Determining appropriate staffing levels for the EMCC

J.G.B. Besenbruch (SN:2546851) and G.C. Chouliaras (SN:2592496)

ABSTRACT

In this paper, the selection of the appropriate staffing levels for the call center of an emergency medical service (EMS) is investigated. The time-varying nature of the arrival rate in call centers during different time intervals, makes the modeling of such systems a demanding task. Thus, after a detailed analysis of the data three possible approaches / models to this problem are investigated and presented, along with a discussion about the efficiency and the drawbacks of each model. The most suitable approach is then selected for implementation and the results are presented in order to evaluate the model's performance and possible limitations.

Key words.
Call Center – Erlang C – Non-homogeneous Poisson process – Queueing theory – Staffing levels – Time-varying arrivals

1. Introduction

The EMS call center can be described as a typical queueing system in which the customers are callers (patients) who seek support (service) given by the EMS dispatchers (agents), and they are queued in virtual queues waiting to be served. One of the most significant tasks in this particular context, is how to optimally manage the EMS dispatchers to serve the arriving calls in order to meet targets in certain performance measures by minimizing the agents cost. Thus, the problem is to specify the number of agents in the call center for each staffing interval, something that is specified according to the arrival rate, that is the frequency of the receiving calls. However, the arrival rate shows a time-varying nature while it experiences significant variations with time during different hours, days and also weeks. Thus, the development of efficient queuing models which take the time-varying arrival rate into account, is a key requirement in order to make accurate staffing decisions which will not lead to inadequate performance (*under-staffing*) neither to high agents' costs (*over-staffing*).
This paper is organized as follows. Section 2 includes the analysis of the given data along with informative plots and discussion about peculiarities in the figures. Based on this analysis, possible approximations are being suggested. Section 3 provides suggestions about possible methods and approaches in order to deal with the time-varying arrival process and the problem of determining appropriate staffing levels. Motivation about the selection of each candidate model is also provided, along with specifications about the actual implementation of each model. Furthermore, drawbacks of each proposed approach are mentioned, along with limitations which might occur due to various approximations and assumptions. In section 4 the implementation of the most appropriate model takes place and results regarding the staffing intervals are provided. Important performance measures are also calculated and key figures are produced in order to evaluate the efficiency of the implemented approach. Finally, section 5 provides a conclusion about the findings and some final remarks are made concerning strengths and drawbacks of the selected approach.

2. Data Analysis

In order to determine the appropriate staffing levels for the EMCC, it is of utmost importance to first analyze the data, so as to draw conclusions about the appropriate model to be used. The first thing that needs to be analyzed is whether the *arrival rate* is constant, or time dependent. In order to check the dependence of the arrival rate in the hour of the day, the total number of arriving calls per hour for the whole observed period has been plotted and can be seen in figure (1).

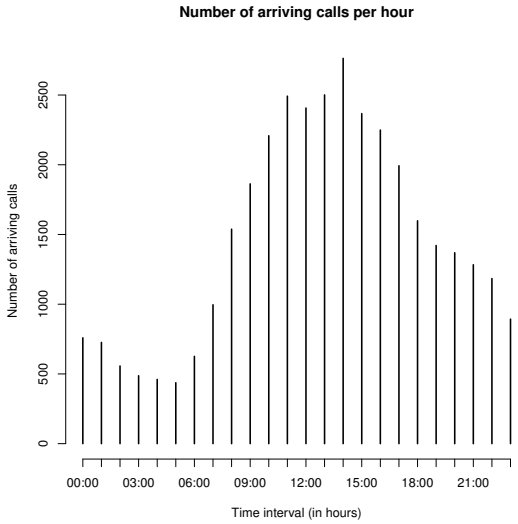

Fig. 1: Total number of arrivals per hour for the whole period.

From this figure it is easily observable that the number of arriving calls clearly depends on the hour of the day, as it has very low values for the first hours and peaks between 12:00 and 14:00. Another kind of dependence which needs to be investigated is whether the number of arriving calls depends also on the day of the week. For that purpose, the mean number of arrivals for every day and for every hour of the day have been plotted and can be seen in figure (2).

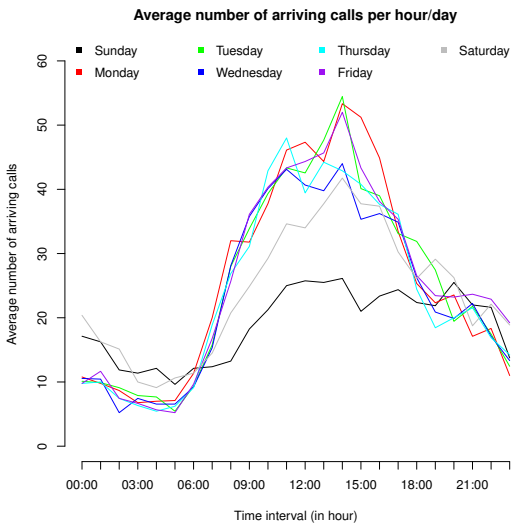

Fig. 2: Mean number of arrivals per hour and per day.

From this figure it can be observed that indeed, there is a difference in the mean number of arrivals for each day. While the weekdays have approximately the same number of arriving calls, Saturday and Sunday have on average less arrivals with Sunday being the day with the lowest number of arrivals throughout the week. From this analysis it can be concluded that the arrival process is time-dependent and this crucial observation should be taken into consideration when fitting a model to the data.
The next parameter that should be investigated is the *service time distribution*. This can be done by counting the frequency of the call duration and the plot can be seen in figure (3).

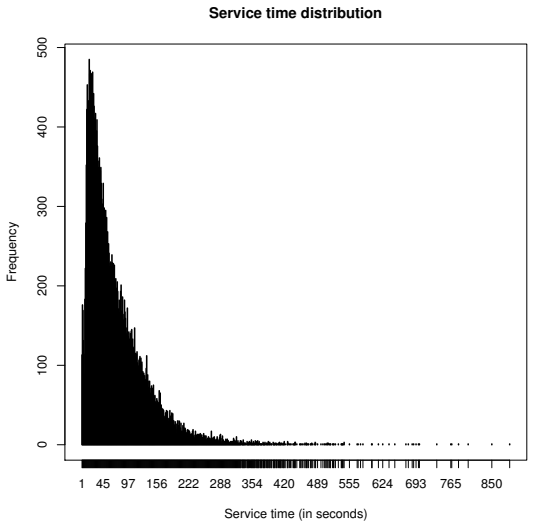

Fig. 3: Probability density function of the service time distribution.

As it can be seen in figure (3) the form of the probability density function is very similar to that of an exponential distribution, something that indicates that the service time distribution can be approximately considered as *exponential*. Another possible approach could be to consider the service time as *log-normal* in the same way as Lawrence et al. in [2], however this approach is more complicated and could not yield much better results since the form of the PDF in figure (3) is very close to the exponential PDF. In order to check whether the service time distribution can be approximated better as exponential or log-normal, the exponential as well as the log-normal QQ-plots of the service time have been plotted and can be seen in figure (4).

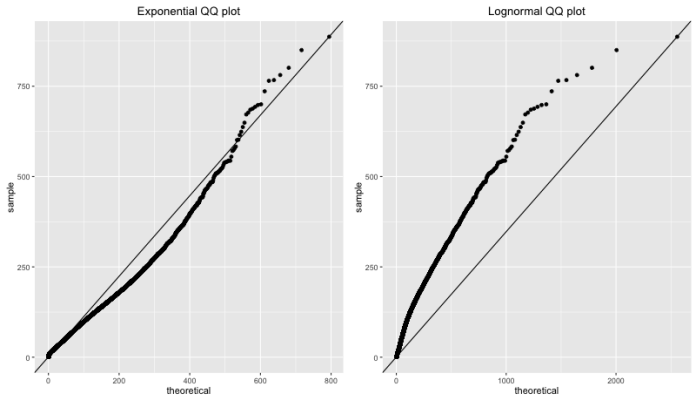

Fig. 4: Exponential (left) and Log-normal (right) QQ-plots of the service time.

In figure (4) it is easily observable that the sample against the theoretical quantiles in the exponential QQ-plot have a better linear relationship than those in the log-normal QQ-plot. That is, the majority of the points in the exponential QQ-plot follows the $y = x$ line, with only a small deviation from the line. On the other hand the points in the log-normal QQ-plot deviate from the $y = x$ line in a greater extent, something that indicates that the exponential distribution is more appropriate for modeling the service time distribution. Another indication that the service time can be considered exponential is the *squared coefficient of variation* (SCV, variance divided by the square of the mean) is approximately 1 and in particular:

$$c_s^2 = \frac{\text{Variance service time}}{(\text{mean service time})^2} = 0.82$$

As Green et al. state in [4], the squared coefficient of variation is one of the most significant parameters which should be taken into account when considering which service time distribution to use. A rough guideline is if $c_s^2 \leq 2$ then the exponential distribution assumption for the service time tends to be a reasonable approximation. In this specific case where $c_s^2 = 0.82 \leq 1$ it is clear that choosing the service time distribution as exponential is a safe choice. A further analysis should be conducted in order to check whether the service time also fluctuates through time. For this reason the mean service time for every hour of the day and for every day has been plotted in figure (5).

From figure (5) it is observed that the mean service time is indeed different for every hour and for every day and fluctuates between 60 and 120 seconds with the peak of 120 seconds in the interval 01:00-02:00 on Wednesday. This fact is also crucial and needs to be taken into account before implementing a model in the data.
Since the goal of this research is to determine the appropriate staffing levels in order for the mean waiting time to be less than 6 seconds it is reasonable to calculate and plot the mean waiting time (mean response time) which corresponds to the given data. The overall mean waiting time according to the data was calculated as:

$$\mathbb{E}W_q = 8.089601s$$

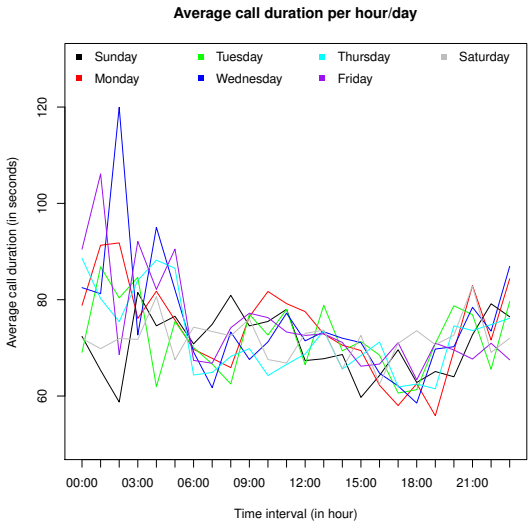

Fig. 5: Mean service time (in seconds) per hour and per day.

This value is 2.089 seconds larger than the target level, so it needs to be reduced in order to be less than 6 seconds. In order to have a better picture of the problem, the mean waiting time per hour and per day has been plotted in figure (6) along with the target level of 6 seconds and the mean waiting time of 8.089 seconds.

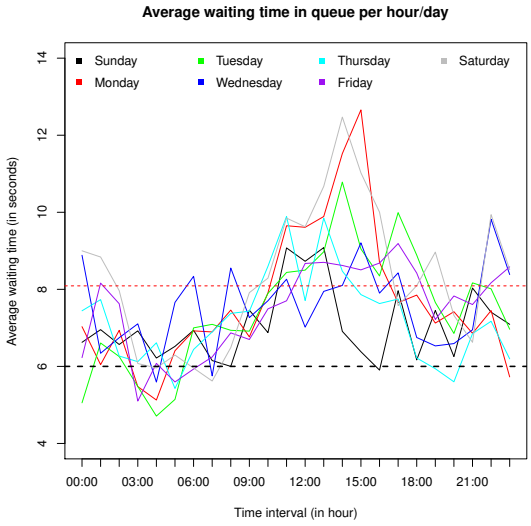

Fig. 6: Mean waiting time (in seconds) per hour and per day.

From this figure it can be seen that the mean waiting time differs from hour to hour and from day to day. The mean waiting time fluctuates between 5 and 12.5 seconds and almost all the values are higher than the target level of 6 seconds. A surprising observation is that while on Saturday the number of arrivals is not so large compared to the other days, the mean response times for this specific day are quite large.

3. Model Considerations

The selection of the appropriate approach for modelling data with time-varying arrivals is a demanding task since there must be a balance between complexity and efficiency. Fortunately, there is a variety of approaches in the literature, hence one has the opportunity to choose the appropriate model depending on the nature of the data at hand. Specifically for call centers, there are three modeling approaches [1] which can be implemented and are provided below. The *descriptive models* which summarize empirical data obtained from the real system and present it in the form of tables and histograms in order to get insight and quantify relationships in data. *Explanatory models* which use time series and regression in order to describe and get inference for the desired parameters using explanatory variables. *Theoretical models* which utilize theoretical principles and statistical distributions in order to provide a mathematical representation to fit the data. The approaches that will be discussed in this paper fall mainly in the third category. Call centers are technology-intensive operations which adhere to a sharply defined balance between agent efficiency and service quality [2]. In order to achieve this, they utilize theoretical queueing models which take as inputs, statistics regarding system primitives, such as the number of working agents, the call arrival rate and the mean waiting time until a customer is served and they calculate performance measures. In practice, determining the appropriate number of working agents which can be changed in order to attain the desired efficiency-quality balance is a crucial and complex task, as many parameters need to be taken in to account and many constraints to be set. The selection of the appropriate model requires first some assumptions that need to be made concerning mainly the arrival process and the service time distribution. Below, three possible models are suggested along with a short description of their actual implementation as well as the drawbacks for each of the suggested models.

3.1 The $M_t/G/s_t$ model

The first proposed model is a single basic queuing model, the $M_t/G/s_t$ model, in which the arrival process is a *non-homogeneous Poisson process* with a time-varying arrival rate function $\lambda(t)$, the service times are independent and identically distributed (iid) and follow a *general* probability distribution, the waiting space is *unlimited* and the queueing discipline is *first-come first-served* (FCFS). In the case of the emergency medical service call center, the arrival process can be assumed to be non-homogeneous Poisson process, due to the fact that there is an infinite pool of customers who arrive in the system with a small probability, and also the arrival rate depends on time, as it was shown in figure (2). According to the Palm-Khintchine theorem, the counting process of events occurring from a large number of independent sources, where anyone of the sources contribution to the total number of events is small, behaves asymptotically as a Poisson process [3].
When it comes to the service time distribution, as it can be seen in figures (3) and (4) it can be considered to be approximately exponential, however making an approximation for the service time distribution might affect the results. Nevertheless, in most cases approximations are necessary in order to use specific formulas and calculate desired quantities. In general queueing formulas, the service time often affects performance measures through its squared coefficient of variation c_s^2 and a common useful approximation for the average waiting time in an $M/G/s$ model is given by

$$\mathbb{E}[\textit{waiting time } M/G/s] = \mathbb{E}[\textit{waiting time } M/M/s] \times \frac{(1+c_s^2)}{2} \tag{1}$$

The $Mt/G/s_t$ model is flexible in the sense that it allows for a general service time distribution and hence it can be applied to all kinds of service time distributions, however it is more complicated to implement than other models, such as the $Mt/M/s_t$ model which assumes an exponential service time distribution. In order to calculate performance measures in the $Mt/G/s_t$ queue firstly the time-varying arrivals should be handled. However, from a mathematical perspective, the finite server $M_t/G/s$ model is analytically intractable and thus, in order to calculate performance measures, this model can be approximated by an infinite-server $M_t/G/\infty$ model having the same arrival rate and the same service-time distribution [4]. This approach is called the *infinite-server approximation* and the reason that it is used is due to the fact that the infinite-server model can be used to show the amount of capacity that would actually be used (and is thus needed) if there were no capacity constraints, such as limited number of servers [5]. The number of customers in the $Mt/G/\infty$ queue has a Poisson distribution with a time-dependent mean $m(t)$ [6]

$$\mathbb{P}(Q^{\infty}(t) = k) = \frac{m(t)^k}{k!} e^{-m(t)} \tag{2}$$

where $Q^s(t)$ denotes the number of customers at time t when the number of servers (agents) is s (s = ∞ in this case). The mean number of customers in the system at time t can be expressed as

$$m(t) = \mathbb{E}\left[\int_{t-S}^{t} \lambda(u)du\right] = \mathbb{E}[\lambda(t-S_e)]\mathbb{E}[S] \tag{3}$$

where S_e denotes the random variable with the stationary excess distribution of S, i.e.,

$$\mathbb{P}(S_e \le t) = \frac{1}{\mathbb{E}[S]} \int_0^t \mathbb{P}(S > u)du \tag{4}$$

Using this, equation (3) can be rewritten as:

$$m(t) = \int_{v=0}^{\infty} \lambda(t-v)\mathbb{P}(S > v)dv \tag{5}$$

After having specified the load $m(t)$, then for every point in time the mean waiting time for the $M/G/s$ queue can be specified using equation (1). Utilizing a recursive procedure, one can start with 1 server which corresponds to an $M/G/1$ queue and calculate the mean waiting time. If it is larger than 6 seconds (the target level), then one more server can be added and the mean waiting time can be computed again. The optimal s is the one for which the mean waiting time is smaller than 6 seconds. One of the biggest drawbacks of this approach is the difficulty to determine the precise form of the arrival rate function $\lambda(t)$. An approximation which can be used here is to assume the arrival rate as piecewise constant, that is to divide the periodic cycle into n intervals where the arrival rate is constant. However, it is clear that more approximations lead to less precise results, hence there has to be a balance between the model's efficiency and complexity.

3.2 The time-varying Erlang C model $M_t/M/s_t$

A more suitable for the given problem and easier to implement model is the *time-varying Erlang C model* ($M_t/M/s_t$) which has *non-homogeneous Poisson* arrival process with a time-varying arrival rate function $\lambda(t)$, an exponential service time distribution with iid service times, unlimited waiting space and FCFS queueing discipline. The exponential service time distribution provides the necessary machinery to compute the desired target performance measures and this is the reason that this model is simpler than the $M_t/G/s_t$ model. The reason that the service time distribution can be well approximated by the exponential distribution is due to the fact that the squared coefficient of variation is $c_s^2 = 0.82 \approx 1$ and also the plot of the probability density function of the service time distribution approximates the exponential probability density function as it can be seen in figure (3). Furthermore, the QQ plot of the sample quantiles and the theoretical exponential quantiles in figure (4), shows a linear relationship and this is another indication that the service time distribution can be well approximated by exponential distribution. Despite the fact that the service time distribution in this model is considered exponential, the $M_t/M/s_t$ model is still a complex model which cannot be implemented directly and thus, it has to be approached by stationary Erlang C models ($M/M/s$). However, it is inappropriate to staff to the overall average arrival rate over the entire day [7]. A way to deal with the time-varying arrival process is to approximate the arrival rate function as *piecewise constant*, that is to chop time into segments and use a stationary $M/M/s$ model in each segment. By denoting the hour of the day as $i = 0, \ldots, 23$, with $i = 0$ to be the interval $00:00 - 01:00$, and the day of the week as $j = 1, \ldots, 7$ with $j = 1$ to be Sunday, the arrival rate $\lambda_{i,j}$ equals the mean arrival rate for this specific interval. In an attempt to check whether the arrival rate can be assumed constant in those intervals, for every day of the week, the arrival rate per hour and per week was calculated, in order to see the range of the fluctuations of the arrival rate. Indicatively, since Monday is a busy day, the number of arrivals for the peak hours per week is plotted and is presented in figure (7).

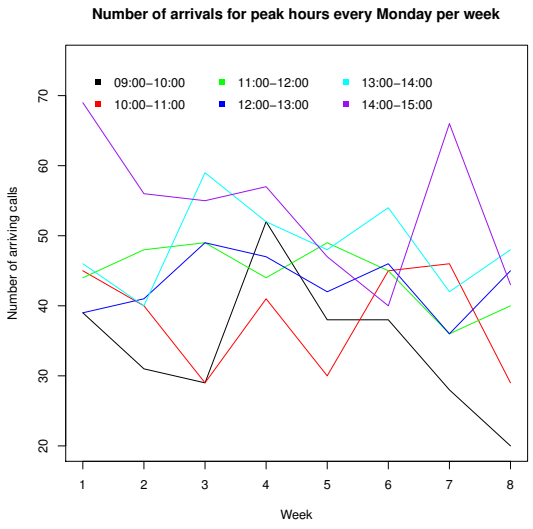

Fig. 7: Number of arrivals for the peak hours of Monday, for all the eight weeks.

As it can be observed from figure (7) the arrivals for each time interval do not fluctuate much throughout the eight weeks period and hence it can be assumed as constant by using the

average of all the eight weeks. This approach is referred as *stationary independent period-by-period* (SIPP) approach in [8]. It has to be noted here, that the fluctuation of the number of arriving calls for non-peak hours was much smaller than the peak hours, something that mitigates the approximation's adverse effect on the results. Furthermore, the fluctuation of the number of arrivals was also smaller for less busy days than Monday. Hence, in order to approximate the arrival rate function $\lambda(t)$ the average arrival rate was calculated for every time interval and for every day and it was stored in an 24×7 matrix. This matrix includes the $\lambda_{i,j}$ for every hour of the day ($i = 0, \ldots, 23$) and for every day of the week ($j = 1, \ldots, 7$).
Concerning the service time, as it was shown in figure (5) it also fluctuates for different hours and days hence a similar approach was also conducted to the mean service time $\mathbb{E}[S]$. In other words, the service time for each time interval and for each day was averaged and stored again in an 24×7 matrix in which every element corresponds to the mean service time for a specific day and a specific hour of the day. Due to the fact that the service time can be assumed exponential, the parameter $\mu_{i,j}$ for hour i and day j can be easily computed by the well-known formula:

$$\mu_{i,j} = \frac{1}{\mathbb{E}[S]_{i,j}} \tag{6}$$

Then, the *offered load* $\alpha_{i,j}$ can also be determined via

$$\alpha_{i,j} = \frac{\lambda_{i,j}}{\mu_{i,j}} \tag{7}$$

After having determined all the necessary quantities, an $M/M/s$ model can be applied for every hour and for every day using as constraints for every of these models that the *load* has to be smaller than one (stable system) and the mean waiting time in queue less than 6 seconds

$$\rho_{i,j} = \frac{\alpha_{i,j}}{s} < 1 \qquad \textit{and} \qquad \mathbb{E}[W_q]_{i,j} < 6s \tag{8}$$

Then, for every of these $M/M/s$ models that satisfy both constraints, the s parameter indicates the optimal number of servers for this specific hour and day, in order to meet the target mean response time of 6 seconds. After the implementation of this model, one can expect that the mean waiting times that were plotted in figure (6) will all be under the target limit of 6 seconds.
Despite the fact that this model is simpler than the $M_t/G/s_t$ model, the most important drawback of this model is that uses many approximations and thus, there will always be an error in the results. For example, as it was shown in figure (7) the number of arrivals for each time interval and for each day fluctuate over the period of 8 weeks, hence by considering them as constant produces an error in the results. This error might be translated as over-staffing or under-staffing depending on the various approximations that take place. However, it has to be noted that due to the fact that the problem concerns the emergency medical call center, over-staffing is preferable than under-staffing, as there are human lives involved.

3.3 Simulation

Besides the Erlang C model an alternative way to model staffing levels in call centers is by using *simulation*. This method gives the opportunity for real world processes, like arrival rate, to be simulated. This is useful when the call center structure becomes too complex to model with analytical (queueing) methods. With this approach, it is not only possible to simulate staffing levels for call agents (operators), but also to model full routing processes which might be encountered in multi-skill call centers. Despite the fact that this method can provide flexibility and improve modeling in a great extent, there are also some (minor) disadvantages. This technique requires the development of simulation models in order to produce simulated data sets similar to the real one. However, the consequence of this is having a different outcome with every simulation. More importantly, the required computer time to run and implement the simulation grows as the square of the required precision [11]. However the computer time can be substantial, so one has to take this into account and choose accordingly between simulation and the Erlang based models, which run faster and are easier to implement.

Nowadays 60% - 70% of call center costs are related to staffing / human-resources. A known problem of Erlang-C is that it tends to overestimate the staffing needs [9]. On the other hand, with simulation all the interrelationships between callers, agents, skills, technology, call management algorithms and techniques can be explicitly defined [10]. This ensures the best possible staffing decisions and also allows the analyst to tweak parameters in the simulated call center in order to answer questions concerning various issues which might be encountered. The real value of Erlang- based calculations then comes in providing an initial input data set required to feed a model [9]. Furthermore, with a simulation the number of call agents could be minimized while the waiting time can be reduced to six seconds or lower, and at the same time the service level can be increased. Despite the fact that there are some variations, the essential structure in simulating is as follows [12]:

1. *Forecasting*: Obtain data about the expected number of calls and their expected service time in the planning horizon, divided in periods of one hour. (Under the assumption that shifts start and end every hour.)
2. *Service level requirements*: Determine the minimum number of call operators, during the period, to achieve the service level and waiting time restrictions.
3. *Shift construction*: As operators work in shifts and are entitled to have a break, it is necessary to make a schedule that satisfies these restrictions.
4. *Rostering*: Obtain the final schedule where an minimum of agents is used and all shifts are satisfied.

Some parameters that can be used in a simulation of the staffing levels for the EMCC are: Arrival rates of calls, service time to handle a call by the call taker, time needed to forward the call to a dispatcher, et cetera. Moreover, calls can be analyzed based on priority level. Also, it has to be noted that the time and date is important in a simulation, in order to simulate the right amount of arriving calls and their distribution.

Using this approach there are several possibilities to make advanced forecasts and more adequate staffing schedules. The risk of making poor decisions in terms of over-staffing or under-staffing will also be minimized, while it is possible to reduce the waiting time and at the same time increase the service level.

4. Calculate staffing levels

In order to determine the appropriate staffing levels throughout the week, the approximation of the $M_t/M/s_t$ model was chosen.

The model has been implemented in the same way as it is described in section 3.2. For every different λ and its corresponding μ, $M/M/s$ models were fit recursively, starting from $M/M/1$ and adding one server each time, until the mean waiting time to be less than the target level of 6 seconds. For the first of the $M/M/s$ models that produced a mean waiting time less than 6 seconds, the parameter s indicates the optimal number of servers for this specific day and hour. All these optimal parameters s were obtained and are presented in table (1).

	S	M	T	W	T	F	S
00:00-01:00	2	2	2	2	2	2	2
01:00-02:00	2	2	2	2	2	2	2
02:00-03:00	2	2	2	2	2	2	2
03:00-04:00	2	2	2	2	2	2	2
04:00-05:00	2	2	2	2	2	2	2
05:00-06:00	2	2	2	2	2	2	2
06:00-07:00	2	2	2	2	2	2	2
07:00-08:00	2	2	2	2	2	2	2
08:00-09:00	2	3	2	3	2	2	2
09:00-10:00	2	3	3	3	3	3	2
10:00-11:00	2	3	3	3	3	3	2
11:00-12:00	3	3	3	3	3	3	3
12:00-13:00	2	3	3	3	3	3	3
13:00-14:00	2	3	3	3	3	3	3
14:00-15:00	2	3	3	3	3	3	3
15:00-16:00	2	3	3	3	3	3	3
16:00-17:00	2	3	3	3	3	3	3
17:00-18:00	2	2	2	3	3	3	3
18:00-19:00	2	2	2	2	2	2	2
19:00-20:00	2	2	2	2	2	2	3
20:00-21:00	2	2	2	2	2	2	2
21:00-22:00	2	2	2	2	2	2	2
22:00-23:00	2	2	2	2	2	2	2
23:00-24:00	2	2	2	2	2	2	2

Table 1: Optimal Number of servers per hour of the day

From a comparison between the number of servers for every hour of the day in figure (8) and the optimal number of servers for the corresponding hour of the day in table (1), it can be seen that for the hours that the number of arrivals is relatively small it is suggested that two agents are enough in order for the mean waiting time to be less than 6 seconds. On the contrary in the peak hours where the number of arrivals is large, it is suggested that 3 agents are needed in order for the target mean waiting time to be met. Moreover, for Sunday which is not a busy day as it can be seen from figure (2), the proposed number of agents is 2 for all the hours, except for the interval 11:00-12:00 in which 3 agents are assigned. This interval corresponds to the peak of this specific day in figure (2). These facts indicate that the model adjusts the optimal number of servers depending on the arrival rate for each specific day and hour. In order to check whether the mean waiting times are now actually smaller than 6 seconds, the mean waiting times in queue for each hour and each day were plotted and can be seen in figure (8).

Through a comparison between figure (6) and figure (8) it can be observed that there is a significant reduction in the mean waiting times. More precisely for the first hours with only two agents it is achieved a reduction of 3-4 seconds. It can also be observed that there are some peaks where the mean waiting time is close to the limit of 6 seconds, however the majority of the hours is far from that limit something that verifies the efficiency of the model. It should be noted that due to the fact

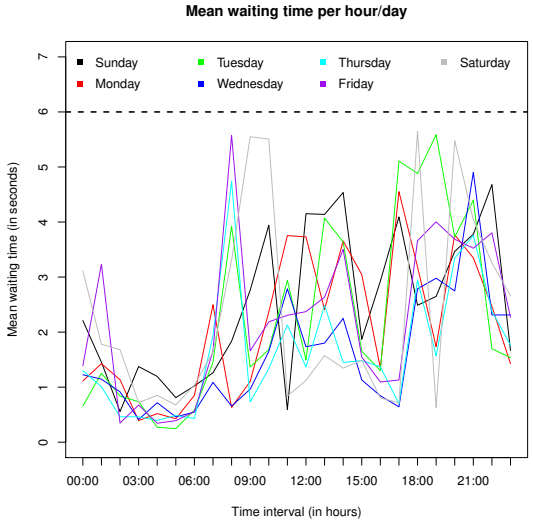

Fig. 8: Mean waiting time (in seconds) per hour and per day after the implementation of the $M_t/M/s_t$ model.

that the modelling concerns emergency calls and that it is preferable to do some over-staffing than make a patient wait when he is in need, it is advisable to add one more agent for the hours whose peak is close to the target level of 6 seconds. This will decrease significantly the mean waiting time, making the process more safe. Furthermore, the overall mean waiting time was reduced from $\mathbb{E}[W_q] = 8.089601s$ to $\mathbb{E}[W_q'] = 2.1036s$ which is a significant reduction of almost 6 seconds. Moreover, for every selected $M/M/s$ model, other performance measures were also calculated. Indicatively, the performance measures and the parameters for the $M/M/2$ model which was applied for Friday at 07:00-08:00 (where there is a peak in the mean waiting times) are presented to the following table.

Performance measure	Value
s	2
λ	0.0071 *calls per second*
μ	0.0134
$\alpha = \lambda/\mu$	0.5299
$\rho = \alpha/s$	0.2644
$\mathbb{P}(N = 0)$	0.5817
$C(s, \alpha) = \mathbb{P}(W_q > t)$	0.1106 s
$\mathbb{E}L$	0.5686
$\mathbb{E}L_q$	0.040
$\mathbb{E}W$	79.7488 s
$\mathbb{E}W_q$	5.5756 s
$\mathbb{E}(W_q \mid W_q > 0)$	50.4177 s

Table 2: Parameters and performance measures for the $M/M/2$ model implemented for Friday at 07:00-08:00 time interval.

From table (2) it can be seen that the mean waiting time in the queue $\mathbb{E}W_q$ is 5.5756 seconds, something that justifies the peak for that specific point in figure (8).

5. Discussion

The implemented approximation of the $M_t/M/s_t$ model produced satisfying results and achieved a significant reduction to the mean waiting time using a time-varying number of agents.

The model proposes more agents for the hours that the arrival rate is high and less agents when the arrival rate is relatively low. This indicates that the model takes into account the time-varying nature of the arrival rate, and hence the model is appropriate for modeling call centers in a real environment where the arrival rate is almost never constant. The approximation of the service time distribution as exponential as well as the approximation of the non-stationary $M_t/M/s_t$ model with many stationary Erlang C $M/M/s$ models make the specification of the staffing levels an easy task, as the convenient formulas for the $M/M/s$ model can be used. However, the more the approximations the less precise the results and for this reason, the alternative approach of the $M_t/G/s_t$ model is being suggested, in which less approximations are being made something that leads to more accurate results but also more complexity. Due to the fact that Erlang calculations tend to over-estimate staffing needs, one can also rely on simulation in order to determine the appropriate staffing intervals. Simulation can be utilized even for call centers with a high complexity, however in order to achieve a high precision it requires enough computer time, something that might be a significant restriction in cases when the time is limited. Conclusively, every method has its advantages and its drawbacks and it is the analyst's decision which approach to implement, depending on the available time horizon he has, as well as the precision he wants to achieve.

References

[1] M. S. Shafae, M. H. Elwany, M. N. Fors and Y. Abouelseoud. *Stochastic Data Analysis and Modeling of a Telephone Call Center*. Proceedings of the International Conference on Industrial Engine, pp 1340, 2012.

[2] Lawrence Brown and Noah Gans and Avishai Mandelbaum and Anat Sakov and Haipeng Shen and Sergey Zeltyn and Linda Zhao. *Statistical Analysis of a Telephone Call Center*. Journal of the American Statistical Association, vol. 100, no 469, pp. 36-50, 2005.

[3] Heyman, D. and Sobel, M. *Stochastic Models in Operations Research:Stochastic Processes and Operating Characteristics.*. 1st Edition. Vol. I, Mineola, NY: Dover Publications, 2004.

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

[5] Z. Feldman, A.Mandelbaum, W. A. Massey, W. Whitt. *Staffing of Time-Varying Queues to Achieve Time-Stable Performance*. Management Science manuscript, pp. 324 - 338, 2004.

[6] R. Bekker, A.M. de Bruin *Time-dependent analysis for refused admissions in clinical wards*. A.M. Annals of Operations Research, Vol. 178, pp. 45 - 65, 2010.

[7] L.V. Green, P.J. Kolesar, A. Svoronos *Some effects of non-stationarity on multiserver Markovian queueing systems*. Operations Research, Vol. 39, no 3, 1991.

[8] L.V. Green, P.J. Kolesar, J. Soares *Improving the SIPP approach for staffing service systems that have cyclic demands*. Operations Research, Vol. 41, no 4, pp. 549-564 , 1991.

[9] V. Bapat, E.B. Pruitte *Using simulation in call centers*. Simulation Conference Proceedings, 1998

[10] D. Pegden , R. Shannon, and R. Sadowski. *Introduction to SIMAN using Simulation.*, New York: McGraw-Hill, 1995

[11] M. Rothkopf, S. Oren, *A closure approximation for the nonstationary $M/M/s$ queue*, Management Science, Vol.25, No. 6, 1979

[12] J. Atlason, M. Epelman, S. Henderson, *Optimizing Call Center Staffing Using Simulation and Analytic Center Cutting-Plane Methods*, Management Science, Vol. 54, no 2, pp. 295-309, 2008