THE FLAGSHIP RESEARCH JOURNAL OF INTERNATIONAL CONFERENCE
OF THE PRODUCTION AND OPERATIONS MANAGEMENT SOCIETY
Volume 2• Number 2 • July - December 2009

Joscm

Dimensioning a Call Center: Simulation or Queue Theory?

Marco Aurélio Carino Bouzada, PhD
Universidade Estácio de Sá (MADE)
marco.bouzada@estacio.br

ABSTRACT: The objective of this paper is to establish a dichotomy – opposing analytical methods (such as Queue Theory) to experimental methods (such as Simulation) and discussing their adequateness to complex operations – set up in the matter of dimensioning the handling capacity of a large brazilian call centers company. The literature related to the application of such methods at call centers is reviewed, and the way the question is treated nowadays by the company is described. Then an experimental approach is suggested to be implemented as an alternative methodology to deal with the issue, instead of the analytical method in use. The results obtained are used to justify the adequacy of the experimental approach to the modern call centers operation, as long as it is possible to have the model closer to reality. The main implication points to a better understanding of the operation achieved with the new approach.

KEY WORDS: Call center, Dimensioning, Simulation, Queue Theory.

1. INTRODUCTION

Tens of billions of dollars were spent on call centers during the last years of the 90's. The growth of this industry on that decade was around 20% per year and it is expected that this rate will be maintained at this level during the beginning of this century. In particular in Brazil, the call center industry has grown a great deal during the last years, the domestic market representing high financial amounts. In todays' economy, the call centers not only became the primary contact points among clients and companies, but also a great investment for many organizations as well. Labor costs represent approximately 70% of the total industry, justifying the need for an efficient management and the importance of a quantitative approach for the dimensioning of a service handling capacity, which consists on a trade-off between this cost and the determination of the right service level; in other words, in having the right number of qualified people and resources at the right moment, in order to work with the forecast working cargo, maintaining the quality patterns and the required service level. This way, and according to the call centers' day to day more and more complex operations, the use of better accurate models on the dimensioning of the size of the personnel team, of the industry that works with great financial volumes, has been considered as more important than ever (HALL; ANTON, 1998; GROSSMAN et al., 2001; ALAM, 2002; WEINBERG; BROWN; STROUD, 2006; BOUZADA, 2006).

Hall and Anton (1998) said that call centers may use the Simulation tool to test (and eventually justify its implementation) whether some changes can prove or not to be able to improve the system before its implementation. The best call centers use this tool effectively to design the system, manage the operation and plan ahead, in the face of potential scenarios.

Bouzada (2006) explains that this happens because, amongst other reasons, the handling capacity dimensioning consists of a critical activity in the reaching of the efficiency and effectiveness of the operation. And the simulation tool usually fits better to the dimensioning of more complex operations (as that related to modern call centers, for instance), since it can model very well the real world, presenting more accurate and relatively precise results. It is true that these results are not as precise as the theoretical ones

obtained by analytical methods, but they are usually pretty close to them. The precept of the Simulation says that "it is better to have a rough solution for a very realistic model than an exact solution for a model with several approximations".

As studied by Mehrotra and Fama (2003), and Hall and Anton (1998), the call centers are interesting objects for the simulation studies, because: (i) they cope with more than one type of call, where each type represents a line; (ii) the calls received in each line arrive by chance – as time goes by; (iii) in a few cases, agents make calls proactively (especially in telemarketing or charging calls), or as a return for a call received; (iv) the duration of each call is random, as well as the work that the agent executes after the call (collecting of data, documentation, research...); (v) the progress on the systems which route the calls for the agents, groups or locals, make the logics behind the call center even more sophisticated; (vi) agents can be trained to answer only one type of call, several types of calls or all types of calls with different priorities and preferences specified for the routing logics; and (vii) the great amount of money invested in call centers, on both forms, capital and work, is capable to justify the use of this so powerful tool.

Thus, this paper intends to describe the handling capacity dimensioning matter of a large brazilian call centers company in order to approach it using an alternative methodology: the Simulation. The objective is to use the case of the highlighted company as an empirical scenario for the theoretical discussion on the adequateness of experimental methods (such as Simulation) to complex operations (such as modern call centers), in detriment of analytical methods (such as the Queue Theory).

2. LITERATURE REVIEW

2.1 Queue Theory applied to Call Centers

In a call center system, a queue occurs when there is no agent available to handle a client, which waits on a virtual line from which he will leave only when an operator is set to attend him or when he disconnects the call. As observed by Brown et al. (2002), in the case of call centers, the virtual queue is invisible among the clients and among the clients and the operators.

In the call centers scenario, Araujo, Araujo and Adissi (2004) say that the queues discipline, when well managed, is a strong ally for the call centers production planning and controlling area, which have as a goal to achieve the expected results with scarce resources, turning this area more and more important for these companies. The queues discipline, when well managed, can bring a significant reduction to the clients waiting time.

A few call center characteristics make it difficult to apply analytical formulas from the Queue Theory for its modeling, including: generic distribution for the handling time, time-varying arrival rates, temporary overflows and abandonment. The model introduced by Chassioti and Worthington (2004) consists of a practical approach capable of incorporating most of these features.

According to Bapat and Pruitte Jr. (1998), the premises adopted by the studies based on Queue Theory analytical models are extremely limited when based on call centers current context because: (i) the incoming calls are all of the same kind; (ii) from the moment a call enters a queue, it never leaves it, and this usually overestimates the labor needed, increasing the personnel costs for the company; (iii) the attendants handle the calls following the FIFO ("first in, first out") discipline; and (iv) each operator handles all calls the same way.

These premises rarely work at the environment in which call centers are inserted, since, according to the mentioned authors – depending on the individual tolerance for waiting his turn to be handled – a client may disconnect the call, if queued. Furthermore, the operators normally differ in relation to their own skills and to the handling time. Additionally, the clients' needs are very different and, sometimes, a priorization that can offer a better service might be necessary. Nevertheless, many companies continue to support the usually complex decisions related to the resources allocation by means of Queue Theory analytical models driven by the approach easiness and quickness.

Many call centers present a generic distribution (lognormal, for instance) for their handling times and not necessarily a negative exponential distribution (BROWN et al., 2002). The exponential format is used in most of the Queue Theory literature, not only for the time between clients arrival, but also for the handling time. This is due to the fact that there are analytical solutions for the system stationary state when these times are considered as following an exponential distribution.

But within the real world call centers, at least the incoming clients rate varies as time goes by. This vari-

ation is driven by advertisements, work shifts etc.. The attendants fatigue can also generate a variation on the handling time as the day goes by, but it is insignificant when compared to the arrival rate variation. The solutions found in the literature to deal with the time-varying arrival rates are not so useful because they involve Bessel's functions, of difficult application (CHASSIOTI; WORTHINGTON, 2004).

2.2 Simulation in Call Centers

Chokshi (1999), Klungle and Maluchnik (1997), Hall and Anton (1998), Mehrotra and Fama (2003), Avramidis and L'Ecuyer (2005), Klungle (1999) and Bapat and Pruitte Jr. (1998) go beyond a few recent factors that contributed to the increase on the demand for the use of the simulation tool in the call centers sector: (i) the increasing importance of the call centers for a good number of corporations, due to the fast increase of information, communication and technological gadgets, increasing the need to use scientific methodologies on decision makings and tools for its strategic management instead of using the intuition, only; (ii) the increasing complexity of call traffic along with rules more and more viewed on the skill-based routing; (iii) the uncertainty more and more predominant at the decision problems usually found on the operational management of call centers phone desks; (iv) fast changes on the operations and improvement of the re-engineering activities resulting from the increase of joint-ventures and acquisitions, business volatility, outsourcing options and the utilization of different channels in order to reach the consumer (telephone, e-mail, chat...); and (v) the availability and accessible price of the computers, together with a range of Simulation applications in call centers, available in an everyday market less and less complex, intuitive and easier to be assimilated and used.

Simulation, according to Mehrotra (1997), explicitly shapes the interaction between calls, routes and agents, as well as the random individual incoming calls and the also random duration of the handling service. Through the use of Simulation, managers and analysts translate the call centers gross data (call forecast, distribution of the handling times, schedule hours and the agents abilities, call route vectors, etc.), in handling information on the service levels, clients abandonment, use of agents, costs and other important performance measures of a call center.

According to Chokshi (1999) and Klungle and Maluchnik (1997), the use of Simulation to help management decisions in a call center allows the following benefits: (i) to visualize future processes and be used as a communication tool; (ii) to validate the processes premises before its implementation; (iii) to analyze the impact of the changes (scenario studies) in detail; (iv) to foresee the aggregated needs of resources and to schedule the working team; (v) to measure the performance indicators; and (vi) to estimate impacts on costs and economies.

One of the usages of the Simulation in a call center, as said by Hall and Anton (1998), is the evaluation when one may verify "where the call center is". The key-question is "how efficient is the operation nowadays?" The goal of this evaluation is to establish a point of departure (and reference) for the change.

In accordance to Mehrotra, Profozich and Bapat (1997), Yonamine (2006), Gulati and Malcolm (2001), Bapat and Pruitte Jr. (1998) and PARAGON (2005), a simulation model can be used (and has been used more frequently than ever) – besides normally allowing graphics and animations – to contemplate a few other critical aspects of the modern receptive centers of all sizes and types, such as: (i) a specific service level; (ii) flexibility on the distribution of time between incoming calls and of handling time; (iii) consolidation of the central offices; (iv) skill-based routing; (v) multiple types of calls; (vi) simultaneous lines; (vii) call disconnect patterns; (viii) call returns; (ix) overflow and filling of capacity; (x) waiting lines prioritization; (xi) call transference and teleconferences; (xii) operators preferences, proficiency, time learning and schedule. The outputs model can emerge in shape of waiting time, call disconnecting average amount, (both with the possibility of differentiation on the call types) and level of the operators utilization (with possibility of the operator types differentiation). And, due to the applicability of this approach to the real and complex characteristics of call centers, the Simulation can make its dimensioning and management more reliable.

In accordance to Mehrotra, Profozich and Bapat (1997), Steckley, Henderson and Mehrotra (2005), PARAGON (2005), Mehrotra (1997), Klungle and Maluchnik (1997), Pidd (1998) and Tanir and Booth (1999), the traditional methods most often used to manage and size a call center (intuitive estimatives, unprepared computations, worksheets and Erlang queue theoretical models) are becoming significantly limited due to the variability of the incoming calls, routes and handling time, to the operators skills and priorities, to the call heterogeneity and the interac-

tion among them and the line trunks, to the dynamic of the call disconnections, to the recent tendencies (such as the skill-based routing, electronic channels and interactive calls handling) and, in general, to the sophistication and complexity more and more evidently noticed in the call centers systems. For example: analytical models usually assume that the clients arrival follows a Poisson process when, as a matter of fact, the call centers' data constantly reject this premise. In addition, worksheets and Erlang models overestimate the number of agents, besides having not much precision for call centers with different handling for each kind of client.

The Simulation enlarges the capacity of the analytical tools and consists of a superior approach when there is no workable theoretical model capable to provide a reasonable system representation and when the means are not sufficient, the accuracy is important, the operation is detailed, the demand varies too much, bottlenecks and processes design changing needs must be identified, or else an animation is necessary to improve the communication of a change to the company's board. The industry recent tendencies demand more sophisticated approaches and the Simulation provides the necessary techniques to acquire the insights about these new tendencies and helps to shape its present and future designs, consisting in the only analysis method able of modeling a call center efficiently and accurately, throughout an approach much more practical, flexible in terms of inputs and outputs, and capable of allowing the inclusion of important details, of representing much better the reality (without great needs of simplifications as theoretical models do), of enabling a better and deeper understanding concerning the call center processes and of generating much more robust results regarding the call center performance, allowing its optimization in a more reliable way (PARAGON, 2005; RILEY, 2005; MEHROTRA, 1997; KLUNGLE; MALUCHNIK, 1997; TANIR; BOOTH, 1999; SALIBY, 1989; HILLIER; LIEBERMAN, 1995; HERTZ, 1980; MEHROTRA; PROFOZICH; BAPAT, 1997; BAPAT; PRUITTE JR., 1998; CHOKSHI, 1999; KLUNGLE, 1999; WORTHINGTON; WALL, 1999; RAGSDALE, 2001; MEHROTRA; FAMA, 2003).

3. THE CASE

3.1 The company

Contax emerged by the end of the year 2000 as a natural extension of Telemar's business, in a branch of the economy which did not invest much in technology and qualification of the customer's service, in order to help its clients on their operational management, aggregating value on the relationship with final customers (CONTAX, 2006).

Presently in Brazil, Contax is the largest growing company in this industry, with a growth of almost 60% in 2005, when it invoiced R$ 1.129 millions. It is known as the largest enterprise of this branch based on the number of attending service positions, and the second largest in terms of sales and work force, inside the national territory (OUTSOURCING, 2005).

The Contax capital is 100% national and today it operates with more than 22.000 positions at the customer service, almost 50.000 employees and more than 40 clients, with Telemar being the largest one (representing approximately 60% of the sales). The main products related to this client are: (i) Velox; (ii) 103; (iii) technical support and repairs; (iv) Oi; and (v) 102, which receives calls of customers that need information from the telephone directory.

3.2 The current dimensioning process of handling capacity

The dimensioning consists of the analysis that may customize physical, technical and personnel structures of a call center towards the objectives of the customer service operation that begins with the forecast of the demand within the days.

The 102 product was chosen to illustrate the dimensioning problem, since its demand is the most foreseeable and, therefore, being possible to measure the quality of a dimensioning process independently i.e., departing from the premise that the input – demand forecast – presents a good quality.

The service level for this product is related to the waiting time of the final client at the telephone line, from the moment the incoming call arrives to when it is answered. In other words, it is the time which the client remains waiting in line, listening to the background song and waiting for the operator. More precisely, the level of the service consists of the percentage of calls – amongst the completed ones, only – that wait no more than 10 seconds to be answered.

As only the calls answered count in the computation of the service level, the disconnections are not accounted (and, therefore, not punished), for effects of the service level. Nevertheless, they are measured through another indicator (abandonment rate) and

Contax pays fines when this rate exceeds 2% in a month. As this may happen, to avoid the disconnect is seen as a priority, to the detriment of the service level, as long as this is maintained above a minimum value. The service level does not involve formal requirements of the contract (as the abandonment rate), but does influence the commercial relationship and dignity; i.e., it is interesting to not prioritize only the abandonment and, as a consequence, not worry with the maintenance of the service level in decent values.

The dimensioning routine – isolated for each product (basic and plus – due to the priority of the last over the first) – begins by computation of the daily needs of operators, departing from the forecast amount of calls, the average handling time (AHT) and the average time that operators are busy per day. After that, the need of operators (converted to the 6-hours-operators pattern) is compared to todays' resources availability, discounting the losses concerning vacation and absence. The result of this comparison is the balance or the deficit of the labor for each day of the planned month. The output of this first step is the amount of operators that need to be hired or dismissed in the referred month so that the required numbers can be achieved.

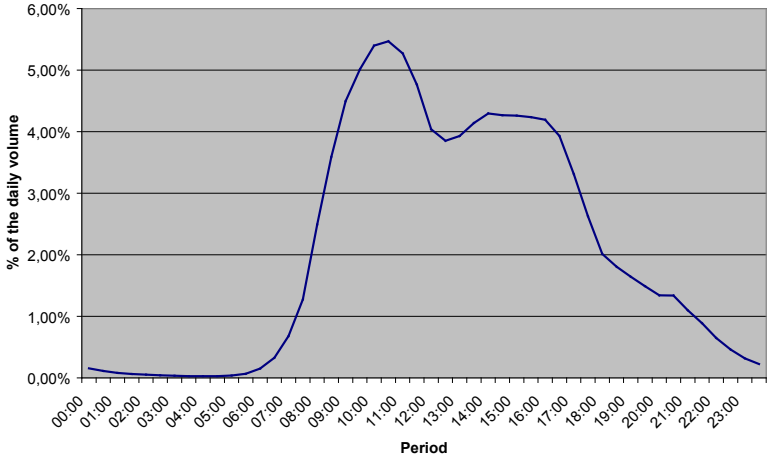
From the moment the contract decision is taken, or the dismissal is decided and implemented, the planning team can look forward to a more detailed analysis – daily dimensioning. This must be done for one day only, and this pattern format must be repeated to the other days of the period, since the scheduled hours of each employee must be the same on everyday of the month.

In conclusion, a volume of calls and an average handling time (necessary numbers for the dimensioning) shall be chosen to be used as a pattern for the dimensioning of all days of the month. The chosen day for the pattern is, usually, the fifth day of higher movement. This way, the dimensioning will guarantee the desired service level for this day and all the not-so-busy days, but not for the four days of higher demand, when there will be a loss in the service level. Nevertheless, this does not represent a problem, because the agreement related to the 102 product involves a monthly level service and not a daily level.

Over the day chosen as a pattern for the dimensioning of the month is applied a curve that shall reflect the daily demanding profile, i.e., which daily volume percentage will happen in the first half hour of the day, in the second half hour of the day, …, and in the last half hour of the day. This curve is shown based on the calls report received at each period of half hour for each day of the week. Concerning the 102 product, the curves of each day of the week are very similar (mainly from Monday to Wednesday, with a little increase of the volume in the afternoon of Thursdays and Fridays), and on Saturdays and Sundays they happen to be a little different. Figure 1, that follows, illustrates the historic curve for Tuesday.

The result of this process is a forecast call demand (volume and AHT for each half hour). With the help of the Excel Supplement Erlang[1] formulas, called Turbotab, it is computed the necessary amount of operators that will be handling the demand of each period with a minimum pre-established service level (normally 85% of the calls being answered before 10 seconds).

Figure 1 – Historic profile for the demand intraday behavior, Tuesday

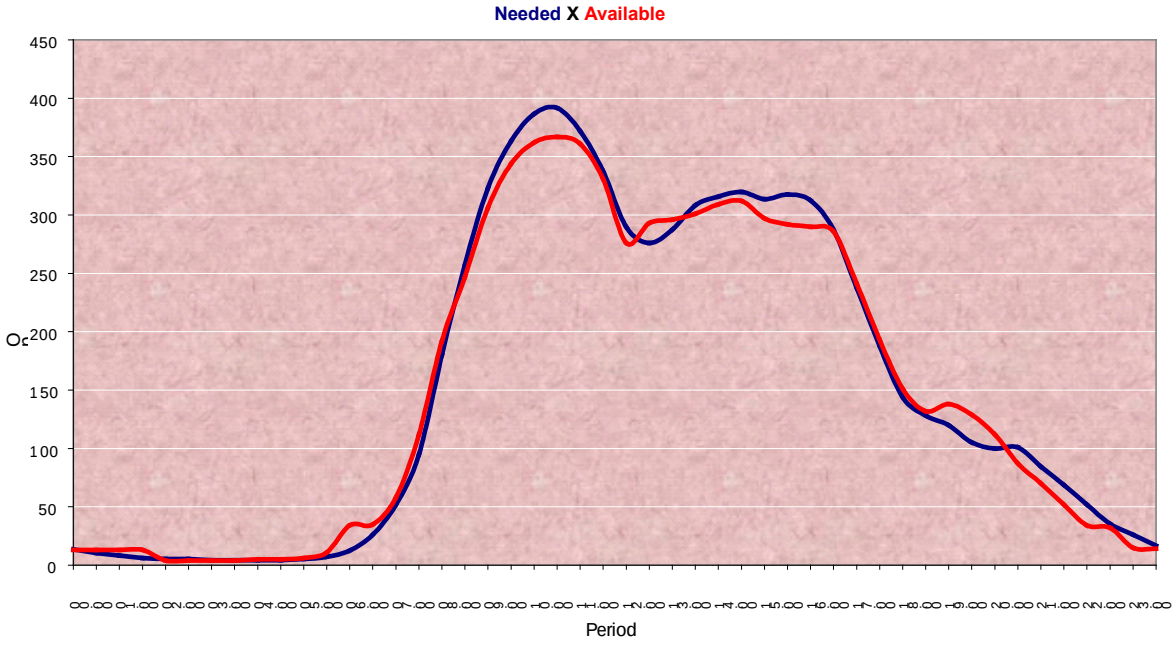


Source: Contax

The last month contingent of operators is then considered. Due to the amount of operators that are initiating their work at each day period and the daily work load of each one of them (4 or 6 hours), a sheet computes how many operators will be available for each period of half hour. This information is then compared to the operators need for each period of 30 minutes, previously calculated. Such comparison is summarized in a graphic way as exemplified in the following figure 2.

Over the actual operators scale, the planning team will work on the changing of the operators' availability for each period of the day, in order to achieve the desired service level. The objective is to assure a certain amount of people in each scheduled hour, throughout a trial-and-error process, during which it will be necessary to analyze several factors, such as daily working hours load, working laws aspects, union agreements and available physical space. In the case of the 102 product, the balanced scale (alternating times with the operational contingent over or under the requirements) can be used, since what really matters for commercial purposes is the daily average level service.

Figure 2 – Operators need and availability by period, Aug/06



Source: Contax

During the staffing process, the planning team makes experiments by modifying the quantity of operators that begin to work at each period of time. These changes consequently alter the quantity of operators available in each period of half hour. The sheet containing the Erlang formulas uses this information then to estimate the service level for each period of half hour and for the day, which depends also on the forecast demand.

At this interactive process, the principal motivation of the analyst is to maximize the day's average service level. The level of the service in each hour band, itself, does not present a great concern to the analyst who, nevertheless, tries to avoid great deficits of operators assigned in relation to the demanded within the hour bands of the day.

The concern about daily deficits does exist because, in hours with a higher deficiency of operators it is possible to register a great incidence of abandonments. And this could be very bad for two reasons: fines for excess of call abandonments and the possibility that the client left without an answer returns the call later on and waits until getting an answer, therefore deteriorating the service level.

This dimensioning effort main goal is to provide a better adjustment between the demanded and offered capacity during the day. An example of the

changes made in order to achieve a better dimensioning can be visualized on figure 3 that follows.

On the last part of the dimensioning and staffing processes, the analyst tries to estimate how the operation level service will be (percentage of calls answered in less than 10 seconds), on all days of the month (until here the computation was based on the fifth day of larger movement, only). The intraday distribution of operators elaborated during the past steps is repeated on all days of the month and, along with the daily call demanding forecast as well as with the demand intraday behavior profile, is able, therefore, to estimate – through the Erlang Methodology – the service levels to be obtained for each day and hours, within the month in question.

Figure 3 – Changes made during the resources staffing, Aug/062

Period	06:00	06:30	07:00	07:30	08:00	08:30	09:00	09:30	10:00	10:30
02:30										
03:00										
03:30										
04:00										
04:30										
05:00										
05:30										
06:00	20									
06:30	20	1								
07:00	20	1	21							
07:30	20	1	21	25						
08:00	15	1	21	25	72					
08:30	17	0	15	25	72	40				
09:00	18	1	20	21	72	40	40			
09:30	20	1	21	24	72	40	40	26		
10:00	20	1	21	24	72	40	40	26	18	
10:30	20	1	21	25	72	40	40	26	18	9

Source: Contax

For the planning team, the Erlang formula used to compute the service level is not totally precise, but it is not radically inaccurate to the point of generating situations where the actual service level is far from the one computed. In fact, a few internal questions came out concerning this formula, driven by a few empiric observations, such as in the situation where the service level computed for a time band presenting a deficit of 3 agents was 77% and dropped to 0% when the deficit went up to 12 agents. Therefore, there is not a common agreement about this methodology being the ideal tool to calculate the service level, but the team did not find any other more accurate approach during researches that collected information and analyzed worksheets.

3.3 Suggested methodology for the dimensioning of the handling capacity

For the real world of call centers, the Queue Theory is the best analytical methodology to be used, but there are experimental methods – as Simulation, for instance – that should be even more adequate for an industry with an operational day to day as complex as modern call centers, as suggested by section 2.2 of this paper. For example, the methodology currently used for the service level computation does not take into account the hypothesis of clients abandoning the call or receiving the busy line signal; i.e., the clients that arrive and do not find available agents, await indefinitely in a queue until being attended. This fact creates a tendency to underestimate the service level because, in the real operation, some clients will disconnect their calls (if waiting too long), shortening the queue and making the other calls be handled before the expected moment. In addition, the worksheets used to compute the service level consider the handling time following an exponential distribution. Nevertheless, this rarely occurs effectively, characterizing an unreal premise.

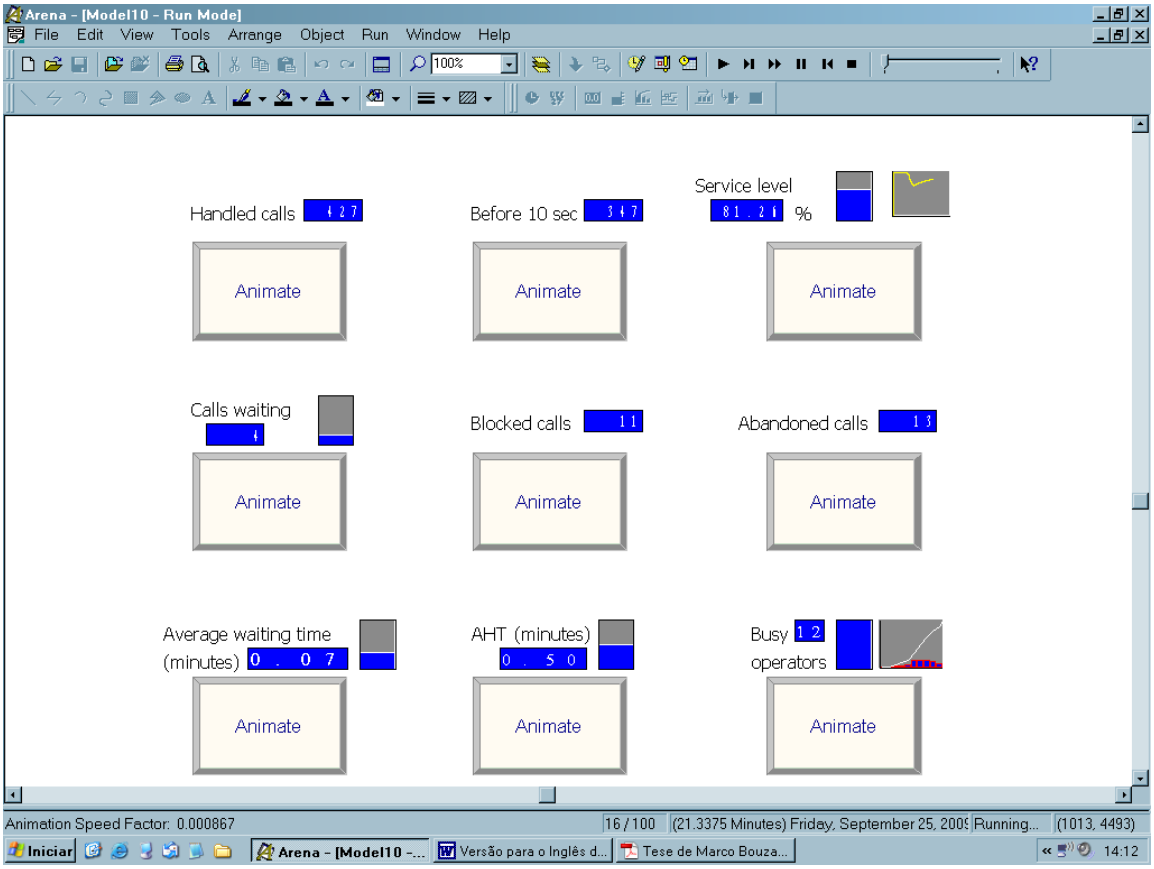
The employment of the Simulation allows us to contemplate the highlighted characteristics of section 2.2, including the abandonment behavior (it is pos-

sible to consider that a percentage of clients that disconnected their calls, will return and try a new contact in a given amount of time, which can be modeled by a statistical distribution) and a flexibility on the definition of the handling time distribution.

In order to verify this better adequateness, the idea consists in simulate by computer and in a few seconds, the call center's operation during periods of 30 minutes, contemplating more realistically its characteristics to obtain more accurate results than the ones generated by the analytical methodologies.

This way, it is possible to visualize the operation itself (with the calls arriving, being sent to the queues and handled afterwards) and what would be going on, in detailed forms (practically as being in loco), in order to understand why a certain period of the day presented a service level so low, for instance (instead of only accepting the number supplied by the analytical formulas). The following figure 4 illustrates the dynamic presentation of the performance indicators that a Simulation software is capable to provide.

Figure 4 – Dynamic indicators of a call center simulation model



Source: Screen captured by the software during simulation of the model

As it may be seen, several indicators can be followed up – while varying dynamically as the simulation is being executed: the amount of clients handled, answered before 10 seconds, the consequent service level (its value at the moment, in numerical and graphical forms – the level bar, besides its behavior along the simulation), the number of clients waiting in the queue (in numerical and graphical forms), of blocked clients as well as of those who disconnected, the average waiting and handling times (in numerical and graphical forms), and the amount of busy agents (the current value, in numerical and graphical forms, besides an histogram showing its behavior along the simulation). At the end of the simulation, the reports summarize the consolidated results.

The use of Simulation renders it possible to compute empirically (instead of estimate analytically)

– through the use (as inputs) of certain historical premises (about clients and system behavior and demand forecast) and the amount of agents assigned – some important performance indicators, such as service level, abandonment rate, average waiting time, agents busy and idle times.

For the dimensioning and staffing of the operators to handle the plus clients of the 102 product, in August of 2006, it was used the premise (originated on the demand forecast) that 586 calls would come to the phone desk with an AHT of 29 seconds in the first half hour of the day (from 00:00 a.m. to 00:30 a.m.). It would be necessary, according to the analytical formulas, to allocate 12 agents for this part of the day, in order to achieve a service level of 85%. The staffing team requested then 12 operators and the analytical formulas forecast 88.04% for the service level during this period.

With the purpose of questioning these numbers, it was built a model in the software Arena Contact Center to simulate how the system would behave in this period, with the same demand premises (volume and AHT) and with the same operational capacity (12 agents).

As the calls come to the phone desk without any kind of control, this process can be considered a random one, the conceptual basis suggesting therefore that the call arrivals rate might be shaped through a Poisson process. The conceived simulation model implemented this process with a mean of, approximately, 0.33 calls arriving per second (or 586 in a 30 minute interval).

In relation to the handling time, the Erlang distribution uses to better shape this process and, therefore, it was used with a mean of 29 seconds. It requires nevertheless an additional parameter (k) related to the variance of the data around the mean. The standard deviation of the distribution is equal to its mean divided by the square root of k. To be able to consider a moderate variance of the data around the mean, the model takes the Erlang distribution with k = 4, resulting on a variation coefficient of 50%.

In order to allow a correct interpretation of the clients' abandonment behavior, it was necessary to perform a research close to the Contax basis, which includes the disconnected calls of the 102 product. The research showed that the waiting time of the calls disconnected historically present a mean of about 2.5 minutes, following a distribution not too far from an exponential one. It was also necessary to model the return behavior of the disconnected calls. To make it possible, it was used the premise that 80% of the disconnected calls are recalled between 1 and 9 minutes after the disconnectment happens (uniform distribution).

3.4 Results analysis

The simulation of the call center operation during 30 minutes was replicated 100 times in the software in a period of 142 seconds, and the first results indicate that, in average, 595 calls were generated in each replication. This number is a little higher than that demand premise of 586 calls, due to the fact that, in the simulation, a few of the disconnected calls were replicated and put on the queue again. From the generated calls, 579 calls in average were effectively handled by the operators in each replication.

From these calls, 541 were handled before 10 seconds, resulting in a service level of 93.31%. This value is fairly higher than 88.4% (the service level forecast by the analytical approach). More, this fact sustains, at a first sight, the expectation of underestimating this variable by the Queue Theory.

After consulting Contax' database, it was possible to recover the information revealing that, on August 22, 2006, 12 agents were operating from 00:00 a.m. to 00:30 a.m.. Within this period of time, 592 calls were handled in 29.4 seconds each, in average. These numbers exceed just a little the demand premises (volume = 586; AHT = 29 seconds) used for the dimensioning process via analytical formulas, as well as in the simulation model presented here. This is the reason why this day and time band were chosen to serve as a comparison base between the results obtained by both approaches.

During these 30 minutes, 549 of the 592 calls were answered before 10 seconds. In other words, the real service level in this interval was of 92.74%; i.e., much closer to the value empirically computed by the simulation (93.31%) than the value analytically estimated by Erlang formulas (88.04%). According to what was said, the underestimation of the service level promoted by the Queue Theory is mainly due to the non-contemplation of the call abandonment carried out – effectively – by some clients, but not considered – theoretically – by the analytical models.

From the 595 calls generated in each replication, 14.5 (in average) were disconnected by the clients, generating an abandonment rate equal to 2.44%. Amongst

the disconnected 14.5 calls, 11.5 (79.41%) returned to the queue a few minutes after the disconnectment.

As may be observed, the Simulation allows several other performance indicators related to calls – besides the service level – to be computed and analyzed. For example, the AHT was 29,35 seconds, but a client came to be handled in 4,70 seconds, and another one in 147,60 seconds! In relation to the waiting time, the calls held for, in average, 1,94 seconds before been answered. Some clients were answered immediately and at least one awaited 38 seconds.

According to the section 2.2 alert, this type of analysis involving the variability and the maximum and minimum values of the performance indicators, cannot be made by means of analytical methods (capable of presenting only average values), becoming feasible only via experimental approaches.

Simulation can also generate indicators related to agents. In the highlighted period, the occupation rate for them was 78.75%, in average. This indicator can guide the management towards a staff increase or decrease, according to the previously determined occupation goals.

It is interesting to also notice the experimental approach accuracy during other time periods, with different service level platforms.

For the operators staffing within the period between 02:00 a.m. to 02:30 a.m. (also in August, 2006), it was used the premise that 196 calls would reach the call center, with an AHT equal to 28 seconds. It would be necessary, according to the analytical formulas, to allocate 5 agents for the period, in order to reach the required service level (85%). Nevertheless, only 4 operators were assigned for that period and this agents deficit (-1) made the analytical formulas forecast 44.05% for the service level.

Again, a model was built, this time to simulate the system behavior at this second time period, with the same demand premises (volume and AHT) and with the same capacity dimensioned (4 agents). The calls arrival and the handling time were modeled in accordance to these average premises and following the exponential and Erlang (k = 4) distributions, respectively; and the abandonment behavior was modeled the same way as before.

In average, 209 calls were generated in each replication; from these, 193 (in average) were actually handled by the operators, 144 from which before 10 seconds. It was obtained a service level equal to 74.44%, a value lower than the goal (85%), but extremely higher than the 44.05% forecast by the analytical formulas that, as a matter of fact, seem to have underestimated this variable.

In accordance to Contax' database, in August 29, 2006, 4 operators handled 192 calls between 02:00 a.m. and 02:30 a.m., with an AHT equal to 27.6 seconds. These numbers are very closer to the demand premises used for the dimensioning process via Erlang formulas, as well as in the simulation model presented here. Within this interval, 136 calls were answered before 10 seconds, generating an actual service level equal to 70.83%.

As it happened in the original scenario (from 00:00 a.m. to 00:30 a.m.), this value was closer (this time, even closer) to the service level computed by the simulation. Thus, it seems to happen an even higher accuracy gain for the service level forecast in scenarios with low values for this variable. One shall now evaluate the behavior of this accuracy when the service level is very high (more dimensioned agents than the amount needed).

In the period between 05:30 a.m. and 06:00 a.m., during August of 2006, it was considered that the call center would receive 253 calls, with an AHT equal to 31 seconds. According to the analytical formulas, it would be necessary to allocate 7 attendants for this period, in order to achieve the service level goal. Eleven agents were assigned (balance of 4) for this time band and the Erlang formulas had forecast 99.78% for this period service level.

With the objective of questioning this estimated service level, a model was built considering the same demand premises (volume and AHT) and the same dimensioned capacity (11 operators), this time for this third period.

The calls arrival and the handling time were modeled in accordance to these average premises and following the exponential and Erlang (k = 4) distributions, respectively; and the abandonment behavior was modeled the same way as before

The simulation of the operation during this period of time was replicated 100 times. In average, 253.31 calls were generated, 253.27 answered and 253.22 before 10 seconds, in each replication. Abandonments almost did not happen, and this is not surprising due to the over dimensioning table done for this scenario. The resulting service level was equal to 99.98%,

an extremely high value (even when compared to the goal of 85%), and very similar to the one forecast by Erlang formulas for this scenario (99.78%).

Contax' database reveals that in August 15, 2006, 11 agents were handling calls during the period between 05:30 a.m. and 06:00 a.m., when 249 calls were handling with an average time equal to 31.5 seconds. These numbers are very close to de demand premises used for the analytical and simulation models.

Within this period, 248 calls were answered before 10 seconds, generating an actual service level equal to 99.60%. This time, the values forecast by both approaches (analytical and experimental) were very similar to the real value. It seems there is no accuracy gain for the service level forecast – in scenarios with very high values for this variable – achieved by the usage of the experimental approach.

At last, it is possible to present a more complete comparison. The accuracy gain behavior – achieved by the usage of the Simulation approach – in relation to the service level platform can be summarized in the following Table 1.

Table 1 – Service level – actual and estimated by Erlang formulas and Simulation, estimation errors and accuracy gain for different service level platforms

Period	02:00 - 02:30	00:00 - 00:30	05:30 - 06:00
Operators balance	1 less (out of 5)	1 less (out of 13)	4 more (out of 7)
	-20%	-8%	+57%
Actual service level	70,83%	92,74%	99,60%
Erlang formula	44,05%	88,04%	99,78%
Error	*26,78%*	*4,70%*	*0,18%*
Simulation	74,44%	93,31%	99,98%
Error	*3,61%*	*0,57%*	*0,38%*
Accuracy gain	**23,17%**	**4,13%**	**-0,20%**

Source: Table elaborated by the author

4. CONCLUSIONS

It was possible to verify, along this research and via the use of simulation models to deal with the handling capacity dimensioning problem faced by the studied company's call center, some advantages of the experimental approach when compared to the analytical ones, mainly in more complex operations: (i) it is possible to include more details of the operation, to use statistical distributions more compatible with the input data and to have the model closer to reality, assuring the collection of more accurate results; (ii) the service level computed by Erlang formulas is usually underestimated, mainly because these formulas ignore the calls abandonment; (iii) other performance indicators (not available while using analytical approaches, as the abandonment rate) can be evaluated, presented and analyzed; (iv) minimum and maximum values of each important indicator can be obtained, the analyst not being restricted to the average values as when using the Queue Theory; (v) a better understanding of the operation is achieved with the adoption of the experimental approach, which provides the possibility to dynamically follow up the system behavior and its performance indicators behavior and, therefore, understand why the queues are being formed and the reason why the waiting time is high, for example, while throughout the Erlang methodology it is possible to see only the generated outputs (numeric indicators) in relation to the provided inputs, making more difficult the complete comprehension of the operation; (vi) the communication can become easier via the use of graphic animations.

This paper also allowed concluding that the accuracy gain for the service level, promoted by Simulation, tends to be higher when this variable shows lower values, according to Table 1, already shown.

4.1 Suggestions and recommendations

Much has been done since the complexity inherent to the management of modern call centers was real-

ized. Obviously, there is much more to be answered, and in a more accurate way, such as questions related to call centers optimization, throughout more discussions and new studies, and the improving of a fundamental aspect for the modeling systems: the proximity to the real world.

An interesting research object would be to map the clients' post-abandonment behavior (the amount which tries to call again and the distribution format for the time period elapsed before the second attempt) and to include it in the model. Additionally, it should also be taken into account the fact that after disconnecting the initial call, a client trying a new contact could be more impatient and decide to stay less time in the queue now, before disconnecting again.

The decret number 6523 (july, 2008) states several issues that must be addressed by brazilian call centers. Some of them are capable to impact this and future studies about call centers. For instance, the initial access to the attendant will no more be subject to the prior provision of data by the consumer; these data may have to be provided during the service, increasing the handling time and impacting the performance measures.

In addition, specific regulations will address the maximum time required for direct contact with operators; this new need will you force brazilian call centers to improve their performance not only due to commercial reasons but henceforth driven by legal issues. These and other kinds of impact may be explored by future and similar works, maybe studying call centers fully adapted to this decree.

Finally, future Simulation works to be applied to the call centers industry could explore other peculiarities present on this kind of operation, which are not used to be well approached via analytical methodologies, such as, for instance: (i) the call transference process during a client attending operation before being handled by the correct agent; (ii) conferences amongst the client and more than one operator at the same time; (iii) conditional call detours towards specialized services; and (iv) other queue disciplines than the traditional FIFO.

Since most of the Simulation softwares address such operational features, it would not be complicated – for a researcher interested in these suggestions and willing to deepen about the used tool functional details – to model these peculiarities and reach interesting conclusions for the industry being focused here.

5. REFERENCES

Alam, M. (2002), "Using Call Centers to Deliver Public Services", House of Commons Paper, London: The Stationery Office Books.

Araujo, M., Araujo, F. and Adissi, P. (2004), "Modelo para segmentação da demanda de um call center em múltiplas prioridades: estudo da implantação em um call center de telecomunicações", Revista Produção On Line, Vol. 4, N. 3, p. 1-20.

Avramidis, A. and L'ecuyer, P. (2005), "Modeling and Simulation of Call Centers", Winter Simulation Conference, p. 144-152.

Bapat, V. and Pruitte Jr, E. (1998), "Using simulation in call centers", Winter Simulation Conference, p. 1395-1399.

Bouzada, M. (2006), O uso de ferramentas quantitativas em call centers: o caso Contax, Thesis (Ph. D. in Business Administration), Rio de Janeiro: UFRJ/COPPEAD.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyin, S. and Zhao, L. (2002), "Statistical analysis of a telephone call center: a queueing-science perspective" (working paper 03-12), Wharton Financial Institutions Center.

Chassioti, E. and Worthington, D. (2004), "A new model for call centre queue management", Journal of the Operational Research Society, Vol. 55, p. 1352-1357.

Chokshi, R. (1999), "Decision support for call center management using simulation", Winter Simulation Conference, p. 1634-1639.

Contax (2006), Contax Contact Center, <www.contax.net.br>.

Grossman, T., Samuelson, D., Oh, S. and Rohleder, T. (2001), Encyclopedia of Operations Research and Management Science, Boston: Kluwer Academic Publishers, p. 73-76.

Gulati, S. and Malcolm, S. (2001), "Call center scheduling technology evaluation using simulation", Winter Simulation Conference, p. 1841-1846.

Hall, B. and Anton, J. (1998), "Optimizing your call center through simulation", Call Center Solutions Magazine, p. 1-10.

Hertz, D. (1980), "Análise de risco em investimentos de capital", Biblioteca Harvard de Administração de Empresas, Vol. 8, N. 3, p. 1-14.

Hillier, F. and Lieberman, G. (1995), Introduction to Operations Research, New York: McGraw-Hill.

Klungle, R. (1999), "Simulation of a claims call center: a success and a failure", Winter Simulation Conference, p. 1648-1653.

Klungle, R. and Maluchnik, J. (1997), "The role of simulation in call center management", MSUG Conference, p. 1-10.

Mehrotra, V. (1997), "Ringing Up Big Business", OR/MS Today, Vol. 24, N. 4, p.18-24.

Mehrotra, V. and Fama, J. (2003), "Call Center Simulation Modeling: Methods, Challenges and Opportunities", Winter Simulation Conference, p. 135-143.

Mehrotra, V., Profozich, D. and Bapat, V. (1997), "Simulation: the best way to design your call center", Telemarketing & Call Center Solutions, p. 1-5.

Outsourcing (2005), Ranking, <www.callcenter.inf.br>.

Paragon (2005), Simulação de Call Center com Arena Contact Center, <www.paragon.com.br>.

Pidd, M. (1998), Computer Simulation in Management Science, New York: Willey.

Ragsdale, C. (2001), Spreadsheet Modeling and Decision Analysis, Tennessee: South-Western.

Riley, D. (2005), "Simulating a Virtual Customer Service Center", Winter Simulation Conference, p. 56-61.

Saliby, E. (1989), Repensando a Simulação: a Amostragem Descritiva, São Paulo: Atlas.

Steckley, S., Henderson, S. and Mehrotra, V. (2005), "Performance Measures for Service Systems with a Random Arrival Rate", Winter Simulation Conference, p. 566-575.

Tanir, O. and Booth, R. (1999), "Call center simulation in Bell Canada", Winter Simulation Conference, p. 1640-1647.

Weinberg, J., Brown, L. and Stroud, J. (2006), "Bayesian Forecasting of an Inhomogeneous Poisson Process with Applications to Call Center Data (white paper), University of Pennsylvania.

Worthington, D. and Wall, A. (1999), "Using the discrete time modeling approach to evaluate the time-dependent behavior of queueing systems", Journal of the Operational Research Society, Vol. 50, p. 777-788.

Yonamine, J. (2006), O Setor de Call Centers e Métodos Quantitativos: uma Aplicação da Simulação, Dissertation (M. Sc. in Business Administration), Rio de Janeiro: UFRJ/COPPEAD.

Endnotes

1 The Queue Theory is used for the agents dimensioning at Contax. The clients that arrive and do not find available operators await indefinitely in a queue until being handled (there is no abandonment or busy line signal), according to the theory. The waiting times are forecast: (i) using the premise that the interval between arrivals and the handling time follow exponential distributions; and (ii) based on the agents amount, the number of clients in the queue and the average handling time.

2 The available agents amounts highlighted in blue are smaller than those of operators beginning their workday in each period, due to the break time for part of the employees during these periods.

AUTHOR'S BIOGRAPHY

Marco Aurélio Carino Bouzada is a Production Engineer (UFRJ, 1998), M. Sc. in Business Administration (COPPEAD/UFRJ, 2001) and Ph. D. in Business Administration (COPPEAD/UFRJ, 2006). Currently a professor belonging to the permanent staff of the Estacio de Sá University M. Sc. in Business Administration Program, to the Escola Superior de Propaganda e Marketing Graduation in Business Administration Program and to the COPPEAD/UFRJ Finance Specialization Program, with experience on topics like Statistics, Quantitative Methods, Operational Research and Business Games.