with arrival rate \( \lambda_{SIPP} \equiv \frac{1}{x_2-x_1} \int_{x_1}^{x_2} \lambda(t) dt \)
where \([x_1,x_2]\) represents the \( i^{th} \) staffing interval

Some results on the accuracy of the SIPP approach in an M(t)/M/s(t) model for varying model parameters, are given in Green et al. (2001). The reliability of the SIPP method as a function of arrival rate, service rate, relative amplitude, target delay probability and the length of the staffing interval is examined. All considered performance measures are related to the delay probability. In the SIPP method, arrival rates are averaged over the staffing interval, therefore it can be expected that its accuracy depends upon the length of this interval. Indeed, the main result is that SIPP should be avoided when the arrival rate changes substantially over the staffing interval. This implies that SIPP performance can be expected to be poor for long staffing intervals or arrival rates varying with a large (relative) amplitude. Moreover, SIPP is most appropriate for short service times (as is also confirmed in Green et al. (2007), who provide an example with an average service time of 300 minutes where the SIPP method fails to achieve reasonable performance). This effect is probably related to the assumption of independence between consecutive staffing intervals: within each staffing interval, steady state is reached faster for larger service rates. Results concerning the effect of changes in the average arrival rate are not consistent, although SIPP performance tends to deteriorate slightly as the average arrival rate increases. Furthermore, SIPP seems most appropriate for small systems, and only works well if no overloading is present, because this would cause instability in the stationary models (Green et al., 2007).

However, the SIPP approach has some downsides, for example the underestimation of staffing requirements in staffing intervals where the arrival rate decreases (this is due to the use of the averaged arrival rate over an interval). Therefore, to further improve the method, three refinements to the SIPP approach are provided in Green et al. (2001).

In the *SIPP Max* approach, the maximum arrival rate over staffing interval is used, instead of the average arrival rate. As can be expected, this approach is always more reliable, but unfortunately at the expense of higher staffing costs, as the staffing function will always be higher than SIPP staffing. In an attempt to combine the reliability of SIPP Max while keeping staffing costs as low as possible, the *SIPP Mix* approach was developed. Here, SIPP is applied if the arrival rate is strictly increasing over interval and SIPP Max is used otherwise. Performance turned out to be better than SIPP in some settings, although in general, it does not outperform the original SIPP approach. A third extension of SIPP, *Lagged SIPP*, was presented as well but this will be discussed in a next subsection.

**Segmented PSA**
Another way to keep staffing levels constant during a staffing interval is to set the staffing level equal to the maximum of the PSA staffing requirements over that interval, which is referred to as the \textit{Segmented PSA} (Green et al., 2007). This way, staffing levels may turn out to be too high, but these can be improved afterwards, for instance by means of simulation. Segmented PSA is equivalent to the SIPP Max approach discussed in the previous subsection.

\begin{center}
\textbf{Segmented PSA :}
\begin{enumerate}
\item PSA: apply a M/G/s+G model on each time instant \( t \), with arrival rate \( \lambda_{\text{PSA}} \equiv \lambda(t) \)
\item Staffing level \( s_{\text{Segmented PSA}}(t) \) in interval \( i = \max_{t \in \text{interval}}(s_{\text{PSA}}(t)) \)
\end{enumerate}
\end{center}

\section*{5. Models with a time lag}

When comparing the arrival rate to the actual offered load, it becomes clear that the peaks in actual offered load lag the arrival rate peaks (Green and Kolesar, 1995). This is illustrated in Figure 3, where the actual offered load, the PSA offered load, and the arrival rate are plotted for an example setting.\footnote{Notice that the actual offered load curve is intersected by the PSA offered load at its peak value. Empirical findings hereon were reported by Green and Kolesar (1997) and proved to be valid for the infinite server M/G/\( \infty \) queue in Eick et al. (1993a).} Apparently, the size of the time lag is proportional to the expected service time, as can be deduced by comparing Figure 4 with Figure 5, where arrival rate and offered loads are plotted for small and large service rates respectively. Green et al. (2001) describe the concept of such a \textit{time lag} intuitively as \textit{“the amount of time during which the arrival rate at a given epoch will continue to impact system congestion”} (Green et al., 2001). As each customer stays in the system for \( E[S] = \frac{1}{\mu} \) time units on average, the presence of a link between this time lag and the service time seems reasonable. As the PSA offered load is in phase with the arrival rate, this implies that the actual offered load lags the PSA too, resulting in a decreased effectiveness as service times increase. This implies that taking this time lag into account is particularly useful when service times are long.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{time_lag.png}
\caption{Illustration of time lag between arrival rate and actual offered load (for large expected service times)}
\end{figure}
Evidence supporting this intuition was given in Eick et al. (1993a), who derive a quadratic approximation of the infinite server offered load\(^6\). Besides a service time random variable \(S\), this approximation utilizes the concept of a service time stationary-excess random variable \(S_e(t)\), with cumulative distribution function

\[
P(S_e \leq x) = \frac{1}{E[S]} \int_0^x P(S \leq u) \, du
\]  

The quadratic approximation is then represented as

\[
m_{\infty}^{\text{QUAD}}(t) = \lambda(t - E[S_e]) E[S] + \frac{\lambda'(t)}{2} \text{Var}[S_e] E[S]
\]  

Which approximately equals the PSA offered load \(\lambda(t) E[S]\), although a time lag in the arrival rate as well as a space shift are present, defined as

\[
\text{Time lag: } -E[S_e]
\]

\[
\text{Space shift: } \frac{\lambda'(t)}{2} \text{Var}[S_e] E[S]
\]

Thus, the quadratic approximation indicates that the PSA is approximately correct if this time lag and space shift are negligible. The presence of a time lag again can be visualized: the actual offered load lags both the arrival rate and the PSA offered load (the latter two are in phase). This observation supports the use of a lagged variant of the PSA (denoted \(\text{LagPSA}\)), which would be represented graphically by shifting the PSA curve in Figure 3 and Figure 4 to the right. So, if one would be able to determine an appropriate value for this lag (denote it as \(L\)), then a better approximation entails the application of the PSA logic with arrival rate \(\lambda(t - L)\) for all \(t\), instead of \(\lambda(t)\). The lag is commonly approximated by the expected service time \(E[S]\). Theoretical background for this choice can be found in Eick et al. (1993a), where it is shown that this provides good results if the arrival rate is approximately quadratic before time \(t\) and the squared coefficient of variation lies around 1.

---

\(^6\) The infinite server offered load can be described as the time-varying expected number of customers in system if infinitely many servers would be available. In an infinite server system, this equals the expected number of busy servers.
Similarly, a time lag can be added when applying the SIPP or SIPP Max method (Green et al. (2001), Green et al. (2003)). In the Lag SIPP approach, performance in interval \([x_1, x_2]\) is approximated by a stationary model with arrival rate 
\[ \lambda_{LagSIPP} = \frac{1}{x_2 - x_1} \int_{x_1}^{x_2} \lambda(t - E[S])dt \] 
where interval \(i\) starts at \(x_1\) and ends at \(x_2\). In Green et al. (2003), an evaluation of staffing levels generation by these methods indicates that adding a lag always leads to better results. Roughly speaking, they suggest to use Lag SIPP when staffing intervals are short and the amplitude of the arrival rate is rather small. In all other situations, Lag SIPP Max is more appropriate.

6. Infinite server (IS) and modified offered load (MOL) approximations

If service times are long and the time lag and space shifts can no longer be neglected, the need arises for a better method to translate nonstationarity into input parameters for the stationary model. This is where the Modified Offered Load (MOL) approach comes in: based on convenient closed-form results for infinite server queues, a modified arrival rate function is derived which proves to be a better approximation of the offered load in the nonstationary system. Similar to the previous methods discussed in this section, this modified arrival rate is then again used in a series of stationary models.

Instead of using MOL, one could use an Infinite Server (IS) approximation too. The difference between IS approximations and MOL approximations is not always stated clearly in the related literature, yet it is important to differentiate between both methods. In the IS approximation on the one hand, the nonstationary \(M(t)/G/s(t)+G\) model is approximated by a nonstationary \(M(t)/G/\infty\) model, which can be solved by numerical integration of ODE’s or by means of convenient analytical expressions for the infinite server offered load. The number in system is then approximated by the infinite server number in system, or equivalently, the number of busy servers that would prevail if infinitely many servers would be available. In the MOL approach on the other hand, the analytically tractable results on infinite server queues are used to obtain an adapted arrival rate that leads to better results when plugged into a stationary model. In a strict sense, IS approximations do not fall in the category of stationary approximations, but because IS and MOL approximations are frequently discussed simultaneously in the literature, some results concerning the IS approach will be given here as well.

The MOL concept was first introduced in Jagerman (1975), who applied MOL for time-dependent multiserver Markovian Erlang loss models\(^7\) and further explored in Massey and Whitt (1994), where MOL was analysed from a more mathematical perspective and error bounds were provided. In Massey and Whitt (1997), the accuracy of MOL for predicting peak congestion in \(M(t)/M/s\) models is examined. In the literature, the main focus lies on Markovian systems, aiming at the realization of a certain delay probability target. Nonetheless, the MOL approach should be extendable to both more general \(M(t)/G/s(t)+G\) models (Whitt, 2007) and other performance measures (results can be found Liu and Whitt (2009) and Feldman et al. (2008)), which make this a promising approach for application in the context of an emergency department.

\(^7\) Loss models can be described as systems where no waiting room is available and customers leave (instead of entering a queue) if no idle servers are available upon arrival (Gross 2008).
**IS**: approximate the $M(t)/G/s(t)+G$ model by a $M(t)/G/\infty$ model

**MOL**: apply a $M/G/s+G$ model on each time instant $t$, with arrival rate $\lambda_{MOL}(t) \equiv \mu m_{\infty}(t)$

The idea behind the MOL approach is fairly simple: use a stationary model on each moment in time (similar to the PSA), but replace the nonstationary arrival rate function by the product of the service rate and the infinite server offered load function prevailing at that time (denoted by $m_{\infty}(t)$). The latter represents the number of servers that would be used on each moment in time if infinitely many servers would be available. The convenience of this approach lies in the fact that the IS-model is analytically tractable: closed form expressions for the infinite-server offered load are available. Eick et al. (1993a) prove that number of customers in an $M(t)/G/\infty$ queue follows a Poisson distribution, with mean

$$m_{\infty}(t) = \int_{-\infty}^{t} (1 - G_u(t-u)) \lambda(u) \, du$$

With $G_u(t) = P(S(u) \leq t)$ the cumulative distribution function related to the service process service $S(u)$ the service time of an arrival at time $u$. In the case of a Markovian service process, simply solving the ordinary differential equations to obtain $m_{\infty}(t)$ is an option, whereas under phase-type service time distributions, a set of differential equations can be solved as is done in Davis et al. (1995) and Massey and Whitt (1994). However, closed-form expressions for $m_{\infty}(t)$ exist (see Eick et al. (1993b) for the case of sinusoidal arrival rates) and several efficient algorithms to determine $m_{\infty}(t)$ are provided in Jennings et al. (1996), who also evaluate the accuracy of the IS approximation for $M(t)/M/s(t)$ models.

Jennings et al. (1996) state that the MOL approximation results in more accurate performance than the IS approximation, although in their opinion, the improvements have only little influence on the resulting staffing levels. Green et al. (2007) experienced that MOL becomes less effective if the system is overloaded, because of instability in the corresponding stationary models. Moreover, the MOL approximation performs well if the number of servers is sufficiently high (Ingolfsson et al., 2007) because then, the infinite server approximation proves to be accurate. The error bounds on the MOL approximation provided in Massey and Whitt (1994) lead to the conclusion that MOL is asymptotically correct as the arrival rate changes more slowly and thus its derivative approaches zero. In contrast, examples given in Jennings et al. (1996), show that the IS approximation should work well for both slowly and rapidly changing arrival rates (contrary to the PSA approach, which is in general more suitable for slowly varying arrival rates). As the number of servers declines, the MOL approximation becomes less accurate because of a lower resemblance to the infinite server system (Massey and Whitt, 1997). However, MOL still performs surprisingly well in small systems because the IS approximation remains fairly accurate even when system size is small (Jennings et al., 1996). Also, IS seems more advisable for small target delay probability, although large delay probabilities do not substantially worsen performance (Jennings et al., 1996). In comparison with other methods, Ingolfsson et al. (2007) report that both perform better than Lag SIPP (at the expense of somewhat higher computation times), although IS is less accurate than MOL.

Recently, Liu and Whitt (2009) further explored the possibilities of this MOL approximation, resulting in two new approximation methods for $M(t)/G/s(t)+G$ queues. Here, instead of targeting a particular delay probability, other performance measures are considered, namely the abandonment probability and expected waiting time.

When an infinite server approximation is used, the effect of abandonments is neglected. To account for this, the *delayed infinite-server offered load* (DIS-OL) model is proposed. Two infinite server models are applied, the first of which represents a virtual waiting room where all patients are delayed for a fixed amount of time, depending on the desired expected waiting time or abandonment.
probability. After departing from this virtual space, all non-abandoned customers enter a second infinite server system, that will provide the infinite server offered load to be used for making staffing choices, similarly to the regular MOL approach.

Comparing the DIS-OL and MOL, it appears that DIS-OL succeeds in keeping abandonment probability stable over time, independent of the load, whereas in MOL the delay probability was only stabilised under lighter loads (or for small delay probabilities). Moreover, results indicate that DIS-MOL works well even if no Markovian assumptions are made, although results depend on the system size (as an indication, arrival rates of approximately 100 customers are mentioned). In addition, the authors address issues such as discretization of the staffing level function (because the model returns continuous values), the exhaustiveness of the service process and accounting for staffing intervals.

7. Effective arrival rate approximation

In the MOL approach, arrival rates are altered such that the time-varying system characteristics are approximated more accurately. In Thompson (1993), a similar approach is used, but now a different calculation method is used to specify the adapted arrival rate that is to be used in a stationary model. Waiting times are assumed to be deterministic and are approximated by the expected waiting time (denoted by $W$); service times are assumed to be equal to $1/\mu$. The arrival rate $\lambda(t)$ is then determined as a moving average over the interval $[t-W-1/\mu, t-W]$, as this interval will determine the number of customers in service at time $t$. Results on $M(t)/M/s(t)$ models by Ingolfsson et al. (2007) indicated that this approach is similar to MOL in terms of accuracy, but systematically faster (although differences are rather small).

**Effective arrival rate approximation:** apply a $M/G/s+G$ model on each time instant $t$, with arrival rate $\lambda_{\text{Effective ArrRate}}(t) \equiv \mu \int_{x-W-1/\mu}^{x-W} \lambda(t)dt$

B. Determining staffing levels using a stationary approximation

Once the nonstationary model is translated into a series of related stationary models and the outcome of those models is obtained, staffing levels need to be chosen such that the performance target is satisfied (using as few servers as possible).

A straightforward (yet computationally expensive) way to address this problem is to calculate stationary outcomes for several staffing levels in each stationary model, after which the smallest staffing level that meets the constraint is chosen. This logic is applied in Green et al. (2001), where SIPP staffing levels are determined by repeatedly assessing the Erlang delay equation (Gross et al., 2008) for each interval and then choosing the smallest staffing level that meets the delay probability target. One important drawback lies in the calculation of steady state performance measures: depending on the system characteristics, closed form expressions for the stationary performance are not always available and therefore “applying a stationary model” is not always as straightforward as one might expect. This particularly holds for the difficult $M/G/s+G$ model, which we focus on. Consequently, approximations are often necessary for obtaining the steady state performance measures that are needed to determine appropriate staffing requirements. Research has been done on approximating performance measures of stationary models. In Table 3, references to the relevant literature are given, along with a short description. Reported performance measures include the delay probability (given that the customer is served), abandonment probability (given that the wait exceeds a target), probability that the waiting time exceeds a target, expected waiting time (given that the customer is served or given that the wait exceeds a target).
Sometimes, staffing levels can be determined without explicitly calculating steady state performance measures. One rule of thumb that has received a fair amount of attention because of its simplicity and robustness is the square-root staffing rule (SRS), often also referred to as square root safety staffing (general background and applicability of SRS is provided in Whitt (1992) and Koole and Mandelbaum (2002)). Here, staffing levels \( s(t) \) are chosen to be equal to the offered load \( m \) (i.e. the expected number of customers in the system), augmented with an amount of safety capacity, proportional to the square root of the offered load (cfr. expression (6)). The SRS rule has proven to be remarkably accurate in different settings (e.g. in the PSA approach or IS approximation, cfr. Jennings et al. (1996)).

\[
\text{SRS-rule: } s = [m + \beta \sqrt{m}]
\] 

Again, infinite server approximations prove to be helpful to improve general understanding of this rule (for extensive theoretical background on the SRS rule we refer to Whitt (1992)). Here, we briefly discuss how the SRS-rule can be applied in an M(t)/G/s(t)+G system. First, the M(t)/G/s(t)+G system is approximated by an infinite server M(t)/G/\( \infty \) model. In this latter model, the number of busy servers at time \( t \) follows a Poisson distribution with mean equal to the offered load \( m_\infty(t) \), if the number in system at time 0 is Poisson as well (Eick et al. (1993a), Eick et al. (1993b)). Although this distribution is not valid exactly in the M(t)/G/s(t)+G system\(^8\), it tends to be more suitable as the number of servers increases (Whitt, 1992). Or, the number in system in the M(t)/G/s(t)+G (denoted by \( N(t) \)) is approximated by the infinite server number in system (represented as \( N_\infty(t) \)):

\[
N(t) \approx N_\infty(t)
\] 

with \( N_\infty(t) \rightarrow \text{POISSON} \left( m_\infty(t) \right) \)

This Poisson distribution can in turn be approximated by a normal distribution with both mean and variance equal to the offered load, under the condition that the offered load is not too small and targeted quality of service is high (Green et al., 2007).

\(^8\) In one particular case, the distribution of number in system in M(t)/G/s(t)+G is identical to that of the infinite server model, namely if the specific condition that the abandonment rate is equal to the service rate (Whitt, 2007).
\[
N_\infty(t) \xrightarrow{\text{approx.}} \text{NORMAL}\ (m_\infty(t), m_\infty(t))
\]

Where \( \text{NORMAL}(\mu, \sigma^2) \) represents a normal distribution with mean \( \mu \) and variance \( \sigma^2 \). Now the targeted delay probability \( \alpha \) can be written as

\[
\alpha \equiv P(N_\infty(t) \geq s(t)) \approx 1 - P\left(\frac{N_\infty(t) - m_\infty(t)}{\sqrt{m_\infty(t)}} < \frac{s(t) - m_\infty(t)}{\sqrt{m_\infty(t)}}\right)
\]

\[
\approx 1 - \Phi\left(\frac{s(t) - m_\infty(t)}{\sqrt{m_\infty(t)}}\right)
\]

Where \( \Phi \) denotes the standard normal cdf. Which results in an expression that links the target delay probability to a quality of service parameter \( \beta \):

\[
\beta \equiv \frac{s(t) - m_\infty(t)}{\sqrt{m_\infty(t)}}
\]

This leads to the square root staffing formula for the M(t)/G/s(t)+G model, a rule of thumb that can be applied in order to obtain staffing levels that meet the targeted delay probability.

\[
s(t) = \left| m_\infty(t) + \beta \sqrt{m_\infty(t)} \right|
\]

The factor \( \beta \) is chosen depending on the performance target and evidently, the extent to which SRS will lead to overstaffing or understaffing is directly related to the specification of this parameter. However, the normal approximation might not always be suitable and thus other ways were explored to link the desired performance to a parameter \( \beta \) that can be plugged into the SRS rule. If performance is defined in terms of the delay probability, two well-known expressions can be used, although additional Markovian assumptions might be necessary. Inverting the Halfin-Whitt delay function can be used to obtain the appropriate \( \beta \) in a stationary M/M/s model (Halfin and Whitt, 1981); the Garnett delay function is meant for Markovian M/M/s+M models with abandonments (Garnett et al., 2002) and in Zeltyn and Mandelbaum (2005) the M/M/s+G model is addressed. These functions are derived for many-server heavy-traffic regimes, but should work for small system sizes as well (although accuracy increases as the number of servers and the arrival rates approach infinity); examples supporting this are provided in Jennings et al. (1996). If other performance target are used, these delay functions can be used after reformulating the performance measure of interest in terms of the delay probability.

For instance in M/M/s models, waiting time percentiles can be expressed in terms of the delay probability in the following way, because the conditional waiting time given that all servers are busy is exponential (Whitt, 1992):

\[
\alpha \equiv P(W \leq x)
\]

\[
= P(W > 0) \cdot P(W \leq x|W > 0) = P(W > 0) \cdot e^{-x(1-\rho)}
\]

The symbols \( W \) and \( x \) represent the waiting time and maximal allowed waiting time respectively and \( \alpha \) stands for the targeted probability of a waiting time smaller then the maximal allowed wait (the tilde is used to emphasize the difference with previous targets which were related to the delay probability instead of waiting time service levels). This enables the use of the Halfin-Whitt delay function, as the delay probability \( P(W > 0) \) is linked to \( \beta \) through this expression. This approach can be extended to other performance measures, e.g. Borst et al. (2004) suggest approximate \( \beta \)-values that are related to labor and waiting costs. While approximate values of \( \beta \) are often available, determining exact expressions remains an important challenge. If available delay functions or approximations fall short, an alternative option is to determine the appropriate beta trough simulation (Liu and Whitt, 2009).
IV. CONCLUSIONS AND RECOMMENDATIONS

In this paper, we have discussed the different approaches which have been put forward in the literature in order to (1) approximate the behavior of time-varying systems by traditional (stationary) queuing models, and (2) determine appropriate staffing levels for these systems, in order to meet given waiting time related targets. Two subproblems can be distinguished: on the one hand, staffing levels need to be determined for each moment of the day and on the other hand, the system’s performance resulting from a staffing level function needs to be assessed, to check whether the imposed performance constraints are met.

An ED context is characterized by a small system size (few servers), low abandonment probabilities (most patients are willing or obliged to wait) and periods of overloading. Typically, the aim is to strike a balance between quality of service (in terms of waiting times) and efficiency (in terms of labor costs). In our opinion, the most promising method in view of practical applications is MOL, combined with the square root staffing rule, because of the simplicity and limited computational effort. In certain settings, the even simpler Lag SIPP approach can be considered as well, although in general MOL leads to more accurate results. However, this requires an appropriate specification for the quality of service parameter $\beta$ in the square root staffing rule, to allow for waiting time related performance measures (and preferably general service times and abandonments as well).

The main advantage of the stationary model approach is that the models are usually fast and fairly easy to implement. However, due to the use of approximations, satisfactory performance is not guaranteed. Under certain circumstances stationary approximations indeed fail. These circumstances are linked to the implicit assumptions discussed in section III.A. As highlighted there, stationary models assume that the arrival rate is constant over the time period. While this is obviously not the case in practice, some stationary approximations succeed quite well in translating nonstationary characteristics in constant arrival rates that are then used in stationary models. Consequently, this assumption can be acceptable, especially when the length of the staffing intervals decreases (then, the change in the arrival rate over the staffing interval becomes negligible). The main problems are caused by the assumptions that successive staffing intervals are considered to be independent, and that steady state is reached within each staffing interval (so transient effects can be neglected). The independence assumption is likely to be violated in an emergency department, especially when delays are frequent and waiting time related performance measures are used: e.g. a high excess wait probability in an interval $t$ often results in a higher excess wait probability in the next interval $t+1$ (because of a large number of patients in the waiting room at the start of the interval $t+1$). The validity of the steady state assumption, on the other hand, increases along with the event frequency: the higher the frequency of arrivals and departures per interval, the faster steady state is reached. In an emergency department, though, these event frequencies tend to be low.

In conclusion, the implicit assumptions for stationary approximations are not always met. Nevertheless, these simple approaches can provide a good starting point for setting staffing levels. Their effectiveness can be checked in the specific context of interest (e.g. by means of a simulation model). If stationary approximations fail to capture the system’s dynamics, one can try to improve and fine-tune the corresponding staffing levels (e.g. based on simulation results), or alternatively, one might resort to (often more complex and computationally expensive) methods, which explicitly account for the nonstationarity in the system.

For completeness, we would like to draw attention to the performance measures used in the discussed stationary approximations. These all focus solely on the quality of service (e.g. performance is measured through expected wait or delay probability) and the cost aspect is neglected. From a managerial perspective, labor costs should obviously be accounted for and staffing levels should be set in a way that the performance is obtained whilst keeping labor costs under control.
Acknowledgements

This research was supported by the Research Foundation-Flanders (grant no. G.0547.09).

References


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