Evaluation of service level approximations in call centers

Approximations of the non-stationary M(t)/M/s(t) queue

Research Paper Business Analytics

Fatima Babat

Supervisor: Dr. Alex Roubos

June 27th 2015

VU University
Faculty of Sciences
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands

Preface

This research paper is written as part of the Business Analytics Master program at the VU University Amsterdam. The research paper aims at developing the ability to perform research and to describe a problem clearly for a competent manager with the emphasis on the business part of the study besides mathematics and computer science.

I would like to thank my supervisor dr. Alex Roubos for his help with this research paper.

Abstract

This research paper sets out to explore the methods that are available to assess the quality of agents' shift schedules in call centers. Call centers are best modelled by the time-dependent and non-stationary *M(t)/M/s(t)* queueing model. Unfortunately, exact solutions of this model are only known for a number of specific configurations and numerical solutions require high computation times. Because of this, methods have been developed that approximate the behaviour of this queue.

In this paper we provide the theory behind the six approximation methods that are most discussed in literature and most used in practice. Among these methods are the Pointwise Stationary Approximation, the Stationary Independent Period-by-Period approximation and the Modified Offered Load approximation. We created two different scenarios in which we compared the performance of the methods using two different ways of scheduling. As a performance measure we used the service level, which is the widely used measure for call centers. By using the numerical solution of the *M(t)/M/s(t)* queue as a benchmark, we were able to compare the approximations.

Contents

1. Introduction

Call centers are used by many businesses as the central point for their customer service and communications. Because of the large number of operational call centers today, there is a great interest in optimizing their functioning. Improving the scheduling of the employees, referred to as agents in the call center environment, can help significantly in reducing the workforce costs. From a customer point of view, optimal scheduling should also help ensure that there is enough staff available to limit the waiting time before service to a certain level.

The optimization of employee scheduling is typically divided into four components: (1) forecasting the demand; (2) translating this demand into staffing requirements; (3) shift scheduling; and (4) rostering (Buffa et al., 1976). Ingolfsson et al. (2010) argue that these components should all be taken into account at once to optimally schedule agents. However, most other researchers have separated them; using approaches that are focused on either determining the staffing requirements or solving the actual scheduling problem. This paper will focus on the methods available to calculate the quality of any given shift schedule. The actual shift scheduling, consisting of searching for feasible schedules based on the required capacity derived from the previous steps and complex constraints, is outside of the scope of this paper.

The demand in a call center, the number of incoming calls, is dynamic in nature. As a consequence, the number of agents required also varies throughout the day. To forecast the demand and the accompanying staffing requirements in a call center, queueing models are used. Unfortunately, because classical queueing models focus on the long-run behaviour of queueing systems with stationary arrival rates and server amounts, they cannot be directly applied to model this problem.

Therefore, researchers have looked into developing adjusted schemes to model non-stationary queueing systems. Closed-form solutions have only been found for very particular cases and the computation of these solutions can be numerically challenging and requires high computation times, as demonstrated by Ingolfsson et al (2007). Consequently, various approximation methods have been developed and are being used in practice.

Research indicates that these approximation methods vary widely in applicability, speed and accuracy. Also, some of the assumptions made in these models might not be justified. This research paper will set out to review the most applied approximation methods through an extensive literature study and comparative (numerical) experiments in certain scenarios. By doing this, the research question that we aim to answer is:

"What are the strengths and limitations of currently used methods for determining the quality of staffing schedules in time-dependent and non-stationary queueing systems, such as call centers?"

This paper starts by a definition of the queueing model that is best fit to model a call center in Chapter 2. In Chapter 3 the theory behind approximation methods of this model is discussed. Chapter 4 provides the result of an implementation of the discussed models, after which the discussion follows in Chapter 5.

2. M(t)/M/s(t) queue

The queueing model that might seem fitting to model a call center is the M/M/s queue. This model, better known as the Erlang C model, can be used to describe queueing systems with multiple servers in which the arrivals can be described by a Poisson process and the service time is distributed exponentially. Because of the simplicity of applying the Erlang C model, it is widely used.

However, the Erlang C model assumes all three of its parameters to be time-independent and remain constant. In a call center, we know that the demand fluctuates in a certain pattern throughout the day and we want to be able to vary the staffing level based on these fluctuations. Therefore, the non-stationary M(t)/M/s(t) queue is better fit to model the dynamic conditions of a call center. The paragraphs in this section each discuss a certain aspect of this queueing model.

2.1 Arrival process

The arrivals in the M(t)/M/s(t) queue are assumed to happen according to an inhomogeneous Poisson process. To account for the inhomogeneity, the constant arrival rate λ in an M/M/s queue is replaced by a time-dependent function λ(t). This function can be modelled in many different ways to replicate the arrival behaviour in a call center. For example, Feldman et al. (2008), Green et al. (1991, 2001) and Ingolfsson et al. (2010) use a sinusoid to reflect the behaviour of the demand throughout the day. Jongbloed and Koole (2001) use Poisson mixtures to model the uncertainty in the demand. For practical purposes, the arrival rate is often specified to be piecewise constant. This means that the arrival rate does not vary over time continuously when modelling, but remains constant over a fixed (short) period. Figure 2.1 illustrates the daily pattern of incoming calls to a call center as shown in Green et al. (2007).

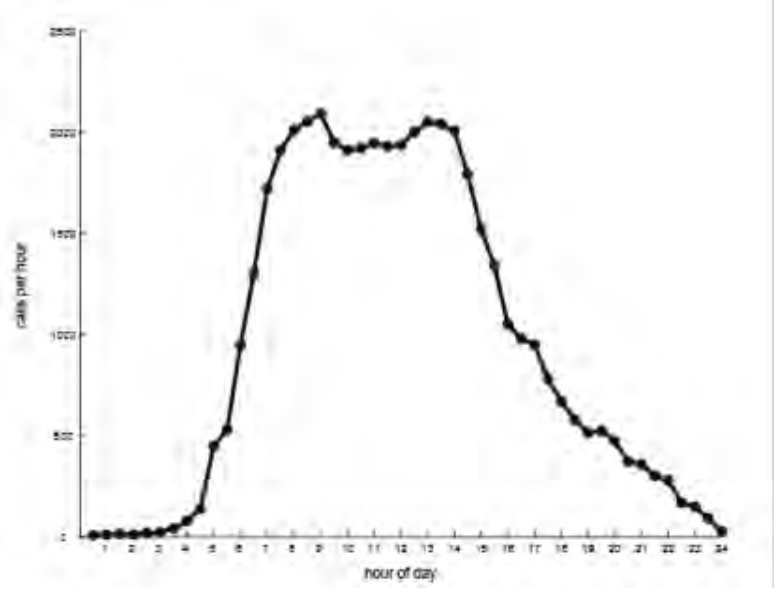


Figure 2.1: Arrivals per hour to a medium-sized financial-services call center (from Green et al., 2007, p. 14)

2.2 Service times

As can be seen from the notation of the M(t)/M/s(t) queue, the service times of the queue are assumed to be exponentially distributed with the same average length over time. One could argue that the average call durations in a call center might also be subjected to time-dependency. However, in modelling call centers it is reasonable to model the average call handling times as constant throughout the day, because of the fact that the service rate changes far slower than the arrival rate (Ingolfsson et al., 2007).

2.3 Number of servers

The number of servers is allowed to change over time in the M(t)/M/s(t) queue. As with the arrival rates, it is not practical to allow the number of servers to change in a continuous manner. For example in a call center, it is not desirable that the start or end of an agent's shift is scheduled at 14:08:56. Instead, the number of shifts should be allowed to only change in fixed and practical time steps, such as 15 or 30 minutes.

2.4 Queueing discipline

We assume that customers are served in a call center on a First-Come-First-Served basis. Another important aspect in the application of the queueing model is what happens when an agent is scheduled to leave at the end of a planning period. The structure in which a leaving agent finishes the current service before stopping is known as the exhaustive discipline. The alternative, where the customer is sent back to the queue and the agent immediately stops when his shift is over, is called the pre-emptive discipline. As Ingolfsson (2005) states, when the customers are humans, the pre-emptive discipline is often unrealistic and from a service point of view undesirable. This is of course also the case for a call center.

2.5 Abandonment

This queueing model does not take abandonments into account. This means that customers are assumed to have infinite patience and do not leave the queue before they have received service. In a call center setting this is highly unlikely. It is possible to extend the model into the M(t)/M/s(t) + M or M(t)/M/s(t) + G model, depending on the distribution of the abandonments, but this extension is outside the scope of this research paper.

2.6 Exact solutions

The exact behaviour of the M(t)/M/s(t) queueing model can be described by the following set of forward differential equations, as specified in Gross and Harris (1974):

$$\frac{dp_0(t)}{dt} = -\lambda(t)p_0(t) + \mu p_1(t)$$

$$\frac{dp_n(t)}{dt} = -(\lambda(t) + n\mu)p_n(t) + \lambda(t)p_{n-1}(t) + (n+1)\mu p_{n+1}(t) \qquad \text{for } 0 < n < s(t)$$

$$\frac{dp_n(t)}{dt} = -(\lambda(t) + s(t)\mu)p_n(t) + \lambda(t)p_{n-1}(t) + s(t)\mu p_{n+1}(t) \qquad \text{for } n \geq s(t)$$

In these equations $\lambda(t)$ stands for the arrival rate at time t, μ stands for the service rate, s(t) for the number of servers at time t, and $p_n(t)$represents the time-dependent probability of n customers being in the system at time t.

Unfortunately, closed form solutions of these equations for $p_n(t)$are only known for a limited number of cases, as stated by Stolletz (2008) and others. Numerically, the differential equations can be evaluated by limiting the size of the system from infinity to a fixed number K. This entails that the last equation is satisfied for $K > n \geq s(t)$ and the following equation should be added:

$$\frac{dp_K(t)}{dt} = -s(t)\mu p_K(t) + \lambda(t)p_{K-1}(t) \qquad \text{for } K > n \geq s(t)$$

Despite the seemingly simple structure of these differential equations, it can be numerically challenging to solve them. As Ingolfsson et al. (2007) found in their research, the calculation time of solving the differential equations increases as the service rate or the system load increases. This is explained by the fact that as one of these system properties increases, the system capacity K needs to increase as well to sufficiently approximate the infinite M(t)/M/s(t) system.

Because of the required computational effort to solve the M(t)/M/s(t) system in an exact manner, many other methods have been developed and are being used in practice to approximate its behaviour. A detailed overview of the most practiced methods is given in section 3.

3. Methods

Because of the simplicity of the M/M/s queueing model, many approximation methods divide the working day into planning periods, during which the mentioned variables remain constant, and construct an Erlang C model for each of the periods. The sections in this chapter provide an in-depth description of these methods as found in literature.

3.1 Pointwise Stationary Approximation

The pointwise stationary approximation method (PSA) was first proposed by Green and Kolesar in 1991 as a method to determine the long run performance of queueing systems that fit the M(t)/M/s(t) model. As its name partly suggests, the PSA method approximates the non-stationary system by pointwise fixing of the arrival rate $\lambda(t)$ and integrating over time of the stationary Erlang C performance measures. Green and Kolesar defined the PSA method as described in section 3.1.1.

3.1.1 Definition

Green and Kolesar begin defining the PSA method by assuming that the arrival rate $\lambda(t)$ varies over time according to a sinusoidal function with period T (24 hours). Another assumption they make is that the average arrival rate $\bar{\lambda}$ over T is smaller than the overall service rate:

$$\bar{\lambda} = \frac{1}{T}\int_0^T \lambda(t)\,dt < s(t)\mu$$

Then, the performance measures L_q, W_q, p_d, p_b of the M(t)/M/s(t) system are defined as follows:

$$L_q = \text{daily average queue length} = \frac{1}{T}\int_0^T \left(\sum_{n=s(t)}^{\infty} (n - s(t))p_n(t)\right) dt$$

$$W_q = \text{daily expected delay} = \frac{L_q}{\bar{\lambda}}$$

$$p_d = \text{daily probability of delay} = \frac{1}{\bar{\lambda}T}\int_0^T \lambda(t)\left(1 - \sum_{n=0}^{s(t)-1} p_n(t)\right) dt$$

$$p_b = \text{ daily probability of all servers busy} = \frac{1}{T}\int_0^T \left(1 - \sum_{n=0}^{s(t)-1} p_n(t)\right) dt$$

Now, the pointwise stationary approximations for each of these performance measures are defined as follows:

$$L_q^{\infty} = \frac{1}{T}\int_0^T L_q(\lambda(t))dt$$

$$W_q^{\infty} = \frac{1}{\bar{\lambda}T}\int_0^T \lambda(t) W_q(\lambda(t))dt$$

$$p_d^{\infty} = \frac{1}{\bar{\lambda}T}\int_0^T \lambda(t)p_d(\lambda(t))dt$$

$$p_b^{\infty} = \frac{1}{T}\int_0^T p_b(\lambda(t))dt$$

For each of the measures, when $\lambda(t)$, μ and $s(t)$ are assumed as given at a certain time t, the formulas for a stationary M/M/s queue are used to calculate $L_q(\lambda(t))$, $W_q(\lambda(t))$, $p_d(\lambda(t))$ and $p_b(\lambda(t))$.

Green et al. (1991) found that when the maximum load remains less than one throughout the day, the following is true:

$$W_q^{\infty} \leq W_q$$

$$L_q^{\infty} \leq L_q$$

Also, they determined that p_d^{∞} and p_b^{∞} give a finite upper bound for p_d and p_b, even when the maximum load exceeded one.

3.1.2 Applicability

The strength of the PSA approach lies in the fact that it can be easily implemented and in that it requires not a lot of computational effort, as noted by Ingolfsson et al. (2007). Because of its use of the M/M/s formulas, the PSA values can be calculated directly for any given parameters. It also provides a tight upper bound for the expected delay and the expected delay as noted in section 3.1.1.

However, one of the consequences of using this model is that when $\lambda(t) \geq s(t)\mu$, the formulas for the daily probability of delay and the probability of all servers being busy, will provide a value higher than one.

Furthermore, since the model uses the formulas of the stationary M/M/s system, the assumption is made that stationarity is achieved for each time epoch t. Another problem with the PSA approach is
that it fails to take into account that the number of servers should be restricted to change in fixed time steps, as noted by Green et al. (2007).

Also, for the Erlang C model to be applicable, the following stability condition should hold:

$$\rho = \frac{\lambda}{s\mu} < 1$$

Since the PSA consists of applying the Erlang C model in each time slot of the day, the stability condition

$$\rho(t) = \frac{\lambda(t)}{s(t)\mu} < 1$$

should hold for each period t = 1,…,T. This means that it is never allowed to have a higher arrival rate than the service rate.

Hence, because the PSA depends on the use of stationary models in each of interval, overloaded situations cannot occur. In a call center, an overloaded system might however be temporarily desirable, because it means that there are no agents idle at that moment and the capacity is being well used.

3.2 Stationary Independent Period-by-Period

3.2.1 Definition

The stationary independent period-by-period (SIPP) method is widely used for determining staffing requirements in service environments. This approach encompasses the division of a workday into staffing periods, e.g. 30 minute intervals, and the construction of a series of M/M/s models, one for each staffing period, as described in Green et al. (2001). Each of the M/M/s models is then independently solved for the minimum number of servers needed in the period to meet a specified service level.

The SIPP method averages the arrival rate over a staffing period, instead of using a pointwise arrival rate. Another difference between the SIPP method and the PSA approach as described in the previous section is that the staffing at fixed lengths of intervals can now be taken into account, while maintaining the low computation times.

Green et al. (2001) were able to show that using the SIPP method often leads to unreliability and understaffing. Therefore, they investigated an improvement to this method and came up with two alternatives. These alternatives were found based on their belief that a lot of the unreliability was due to the fact that SIPP uses an average arrival rate over the entire planning period. Green et al. empirically proved that these alternatives provide more accurate results when used for staffing service systems that have cyclic demands. The performance measure that they aimed to approximate in their paper was the probability of delay.

One of the alternatives they proposed was the Lagged Avg SIPP. This approach involves backwards shifting of the planning period by one average service time and taking the average of the arrival rate during the resulting interval. The other alternative was the Lagged Max SIPP approach. This approach encompasses the modification of the use of the average arrival rate into the maximum arrival rate of the previous service period instead.

The formulas for approximating the arrival rate of each period i = 1,…,T are the following:

$$\text{Regular SIPP:} \qquad \lambda_i = \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}}^{t_i} \lambda(u)du$$

$$\text{Lagged Avg SIPP:} \qquad \lambda_i^{lagavg} = \frac{1}{t_i - t_{i-1}} \int_{t_{i-1}-\mu^{-1}}^{t_i-\mu^{-1}} \lambda(u)du$$

$$\text{Lagged Max SIPP:} \qquad \lambda_i^{lagmax} = \max\{\lambda(u) | t_{i-1} - \mu^{-1} \leq u \leq t_i - \mu^{-1}\}$$

Green et al. recommend that the best results are found when using the following guidelines:

- The Lagged Avg SIPP method should be used when the planning periods are short and the maximum arrival rate is relatively not much higher than the average arrival rate.
- The Lagged Max SIPP method should be used in all other cases.

3.2.2 Applicability

Because the SIPP approach determines stationary performance measures for each period separately, it faces the same problems as the pointwise stationary approximation method. Because of its use of the stationary M/M/s measures, (temporarily) overloading is again prohibited. And the inaccurate assumption is made that stationarity is realized during each of the planning periods, independent of their lengths.

Furthermore, the planning periods, even those that are consecutive, are assumed by the SIPP method to be statistically independent from each other. This means that queues that might have been built up in previous periods are ignored and it further clarifies the understaffing issue of this model.

3.3 Stationary Backlog-Carryover

One of the approximation methods for the M(t)/M/s(t) system that does take dependence of periods into account is the stationary backlog-carryover approach (SBC), as proposed by Stolletz (2008).

The SBC approximation starts by the division of the day into staffing periods, just like the SIPP method. The difference between these two methods lies in the measuring of a certain backlog b_i in each period i and carrying it over into future periods j > i. The benefit of this approach is that it allows for queues to build up in busy periods and the transfer of waiting customers into the following periods.

For the approximation of the arrival rate in a period, the same approach is used as for the regular SIPP method, being the average arrival rate λ_i. However, instead of determining the performance in each period by applying the M/M/s model, the Erlang B system M/M/s/s is used instead. This model is designed to describe systems in which there are no queues; if a customer arrives when all servers are busy, he is blocked from entry. The probability of customers being blocked in each period i is given by $P_i(B)$ and calculated from the Erlang B blocking formula.

The arrival rate that is used for this system is $\tilde{\lambda}_i = \lambda_i + b_{i-1}$, which consists of the average arrival rate in the current period and the backlog rate of the previous period. Assuming a service rate of μ_i and s_i available servers in period i, the backlog rate is calculated as follows

$$\tilde{\lambda}_1 = \lambda_1 \qquad b_0 = 0$$

$$P_i(B) = \tilde{\lambda}_i \frac{\left(\tilde{\lambda}_i/\mu\right)^{c_i}}{c_i! \sum_{k=0}^{s_i} \frac{\left(\tilde{\lambda}_i/\mu\right)^k}{k!}}$$

$$b_i = \tilde{\lambda}_i P_i(B)$$

Another added advantage of the SBC method, according to Stolletz, is that it works more accurately than the SIPP method in not just underloaded situations, but also when approximating systems that are temporarily overloaded.

Because the Erlang B model does not allow for queues to build up, it cannot be applied directly to approximate measures such as the expected waiting time in queues or the number of customers waiting.

Therefore, Stolletz defined a modified arrival rate (MAR) for each period i as λ_i^{MAR}. For this rate, the expected utilization of the loss model is calculated first, as follows:

$$E[U_i] = \frac{\tilde{\lambda}_i(1 - P_i(B))}{s_i\mu} = \frac{\lambda_i + b_{i-1} - b_i}{s_i\mu}$$

In this equation, the portion of incoming and backlogged calls that is not rejected in period i, is represented by $\tilde{\lambda}_i(1 - P_i(B))$. Then, the MAR is defined as $\lambda_i^{MAR} = E[U_i]s_i\mu = \lambda_i + b_{i-1} - b_i$. This definition allows for the stationary delay model, the M/M/s model to be applied with MAR as its arrival rate. This follows from the fact that if the assumption is made that $E[U_i]$ is the steady-state utilization of an M/M/s model, this measure can be calculated by dividing the arrival rate over the number of agents times the service rate parameter.

3.4 Modified-Offered-Load Approximation

The final approximation method that we will discuss is the Modified-Offered-Load approximation (MOL), as introduced by Jennings et al. (1996). This method uses the infinite-server model M(t)/G/∞ to approximate the behaviour of the non-stationary M(t)/M/s(t)queueing system. Eick et al. (1993) found that the M(t)/G/∞ queue is easier to analyse than almost every other queueing models with time-dependent arrival rates and they were able to come up with a number of formulas to describe its behaviour. Initially, the MOL approximation was designed for non-stationary loss systems, such as the Erlang Loss Model with a nonhomogeneous Poisson arrival process, as in Davis et al. (1995). However, Massey and Whitt (1997) showed that it can also be applied to delay queueing systems such as the M(t)/M/s queue.

The MOL approximation uses a different approach from the previously discussed methods for calculating the offered load to the system. The time-dependent expected number of busy servers in the M(t)/G/∞ queue is given by the following formula, with G being the cumulative service time distribution:

$$E[B(t)] = \int_0^t \lambda(u)\,\big(1 - G(t-u)\big)du$$

This formula can then be used to calculate the approximate arrival rate for the M(t)/M/s(t) in each period t as follows:

$$\lambda_t^{MOL} = \mu E[B(t)] = \int_0^t \lambda(u)\mu e^{-\mu(t-u)}du$$

Ingolfsson et al. (2007) refer to λ_t^{MOL} as being an effective arrival rate that is an "exponentially weighted moving average of the arrival rate". Eick et al. (1993) showed that the use of this approach for the offered load can be expected to give a better approximation of non-stationary queues than the approach used by the PSA method, because the service time distribution is taken into account beyond its mean. The MOL approximation was found by Ingolfsson et al. (2007) to perform significantly better than the Lagged SIPP method in most tested cases. The required computation time was however higher for all tested cases.

First, the time - dependent number of busy servers of a related infinite server model is estimated, such as at the infinite server approximation . And then, the obtained number of busy servers is used as the offered load in a steady state model at each point in time . Since the offered load is

defined as the ratio of the arrival to the service rate, the approach can be regarded as a pointwise stationary approximation with a modified arrival rate. A shortcoming that the MOL approximation shares with the other approximations, except for the SBC approximation, is that it requires that the system is never overloaded. This means that the arrival rate calculated by the MOL approximation should never be bigger than the product of the number of agents available and the service rate.

4. Results

4.1 Data

To test the performance of the methods in a specific setting, we used a simulated data set for a call center. For the ease of use, we only considered the actual arrivals to the call center, the average handling times of the calls and the number of agents scheduled to work. The data has the following characteristics:

- The call center is assumed to be open during work days only (Monday through Sunday). The opening times are 08:30 – 16:30.
- Data was simulated for one month of 30 days.
- The calls are aggregated over a period of 10 minutes, resulting in 30 * 49 = 1470 intervals that are considered in total.
- The incoming number of calls throughout the day follows a Poisson inhomogeneous process. We averaged the number of arrivals from a month of real data from a call center, and took these values as arrival intensities to simulate new data. The average arrival pattern that was used is shown in Figure 4.1. The simulation was done to create fluctuation in the data. The arrivals were simulated by drawing from a Poisson distribution with a time-dependent parameter value at each time interval. Because of the irrelevance to the experiments, no kind of seasonality or day-of-week dependence was taken into consideration in the simulation of the arrivals.
- The average handling time (AHT) of the calls is assumed to be independent of the agent taking the call and to have an expected value of 12 minutes. The assumption is made that the AHT is also not time-dependent and exponentially distributed.
- The assumption is made that agent shifts can only start and end at fixed half hour points.

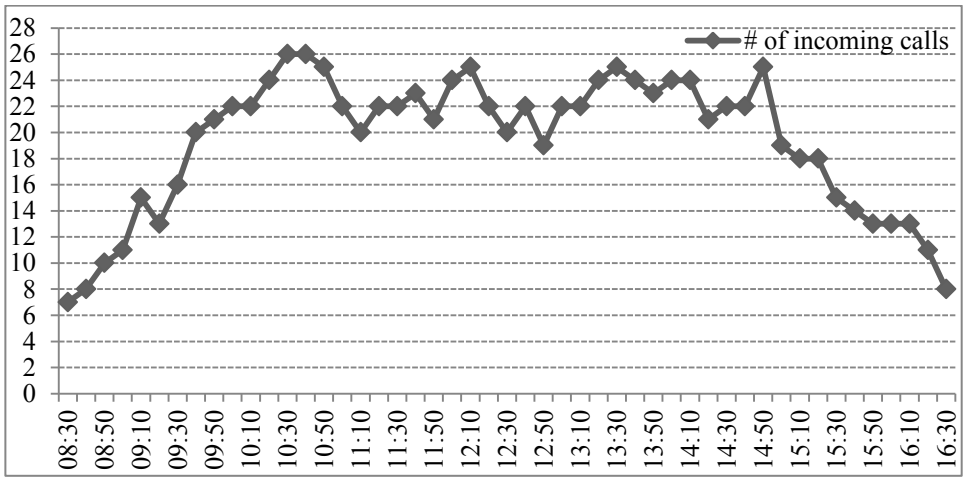


Figure 4.1: Average arrival pattern per 10 minutes to a fictive call center

4.2 Performance measure

For the comparison of the methods that we have implemented we need to define an appropriate performance measure. The service level is a performance measure (and target) that is widely used in call centers. This measure is defined as the fraction or percentage of calls that wait no more than a certain time before proceeding into service.

In the case of stationarity, a constant arrival rate and a constant number of available agents, the service level can be calculated using the Erlang C formula, as below:

$$E[SL] = P(W_Q \leq \tau) = 1 - C(s,a)e^{-(s\mu-\lambda)\tau}$$

$$\text{with} \quad C(s,a) = \frac{a^s}{(s-1)!\,(s-a)}\left[\sum_{j=0}^{s-1}\frac{a^j}{j!} + \frac{a^s}{(s-1)!\,(s-a)}\right]^{-1}$$

Here, W_Q stands for the average waiting time in the queue, τ stands for the target answering time and $C(s,a)$ stands for the probability of a caller having to wait before service. The a in this formula stands for the load in the system, which is equal to λ/μ.

The approximation methods that we have described in the previous section all determine the service level by assuming a stationary situation, after the calculation of their modified arrival rate and therefore, their approximation of the service level can be obtained by the formula above. However, because of the non-stationarity of the M(t)/M/s(t) queue, the formula above cannot be used to determine the service level in an exact manner.

Ingolfsson et al. (2007) and Green and Soares (2007) determined the following expression for the calculation of the time-dependent service level of the M(t)/M/s(t) system:

$$SL(t) = P\big(W_Q(t) \leq \tau\big) = 1 - \sum_{i=0}^{\infty} p_{s(t)+i}(t) \sum_{j=0}^{i} \frac{(\mu\tau s(t))^j e^{-(\mu\tau s(t))}}{j!}$$

In this equation $p_{s(t)+i}(t)$ stands for the probability of there being $s(t)+i$ calls in the system at time t. As a benchmark, to assess the quality of the approximation methods, we calculated the service level for each day in our data by numerically solving the system of differential equations from section 2.6, obtaining the state probabilities and using those to solve the equation above.

In our experiments we will be assuming that the call center has a service level of 80% of the calls being answered within 20 seconds in each staffing period of half an hour. For solving the system of differential equations, we used the ode45 solver in Matlab and divided the day in 490 periods. (Because we assume a 8 hour working day with 10 minute intervals, there are 490 minutes in one day.) Therefore, we have an estimate of the 'exact' service level at each minute of the day.

4.3 Approximation results

We implemented the approximation methods into a spreadsheet in Microsoft Excel. To measure the performance of the approximation methods in assessing the quality of shift schedules, we created a number of different schedules.

4.3.1 Schedule with overloading

First, we used an Erlang C Excel add in[1] to calculate the number of agents needed to achieve the desired service level. However, we assumed that the staffing level is only allowed to change at the beginning of a staffing period. Because of this, when the number of incoming calls is higher than the first part in the second part of the interval, this means that the service level will probably not reach its target in the

[1] http://www.gerkoole.com/CCO/downloads/tools.xls

staffing period. To illustrate this, Figure 4.2 shows an example of a situation in which this is the case for the staffing period between 09:30 and 10:00.

Calculations show that scheduling in this way leads to over- and understaffing and the system being overloaded at multiple times during the day. Despite the fact that the approximation methods are known to function best when there is no overload, we would still like to find out which methods functions best in these situations, because as mentioned in section 3.1.2, (temporarily) overloaded situations are not always undesirable in call centers. We used this way of staffing and tried to approximate the service level at each 10 minute point for all 30 days in the call center individually.

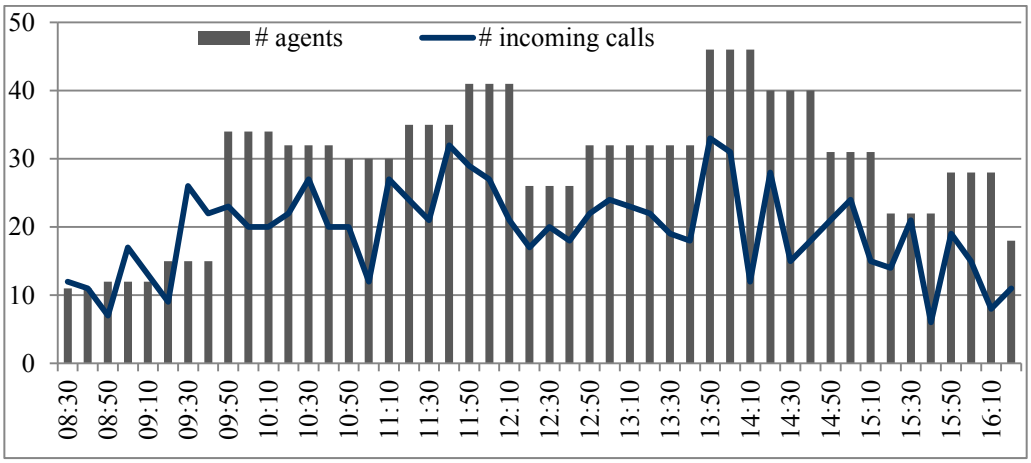


Figure 4.2: Number of incoming calls versus the number of scheduled agents (overloaded case)

Now, if we calculate the exact service level at each minute by solving the differential equations, we can use this as a benchmark for the performance of the approximations. Doing this for the configuration in Figure 4.2, we get the service level behaviour as shown in Figure 4.3.

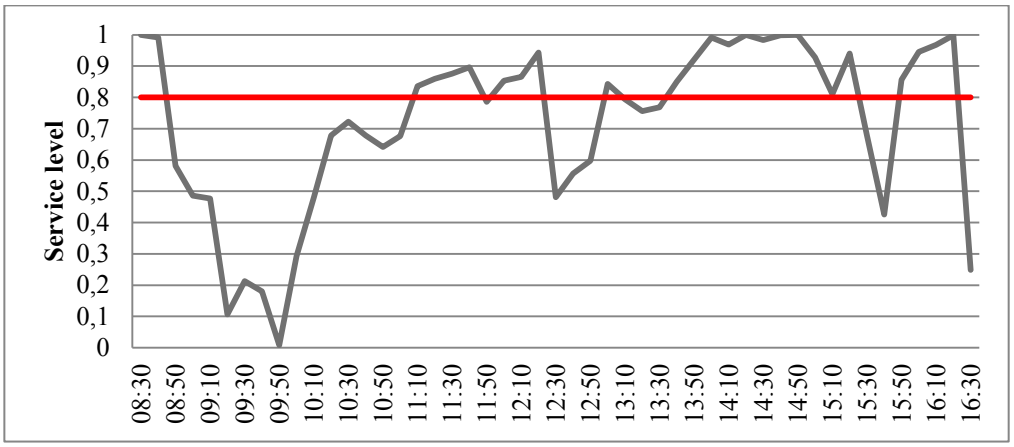


Figure 4.3: Exact behaviour of service level for example in Figure 4.2

From Figure 4.3 it is obvious that the system is overloaded at multiple times during the day, because the service level drops to very low peaks. This means that this staffing schedule needs to be adjusted if the performance goal is to be met.

Now, when we try to approximate this behaviour of the service by the previously discussed approximation methods, we get the figures as shown in the appendix, section 7.1. The results for this example clearly indicate that some methods perform better than others. For example, the SIPP Lag Avg

and the SBC approximations show a good fit. The PSA and SIPP performance is evidently influenced by the periods in which the system is overloaded.

For the MOL method, we see a quite good fit at the end of the day, but a really bad fit in the morning. This might be the result of the method needing a certain 'warm up period'. To assure that the method is implemented correctly, we validated it with a situation in which the arrival rate remains constant and calculated the resulting integral.

To compare the results for all 30 days, we calculate their differences with the exact service level by using the sum of squared errors (SSE) measure:

$$SSE = \sum_{i=1}^{N} (x_i - \hat{x}_i)^2$$

Table 4.4 shows the error values of each of the methods:

Method	SSE
PSA	169.38
SIPP	141.83
SIPP Lag Avg	151.27
SIPP Lag Max	351.70
SBC	83.01
MOL	225.22

Table 4.4: SSE results for schedule with overload

From these results it is obvious that the SBC method is far better in approximating the service level in this case where we allow for the system to be overloaded. The SIPP Lag Max is the worst performing method. Also, where the computation time of solving the differential equations was equal to 306 seconds, the computation time of the approximations methods is all negligible.

4.3.2 Schedule without overloading

To test the methods in the situations in which they are assumed to work best, we created a schedule in which the traffic intensity was not allowed to exceed one, and hence overloading was forbidden. We created this schedule by increasing the number of agents in the overloaded intervals of the first one. Figure 4.5 shows the exact behaviour of the service level for an example arising from this schedule.

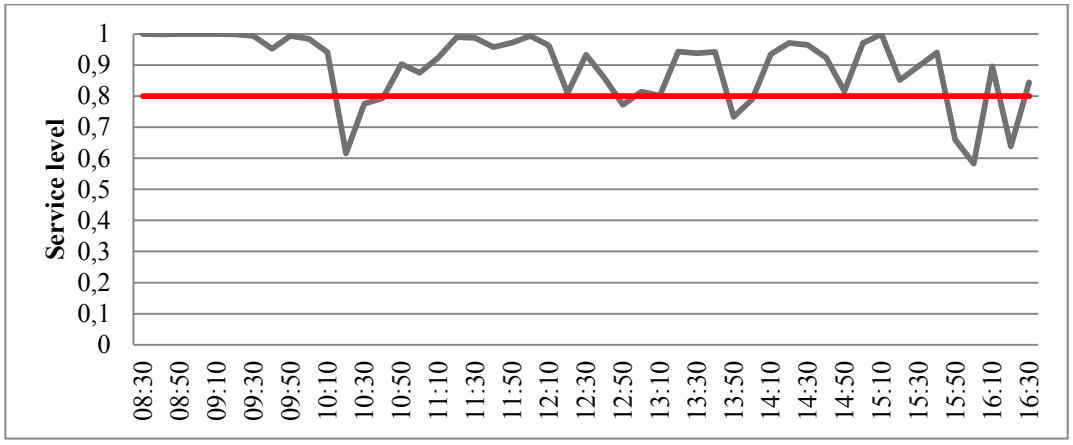


Figure 4.5: Exact behaviour of service level for example without overloading

The approximations of this behaviour are shown in the appendix, section 7.2. What is noticeable right away is that most methods perform far better than they did in the overloaded situation. Except for the SIPP Lag Max and the PSA approximations, the methods seem all to give a quite reasonable fit. The SBC approximations looks like the most suiting one. To validate these observations we calculated the error values of these fits for all 30 days. Table 4.6 shows the SSE measures for all six methods.

Method	**SSE without overloading**	**SSE with overloading**
PSA	105.62	169.38
SIPP	67.08	141.83
SIPP Lag Avg	118.42	151.27
SIPP Lag Max	467.62	351.70
SBC	26.12	83.01
MOL	71.63	225.22

Table 4.6: SSE results for schedule without overload

Except for the SIPP Lag Avg and the SIPP Lag Max, all other methods produce a significantly lower error than in the schedules where we allowed for overloading. The SBC approximation again comes out on top. Furthermore, we see that the MOL method has an error that is three times smaller than in our previous results, which indicates that this method is highly susceptible to overloading.

5. Discussion

The aim of this paper was to explore and evaluate the methods available to assess the quality of agents' shifts schedules in call centers. Based on literature and our own implementation we were able to compare the performance of the Pointwise Stationary Approximation, the different variants of the Stationary Independent Period-by-Period method, the Stationary Backlog Carryover approximation and the Modified Offered Load in approximating the service level.

All approximation methods were found to have a computation time that is negligible compared to the calculation of the numerical solutions. Implementing the Pointwise Stationary Approximation and the Stationary Independent Period-by-Period methods was quite simple. However, implementation of the Stationary Backlog Carryover and the Modified Offered Load approximations took quite some time and effort.

The Stationary Backlog Carryover approach was found to give the best fitting approximation in both overloaded and regular situations. The extensive comparison of Ingolfsson et al. (2007) was completed before the paper on the Stationary Backlog Carryover was published by Stolletz in 2008. This method, together with the Modified Offered Load approximation, does not ignore the dependence between successive periods. The Modified Offered Load method was found to perform second best to the Stationary Backlog Carryover method.

In accordance with previous research we found that most of the methods perform better when the schedules were constructed in a way that does not allow for overloading. Because of their use of the Erlang C model to calculate the service level in a certain staffing period, all investigated methods have the underlying assumption that stationarity is reached within the staffing period. However, when queues build up between periods, the transient period might not be small enough for stationarity to be reached in each of the periods. Unfortunately, no method was found in which this assumption was totally absent.

Because of the time limitations, the number of tested scenarios was not as extensive as we would have liked. Furthermore, not all available methods could have been evaluated. With more time open, fluid model approximations could have provided for a fascinating added value to the comparison.

6. References

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

Eick, Stephen G., William A. Massey, Ward Whitt (1993b). The physics of the Mt/G/1 queue. Operations Research 41(4), 731-742.

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

Green, L., Kolesar, P. (1991). The pointwise stationary approximation for queues with non-stationary arrivals. Management Science, 37 (1), 84-97.

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

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

Gross, D., Harris, C.M. (1974). Fundamentals of Queueing Theory, Wiley, New York.

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

Ingolfsson, A. 2005. Modeling the M(t)/M/s(t) queue with an exhaustive discipline. Working paper, Department of Finance and Management Science, School of Business, University of Alberta, Edmonton, Alberta, Canada, http://www.bus.ualberta.ca/aingolfsson/working_papers.htm.

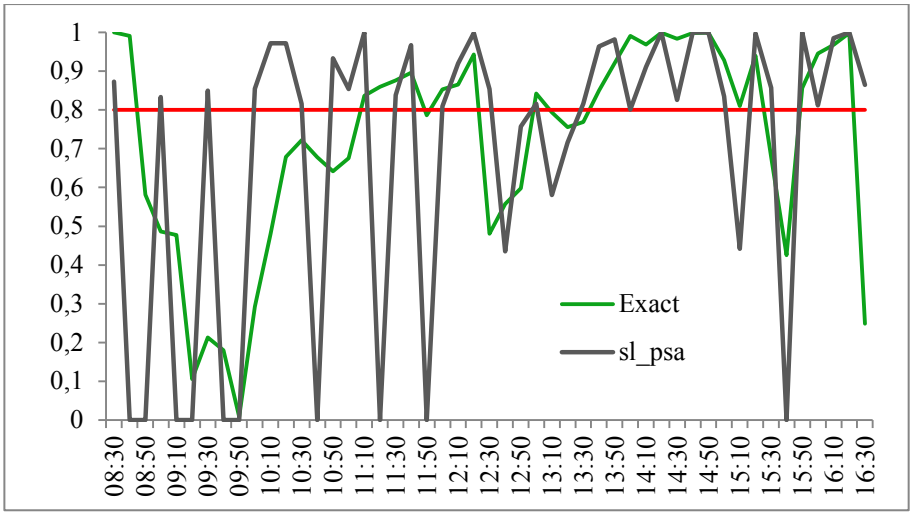
Jongbloed, G., Koole, G. (2001). Managing uncertainty in call centers using Poisson mixtures. Applied Stochastic Models in Business and Industry, 17, 307–318.

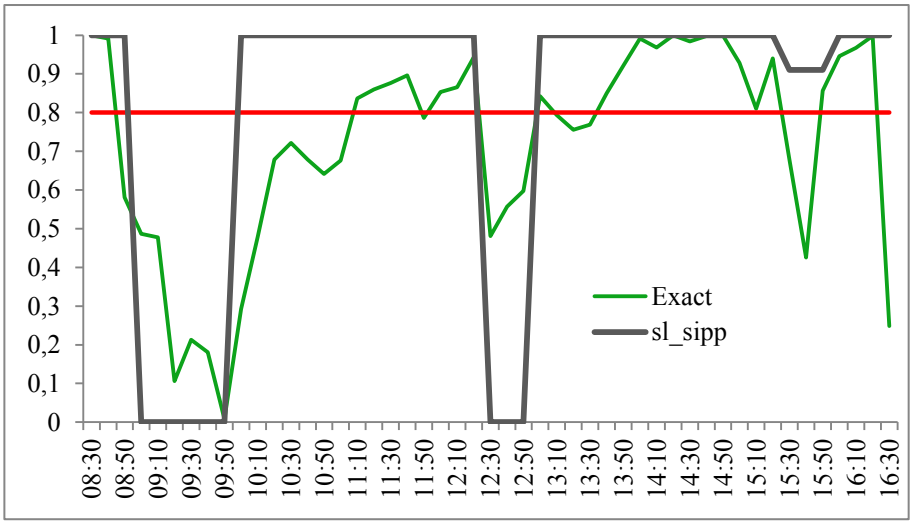
Massey, William A., Whitt, W. (1997). Peak congestion in multi-server service systems with slowly varying arrival rates. Queueing Systems, 25(1-4), 157-172.

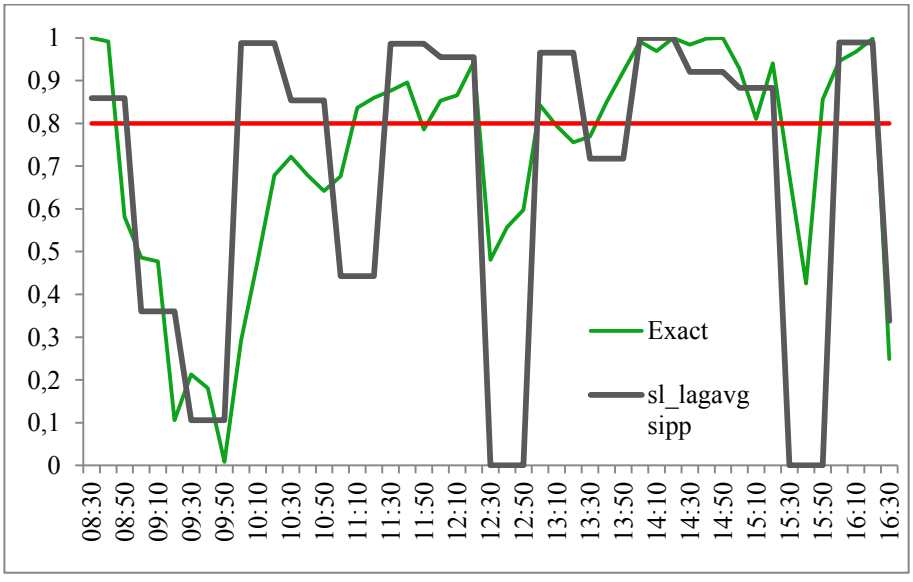
Stolletz, R. (2008). Approximation of the non-stationary M(t)/M(t)/c(t)-queue using stationary queueing models: The stationary backlog-carryover approach. European Journal of Operational Research 190 (2), 478-493.

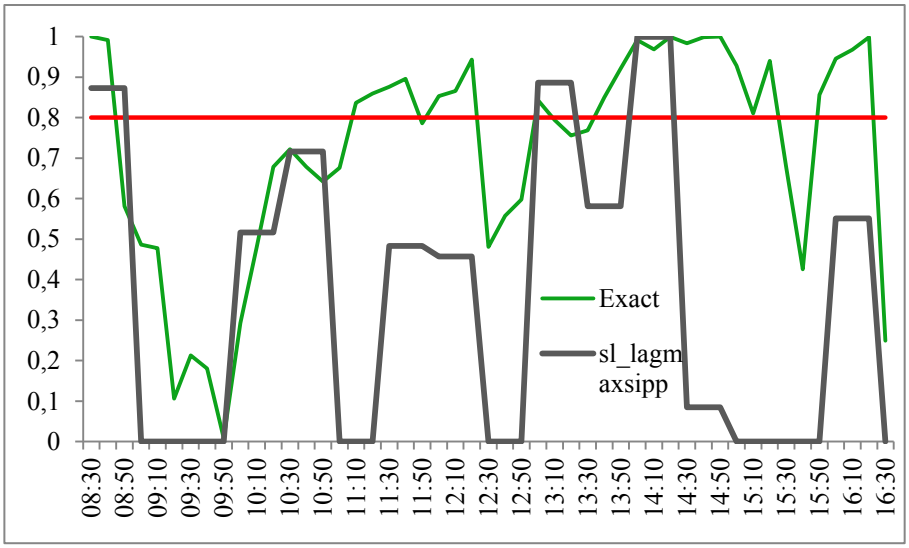
7. Appendix

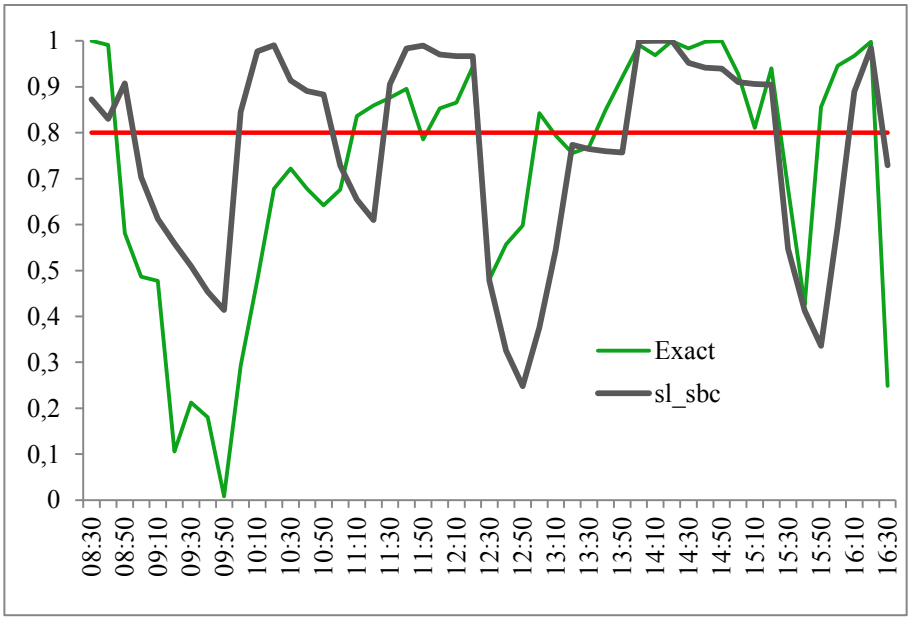
7.1 Service level approximations for overloaded example

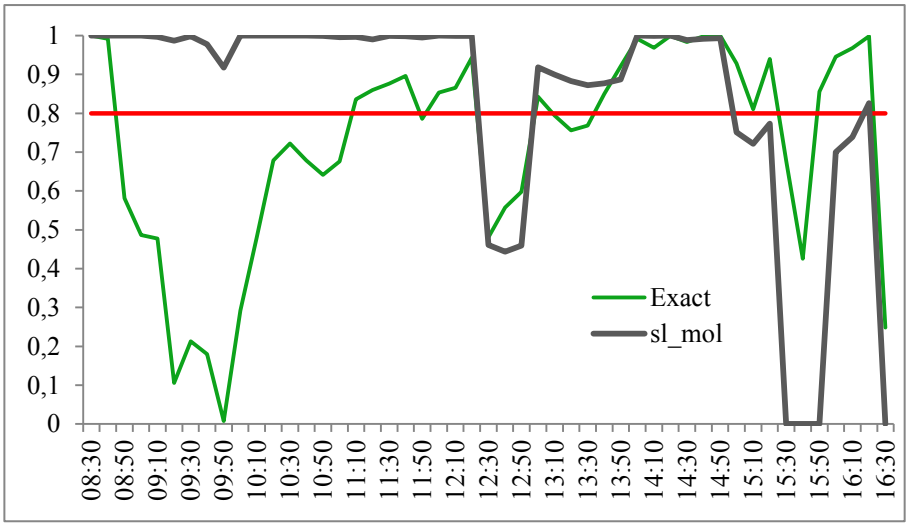




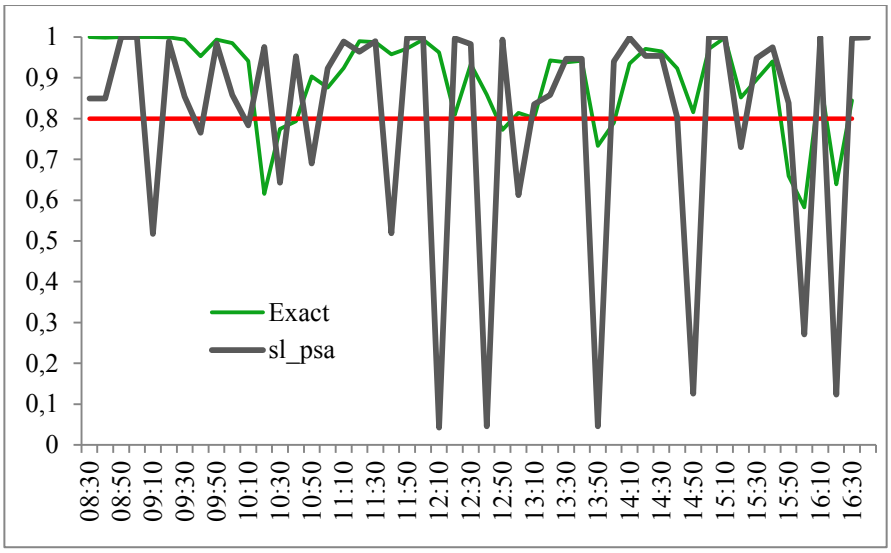


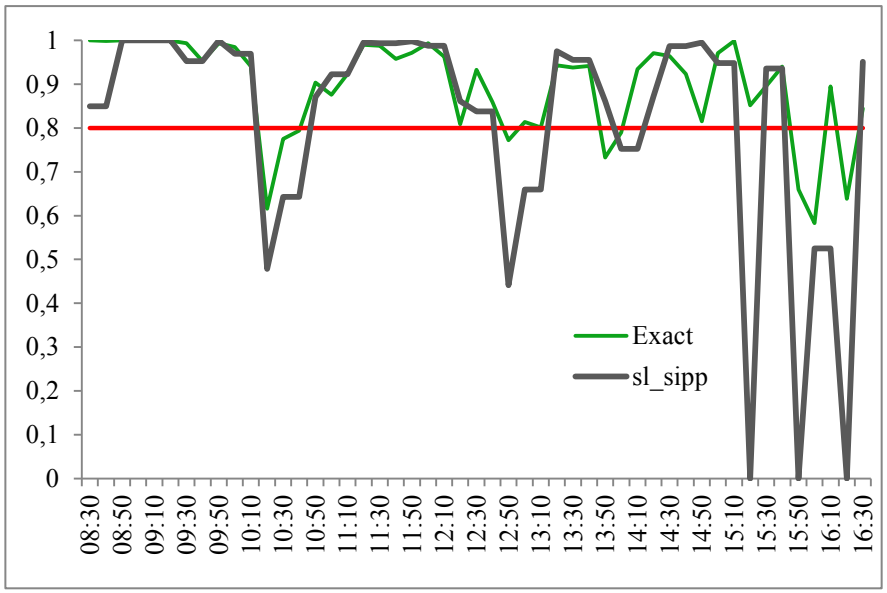

1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
08:30
08:50
09:10
09:30
09:50
10:10
10:30
10:50
11:10
11:30
11:50
12:10
12:30
12:50
13:10
13:30
13:50
14:10
14:30
14:50
15:10
15:30
15:50
16:10
16:30
Exact
sl_lagmaxsipp

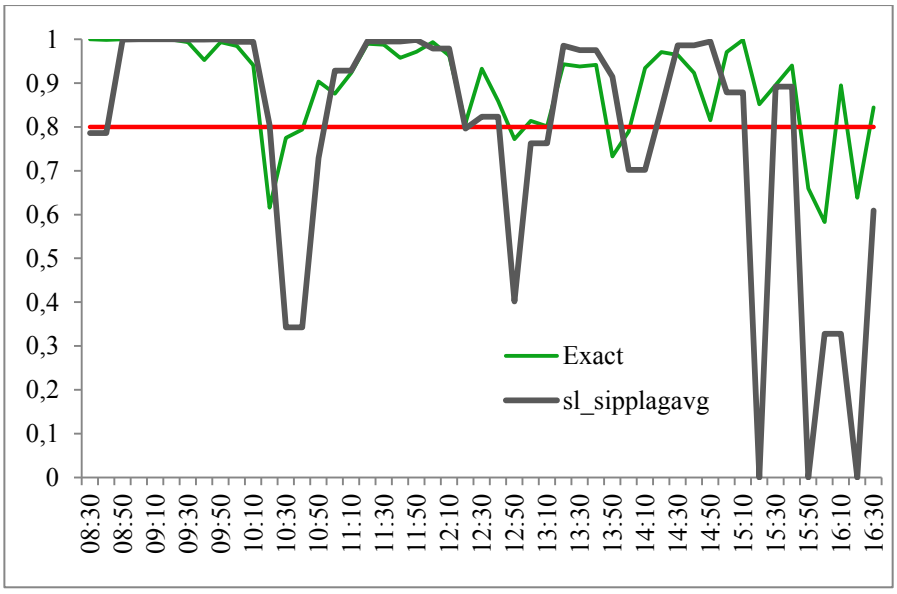

1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
08:30
08:50
09:10
09:30
09:50
10:10
10:30
10:50
11:10
11:30
11:50
12:10
12:30
12:50
13:10
13:30
13:50
14:10
14:30
14:50
15:10
15:30
15:50
16:10
16:30
Exact
sl_sbc

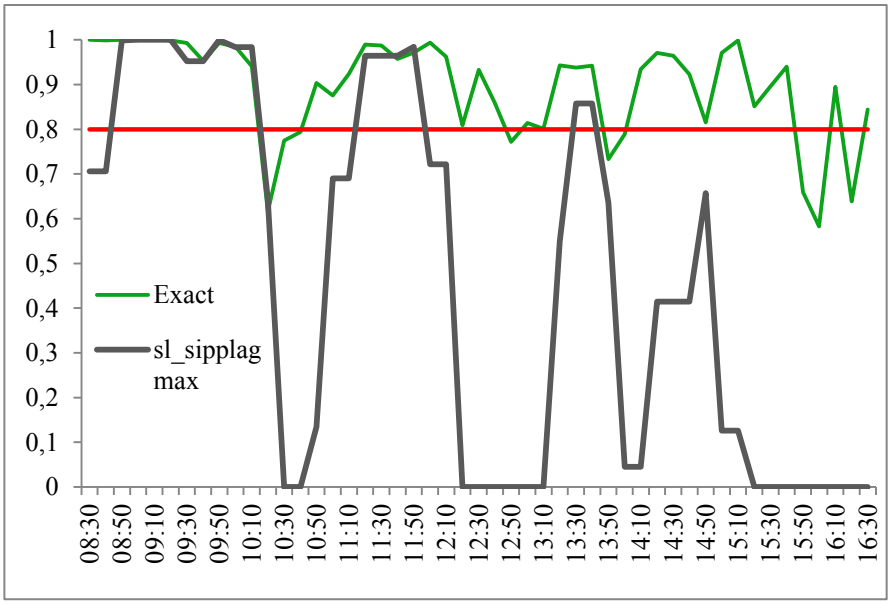

1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
08:30
08:50
09:10
09:30
09:50
10:10
10:30
10:50
11:10
11:30
11:50
12:10
12:30
12:50
13:10
13:30
13:50
14:10
14:30
14:50
15:10
15:30
15:50
16:10
16:30
Exact
sl_mol

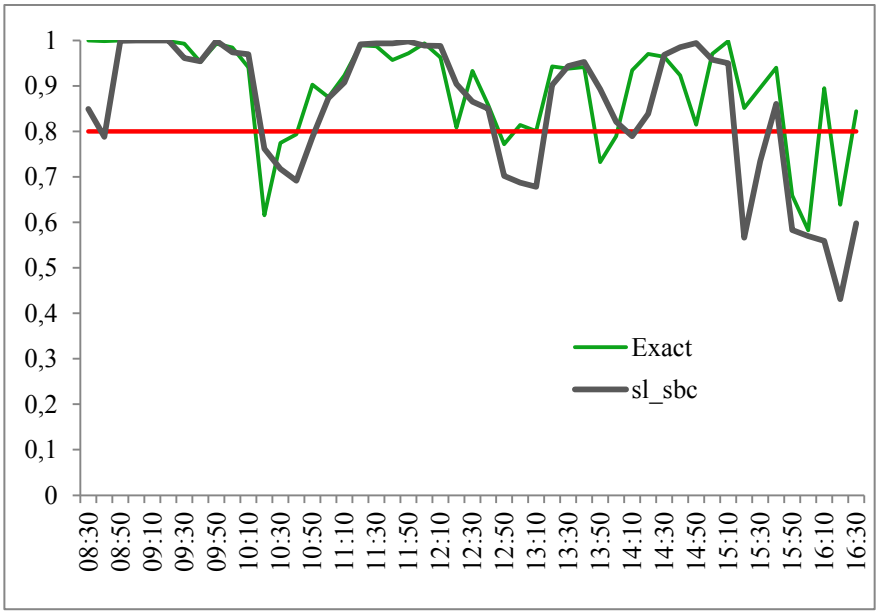
6.2 Service level approximations for example without overload

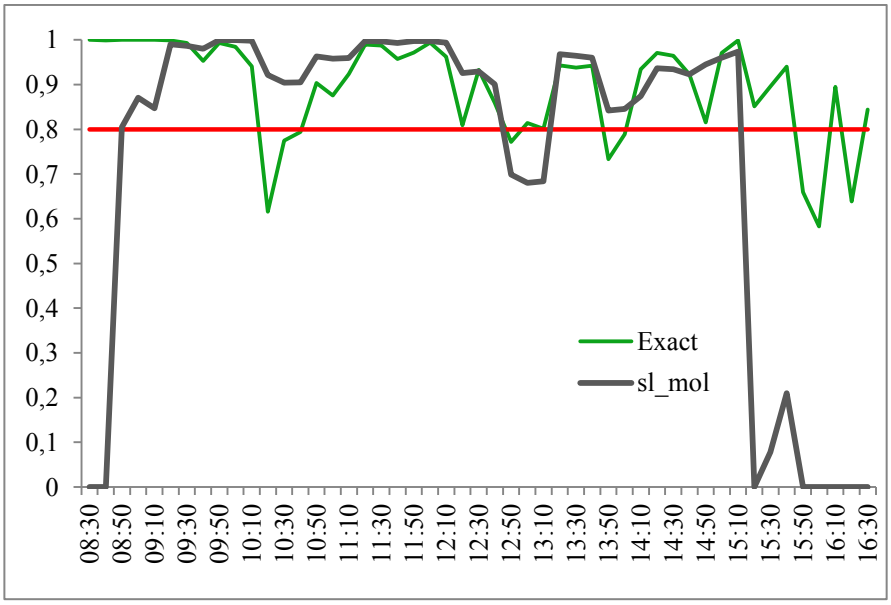







1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Exact
sl_sipplag max
08:30
08:50
09:10
09:30
09:50
10:10
10:30
10:50
11:10
11:30
11:50
12:10
12:30
12:50
13:10
13:30
13:50
14:10
14:30
14:50
15:10
15:30
15:50
16:10
16:30


1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Exact
sl_sbc
08:30
08:50
09:10
09:30
09:50
10:10
10:30
10:50
11:10
11:30
11:50
12:10
12:30
12:50
13:10
13:30
13:50
14:10
14:30
14:50
15:10
15:30
15:50
16:10
16:30


1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
Exact
sl_mol
08:30
08:50
09:10
09:30
09:50
10:10
10:30
10:50
11:10
11:30
11:50
12:10
12:30
12:50
13:10
13:30
13:50
14:10
14:30
14:50
15:10
15:30
15:50
16:10
16:30