Nov 7th, 9:00 AM - 5:00 PM

Performance Indicators Analysis inside a Call Center Using a Simulation Program

Ditila Ekmekçiu
University of Tirana, ditila.ekmekciu@gmail.com

Markela Muça
University of Tirana, markela.moutsa@yahoo.com

Adrian Naço
University of Tirana

Performance Indicators Analysis inside a Call Center Using a Simulation Program

Ditila Ekmekçiu[1], Markela Muça[2], Adrian Naço[3]

[1,2]University of Tirana, Faculty of Natural Sciences, Albania
[3]Faculty of Engineering Mathematics and Physics
ditila.ekmekciu@gmail.com[1], markela.moutsa@yahoo.com[2]

Abstract. This paper deals with and shows the results of different performance indicators analyses made utilizing the help of Simulation and concentrated on dimensioning problems of handling calls capacity in a call center. The goal is to measure the reactivity of the call center's performance to potential changes of critical variables. The literature related to the employment of this kind of instrument in call centers is reviewed, and the method that this problem is treated momentarily is precisely described. The technique used to obtain this paper's goal implicated a simulation model using Arena Contact Center software that worked as a key case at the time where the events analyses could be executed. This article comes to the conclusion that Simulation is a completely suitable instrument to accomplish its purpose since it could be adequate to demonstrate, for the call center taken into consideration principally that: (a) it is feasible to reduce the agent contingent; (b) reasonable variations on the demand design can impact very much the key performance indicators of the Call Centers' project; and (c) it is feasible to increase the service level if a combined handling format is followed.

Keywords: Call center, Simulation, Queueing Models

1. Introduction

Call centers are operational centers created with the purpose of using both Communication and IT resources, in order to make automatic very big volumes of different activities and telephone services, not only of incoming calls, but also those generated by the center as well (outbound activities). The inbound call centers, where we have incoming calls (so are the clients who call the call center), are characterized by a system compounded by many call attendants that receive calls from other persons, usually clients of that service, or even potential clients that wish to get information regarding the specific subject, buy a specific product, look for technical assistance, answer to a certain research, update known data, register incidents or place a complaint, between other requests (GROSSMAN et al., 2001 [1]; HAWKINS et al., 2001 [2]).

As reported by Mehrotra, Profozich and Bapat (1997) [3], managers of call centers have a much more difficult job today than they used to have in the past. With numerous products and services being particularly created, made available and ready to use, sold and supported by technicians out in the market, the call centers have struggled to provide various service levels for different types of clients, which have different needs. At the moment, the telephone systems allow a high flexibility and can decide how the calls can be routed and put on line. However, at the same time, this makes the planning and analyses even more sophisticated because of the fact that they make possible a link among multiple call centers, prioritization of certain specific calls, existence of various abilities between agents and customization of calls routing.

Today managers must be able to know completely what is happening at call centers so they can know how calls, routes, priorities, agents and their capacities, peak periods and other aspects that affect the service level and the utilization rates (BOUZADA, 2006 [4]).

Labor costs represent almost 70% of the total costs of the industry, justifying the requirement for an efficient management and the great importance of a quantitative access for the dimensioning of a service handling capacity that consists on a trade-off between this cost and the establishing of the

right service level; Saying it differently, in having the right number of qualified people and resources at the right moment, in order to work with the forecast working load, keeping the quality arrangements and the required service level. This way the use of better accurate models on the dimensioning of the size of the staff, of the industry that works with big financial volumes, is considered as more relevant than ever (HALL; ANTON, 1998 [5]).

As studied by Mehrotra and Fama (2003) [6], and Hall and Anton (1998) [5], call centers are interesting objects for simulation studies, because: (a) they are faced with more than one type of call, where every type represents a line; (b) the calls received in each line arrive at random, as time passes; (c) in some cases, agents make calls (particularly in telemarketing), or as a return for a call received; (d) the duration of every call is random, as the after call work of the agent (data collection, documentation etc.); (e) the evolution on the systems that route the calls for the agents make the philosophy behind the call center even more sophisticated; (f) agents can be disciplined to answer only one kind of call, several kind of calls or all kinds of calls with different priorities and/or preferences described for the routing logics; and (g) the high amount of money invested in call centers on capital and work, is able to justify the application of this strong tool.

According to Hall and Anton (1998) [5], call centers can use Simulation to test (and eventually legitimize its implementation), if particular alterations can improve the system before its implementation. The most important call centers use this instrument effectively and efficiently, making possible to project the system, to manage the operation and to plan the future, regardless of possible scenarios.

Based to this fact, this paper describes the dimensioning problem of the handling capacity of a large Albanian call center, looking for an immediate proposal presentation of different scenario analysis, altering some key parameters of the system. Such analyses are made possible by the use of Simulation during work, whose goal is to calculate the sensitivity of the call center's performance to eventual changes of critical variables.

2. Literature Review

In accordance to Anton (2005) [7], the main cost in a typical call center is due to human resources (almost 70% of the total costs), far above those regarding the technology (16%), the second in the expensive classification.

For this reason, and in order to reduce staff requirements, one of the most important duties consists in managing the call centers queues that occur when there is no agent available to handle a client, who waits on a virtual line from which he will leave only when an agent is free to attend him or when he disconnects the call. As showed by Brown et al. (2002) [8], in the case of call centers, the virtual queue is invisible between the clients and between the clients and the agents.

In the call centers scenario, Araujo et al. (2004) [9] say that the queues discipline, when well - managed, is an important partner of the call centers production workforce management area, which have as a goal to achieve the wonted results with finite resources, turning this area very important for these companies. In accession, a significant reduction in the clients waiting time can be achieved.

Some of the call center's characteristics make it difficult to apply analytical formulas of the Queuing Theory for its modeling, including: generic distribution for the handling time, time-varying arrival rates, temporary floods and calls abandonment (CHASSIOTI; WORTHINGTON, 2004 [10]).

Chokshi (1999) [11], Klungle and Maluchnik (1997) [12], Hall and Anton (1998) [5], Mehrotra and Fama (2003) [6], Avramidis and L'Ecuyer (2005) [13], Klungle (1999) [14] and Bapat and Pruitte Jr. (1998) [15] go further a few current factors that contributed in the increase on the demand for the use of the Simulation tool in the call centers sector: (a) the growing importance of the call centers for a big number of corporations, due to the fast increase of information, communication and technological devices, increasing the need of using scientific methods in decision makings and instruments for its strategic management rather than using the intuition, only; (b) the increasing complexity of the traffic of the calls ahead with rules more viewed on the skill-based routing; (c) the uncertainty that is more predominant in the decision problems commonly found in the operational management of call centers phone desks; (d) fast changes in the operations of the company and the improvement in the re-engineering projects resulting from the growth of mixed companies and of acquisitions, business unpredictability, outsourcing options and the utilization of different channels in order to reach the

costumer (telephone, e-mail, chat etc.); and (e) the availability and accessible price of the computers available in an everyday market less complicated, intuitive and easier to be absorbed and used.

The Simulation, according to Mehrotra (1997) [16], certainly shapes the interaction among calls, routes and agents, the random individual incoming calls and the random duration of the handling service, too. Through the use of Simulation, managers and analysts interpret the call centers gross data (call forecast, distribution of the handling times, schedule hours and the agents skills, call route vectors, etc.), in handling information on service levels, clients call abandonment, use of agents, costs and other valuable performance measures of a call center.

In accordance to Chokshi (1999) [11] and Klungle and Maluchnik (1997) [12], the use of Simulation to help management in decisions making in a call center permits the coming benefits: (a) visualize future processes and use it as a communication tool; (b) validate the processes premises before its implementation; (c) analyze the impact of the changes in details; (d) predict the aggregated needs of resources and schedule the working staff; (e) measure the performance KPIs; and (f) estimate the impacts on costs and economies.

The first use of the Simulation in a call center, as mentioned by Hall and Anton (1998) [5], is the evaluation when one can confirm "where the call center is". The basic argument is "how efficient is the operation today?" The purpose of this evaluation is to organize a point of start (and reference) for the change.

Mehrotra, Profozich and Bapat (1997) [3] and Yonamine (2006) [17] discuss about the other utility of Simulation in a call center: the Simulation permits fast and accurate understanding of how the operational performance of the call center would work when confronting specific scenarios (based on modifications caused by several initiatives such as the adoption of a new technology, a new business strategy or the increase of the volumes of work), before any change is truly made, but not intervening on the operation of the call center's phone workstations and not impacting its budget, too. This way, a few questions might be answered among others: (a) which is the impact of a call overflow? (b) Which are the compromising in the act of prioritizing special clients? (c) Will the service improve if agents provide main pieces of information to clients? (d) Which are the potential achievements related to the adoption of a predictor dial?

According to these authors and to Gulati and Malcolm (2001) [18], Bapat and Pruitte Jr. (1998) [15] and PARA¬GON (2005) [19], a simulation model can be used and has been used more commonly than ever to plan a few other critical aspects of the modern call centers of all sizes and types, such as: (a) a particular service level; (b) flexibility on the time distribution between incoming calls and of handling time; (c) consolidation of the central offices; (d) skill-based routing; (e) different types of calls; (f) simultaneous lines; (g) call disconnect patterns; (h) call returns; (i) overflow and filling of capacity; (j) waiting lines prioritization; (k) call transference and teleconferences; (l) agents preferences, proficiency, time learning and schedule. The outputs model can emerge in format of waiting time, call disconnecting average amount, (both with the possibility of distinction on the call types) and level of the agents employment (with possibility of the agent types distinction). And, due to the application of this approach to the real and complicated characteristics of call centers, the Simulation can make its dimensioning and management more trustworthy.

Mehrotra and Fama (2003) [6] and Klungle (1999) [14] imagined future propensities capable to impact the simulation of call centers, such as: (a) the operational complexity, which will increase continuously – more waiting lines, more variation on the agents scale and combination diversity among skills and route rules – forcing the analysts to create more powerful models; (b) emerging of more Simulation software's specialized in call centers, whose importance tends to follow the role that the Simulation will assume in the process of remodeling the central offices that are necessary to the dealing with the new complexities; and (c) a greater understanding by the executive managers that the call centers are main components of the clients' value chain, discharging a wish to understand the main risks of any operational configuration and the resultant improvement of the quality of the collected data and precision of the parameters (such as distribution of time between incoming calls, handling time, waiting time, average of disconnecting etc.), holding more healthy results.

Gulati and Malcom (2001) [18] used Simulation to compare the performance for three different calls programming approaches (heuristic, daily batches optimization and dynamic optimization), displaying opportunities for the improvement of the outbound call center process within the studied bank. The model outputs gave a way to check the system's performance compared to the management objectives and showed that the non-heuristic approaches obtained better results, but not during the whole day.

Miller and Bapat (1999) [20] described how Simulation was used to project the Return Of Interest related to the acquisition and utilization of a new call routing technology for 25 call centers. Requesting US$ 17 million of investments and an operation cost of US$ 8 million per year, it was demanded to check if the technology would cause enough benefits (cost reduction, agents productivity increase and possibility to handle more calls) in order to assure its implementation on a national range.
Lam and Lau (2004) [21] wrote about a restructuring effort of a Hong Kong company which supplies service for computer and office equipment. As long as there were many opportunities accessible to improve the process, Simulation was used to explore the different options and judge the results of the existing call centers restructuring. The simulated results analysis confirmed that the great improve opportunity consisted of the juncture of the current resources in a singular call center.
Saltzman and Mehrotra (2001) [22] showed a study where Simulation was used by a software company which meant to visualize its call center operating before the launching of a new paid support service program. They wanted to verify if the objective– the paying customers waiting less than 1 minute before being handled – would be achieved. The management also wondered which would be the new program impact to the service given to the non-paying customers' regular basis.

3. The Case

3.1 The company

The Call Center taken into consideration is Teleperformance Albana, part of the corporate Teleperformance, a leader in this sector in the world. It is present in 62 countries, with 270 call centers, created 36 years ago, that invoiced in 2014 $3.7 billion. Teleperformance Albania was born in 2008 as part of Teleperformance Italy, operates in Tirana and Durres and has 2000 employees. It was rated the sixth company with the highest number of employees in the country. Being part of the Italian company, it operates in the Italian market with very important clients. The dimensioning process of handling capacity The dimensioning consists in the analysis that may customize physical, technical and staff structures of a call center against the objectives of the customer service operation that begins with the forecast of the demand inside the days. The Lottery campaign was chosen to demonstrate the dimensioning problem, since its demand is the most predictable and, as a result, being possible to measure the quality of a dimensioning process independently i.e., starting from the assumption that the input – demand forecast – introduces a good quality. In addition to that, there are only two types of clients of the Lottery campaign: extra (a paid service) and main (a service free of charge). The service level for this project is related to the waiting time of the final client in line, from the moment the incoming call arrives in the system to when it is answered from the agent. Saying it differently, it is the time which the client remains in line, listening to the background song and waiting for the agent. More specifically, the service level consists in the percentage of calls that wait no more than 10 seconds to be answered. Since only the calls answered count in this computation of the service level, the disconnections are not taken into consideration (and as a consequence not punished), for effects of the service level.
However, they are measured through another indicator that is the abandonment rate and our Call Center pays penalties when this rate surpasses 2% in a month. As this can happen, to avoid the disconnection is considered as a priority, to the damage of the service level, as long as it is kept above a minimum value. The service level does not include legal requirements of the contract (like the abandonment rate), but does affect the commercial relationship; i.e., it is interesting to not give priority only to the abandonment and, as a result, not consider the maintenance of the service level in decent values.
The dimensioning routine – isolated for each product (main and extra − due to the priority of the last over the first) – starts with the computation of the daily needs of the agents, departing from the forecasted calls, the average handling time (AHT) and the average time during which the agents are busy per day. After that, the need of the agents (adapted to the 6-hours-agents pattern) is compared to the resources availability, discounting the losses regarding absenteeism (vacations, sicknesses or not justified absences). The result of this comparison is the balance or the deficit of the work for each day of the projected month. The output of this first step is the amount of agents that have to be hired or discharged in the indicated month so that the required numbers can be obtained.

From the moment the contract decision is taken, or the discharge is decided and implemented, the planning staff can go through a more detailed analysis – the daily dimensioning. This must be done for a day only, and this designed format should be repeated for the other days of the considered period, as long as the scheduled hours of each agent should be the same every single day of the specific month.

Concluding, a volume of calls and an average handling time (necessary numbers for the dimensioning) should be chosen to be used as a diagram for the dimensioning of all days of the month. The chosen day for the diagram is, generally, the fifth day of higher movement. Acting like this, the dimensioning will guarantee the desired service level for this day and all the other days with lower movement, but not for the four days of higher demand, when there will be a loss in the service level. Although, this doesn't introduce a problem, because the agreement related to the Lottery includes a monthly service level and not a daily service level.

For the day chosen as a pattern for the dimensioning of the month we applied a curve that should indicate the daily demanding profile, i.e., what daily volume percentage will happen during the first half hour of the day, during the second half hour of the day, …, and during the last half hour of the day. This kind of curve is shown based on the calls report taken at each period of half hour for each day of the week. Regarding the Lottery, the curves of every day of the week are very much alike (principally from Monday to Wednesday, with a little increase of the volumes during afternoon of Thursdays and Fridays), and during Saturdays and Sundays they are a little different.

The result of this operation is a forecast call demand (volume and AHT for each half hour). Using the concepts of the Queuing Theory and with the support of the Excel Supplement Erlang formulas, called Turbo Tab, it is calculated the necessary amount of agents who will be handling the demand of each period with a minimum pre-settled service level (usually 85% of the calls being answered before 10 seconds).

The last month contingent of agents is then taken into consideration. Because of the amount of agents that are starting their work at each day period and the daily work load of each one of them (4 or 6 hours), a sheet calculates how many agents will be available for each period of half hour. This information is then compared to the agents' requirement for each period of 30 minutes, formerly calculated. Over the actual agents scale, the planning team will work on modifying the agents' availability for each period of the day, in order to obtain the desired service level. The goal is to persuade a specific amount of people in each scheduled hour, during a trial-and-error process, over which it will be necessary to analyze several factors, like daily working hours load, working laws aspects and available workstations. In the case of the Lottery, the balanced scale (varying times with the operational staff over or under the requirements) can be utilized, since what really matters for commercial scopes is the daily average level service. Throughout the staffing process, the planning department makes experiments by changing the quantity of agents that start working at each period of time. These changes therefore modify the quantity of agents available in each half hour period. The sheet containing the Erlang formulas uses then this information to predict the service level for each half hour period and for the whole day that depends on the forecast demand, too.

During this interactive process, the principal motivation of the analyst is to maximize the day's average service level. The service level during each hour band, itself, doesn't present a big worry to the analyst who, however, tries to avoid great deficits of agents assigned in comparison to the demanded within the hour bands of the day. The worry about daily deficits exists because, during the hours with a higher deficiency of agents it is possible to register a great occurrence of abandonments. And of course this can be very bad for two main reasons: penalties for surpassing the call abandonments and the possible fact that the client that didn't get an answer returns the call later on and waits until getting an answer, as a result degenerating the service level. This dimensioning effort main objective is to provide a better adjustment between the demanded and offered capacity during the day. During the last part of the dimensioning and staffing processes, the analyst tries to estimate how the operation service level will be (percentage of calls answered in less than 10 seconds), on all days of the month (until here the calculation was based on the fifth day of higher movement, only). The distribution within the day of the agents elaborated during the past steps is repeated on all days of that month and, ahead with the daily call demanding forecast as well as with the demand within the day behavior profile, is capable, as a result, to estimate − through the Erlang Methodology − the service levels to be achieved for each day and hour, within the specific month.

Methodology applied to perform the analyses (scenario and sensitivity)

For the real world of call centers, the Queuing Theory is the best analytical methodology to be applied, but there are experimental methods – such as Simulation, for example − that should be even more satisfactory for an industry with an operational day to day as complicated as modern call centers, as recommended by section 2 of this paper.

The utilization of the Simulation permits us to consider the displayed characteristics of the same section, including the abandonment behavior (it is possible to consider that a percentage of clients who disconnected their calls, will return and try a new contact within a given quantity of time that can be modeled using a statistical distribution) and a flexibility on the definition of the handling time distribution.The concept consists in simulating by computer and in a little time, the call center's operation work during periods of 30 minutes. Acting like this, it is not necessary to experience in practice some of the dimensioning alternatives so that we can know the consequences because the experimentation is made in a virtual and not physical environment. However, it is possible to see the operation (with the calls arriving, being sent to the queues and then handled) and what would happen, in detailed forms, so that to understand why a specific period of the day presented a service level so low/high, for example (instead of only accepting the number provided by the analytical formulas).

For the dimensioning and staffing of the agents to handle the extra clients of the Lottery, in September 2015 it was utilized the assumption (originated on the demand forecast) that 586 calls would come to the phone workstation with an AHT of 29 seconds in the first half hour of the day (from 00:00 a.m. to 00:30 a.m.). The staffing team requested then 12 agents for this period.

In the software Arena Contact Center it was built a model to simulate how the system would behave in this time, with the same demand assumptions (volume and AHT) and with the same operational capacity (12 agents).As the calls come to the phone workstation without any type of control, this process can be considered as a random one, the conceptual basis suggesting as a result that the call arrivals rate could be shaped through a Poisson process. The imagined simulation model implemented this process with a mean of, approximately, 0.33 calls arriving per second (or 586 in a 30 minute interval).In regard to the handling time, it is used the Erlang distribution to better shape this process, and, as a result, it was considered with a mean of 29 seconds. However, it requires an additional parameter (k) linked 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 capable 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 right interpretation of the clients' abandonment behavior, it was essential to perform a research close to the Teleperformance Albania basis that includes the disconnected calls of the Lottery. The research demonstrated that the waiting time of the calls disconnected historically introduce a mean of about 2.5 minutes, keeping a distribution not very far from an exponential one. It was also fundamental to model the return behavior of the disconnected calls. In order to do this, it was used the premise that 80% of the disconnected calls are recalled between 1 and 9 minutes after the disconnection (a uniform distribution).

The simulation of the call centers' operation during 30 minutes was replicated 100 times in the software in a period of 142 seconds, and the first results show that, in average, 595 calls were generated in each replication. This number is a little bit higher than that demand premise of 586 calls because in the simulation, some of the disconnected calls were replicated and put into the queue again. From the generated calls, 579 calls in average were effectively handled by the agents in each replication, generating an AHT of 29.35 seconds. From these calls, 541 were handled before 10 seconds, giving a service level of 93.31%. From the 595 calls generated in each replication, 14.5 (in average) were disconnected by the clients, producing an abandonment rate equal to 2.44%. Between the disconnected 14.5 calls, 11.5 (79.41%) returned to the queue a few minutes after the disconnection. In average, the agents were busy 78.75% of the time, during this period.

From this key scenario, different scenario and sensitivity analyses were studied for a better comprehension of the system operational behavior dealing with potential changes of its main parameters. This methodological experiment is very much alike to the ones used and described by Miller and Bapat (1999) [20], Gulati and Malcom (2001) [18], Saltzman and Mehrotra (2001) [22], Lam and Lau (2004) [21] and Yonamine (2006) [17], whose works were indicated during section 2 of this paper.

Analysis of scenarios, sensitivity and results

All previous results start from the premise that the handling time follows an Erlang distribution (with a variation coefficient equals to 50%). However, it is possible to have the handling time demonstrating a different variance and that this parameter can cause an impact on the most important results.
The sensitivity analysis regarding the variance of the handling time tries to measure this impact. The same simulation was replicated a few times using the software, every time with the same mean on the handling time (29 seconds), but with different values for k (and, as a result, for the variation coefficient), from where the appropriate outputs were collected, and the principal performance indicators (service level and abandonment rate) could be achieved, which are shown on the following Table 1.
As *k* increases, the variance of the handling times reduces. Due to the uniformity of these times, the system becomes more stable, presenting as the most evident consequence an increase of the service level. But the abandonment rate has not a clear tendency, even though it can look like falling as the variance of the handling time decreases (*k* increases). The variation on these outputs is not very large, but is far from being insignificant, demonstrating a significant potential impact of this parameter on the most important results. However, the right consideration of the handling time variance − and not only of its mean – shows to be extremely necessary in order to achieve accurate results.

Table 5. Service level and abandonment rate for various values for the parameter *k* of the Erlang distribution, from 00:00 a.m. to 00:30 a.m., Sept/15, 12 agents

k	Variation Coefficient	Calls				Service level	Abandonment rate
		created	handled	before 10 sec	abandoned		
1	100%	597	580	528	16,17	91,03%	2,71%
2	71%	602	585	539	16,09	92,12%	2,67%
3	58%	598	584	539	13,27	92,31%	2,22%
4	50%	595	579	541	14,52	93,31%	2,44%
6	41%	599	583	545	15,46	93,44%	2,58%
9	33%	597	583	546	13,92	93,77%	2,33%

Source: Table elaborated from the results achieved by the software.

The worst performance (service level = 91.03% and abandonment rate = 2.71%) happened in a situation in which the variance was the largest possible (*k* = 1; variation coefficient = 100%). This is an Erlang distribution case that corresponds with the exponential distribution, the same format used to model the handling time in the analytical methodology employed by Teleperformance Albania to estimate the indicators.
This evidence awakens a curiosity related to the verification of the results of a simulation that takes into account another kind of distribution − since, according to what was indicated on section 2, the behavior of the handling time in call centers can show different formats − generally used, as well, to model this variable: the lognormal. Adopting the same original simulation model, but changing this variable distribution to fit the highlighted format (with the same mean – 29 seconds, as well as maintaining the same variation coefficient of 50%), 100 replications were run in the Arena Contact Center software.
In average, 597 calls were generated, 583 handled (544 before 10 seconds) and 13.95 abandoned. The resulting service level and abandonment rate were respectively 93.40% and 2.34%. These indicators are very close to those achieved with the Erlang distribution where k = 4 (respectively 93.31% e 2.44%), giving a certain accuracy to these values and suggesting that any of the two formats frequently used to model the handling time can be used without making a distinction.
Simulation also permits the scenario analysis (What-if). At the shown example being studied here, since the employment of 12 agents in the highlighted period generated a service level (93.31%) adequately higher than the minimum objective (85%), what would happen with this indicator after a reduction of 1 agent? Would it be possible to maintain it above the objective?
Using this scenario, an average of 576 calls was handled, from which 491 before 10 seconds. The service level achieved in this scenario with 11 agents was then 85.23%. This value continues to be above the 85% settled for the extra clients. In other words, the 12th agent was not so necessary (in terms of achieving the service level objective), despite his absence lowered the service level in more than 8 percentage points.
But it wouldn't be bad to also know the impact caused by this kind of reduction in the abandonment rate. And, from the 604 calls generated in each replication, 26.2 – in average – were abandoned by

the clients, resulting in rate equal to 4.34%, revealing a high impact on this performance indicator. However, if this indicator was not to be treated as so important, the utilization of Simulation could cause the savings of one agent for this time band that can effectively occur in some scenario in which the abandonment rate can be seen in a lower level.
The agents' reduction impact can also be seen in case of greater deficits of handling availability through a more complete analysis of sensitivity. The same simulation was repeated a few times in the software, always with the same parameters, but changing the number of agents. The important outputs were collected, and from them the principal performance indicators (service level and abandonment rate) could be achieved, which are shown on the following Table 2.

Table 2. Service level and abandonment rate for different amounts of agents, from 00:00 a.m. to 00:30 a.m., Sept/15

Agents	Calls				Service level	Abandonment rate
	created	handled	before 10 sec	abandoned		
9	719	540	160	168,02	29,64%	23,38%
10	639	567	354	67,68	62,46%	10,60%
11	604	576	491	26,21	85,23%	4,34%
12	**595**	**579**	**541**	**14,52**	**93,31%**	**2,44%**

Source: Table elaborated from the results achieved by the software

As it was expected, when the handling availability decreases, the service gets worse, its performance indicators, too, mainly the abandonment rate that quadruplicates after a 2 agents reduction. With this contingent, the service level is much lower than the objective, but still cannot be considered inadmissible, which occurs in the 9 agents scenario, when it becomes 3 times lower than the original value. With this handling availability, the amount of abandoned calls surpasses those ones handled before 10 seconds, making the abandonment rate almost as high as the service level!
That shows a great impact of the reduction of the amount of agents on the system performance (revealing the need of the dimensioning activity to be developed with much care); and suggests that hiring 10 agents for this time band would be the fundamental minimum, characterizing a scenario for which the performance indicators would be bad, but not destructive.
The most delicate decision on how many agents to hire effectively for this time band should take in account the potential costs affected in the inclusion/exclusion of 1 or more agent/s on/from the map of scheduled times. This way, Teleperformance Albania could question if it would be willing to spend one additional monthly labor cost in order to improve the service level and the abandonment rate for the highlighted time band by the amounts indicated in the analysis.
As described in section 3.2, the dimensioning of the operational capacity is made individually for each product – main and extra. However, the Simulation allows – in a very much similar way to what described Saltzman and Mehrotra (2001) [22] – a judgment of what would happen with the operation in a scenario in which different clients − main and extra − could be handled under an accumulated form (by the main and extra agents), but keeping the priority for the extra clients and, proceeding this way, breaking up the queue discipline normally used by analytical models – "First In First Out".
Regarding the sizing and staffing of agents to handle main clients of the Lottery during September 2015, it was used the premise that 399 calls of these clients would come to the phone workstations, with an AHT of 34 seconds on the first half hour of the day (from 00:00 a.m. to 00:30 a.m.), for which was determined a staff of 7 agents.
The idea now reposes on simulating − in order to examine the system behavior – the scenario where these calls would be aggregated to the calls of the extra clients during this same time band (whose premises were described before, in this section), creating a singular queue. The 12 extra agents and the 7 main ones would be capable to handle both kind of calls, but with different abilities and priorities, that characterize the skill-based routing, a mechanism feasible only under empirical approaches, as described before in section 2.
However, main clients do not behave as extra clients do, and their different characteristics must be planned by the model. Their handling time, for example, is, in average, a little bit higher. For model scopes, the same Erlang distribution was used for this time, with a variation coefficient of 50%.
Usually, the main client is more patient before disconnecting the call and this waiting time was modeled as an exponential distribution, too, but with a higher mean of 3.5 minutes (vs. 2.5). It was

also considered that a smaller amount of the clients that disconnect the call (70% instead of 80%) try to recall during a space of time also lower: within 1 and 6 minutes (uniform distribution).
Regarding the priorities, the extra calls are preferential upon main ones and should be handled, as long as it is possible, by extra agents − theoretically with higher skills − if the extra agents would be busy at the moment, then the extra calls would be handled by the main agents. The main calls, instead, would be preferably handled by main agents, theoretically with lower skills, in order to let the best agents free for more important calls.
In the cases where the calls are not handled by their preferential agents, their handling time is alternated. The model takes into consideration the fact that an extra call being handled by a main agent (less capable) lasts 10% more to be handled; on the other hand, if a main call is handled by an extra agent (more capable), this lasts 5% less to be handled.
The simulation was repeated 100 times in the Arena Contact Center software. In average, 413 main calls and 591 extra calls were generated in each replication, respectively 19.8 and 9.5 of them being abandoned. The resulting abandonment rates were 4.80% for the main clients and 1.60% for the extra clients. Between the disconnected main calls, 14.2 (71.46%) returned to the queue a few minutes later, and amongst the extra calls, 7.6 (80.44%) did the same. 393 main calls and 581 extra calls – in average – were effectively handled in each replication. From these, 303 main (77.25%) and 575 extra (98.96%) were handled before 10 seconds (service level).
Comparing the performance indicators of the extra clients with previous values achieved in the scenario with segmented handling (abandonment rate = 2.44% and service level = 93.31%), it is easy to deduce that the aggregated operation became adequately better for these clients. Such like results were expected, since 7 main agents started handling extra calls. It is real that the 19 agents also handled main calls, but only when there was no extra call waiting.
This preference permitted a considerable improvement to extra clients handling and, surely, decreased a little the quality of the service for main clients. But it is interesting to see that, at this scenario, the service level for ignored clients (77.25%) remained still above the objective to be obtained (75%). The abandonment rate (4.80%) became a little high, but it is not considered inadmissible for main clients.
Clear as it looks, the handling aggregated format improved the system performance for extra clients without ignoring very much the quality of the service for the main ones. This occurred because the agents could handle their non-preferential calls while ineffective, permitting the increment of the operation quality as a whole. This type of analyses and conclusions would not be able to be performed/achieved through analytical methodologies, being possible only by means of an experimental approach, like the Simulation.
The AHT for main clients was 32.65 seconds, a little bit lower than the value of 34 seconds, used on the AHT assumption, because a few of them were handled by high-skill agents (extra). For extra clients, the AHT was of 30.37 seconds, a little bit higher than the 29 seconds assumption, due to the fact that a few of them were handled by low-skill agents. The main calls waited, in average, 6.24 seconds before being handled, while the extra calls waited only 1.34 seconds. Such difference is due to the handling preference for the latest calls.
The main agent utilization rate was 89.59%, very close to the extra one (88.84%), mainly because both kind of agents were qualified to handle both types of calls. Both types of agents were most part of the time (51-52%) handling the preferential and more numerous extra clients, spending the extra time answering main clients (37-38%) or ineffective.
The decision on how many agents shall be scheduled and on which assumption they should be segmented or aggregated in each time band, is completely up to the Teleperformance Albania planning management. However, it may be probable that the company is concerned on the analysis of the impact of other variables – that are not under their control (parameters) – on the service level and abandonment rate for the central desk, which would be possible using Simulation, confirmed in the section 2.
The scenario analysis can be utilized to find out what would occur with these performance indicators if − for example − the calls volume during a given time band was 10% higher than the forecast.
Within the simulation of this scenario (but with the same handling staff), 637 calls were handled, in average, from which 540 – in average – before 10 seconds. The service level for this scenario was therefore of 84.68%. This value is hardly lower than the 85% established objective for extra clients and moderately lower than the 93.31%, that would be achieved if the demand had behaved according to the forecast.

From the 672 calls generated in each replication, 32.2 were disconnected, in average, by the clients, indicating an abandonment rate equal to 4.80%, which demonstrates a great impact of the added demand to this performance indicator.
So, a simulation of this scenario demonstrated that the original agents contingent (12) − facing an unexpected 10% demand increase − would be able to practically satisfy the service level objective of 85%, but would also intervene too badly with the abandonment rate.
And what would happen with the service level and the abandonment rate if the demand was underestimated (also in 10%), even though not linked to its volume, but to the AHT?
In this kind of scenario, 575 calls were, in average, handled; from these, 482 (in average) before 10 seconds. The resulting service level was of 83.85%, a value a little bit lower than the goal (85%) and fairly lower than the 93.31% that would have been achieved in case the demand had behaved according to the forecast plan.
From the 606 calls generated in each replication, 29.8 in average were disconnected by the clients. As a result we have an abandonment rate equal to 4.92%, demonstrating a great impact of the AHT on this indicator.
Like the scenario that reproduced an increase on the calls amount, the simulation of this situation revealed that the original contingent of 12 agents − before a suddenly 10% increase on the AHT − would be capable to ensure a service level almost equal to the 85% objective (in this case, a little more far), as well as to increase (a little more) the abandonment rate.
Based on the scenario analysis with a deeper demand than the forecast, it is possible to conclude that a not too large variation (10%) in relation to the forecast values, may impact directly the performance indicators, particularly the abandonment rate. This conclusion needs much care for the calls amount and AHT forecast. Another important discovery is related to the probably unexpected fact that the impact can be even higher when the increase happens with the AHT, when in comparison to a same magnitude difference on the calls amount. This may lead to a certainty on the AHT forecast being even more important than the accuracy on the calls volume forecast.
As a result, it could be interesting to explore the impact of higher variations on the AHT (for more and for less) on the system performance, through a more complete sensitivity analysis. The simulation of the key model was repeated in the software a few times, but with different values for the mean of the handling time (from amounts 30% lower to 30% higher), from where the important results were collected. These results were then organized on Table 3, which follows and also calculates and presents the main performance indicators, i.e., service level and abandonment rate for each scenario

Table 3. Service level and abandonment rate for different values for AHT, from 00:00 a.m. to 00:30 a.m., Sept/15, 12 agents

AHT (sec)	Δ	Calls				Service level	Abandonment rate
		created	handled	before 10 sec	abandoned		
20,5	-30%	583	582	582	0,84	99,92%	0,14%
23,4	-20%	586	583	580	2,68	99,50%	0,46%
26,4	-10%	589	582	571	6,34	98,13%	1,08%
29,3	**-**	**595**	**579**	**541**	**14,52**	**93,31%**	**2,44%**
32,2	+10%	606	575	482	29,82	83,85%	4,92%
35,2	+20%	640	567	356	68,39	62,80%	10,69%
38,1	+30%	683	548	223	128,01	40,77%	18,74%

Source: Table elaborated with results achieved by software

As a result, higher AHT values than the forecasted ones quickly degenerate the system performance in terms of service level and abandonment rate (particularly this last one), demonstrating a huge potential impact of that variable on the most important results. This way, Teleperformance Albania should dedicate its best efforts to prevent the AHT increase, making their agents awake about the destructive consequences of an increase on this value.
Analyzing the superior part of Table 3, it is possible to conclude that performance indicators improve significantly after a reduction of only 10% on the AHT: the service level increases in almost 5 percentage points, surpassing 98%, and the abandonment rate falls to less than its half (1.08%). This advises that it might be worthwhile to invest on agents training, in order to try to reduce a little the handling time. Unluckily, it is also true that higher reductions on this time do not hold advantages so important (particularly for the service level, already found close to its optimum value) for the system.

In other words, the cost involved in decreasing the AHT in 10% may be probably compensated by the benefits resulting from this reduction, but it is difficult to believe that the same would happen with more huge reductions on this variable.
Similarly to this, there should be other questions forwarding the same type of issues: (a) what would happen with the abandonment rate if the client became more unwilling to wait and began to disconnect calls after, for example, 1.5 minutes (instead of 2.5) in average, without being handled? (b) What would be the impact of this alteration on the service level?
At this new proposed and simulated scenario, 580 calls, in average, were handled, from which 551, in average, before 10 seconds. The service level for this kind of scenario was, as a result, of 95.09%. This value is a little bit higher than the 93.31% that would have been achieved with the previous clients' abandonment behavior, as well as adequately higher than the 85% objective. Even growing in 40% "the clients' unwillingness", i.e., decreasing the average waiting time before disconnection from 2.5 minutes to 1.5 minutes, the impact on the service level was low. Perhaps one should suppose a higher increase on this indicator, due to the fact that if there are more clients leaving the queue, it would be more common that the remaining calls could wait less before being handled.
What occurs is that, based to this model, 80% of the disconnected calls return to the queue a few minutes later, overloading the system once again and not permitting the service level to increase that much. This type of analysis and conclusion would be mainly impossible through analytical approaches, which do not consider the abandonment behavior.
If the management is interested only on the service level, it might not become so important to make great struggles in order to forecast with great accuracy the clients' average waiting time before disconnecting the call. This is due to the fact that a not so small change of 40% on this average time is impotent of impacting completely the service level. If, nevertheless, there is an interest on monitoring the abandonment rate, too, it is crucial to analyze the impact of the new scenario on this indicator.
From the 600 calls generated in each replication, 19.3, in average, were disconnected, generating an abandonment rate of 3.21%, adequately higher than the previous 2.44%.
In order to verify in a more complete way the performance indicators sensitivity related to the average waiting time before disconnection, the same simulation was replicated a few times by the software, with different values for this variable. The required outputs were collected to calculate the principal performance indicators (service level and abandonment rate), which are presented on the following Table 4.

Table 4. Service level and abandonment rate for different values for the average waiting time before disconnection, from 00:00 a.m. to 00:30, Sept/15, 12 agents

Average waiting time (minutes)	Calls				Service level	Abandonment rate
	created	handled	before 10 sec	abandoned		
0,5	624	578	562	46,65	97,26%	7,47%
1,5	600	580	551	19,25	95,09%	3,21%
2,5	**595**	**579**	**541**	**14,52**	**93,31%**	**2,44%**
3,5	593	582	540	10,02	92,75%	1,69%
4,5	591	582	534	9,10	91,91%	1,54%

Source: Table elaborated from the results achieved by the software

As one may notice, the abandonment rate is adequately sensitive to the average waiting time before disconnection, particularly at lower levels for this variable. As a result, its right consideration looks to be important on the achievement of accurate results, although the fact that the service level demonstrates a small sensitivity to changes at the same variable.
The planning manager can also be interested on knowing what would happen if the contracting company became more demanding in relation to the service level and came to reconsider its concept, changing it to complement to the percentage amount of clients that waited less than 5 seconds (instead of 10) before being handled.
During the simulation of this scenario, 579 calls in average were handled, from which 489, in average, before 5 seconds. The service level for this scenario was, as a result, 84.44%. This value is adequately lower than the 93.31% achieved with the original definition of service level, and, what is more important, a little bit lower than the 85% objective settled for extra clients.

This comparison shows that a redefinition on the concept of the service level can impact by a not so small way this performance indicator, something logically expected. It is even possible that, just like it occurred with the illustrated example, the actual agents' configuration becomes not sufficient anymore to give a service level according to the pre-established goal. In this case, it may be relevant to find out how many additional agents would be necessary to permit this performance indicator to go back to levels higher than the objective.
In order to figure this out, it is essential to verify if the addition of 1 agent only is good enough to achieve the goal. In the simulation of this scenario with 13 agents 580 calls were handled, in average, 534 of which (in average) before 5 seconds. The resulting service level was 91.99%, a value higher than the objective (85%) and as a result higher than the 84.44% achieved with 12 agents, but a little lower than the original 93.31%.
This demonstrates that the hiring of an additional agent is able of making the service level go back to the expected objective, but its benefits do not compensate the negative impact on this performance indicator created by the redefinition of its concept.
The main results obtained by using these analyses (scenarios and sensitivity ones) are summarized and commented on the following Table 5.

Table 5. Main results achieved by scenarios and sensitivity analyses

Scenario		Impact on performance indicators	
		Service level	Abandonment rate
Handling time variance increases		Small increase	Erratic behavior (no bias)
Handling time distribution: Lognormal instead of Erlang		No changes	No changes
Amount of operators decreases		Huge decrease	Gigantic increase
Handling aggregated format	Plus clientes	Fair increase	Fair decrease
	Basic clients	Above the goal	Acceptable
Calls volume increases		Fair decrease	Big increase
AHT increases		Fair decrease	Big increase
AHT decreases		Fair increase	Big decrease
Clients become more impatient		Small increase	Fair increase
Clients become less impatient		Small decrease	Fair decrease
Service level concept becomes more demanding		Fair decrease	No changes
Service level concept becomes more demanding + 1 more operator		Small decrease	No changes

Source: Table elaborated by the author

Conclusions

During this research, several simulation models were constructed, completing different real call centers features and for different possible alternative scenarios, in order to compute performance indicators and suggest solutions regarding operation sizing. Generally, it was apparent that the Simulation permits an easy judgment of the impact of changes on the original characteristics of the operation on the performance indicators. It also permits one performing several sensitivity analyses related to a few operational parameters.
Particularly, the scenario and sensitivity analyses developed within this paper made us notice how Simulation can give support to decisions concerning the process of dimensioning a call center, since the results primarily demonstrated that: (a) it is possible to reduce the agent contingent in some of the time bands of the day, without much intervention on obtaining the service level objective; (b) not too huge variations on the received call volumes and on the mean and variability of the handling time can impact a big deal on the performance indicators, particularly the abandonment rate, indicating the need to forecast these values with much accuracy; (c) it is possible to improve significantly the handling performance for preferential clients, without much interference on the quality of the service for main clients, if an aggregated handling format, with priorities, is adopted; (d) in case the clients become more impatient and disconnect the calls in a faster way, the impact on the service level would be small, but very important for the abandonment rate.

4.1 Suggestions and Recommendations

In order to structure more correctly the situations to be simulated, it would be interesting if future researches could make an effort on the direction of finding out (based on the calls historical map) the correct statistical distribution and the variability for the time among incoming calls and for the handling time. Several Simulation studies take for granted the key assumption that these times follow an exponential distribution and develop researches related only to these variables means for each time band. But the results achieved can be sensitive to the distribution format and to the variability of these times.

Following this same way of thinking, an empirical research could collect information regarding the impact caused on the handling time when calls are not answered by the type of agent used to do so, in a consolidated handling system for different kind of calls. The correct consideration on this impact tends to generate more accurate results for indicators on call centers with aggregated handling.

Another issue that surely demonstrates a great potential of operational improvement to be analyzed in future researches is related to the multi-product agent (CAUDURO et al., 2002) [23]. There is a feeling about being economically favorable to utilize the same agent to handle two or more different operations at the same time in order to reduce his ineffective time. This could happen in case of approximately similar operations on which the same agent could work and which present complementary demand behaviors along the day, the week, or the month. The obtaining of this supposed economical advantage in terms of cost-benefit could be checked through a well detailed simulation model, as suggested by Bouzada (2006) [4].

References

1. Grossman, T., Samuelson, D., Oh, S. and Rohleder, T. (2001), En¬cyclopedia of Operations Research and Management Science, Bos¬ton: Kluwer Academic Publishers.
2. Hawkins, L., Meier, T., Nainis, W. and James, H. (2001), Planning Guidance Document for US Call Centers, Maryland: Informa¬tion Technology Support Center.
3. Mehrotra, V., Profozich, D. and Bapat, V. (1997), “Simulation: the best way to design your call center”, Telemarketing & Call Cen¬ter Solutions.
4. Bouzada, M. (2006),
5. Hall, B. and Anton, J. (1998), “Optimizing your call center through simulation”, Call Center Solutions Magazine.
6. Mehrotra, V. and Fama, J. (2003), “Call Center Simulation Model¬ing: Methods, Challenges and Opportunities”, Winter Simu¬lation Conference.
7. Anton, J. (2005), “Best-in-Class Call Center Performance: Indus¬try Benchmark Report”, Purdue University.
8. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zel¬tyin, S. and Zhao, L. (2002), “Statistical analysis of a telephone call center: a queueing-science perspective” (working paper 03-12), Wharton Financial Institutions Center.
9. Araujo, M., Araujo, F. and Adissi, P. (2004), “Modelo para seg¬mentação da demanda de um call center em múltiplas prioridades: estudo da implantação em um call center de telecomunicações”, Revista Produção On Line.
10. Chassioti, E. and Worthington, D. (2004), “A new model for call centre queue management”, Journal of the Operational Research Society.
11. Chokshi, R. (1999), “Decision support for call center manage¬ment using simulation”, Winter Simulation Conference.
12. Klungle, R. and Maluchnik, J. (1997), “The role of simulation in call center management”, MSUG Conference.
13. Avramidis, A. and L’ecuyer, P. (2005), “Modeling and Simulation of Call Centers”, Winter Simulation Conference.
14. Klungle, R. (1999), “Simulation of a claims call center: a success and a failure”, Winter Simulation Conference.
15. Bapat, V. and Pruitte Jr, E. (1998), “Using simulation in call cen¬ters”, Winter Simulation Conference, p. 1395-1399.

16. Mehrotra, V. (1997), “Ringing Up Big Business”, OR/MS Today.
17. Yonamine, J. (2006.
18. Gulati, S. and Malcolm, S. (2001), “Call center scheduling tech¬nology evaluation using simulation”, Winter Simulation Con¬ference.
19. Paragon (2005), Simulação de Call Center com Arena Contact Center.
20. Miller, K. and Bapat, V. (1999), “Case study: simulation of the call center environment for comparing competing call rout¬ing technologies for business case ROI projection”, Winter Simulation Conference.
21. Lam, K. and Lau, R. (2004), “A simulation approach to restructur¬ing call centers”, Business Process Management Journal.
22. Saltzman, R. and Mehrotra, V. (2001), “A Call Center Uses Simu¬lation to Drive Strategic Change, Interfaces.
23. Cauduro, F., Gramkow, F., Carvalho, M. and Ruas, R. (2002.