Call Center Optimization-Scheduling Problem

Lubabalo Ndobe (lubabalo@aims.ac.za)
African Institute for Mathematical Sciences (AIMS)

Supervised by: Prof Montaz Ali
Wits, South Africa.

12 October 2018

Submitted in partial fulfillment of a structured masters degree at AIMS South Africa

Abstract

Call centers play a pivotal role in business as many companies use them for customer interactions including resolving billing issues. Credit and risk collections call center experiences massive influx of inbound calls due to collection actions performed against owing customers via workflow system (WFS). Consequently, the standard service level is not met due to limited agents working at the call center. An optimal way of spreading out these activities, particularly SMSs is determined. These are scheduled in such a way that resources (agents) are smoothly utilized. Extended resource levelling model addresses this problem. Examples are used for demonstration on how resources can be spread out in future using estimation from queuing theory, resource viability, and resource consumptions.

Declaration

I, the undersigned, hereby declare that the work contained in this research project is my original work, and that any work done by others or by myself previously has been acknowledged and referenced accordingly.

Lubabalo Ndobe, 12 October 2018

Contents

1. Introduction

Nowadays companies often use call centers in order to connect with their customers Koole (2005). By connection we mean that many companies use centers to interact with customers, that is, selling products and services, giving support for the purchased product, and also for sorting out billing issues. So call center plays a vital role in business, Vuthipadadon (2009). Furthermore, according to Atlason, Epelman, and Henderson (2008) it is usually a significant element of the operations for many organizations and as the expansion of technology with sector grows so does the contribution, Gans et al. (2003). Call centers can handle inbound calls (i.e. incoming calls) or outbound calls (i.e. outgoing calls), they are also referred to as contact centers if they do not only handle telephone calls but also communications via email, fax, internet and other means of communications other than telephone calls, Clarke (2007). We will focus on the single-skill inbound call centers, that is handling only inbound calls. In these call centers telephone (single type) calls are handled by any call center agent or customer service representative unlike in multi-skill call centers where an agent cannot pick any call (only pick those he specialized on).

One of the most significant task in call center is to create agent scheduling, this is because too many agents lead to unnecessary costs and on the other hand fewer lead to substandard service, Abba (2011). This is the call center optimization problem and is dealt with in workforce management (WFM). In WFM we can then find or determine staffing levels (finding the required number of agents to satisfy the certain service level) and for single-skill inbound call centers we apply a queuing model i.e. Erlang-C formula, Koole (2013). This formula is mostly used in telecommunications companies. Therefore in practice call centers are often viewed as queuing systems, Koole and Mandelbaum (2002). Queuing models are used to estimate system performance so that the appropriate staffing level is ascertained in order to achieve a desired metric including the average speed to answer (waiting time) or the abandonment rate, Robbins et al. (2017). Erlang-C formula assumes that we are dealing with patient customers (no abandonment) which is not the case in practise, we can then apply Erlang-A formula to account for such. However in, Robbins et al. (2017) it is shown that in terms of accuracy for these two formulae, on average there is no much difference between the two and this is clearly discussed in Chapter 3.

As mentioned earlier, call centers form an integral part of many enterprises. However they are facing a lot of problems, including over-staffing, under-staffing, massive influx of inbound calls etc. As a result we have idle workers leading to unnecessary cost, and the required standard service level (i.e. handling 80% of inbound calls within 20s) is not met. Another consequence of this problem leads to customer dissatisfaction or abandonment. This standard service level is the objective of most call centers, but obtaining this from Erlang-C formula usually not easy, Angus (2001). Thus managing call center operations efficiently has become an issue of vital economic interest, Vuthipadadon (2009).

Earlier we mentioned that call centers are used by collection agencies i.e. resolving billing issues. Most of the South African telecommunications companies offer or provide customers service products on contract and these customers are required to make payment later (at specific period of time). The credit risk collections call centers within these companies perform collections actions or activities against delinquent accounts (customers who were supposed to pay on the specific agreed upon date). These collection activities are performed via workflow system (WFS). These collection activities include SMS notifications (SMS1, SMS2), and suspension of service (SoftLock). In the credit and risk collection department customers are segmented according to their risk profile (low, medium, and high risk customers) using segmentation models (like scorecards). These activities are performed based on customer risk profile i.e. the collection process in high risk customers is often accelerated. Due to these actions taken against delinquent accounts, massive influx of inbound calls are experienced by the call center and this is a major problem, Stoltz (2018). We then return to the scheduling problem faced by call centers. Scheduling of activities in a project consumes time and resources. The scheduling problem in this essay is finding an optimal way of spreading out SMS1 such that the standard service level is satisfied. Generally in scheduling problem activities in a project are performed in such a way that the makespan of a project is minimized, and resources are smoothly utilized (resource levelling). Critical path method (CPM), and program evaluation and review technique (PERT) are classical examples of scheduling problems mainly concerned with time aspect of a project. In these problems resources availability is assumed to be infinite, Taha (2006). In resource constrained project scheduling problem (RCPSP) and resource Levelling Problem (RLP) scheduling problems resource availability is limited. In our case we are going to be dealing with limited set of resources i.e call center agents.

Our main focus is to schedule collection activities, particularly SMS1, in such a way that the standard service level is met and the limited resources are smoothly utilized. This type of scheduling problem is RLP. However since new accounts enter to the collection environment on daily basis, continual rescheduling of activities must be performed. This is the additional feature to the classical RLP. In Stoltz (2018) a new model was devised incorporating this feature together with RLP in trying to tackle this problem. In our case we focus on spreading out only SMS1 and look at some optimization techniques that could be used in addressing this problem.

1.1 Credit and Risk collections call center problem

South African telecommunications company (for example, Vodacom) offers some service products in contract, that is, a customer is required to make payment later. A telecommunications company deals with various kinds of customers, some make payment at the beginning or end of the month, and others in any day of the month. Telecommunications company, particularly Vodacom, has the credit and risk collections department and it is responsible for resolving billing issues. If a customer fails to make payment this department is responsible for collection of money. The credit and risk collections call center play an important role in interacting with customers. This department uses the workflow system (WFS) to take actions against delinquent accounts. These collection actions or activities performed via WFS are SMS1, SMS2, and SoftLock. Evidently, these activities are not performed concurrently , so if customers fail to make payment SMS1 is sent via WFS to those indebted accounts. After some couple of days SMS2 (harshly worded) is sent, and lastly after some time, harsh action is taken against customers by suspending all services (SoftLock) i.e. they cannot make or receive calls in their cell phone. The collection process is clearly depicted in Figure 1.1.

When these collection activities, particularly SMS1, are performed to delinquent accounts, customers respond by making calls to the call center in order to clear their debt or negotiate payments. Due to these responses the call center experiences massive influx of inbound calls and the required service level is not met. This large number of incoming calls is due to the collection action(s) taken against customers in the previous days and also the action that is taken in the current day. Handling these inbound calls is a problem since the call center has limited agents. This is the scheduling problem namely resource levelling problem (RLP). In RLP, we take into account the limited resources and the objective is to smoothly utilize these resources through out the project (collection process). In this essay we need to find the optimal number of SMS1 to be sent in a day, that is we spread out SMS1 such that the incoming calls do not exceed the available call center capacity. An important feature in our problem is the fact that new accounts enter to the WFS on daily basis, so activities which were scheduled to be executed in the following days are rescheduled together with the new activities. This feature is very

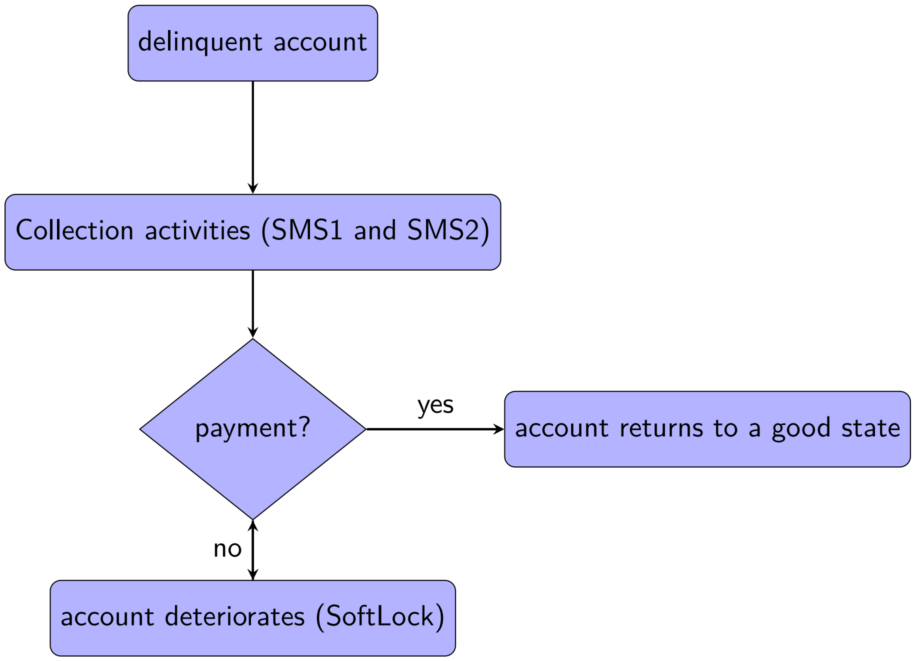

Figure 1.1: Collection process in credit and risk department

significant when addressing our problem.

1.2 Structure of the essay

The main objective of this essay is to present the current problem experienced in credit and risk collections call center and review some of the basic optimization models that can be applied in trying to tackle this novel problem. In Chapter 2.1 we will give some detailed description of some concepts and in Chapter 3 we consider queuing models that are applied in call centers. Our main focus is trying to find the optimal number of SMS1 to be sent in a day, we will not consider the scheduling of SMS2 and SoftLock since that is taken care of by WFS. Then RLP will be considered since we are dealing with limited resource as discussed in Stoltz (2018). Chapter 4 explains more about RLP. Ultimately, the last sections demonstrate some basic idea and methods that can be employed in addressing this problem. We illustrate the idea by use of examples since we do not have data.

2. Attributes of call center

2.1 Call center agents

Call centers are said to be a tool that many companies use for communication with their customers. In our case the call center is handles various customer inquiries with regard to billing issues and typically it comprises centralized pool of trained staff members called agents. When a customer dials telecommunications number the call is received in a center and if agents are available, his call is answered immediately. But if all the staff members are occupied at that instant, the customer will have to wait in queue until a call an agent is available. The call center we considering only communicate exclusively via telephone. In this case every agent can take any incoming calls no specific skill is required. If every agent can take every call, we refer to that center as a single-skilled call center. Where all the agents have received same training and theoretically said to provide homogeneous service level, Molnar et al. (2016). Our problem, as mentioned, is that we are receiving huge number of telephone calls that is difficult to handle since we have limited staff, in our case we have 100 agents. In solving this problem we bear in mind also the fact that our agents should be smoothly utilize through out the collection process. We define utilization as a measure of the amount of time that agents spend working as a ratio of the time available to work, Du Preez (2008).

Generally agents are in one of a number of states during a given interval that they are logged on. Du Preez (2008) showed that the way these states are classified usually differs from center to center, but often include the following:

- Idle - available to take calls

- Active - Handling incoming call (agent is on the phone at this moment)

- Wrap up - Performing after call work and usually not available to any of these calls

- Unavailable - Not handling incoming call and not available to any of these calls. Usually the break time (lunch, tea, restroom) or any other reasons such as formal gatherings etc.

We take note of all these states when addressing our problem. Resource (agents) said to vary with time, thus resource capacity varies with time. This will be discussed in subsequent sections. Agents are said to be renewable resource since they are renewed from period to period (i.e. call center capacity available in a day), Stoltz (2018).

Utilization in Du Preez (2008) is measured as follows:

utilization = (contact time + after call work time)/(total logged in time)

Or, equivalently

utilization = (contact time + after call work time)/(total logged in time - unavailable time)

The second measure takes into account that some of the agents are not available this includes break time. Agent scheduling is a vital activity in any call center without agent schedules the call center cannot function, Koole (2013).

2.2 Maintaining service level

In this essay we will be focussing on one call center performance measure which is the service level. It measures the percentage of inbound calls taken within a specific time limit. It is usually the most common-related metric widely used in call centers. Typically, it is stated as X percent of inbound calls answered within Z seconds, Jouini et al. (2011) and/or, it is measured in terms of customer delay (waiting) times or abandonment rates in the tele-queue. The service is said to be acceptable in a period if the long-run amount of inbound calls that arrive in this period and are taken within a certain time limit Z meets or exceeds a threshold X, Atlason et al. (2008). Defining the grade of service that a call center must achieves is a significant problem. The criteria can be qualitative, i.e. measuring whether or not service was effective. By effective we mean for example, whether or not the customer's problem was solved. However, in our case we limit our attention to the quantitative service levels as we have already discussed at the beginning of this section. We can define the achieved grade of service in our case as the percentage of customers who actually wait less than Z. Queuing models including Erlang-C formula will be discussed in Chapter 3 can be employed in computing the minimum number of agents to satisfy service levels, Koole (2013). One thing that is worth mentioning, is the fact that waiting time is closely related to abandonment rate. If more customers are waiting in queue, there are higher chances that they abandon (drop or stop making) calls. Consequently, this has a negative impact in most companies since maintaining company-customer relationship is of importance. Roughly, service level can be determined by dividing the total number of inbound calls answered within Z seconds by total number of inbound calls, Joia and Oliveira (2010). In this essay, we want to schedule SMS1 in such a way that 80% of inbound calls are answered within 20s. Thus, we want to achieve the standard service level commonly referred to as "80/20 rule".

2.3 Collection process and its strategies

The department of credit and risk collections in a major telecommunications company is responsible for resolving billing issues i.e. collecting on invoices issued to customers and have not been paid on the agreed due date. Customers who missed payment are automatically registered on WFS. This system knows whether or not the payment was made. Indebted accounts enter the system in batches (groups) and appear at unpredictable intervals during course of the month. In fact each and everyday new accounts enter the WFS. Once an account enters the collection environment further actions are taken against customers. Thus, collection activities are performed via WFS on those delinquent accounts. We have three different collection activities performed on delinquent accounts (precedence relation:SMS1<SMS2<SoftLoft). SMS1 is sent to the indebted account to notify about the missed payment, after some couple of days harshly worded SMS2 is sent. If no response is receive still then a harsh treatment is taken by suspending cellular services (SoftLoft). Stoltz (2018) show that the current day SMS1 is sent then we have high response rate meaning large number of calls are received and after two or three days later we have moderate responses. A customer responds to these actions by calling to the call center and resolve or negotiate payment. If a customer resolves or make payment he is removed automatically from the system. Some customers make payment without directly calling to the call center (self cure) this will be discussed shortly. Now as we mentioned earlier that customers are segmented into categories according to their risk profile (low, medium, and high risk customers) and this is done by employing segmentation models. As discussed in Stoltz (2018), each category responds differently to these collection activities. Most high risk customers respond to the last activity (Soft Lock), and medium risk customers respond moderately to SMS notifications, but also respond mostly to SoftLock.

Low risk customers usually respond to SMS notifications and in many cases these customer self cure the owed amount. We should note this when addressing our problem.

This introduces us to collection strategy and it is defined as the set of collections activities performed on delinquent account which has been identified as requiring collection intervention. Following from the discussion above regarding response behaviour of customers, the strategy is different for each category. As an example let us consider a typical strategy applied in medium risk customers:

Day1 of collection strategy,

• On the first day when the payment is missed, SMS1 is sent to that owing customer

Day4 of collection strategy,

• Harshly worded SMS2 in three days later is sent if no response to SMS1 is received (and account is not settled)

and Day7 of collection strategy,

• In this case if there is no response to SMS notifications (i.e account is not settled) harshly treatment is taken (SoftLock).

The problem faced by call center is massive influx of inbound calls due to SMS notifications (also SoftLock) sent today and also others sent in previous days. We have limited renewable resources and at the same time we want to achieve the standard service level. Each and everyday new accounts enter the collection environment so when spreading out or determining the optimal number of SMS1 to be sent we must take all this into account. SMS2 and SoftLoft scheduling is being taken care of by WFS, however for future work scheduling of these activities could also be considered. The study about customer response behaviour shows that we have higher responses in first three days to SMS1, for further discussion see Stoltz (2018).

3. Queuing theory

We begin by explaining some basic concepts in queuing theory before we dive in to the models used. In queuing situation we have two mainly principal aspects which are the customer and server. Primarily, customers are generated from a source and it is either finite or infinite. In a finite source a customer is constrained on arrival at service facility (e.g. machines requesting the service) and infinite source which is of interest, is said to be forever abundant (e.g. calls arriving at a telephone exchange). When customers arrive at service facility two things can happen i.e. either they start service immediately or wait in queue if the facility is occupied at that moment. After completion of a service the facility automatically pulls a waiting customer, if any, from the queue. One must note that a queue is not always packed (or occupied rather) but it may also be empty and in that case it becomes idle until a new customer arrives.

In queuing theory arrival of customers is usually represented by the interarrival time between sucessive customers and for the service, it is described by the service time per interval. A detailed description concerning interarrival and service times will be presented latter. The order in which customers are selected from a queue, is a crucial aspect when analysing queuing models. This ordering and selection of customers is referred to as queuing discipline and first come, first served (FCFS) is the most commonly used queuing discipline. Other disciplines are also used including last come, first served (LCFS) and service in random order (SIRO). In addition, a customer may also be selected from the queue based on some priority. When dealing with queuing models it is very important to consider the behaviour of customers "humans" while they waiting in queues and this plays a crucial role in waiting-line analysis. Customers may jump around from one queue to another in the hope of reducing waiting time and whereas. On the other hand they may resist from joining a queue altogether because of anticipated long delay, or they may just go back on queue since they have been waiting too long. The waiting time is a crucial factor in call centers when discussing service level, Taha (2006).

Having discussed some fundamental elements in queuing system this is a right moment to consider the specialized Poisson queuing situation with c parallel servers as depicted in Figure 3.1 where servers are said to be identical (meaning they have same service time). A customer in waiting queue is selected to commence the service with the first available server. The arrival rate at the system is denoted by λ i.e. customers per unit time. So in the system the total number of customers is given by customers in service plus those in waiting queue. A convenient notation is employed in summarizing the characteristics of specialized Poisson queuing situation as shown in the following:

$A/B/C : D/E/F$

where

A = Arrivals distribution

B = Departures (service time) distribution

C = Number of parallel servers

D = Queuing discipline (usually FCFS)

E = Maximum number allowed in the system

F = Size of the calling source

The above notation is known as Kendall notation which is widely used to classify various queuing systems. Generally A and B are written (standard notation) as M i.e both for arrivals and departures distributions, Taha (2006). The M stands for the Markovian, thus we assume a homogeneous Poisson arrival process and exponential service time. If $E = \infty$ and $F = \infty$, then Kendall notation becomes $A/B/C$ assuming that the queuing discipline is FCFS, Clarke (2007). In the following section we consider the multiple-server models.

3.1 Multiple server models

In these models we have c parallel servers unlike in single-server models where we have only one server. The service rate is denoted by μ and arrival rate is λ as introduced earlier. We begin by considering the most used multi-server model in call centers i.e. $M/M/s$ where both $A\&B$ are written as M and C as s (number of parallel servers) in Kendall notation.

3.1.1 M/M/s queuing model. This stationary or steady-state queuing model is employed when computing the number of agents required to fulfil a particular performance measure. It is often used to estimate the amount of incoming calls that wait less Z seconds. Generally call centers are considered as queuing systems and are often modelled by $M/M/s$ queuing system. When dealing with queuing models in call centers customers are callers, servers are agents, and queues comprise callers waiting for the service (since agents are still occupied at that moment). This convenient and extensively used model is commonly known as Erlang-C formula in call centers (particularly single-skill inbound call centers), Koole and Mandelbaum (2002). Even though it is hailed because of its simplicity, however it does not consider factors like busy signals and abandonments which are usually experienced in practise. Some parameters in our model require forecasting, that is we need to estimate for calling rate and mean service times.

In order to estimate these parameters much operational (historic) data is required and it can be obtained from automatic call distribution (ACD) reports. The ACD is responsible for matching incoming call to the available agent in call center operational systems. If we want to estimate the minimum number of agents required, a call center needs to produce forecasts for the number of incoming calls and an estimate for average handling time in call center. Huge amount of data is generated and it is quite expensive to store it all, hence we use summary statistics over short intervals of half hours (30 minutes). This statistics play a crucial role for predicting likely arrival rates and service time at future times. Call centers usually are not concerned with a long-run, but a short-run (30 minutes) intervals, and in some cases aggregated over full days, Clarke (2007). In most cases service levels fluctuates due to the volatile nature of call centers and with likely incorrect forecasts, and probably the staff levels which are not as planned, Roubos et al. (2012). A steady-state or stationary behaviour in queuing situation is achieved in a long-run, when the system has been operated for a long time, Taha (2006). In a steady state system calls assumed to arrive at a constant rate yet random, Koole and Mandelbaum (2002). This is of course not the case in real world. Hence, service is interpreted as the probability of delay (probability of queue being non-empty) i.e. waiting time in the tele-queue is less than Z seconds (Z as the waiting time usually 20s). Erlang-C formula gives us this probability.

We define this formula as a function of two variables: the number of agents N (which is "s" in our model) and the offered load A, Chromy et al. (2011). Now we can compute the probability:

$$P_C(N,A) = \frac{\dfrac{A^N N}{N!(N-A)}}{\sum_{i=0}^{N-1} \dfrac{A^i}{i!} + \dfrac{A^N N}{N!(N-A)}}, \tag{3.1.1}$$

where,

$$A = \frac{\lambda}{\mu}, \tag{3.1.2}$$

where μ is the service time.

Let us consider another important equation which is going to be useful when discussing call center capacity :

$$\eta = \frac{\lambda}{N\mu}. \tag{3.1.3}$$

Equation (3.1.3) represents the total capacity per agent (load of one agent), where N is the number of agents and μ is the service time.

Also, we may wish to know the average utilization (occupancy) of agents η which is given as (see 3.1.2 ,3.1.3, and 3.1.1) :

$$\eta = 1 - \frac{\sum_{k=0}^{N-1} \dfrac{A^N N}{N!(N-A)}}{\sum_{i=0}^{N-1} \dfrac{A^i}{i!} + \dfrac{A^N N}{N!(N-A)}}. \tag{3.1.4}$$

A crucial factor in terms of a caller is the waiting time in queue and this random variable is given by distribution function as:

$$F_w(Z) = 1 - P_C \cdot e^{\mu(N-AZ)}, \quad Z > 0. \tag{3.1.5}$$

It is also possible to compute average queue waiting time W (mean call waiting time in queue before assigning a call to an agent):

$$W = \frac{P_C}{\mu(N-A)}. \tag{3.1.6}$$

Let us now compute grade of service GoS i.e. finding the percentage of calls answered before the defined threshold AWT (acceptable waiting time), Chromy et al. (2011). We have

$$GoS = 1 - P_C \cdot e^{-\mu(N-A)AWT}. \tag{3.1.7}$$

We have described our equations above as discussed in Chromy et al. (2011). Moreover, the Erlang-C formula (3.1.1) is mostly used in WFM for determining the minimum number of agents required to satisfy the standard service level. This is done by predicting first the service level for fixed number of agents. Then by adjusting number of agents we can get the desired optimal staffing level. Let us consider the following example to illustrate the idea behind staffing level. In this example we concentrate just on the short interval since forecasts vary over the day and so does the staffing levels.

Example: Given forecast FC of 100 (call volume) and average handling time AHT (service time) which is 3.5 (in minutes) for the time interval 10:00 - 10:15. The minimum number of agents can be calculated as

$$FC \times AHT = \frac{100}{15} \times 3.5 = 23.33. \tag{3.1.8}$$

Hence, the minimum number of required agents is 24 (handling incoming calls). By applying (3.1.1) we observe that only 21% of calls will wait less than 20s (the Z parameter). Then by adjusting the number of agents, we find that 28 agents are required to satisfy the standard service level, Koole (2013). These calculations are done by excel (using the above equations) since they involve several computational steps.

Although this model is said to be the basis of parameters and simulation of the call centers its shortcoming is the fact that it neglects busy signals and abandonment factors, Chromy et al. (2011). Customers are assumed to have unlimited patience, which is not the case in real- world. Extensions can be made in $M/M/s$ queuing model to account for abandonment, thus callers abandon if their patience expires. We define abandonment as the maximum time that the caller is willing to wait in the tele-queue. So we have $M/M/s+G$ queue, if we assume the patience is exponentially distributed our model is written as $M/M/s+M$ known as Erlang-A formula ("A" for abandonment), Koole and Mandelbaum (2002). However in the literature it is shown that between the two models $M/M/s$ and $M/M/s+M$ neither model is better in all situations, Robbins, Medeiros, and Harrison (2017). Which is why $M/M/s$ queue is still widely used in call centers. Another extension to our original model is $M/M/s/s$ known as Erlang-B formula ("B" for blocking). In this case a call center can decide to cut out all delays by making the number of trunk-lines to be equal to the number of agents Gans et al. (2003). If all agents are busy the call is blocked automatically.

The aforementioned multi-server models are often used in call centers, in particular $M/M/s$ to find the minimum number of agents required to satisfy the standard service level usually by varying the number of agents until you obtain this grade of service. Moreover, we know that increasing the number of agents does improve the achieved service level. However, since roughly 60% - 70% of all operational costs in a call center are staff costs then achieving the desired grade of service without having to increase number of staff is an important issue in call centers, Clarke (2007). So when the call volume deviates from the expected number changes have to be made either by increasing or reducing staff which is costly for a company. This is a capacity (traffic) management issue and in our case the number of agents is fixed. Data analysis is required when modelling call centers. So, we will use the available data from the WFS and find the expected inbound call volume using queuing models as mentioned earlier and use it as our call center capacity (call center can only handle this expected number of calls). Then we will schedule SMS1 with respect to this capacity constraint. Multi-server model like $M/M/s/s$ (call blocking) could be used to avoid queues but this is not a perfect idea for the business since we want to maintain good relationship with customers. In our problem the massive influx of inbound calls experienced is due to the collection activities performed. Also, even though we want to accomplish the standard service level at the same time resources (agents) must be smoothly utilized which brings us to resource levelling problem (RLP) which is discussed in the next section.

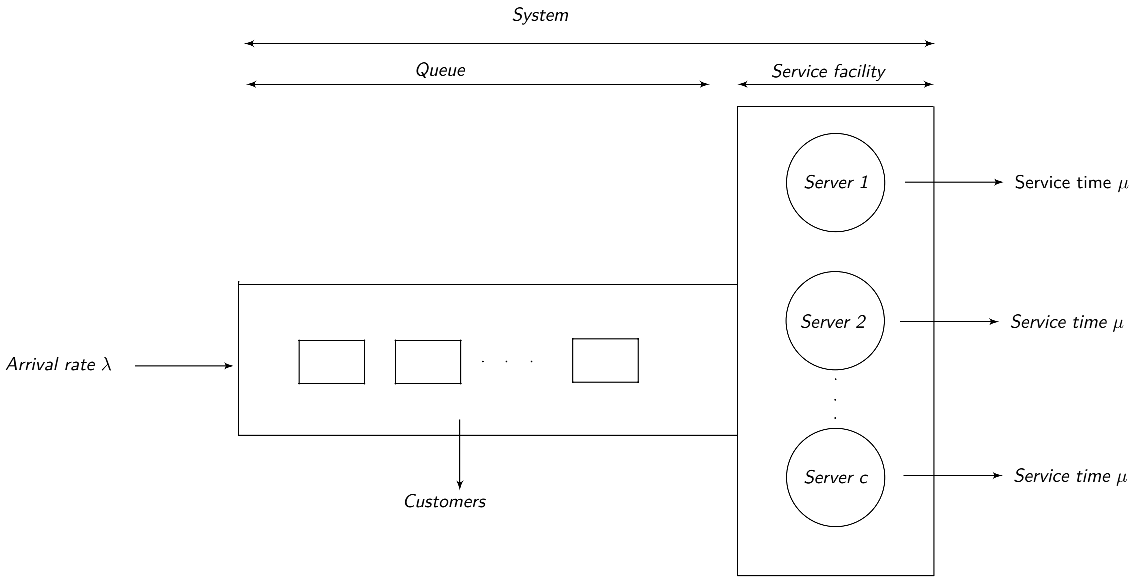

Figure 3.1: Specialized Poisson queuing situation with c parallel servers

4. Resource levelling problem (RLP)

When scheduling an activity in a project (e.g. in a construction) resources are consumed. Various scheduling techniques are employed including critical path method (CPM), Gantt charts, and program evaluation and review technique (PERT) when scheduling activities in projects. These classical techniques are often used, particularly CPM has been widely used since 1950s. However, these scheduling techniques have a shortcoming since they assume that we have an infinite resources which is not the case in real-world. Activities in a project may simultaneously require these limited resources, as a consequence we may have shortages and idleness. This may affect the completion time in a project since some activities might not be executed because of the shortage in resources. Due to these delays companies experience high costs, Damci and Polat (2014). So when dealing with project resources two scheduling techniques are considered, namely resource constrained project scheduling problem (RCPSP) and resource levelling problem (RLP). In both of these techniques we deal with limited resources. RCPSP usually attempts to minimize the project makespan, Hegazy (1999). When formulating RCPSP its objective is to minimize the completion time of a project subject to the resource constraint, Stoltz (2018). However, the RLP's objective is to minimize variations in the utilization of resources without any interference to the duration of a project and minimize the consumption fluctuations of resources. In Damci and Polat (2014) nine possible objective functions for RLP have been reviewed. But the main idea behind these objective functions is very similar i.e. minimizing the resource consumption fluctuations without changing the completion time of a project. Let us formulate the resource levelling problem the same way as Stoltz (2018) in the following :

(a) Let N denote the set of activities to be performed where $\{j_0, j_{n+1}\}$ are dummy variables representing start and end of activities of the project. Thus,

$$N = \{j_0, j_1, ..., j_n, j_{n+1}\}, \tag{4.0.1}$$

where $n \in \mathbb{N}$.

(b) Let D be the set of activity durations. Thus,

$$D = \{d_1, d_2, ..., d_n\}. \tag{4.0.2}$$

(c) Let T be the set of times in which activities must be executed. Thus,

$$T = \{0, 1, ..., t\}, t \leq \bar{T}, \tag{4.0.3}$$

where $\bar{T}$ is the completion time of the project.

(d) Let R be the set of resources. So, we have

$$R = \{r_1, r_2, ..., r_k\}, \tag{4.0.4}$$

where k is the total number of resources.

(e) Let RQ represent resource requirements (k total resources) for each activity (n activities). Thus,

$$RQ = \{\{rq_{11}, rq_{21}, ..., rq_{k1}\}, \{rq_{12}, rq_{22}, ..., rq_{k2}\}, ..., \{rq_{1n}, rq_{2n}, ..., rq_{kn}\}\}, \tag{4.0.5}$$

where rq_{kj} represents the resource requirement of activity j for resource k.

(f) Let U be the set of resource requirement to each time period t. Thus,

$$D = \{u_1, u_2, ..., u_t\}. \tag{4.0.6}$$

(g) Let C be the set of costs associated to each resource. we have,

$$C = \{c_1, c_2, ..., c_k\}, \tag{4.0.7}$$

where k is the total number of resources.

(h) The sequence of starting times for each activity over the project time - line is represented by S_j. Thus,

$$S_j = \{ss_1, ss_2, ..., ss_k\}, es_j \leq ss_j \leq ls_j, \tag{4.0.8}$$

where es_j is the earliest and ls_j latest possible start times for each activity j. The total amount of time which an activity can be delayed without causing delay to the project makespan is k.

(i) RLP objective function can be represented as $c_i f[r_i(S,t)]$, which is the cost c_i of resource r_i times resource consumption of r_i at time t for feasible schedule S, for all resources k. This is one of the objective functions discussed in the succeeding paragraphs.

Mathematically, we have

$$min \sum_{i=1}^{k} c_i f[r_j(S,t)] \tag{4.0.9}$$

subject to :

$$ss_{n+1} \leq \bar{T}, \tag{4.0.10}$$

and

$$ss_i + d_i + \rho_{ij} \leq ss_j, \forall (i,j) \in A. \tag{4.0.11}$$

So ρ_{ij} is the total slack (basic variable) between i and j. Equations (4.0.10) and (4.0.11) represent deadline and precedence relation constraints respectively. The set A is the set of pairs of activities.

At the beginning of this section we mentioned about several objective functions for RLP. These different objective functions are discussed in Table 4.1. The use of these various objective functions usually give the management teams some important idea regarding efficient use of different resource types. However, there is no objective function that provides even resource usage for every project (for example construction project). Basically the idea behind RLP is to try and make resource utilization as smooth as possible. This can be done by choosing the appropriate objective function, Damci and Polat (2014). Despite its pivotal role in scheduling projects by minimizing the fluctuations in resource utilization, applying traditional RLP to the real-world problem in this essay is not enough. Extensions to this scheduling technique have to be made to account for the additional features in our problem. These features include continual rescheduling which is a substantial feature requiring considerable attention. As mentioned earlier new accounts enter the collection environment (WFS) on daily basis. So when creating a schedule of activity such that resources are consumed in a smooth way (consumption levelled) over a day under consideration. A likely situation that will occur is an inadequate resource availability at certain times (days) due to batches of accounts entering WFS daily. In Stoltz (2018) a model was devised to address these additional features to classical RLP. Its goal was to create a schedule such that the resource (call center capacity) usage is levelled within the next few days (including current day). These few days (time periods) were referred to as critical time window. In the objective function of this model, first three days (critical time window) were assigned highest weights and the rest normal weights. The intuition behind these three days was strongly influenced by knowledge about customer response rates further discussed by this paper. The model incorporates critical time window resource levelling and continual rescheduling.

So in this essay we will adopt the same idea and make some changes in the model. Parameters like weight will be removed from this model since we will not focus on the critical time window. Our main objective is to schedule an activity (SMS1) such that the service level is met and the resource usage is levelled. A more detailed discussion is presented in the next section.

Table 4.1: Objective functions for RLP

Objective function	Optimization rule
min $\sum_{i=1}^{\bar{T}} \lvert Rdev_i \rvert$	Mininimization of the sum of the absolute deviation in daily resources, where $Rdev_i$ is the deviation between resources needed from time period i (day) to period $i+1$, i is the time period under consideration , and $\bar{T}$ is the project deadline.
min $\sum_{i=1}^{\bar{T}} \lvert Rinc_i \rvert$	The objective is to minimize the sum of only the increases in daily resource utilization from one time period to the next.
min $\sum_{i=1}^{\bar{T}} \lvert R_i - A_{rr} \rvert$	The objective is to minimize the sum of the absolute deviations between daily resource utilization and average resource utilization. So R_i denote resource needed at time period i and A_{rr} represent average resource utilization.
min $[maxR_i]$	The objective is to minimize the maximum daily resource utilization.
min $[max\lvert Rdev_i \rvert]$	In this case we minimize the maximum deviation in daily resource utilization.
min $[max\lvert R_i - A_{rr} \rvert]$	Similarly, we minimize the maximum absolute deviation between daily resource utilization and the average resource utilization.
min $\sum_{i=1}^{\bar{T}} {R_i}^2$	We minimize the sum of the square of daily resource utilization.
min $\sum_{i=1}^{\bar{T}} (Rdev_i)^2$	Minimizing the sum of the square of daily resource utilization.
min $\sum_{i=1}^{\bar{T}} (R_i - A_{rr})^2$	Minimizing the sum of the square of the deviations between daily resource utilization and the average resource utilization.

5. Optimization

The massive influx of inbound calls is an ongoing process. Various methods have been tried in order to tackle this problem. For example, in Stoltz (2018) SMS notifications were sent at different times of the day. As a consequence high call volumes were experienced in a call center which implied that SMS notifications have an effect or contribute to this major problem. As mentioned in Chapter 3 to address this problem number of agents can be increased until the required service level is met. However, this comes with high costs for a company. Hence, the number of agents cannot be increased any further and we have 100 of them in our case (fixed). Also, the resources must be smoothly utilized during the collection process. We have to find an optimal way of spreading out SMS1. This is the scheduling problem and its objective function is formulated as:

$$min \sum_{k=1}^{K} \sum_{i=1}^{T} [r(t_i) - R(t_i)]^2, \tag{5.0.1}$$

where $r(t_i)$ is the resource consumption for resource k at time (day) t_i and $R(t_i)$ is the resource availability for resource k at time t_i. This objective function is subject to the same constraints as in Chapter 4.This function is a modification of the one presented in Stoltz (2018). Since there is only a single resource (agents) in our real world problem, then $k = 1$. Equation (5.0.1) becomes,

$$min \sum_{i=1}^{T} [r(t_i) - R(t_i)]^2. \tag{5.0.2}$$

In order to compute this function we begin by finding the values of these two parameters, namely $r(t_i)$ and $R(t_i)$. Before we tackle our problem, we consider the collection strategies devised by call center management (a typical example). We do so based on the analysis of response rates described in Stoltz (2018). Suppose that the call center devise the following collection strategies which are applied (respectively) to three owing customer categories segmented according to their risk profile (low, medium, high) using some segmentation models such as scorecards. These collection strategies are shown below:

Table 5.1: Activity days for Collection strategies.

collection strategies	**SMS1**	**SMS2**	**SoftLock**
High risk	1	4	7
Medium risk	1	5	8
Low risk	1	7	12

This is a typical example of collection strategies applied to indebted accounts according to their risk profile. Our main focus is the scheduling of the first activity (SMS1). Therefore, having devised these collection strategies we can now determine two parameters of the objective function.

Finding the resource availability $R(t_i)$ (as the call center capacity)

Capacity of the call center is defined as the ability to handle inbound telephone calls. It is the function of agents at a given time. The total capacity per agent can be calculated using (3.1.3) as presented in Chapter 3. In order to determine this capacity various aspects are considered. For example, number of agents working per hour of the day, the state of an agent as discussed in Section 2.1, and all other factors involve are taken into account. Queuing theory together with its equations as discussed earlier play a vital role in finding call center capacity (including required agents to satisfy certain service level). We compute the inbound call volume as our resource capacity constraint. Suppose we have an available (historic) data from WFS and then we can proceed to find call center capacity.

Table 5.2: Call center planning input metrics.

Parameter	**Value**
Expected incoming (no.call)	4739
Call duration (s)	325
Acceptable waiting time, AWT (s)	20

Table 5.3: Workforce break scheduling inputs

Planning metric	**Tea**	**Lunch**
Break Shifts	10	4
Agents	80	80
People per shift	9	12
People on floor	71	68
Change over efficiency	99.9%	99%
Actual people available	70	67

Table 5.4: Call Centre workforce and capacity (service level) planning tool.

Interval(s)	**Time**	**Calls**	**Agents**	**AWT**	**Traffic intensity**	**Erlang-C**	**Occupancy**	**Service level**
3600	8-9	700	80	20	63.19	0.78	0.02	99
3600	9-10	640	70	20	57.77	0.82	0.08	96.2
3600	10-11	640	70	20	57.77	0.82	0.08	96.2
3600	11-12	520	70	20	46.94	0.67	0.00	100
3600	12-13	679	67	20	61.29	0.91	0.37	73.9
3600	13-14	780	67	20	70.41	1.05	1.59	-96.3
3600	14-15	400	67	20	36.11	0.53	0.00	100
3600	15-16	380	70	20	34.30	0.49	0.00	100

We obtain the following results from the above tables:

Total handled inbound calls $= 4739$ and daily service level $= 71\%$.

We illustrate an implementation of call volume planning in these tables which was developed using Microsoft Excel. The input values are interval, time, AWT, calls, call duration, and number of agents. We assume that these values are given since they can be obtained using data. The assumption behind this is to demonstrate the idea about capacity calculations. On the other hand working values include traffic intensity, occupancy, and Erlang-C computed using equations presented in Chapter 3 (with Excel). Lastly, the output value is the service level. It is very important that we consider Table 5.3 since it contributes to our calculations (capacity management). As we observe from our typical example of call center workforce break scheduling that the actual people available (agents) due to this scheduling has effect on service level. In Table 5.4 this is illustrated by the negative service level in mid-day. The change over efficiency is defined as the loss of production time, for instance sitting, logging in the system etc. One must also note that from Table 5.3 people per shift refers to agents who actually go off together on a break. All these dynamics contribute to the capacity calculations. From our typical example, the call center capacity is around 4739 per day (And we achieve 71% service level as shown above). This is the expected value, but it may vary over a day due to the call center dynamics.This is clearly best discussed in Stoltz (2018). After computing this parameter as demonstrated by this typical example (not the actual data) we can find the resource consumption requirement as the responses due to performed activities.

Finding the resource consumption requirement $r(t_i)$

In this case before we compute the resource requirement of activity (SMS1) for agents at time (day) t. We begin by describing the basic idea behind the scheduling of activities. One must note that when the activities are scheduled in the current day we observe that some which were scheduled in the previous days still have an impact in the current day. Thus, when generating the optimal schedule of SMS1 we take this into account. That is, when performing optimization in the current day initially we have a schedule consisting of activities which were scheduled to occur in the next periods (not yet started). Also, the new activities that enter the system on daily basis. For example, suppose 13000 activities are scheduled but only 2000 were performed (sent) on that day. The remaining 11000 were scheduled to occur in the next days but not yet commenced and hence are rescheduled. Let us illustrate this intuition by an example in the following :

In this example, we consider three collection strategies as shown in Table 5.1 with independent resource consumption profile requiring scheduling. When we schedule activities (optimally spreading out SMS1) we consider the impact of these activities scheduled in three days before the current day and also the next three days (after). Therefore, the schedule time horizon is 6. So let us consider the resource consumption requirement of activity (SMS1) for agents at time t. For example, suppose we are given the resource consumption requirement of activities as 3×4 matrix C. Thus,

$$C = \begin{pmatrix} 0.5 & 0.3 & 0.15 & 0.05 \\ 0.6 & 0.3 & 0.05 & 0.05 \\ 0.7 & 0.1 & 0.1 & 0.1 \end{pmatrix} \tag{5.0.3}$$

where c_{ij} is the resource requirement of activity i (collection strategy as applied to a single account) at time period j. The rows represent three different collection strategies (high, medium, low) and the columns correspond to the time periods . We observe that if one unit of an activity 1 (fist row) was performed at day t, then when customers respond consume 0.5 units of resources (agents) during that day. Consuming 0.3 units of resources in the next day and so on.

Suppose SMS1 was scheduled as shown in matrix S. That is three days ago, on the current day, and for the next three days. Thus,

$$S = \begin{pmatrix} 550 & 500 & 400 & 550 & 150 & 80 & 20 \\ 420 & 350 & 400 & 460 & 70 & 50 & 30 \\ 350 & 200 & 300 & 370 & 50 & 22 & 30 \end{pmatrix} \tag{5.0.4}$$

The rows represent collection strategies (on 3 categories) and columns correspond to the days. Elements of this matrix are the number of SMS1(s) spread out in three days ago, current day, and for the next three days. Column 1-3 represent previous days and column 4 is the current day. Then, the remaining columns are the next few days (column 5-7). In this case when scheduling SMS1 we interest on its impact in the next three days, this is denoted by ϕ (number of periods of response) and in this case $\phi = 3$.

In matrix C and S, we have $\phi + 1$ and $2\phi + 1$ columns respectively. The number of rows as mentioned represent collection strategies and we have three in this case. These two matrices are defined as,

$$C \in \mathbb{R}^{3\times(\phi+1)}, \tag{5.0.5}$$

and evidently from properties of matrix we have

$$C^T \in \mathbb{R}^{(\phi+1)\times 3}. \tag{5.0.6}$$

On the other hand, we have

$$S \in \mathbb{R}^{3\times(2\phi+1)}. \tag{5.0.7}$$

So, we can use the following expression to determine total responses for all activities

$$C^T S \in \mathbb{R}^{(\phi+1)\times(2\phi+1)}. \tag{5.0.8}$$

Now, let us compute the following matrix multiplication :

$$\begin{pmatrix} 0.5 & 0.6 & 0.7 \\ 0.3 & 0.3 & 0.1 \\ 0.15 & 0.05 & 0.1 \\ 0.05 & 0.05 & 0.1 \end{pmatrix} \times \begin{pmatrix} 550 & 500 & 400 & 550 & 150 & 80 & 20 \\ 420 & 350 & 400 & 460 & 70 & 50 & 30 \\ 350 & 200 & 300 & 370 & 50 & 22 & 30 \end{pmatrix} \tag{5.0.9}$$

$$= \begin{pmatrix} 772 & 600 & 650 & 810 & 152 & 85.4 & 49 \\ 326 & 275 & 270 & 340 & 71 & 41.2 & 18 \\ 138.5 & 112.5 & 110 & 142.5 & 31 & 16.7 & 7.5 \\ 83.5 & 62.5 & 70 & 87.5 & 16 & 8.7 & 5.5 \end{pmatrix} \tag{5.0.10}$$

The resulting matrix obtained gives us the total resource consumption requirement. So, we observe the following :

• The total consumption requirement of SMS1 for resource during $t - 3$

From the first column of our matrix we observe that total resource consumption requirement due to SMS1 that was sent is three days ago ($t - 3$) is 772 units during that day, 326 units in two days ago ($t - 2$), 138.5 units a day ago ($t - 1$), and on the current day (t) is 83.5 units.

- The total consumption requirement of SMS1 for resource during $t-2$

In the second column total resource consumption requirement due to SMS1 that was sent in two days ago is 600 units on that day, 275 units one day ago, 112.5 units the current day, and 62.5 units a day later $(t+1)$.

- The total consumption requirement of SMS1 for resource during $t-1$

Third column shows us that total resource consumption requirement due to SMS1 sent one day ago is 650 units on that day, 270 units in the current day, and so on.

- The total consumption requirement of SMS1 for resource during t

In fourth column, the current day total resource requirement due to SMS1 sent is 810 units, 340 units in the following day and so on.

So, we observe that resource consumption during each day rely on the activities (SMS1) sent in the past. Now we can compute the resource consumption on the current day t i.e.

$$r(t_i) = 83.5 + 112.5 + 270 + 810 = 1276. \tag{5.0.11}$$

Since we have determined both parameters then we return to our original function in equation (5.0.2) and solve it. We try to minimize the discrepancy between these two values i.e. number of responses and call center capacity. To find the optimal number of SMS1 to be sent we randomly spread out SMS1 as shown by matrix S and perform the above computations. Then, we optimize our function until we find the optimal number of SMS1 to be sent. For example,

We can perform a random scheduling of SMS1 and compute the resource requirement (responses). This could be done for a long time for instance a year. Thus,

$$min = \sum_{i=1}^{365} [r(t_i) - R(t_i)]^2. \tag{5.0.12}$$

Randomly spreading out we can find $r(t_i)$ that agrees with the expected number of calls $R(t_i)$. This process is computationally expensive. Hence meta-heuristic optimization methods can be employed see, Stoltz (2018). Suggestions have been made regarding the scheduling of SMS1, since we know that the call center capacity is uniform through out the computation. Suppose 4739 is the handling inbound call capacity and we have 5000 accounts requiring collection action. Then, what if we take the difference and send 4739 or less. But, this is the problem since call center capacity varies over a day as shown in Table 5.3 and the activities performed in previous days also contribute.

6. Conclusion

In this essay we have presented the effect of activities performed on indebted accounts. Automatically the workflow system (WFS) that is used identifies customers who miss payment(s). Which results in credit and risk collections actions against them. Most customers will call directly to the collections call center to resolve or negotiate payment. As a consequence of these responses massive influx of inbound calls is experienced since there is a limited number of call center agents. It is important to note that increasing the number of staff to handle the large call influx is costly and at the same time a standard service level has to be met. We discussed in this essay a method to optimize this call center problem by minimizing usage fluctuations of resource (agents). In a nut shell, these massive inbound call volumes in the credit and risk collections call center are positively corrected to the activities performed. We demonstrated this as the scheduling problem.

Since we know that call centers can be viewed as queuing system then queuing theory was discussed in this essay. Fundamental models relevant to our problem were discussed briefly. One of the prominent queuing model i.e. Erlang-C formula was presented along with its contribution in call centers. Although in our case we did not have data but we presented the importance of queuing models in addressing this problem. Since we have limited number of agents it was important for us to determine the call center capacity. All of our computations for determining call center capacity were mainly based on queuing theory. By knowing the ability to handle incoming calls, we can try to schedule activities in such a way that the resulting calls do not violate this capacity constraint.

Maintaining standard service level is an objective of many call center companies. Collections call center wants an optimal way of spreading out SMS1 such the standard service level is met. Also, the agents must be smoothly utilized. Hence, classical resource levelling problem (RLP) was proposed. However, since new customers enter WFS on daily basis this was not sufficient. A model incorporating RLP and continual rescheduling feature was discussed. We demonstrated by use of examples on how this problem can be solved. Our main focus was solely on scheduling SMS1 and the other activities said to be taken care of by WFS. We observed that when scheduling SMS1 in the current day, one have to take into account the previous scheduling phases since they heavily have an effect in current day. In our case when trying to tackle this problem as an example, we used random way of scheduling SMS1 until you find the optimal number of SMS1. This is a novel problem and few studies have been done, and in big scale centers this problem computationally is expensive. So meta-heuristics optimization methods can be employed.

Acknowledgements

I would like to thank the AIMS family for the warmth treatment. Indeed I have learnt a lot. Not to forget my supervisor Prof Montaz Ali for such an exciting topic. My tutor Kenneth he really helped a lot. January-intake , you have been amazing thank you. Ngibonge kuni nonke!

References

Abba, K. *Modelling the queue in a call centre:(a case of IT Helpdesk, Vodafone Ghana)*. PhD thesis, 2011.

Aitkin, M. A., Francis, B., Hinde, J., and Darnell, R. *Statistical modelling in R*. Oxford University Press Oxford, 2009.

Angus, I. An introduction to erlang b and erlang c. *Telemanagement*, 187:6–8, 2001.

Atlason, J., Epelman, M. A., and Henderson, S. G. Optimizing call center staffing using simulation and analytic center cutting-plane methods. *Management Science*, 54(2):295–309, 2008.

Avramidis, A. N., Deslauriers, A., and L'Ecuyer, P. Modeling daily arrivals to a telephone call center. *Management Science*, 50(7):896–908, 2004.

Chromy, E., Misuth, T., and Kavacky, M. Erlang c formula and its use in the call centers. 2011.

Clarke, S. *Robust staff level optimisation in call centres*. PhD thesis, University of Oxford, 2007.

Damci, A. and Polat, G. Impacts of different objective functions on resource leveling in construction projects: a case study. *Journal of civil engineering and management*, 20(4):537–547, 2014.

Du Preez, J. J. *Call centre design, operation and optimisation: A structured and scientific based approach*. PhD thesis, Stellenbosch: University of Stellenbosch, 2008.

Gans, N., Koole, G., and Mandelbaum, A. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

Hegazy, T. Optimization of resource allocation and leveling using genetic algorithms. *Journal of construction engineering and management*, 125(3):167–175, 1999.

Joia, L. A. and Oliveira, A. F. Call center key performance indicators and customer satisfaction. In *AMCIS*, page 31, 2010.

Jouini, O., Koole, G., and Roubos, A. Performance indicators for call centers with impatience. *Submitted for publication*, 3(1), 2011.

Koole, G. Call center mathematics. *Version of January*, 26, 2005.

Koole, G. *Call center optimization*. MG books, 2013.

Koole, G. and Mandelbaum, A. Queueing models of call centers: An introduction. *Annals of Operations Research*, 113(1-4):41–59, 2002.

Molnar, G., Jakobović, D., and Pavelić, M. Workforce scheduling in inbound customer call centres with a case study. In *European Conference on the Applications of Evolutionary Computation*, pages 831–846. Springer, 2016.

Robbins, T. R., Medeiros, D., and Harrison, T. P. Evaluating the erlang c and erlang a models for call center modeling. 2017.

Roubos, A., Koole, G., and Stolletz, R. Service-level variability of inbound call centers. *Manufacturing & Service Operations Management*, 14(3):402–413, 2012.

Stoltz, A. R. *An approach to solving a scheduling problem arising in industry*. PhD thesis, 2018.

Taha, H. *Operations Research an Introduction*. Prentice Hall, 8th edition, 2006.

Vuthipadadon, S. Scheduling inbound calls in call centers. 2009.