QUEUE THEORY VS. SIMULATION IN CALL CENTERS

Ditila Ekmekçiu[1] Markela Muça[2]

[1]Department of Finance, Teleperformance Albania, AMS shpk, Tirana, Albania
[2] Department of Applied Mathematics, Faculty of Natural Science, University of Tirana, Albania

e-mail: ditila.ekmekciu@gmail.com

Abstract: The world of call centers is a rapidly increasing reality these days and helping the decision making of the operations management is very important for this industry. The objective of this paper is to see how to apply the Queuing Theory and how to use Simulation in order to optimize the Service Levels, which include staffing and scheduling of agents, in call centers. The experiment is done in one of the Albanian call centers, which is an Inbound **-** Oriented one. The literature related to the utilization of these approaches at call centers is reviewed, and also the technique applied at this moment by the company is described. After that the experimental approach is proposed to be implemented as a substitute methodology to handle the issue rather than the analytical method in use. The results will show us how important the experimental approach is and will justify the adequacy of applying such method in complex call centers giving results closer to reality. The study will also show other benefits of the new approach and its main implication points to a better understanding of the operations.

Key words: *Call Center, Queue Theory, Simulation*

1. Introduction

During the last years the growth of the call center’s industry has been very high. In particular in Albania, the number of outsourcing call centers is more than 300 with 20.000 employees (Mapo online, Call-center: Biznesi që vijon lulëzimin) 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. Call centers can use Simulation to test (and eventually legitimize its implementation), if particular alterations can improve the system before its implementation (Hall B. & Anton J., 1998). The most important call centers use this instrument effectively and efficiently, making possible to design the system, to manage the operation and to predict the future, regardless of possible scenarios. This occurs because, between other reasons, the handling capacity dimensioning consists in a crucial process built to reach the efficiency and effectiveness of the operation. According to the Simulation “law”, it is better having an approximate solution for a very reasonable model than a precise solution for a model with many approximations. As studied by Mehrotra and Fama (Mehrotra, V. & Fama, J., 2003) and Hall and Anton (Hall B & Anton J, 1998), 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) the duration of every call is random, as the after call work of the agent (data collection, documentation etc.); (d) 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, etc. Based to this fact, this paper describes the dimensioning problem of the handling capacity of a large Albanian call center, so that it can use a different tool: the Simulation. The goal is to use the case of this company as an experimental scenario for the theoretical discussion on the suitableness of empirical methods (like the Simulation) to complicated operations (like the modern call centers), in disadvantage of analytical methods (like the Queue Theory).

2. Queue Theory applied to Call Centers

In a call center, a queue happens when there is no agent available to handle a call of a client who waits on a virtual line from which he will leave only when an agent is positioned to attend him or when he decides to disconnect the call. As examined by Brown et al. (Brown L. et al., 2002), in the case of call centers, the queue is virtual and therefore invisible among the clients and between the clients and agents. In the call centers scheme, according to Araujo, Araujo and Adissi (Araujo M. et al., 2004), the queuing discipline, when handled as it should, is a powerful accomplice for the call centers controlling area, which has the objective to achieve the wonted results with the minimum resources, giving this area more importance for these companies. The queuing discipline, when handled as it should, can give a significant reduction to the clients waiting time. Some of the call center’s characteristics make the application of the analytical formulas from the Queuing Theory hard for their modeling, which include: common distribution for the handling time, arrival rates that are time-varying, temporary overflows and the abandonment rate. The model presented by Chassioti and Worthington (Chassioti E. et al., 2004) consists in a practical approach able to incorporate most of these characteristics. However, different companies continue supporting the commonly complex decisions related to the resources allocation via the Queuing Theory analytical models directed by the easiness and quickness approach. Many call centers present a generic distribution (like lognormal) for their handling times and not necessarily a negative exponential distribution (Brown L. et al., 2002). The exponential distribution is used in most of the

Queuing Theory literature, for the time between client's arrival, and for the handling time, too. This is because of the fact that there are different analytical solutions for the system stationary state when these times are considered like following an exponential distribution.

But in the real world call centers, at least the incoming client's rate differs as time goes by. The agent's tiredness can also generate a variation on the handling time as the day goes by, but it is not significant when confronted to the arrival rate variation. The solutions found in the literature to handle the time-varying arrival rates are not very appropriate as they include Bessel's functions, that have difficult application (Chassioti E. et al, 2004).

3. Simulation in Call Centers

The first use of the Simulation in a call center, is the evaluation when one can confirm "where the call center is". The essential 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. According to Gulati and Malcolm (2001), Paragon (2005), 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. According to some authors (Mehrotra V. et al., 1997, Steckley S. et al., 2005, Tanir O. & Booth R, 1999), the traditional methods usually used to handle and dimension a call center are becoming highly limited because of the variance of the incoming calls, routes and handling time, to the agents skills and priorities, to the call diversity, to the changing of the call disconnections, to the current tendencies and, generally, to the sophistication and complexity clearly noticed in the call center's systems. The industry's current tendencies require more complex approaches and the Simulation provides the necessary methods to achieve the observations about these new tendencies and helps to design its present and future, which consist in the only analysis methodology capable of modeling a call center efficiently and accurately, within a more practical approach, flexible in terms of inputs and outputs, and able to allow the involvement of important details, of representing better the reality, of making possible a better and deeper understanding regarding the call center processes and of generating much more healthy results regarding the call center performance, permitting its optimization in a more trustworthy way (Paragon, 2005, Riley D., 2005 & Mehrotra V. & Fama J., 2003).

4. The case study

4.1. 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.com).

4.2. The dimensioning process of handling capacity currently in use

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 that starts 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 feasible to quantify the quality of a dimensioning process independently i.e., starting from the assumption that the input – demand forecast – introduces a good quality. 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. 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 calculation of the service level, the disconnections are not taken into consideration, for effects of the service level. While, the disconnections 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.

The dimensioning process – 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 number of calls and an average handling time (essential 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 secure the required service level for this day and all the other days with lower movement, but not for the 4 days of bigger demand, when there will be a decrease 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 model 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. As can be seen by Figure 1 curves of every day of the week are very much alike. Therefore we chose Tuesday for our analysis (Figure 2). Using the concepts of the Queuing Theory and with the support of Turbo Tab (that is an Excel Supplement), it is computed the necessary number of agents who will handle the demand of each period with a minimum pre-settled service level (usually 85% of the calls that are answered by the agents before 10 seconds).

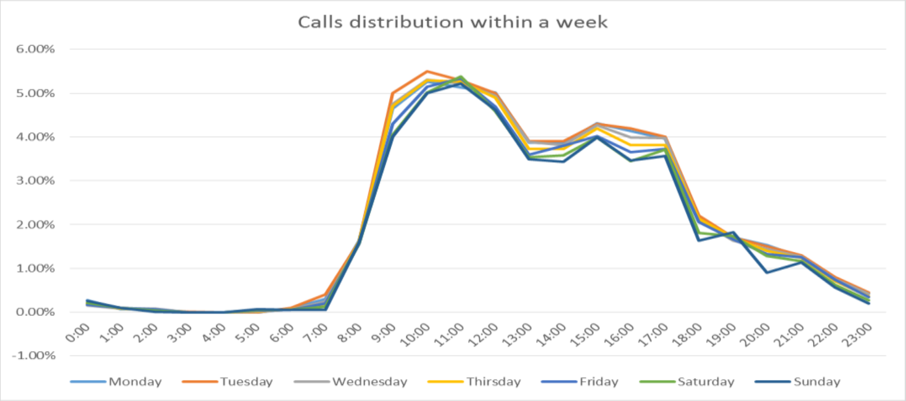


Figure 2. Calls distribution within a week

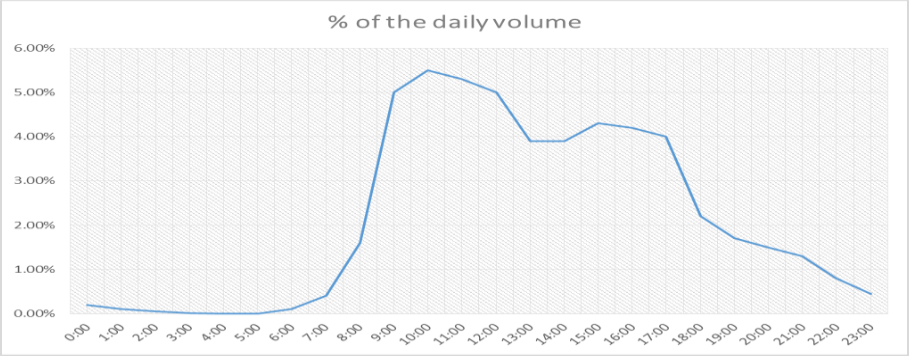


Figure 2. Historic profile for the demand within a day behavior (Tuesday)

The last month contingent of agents is then taken into consideration. Because of the amount of agents that are starting their work at every day period and the daily work cargo of every one of them (4 or 6 hours), we calculate how many agents will be available for every period of 30 minutes. This information is then compared to the agents' requirement for each period of 30 minutes, formerly calculated. Figure 2 is displaying this kind of comparison. 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.

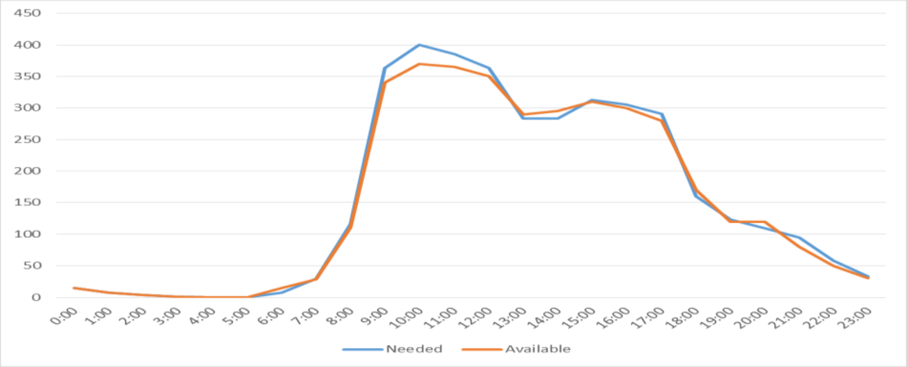


Figure 3. Agents need and availability by period, Sept/15

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. According to the Erlang formulas [1]:

$$W(t) = \text{Service Level} = \text{Prob (waiting time} \leq t) = 1 - E_c(m,u)*e^{-(m-u)\frac{t}{T_S}} \qquad (1)$$

where,

t= target answer time (in our case 10 seconds)

m = number of agents

u = $\lambda * T_S$ = traffic intensity

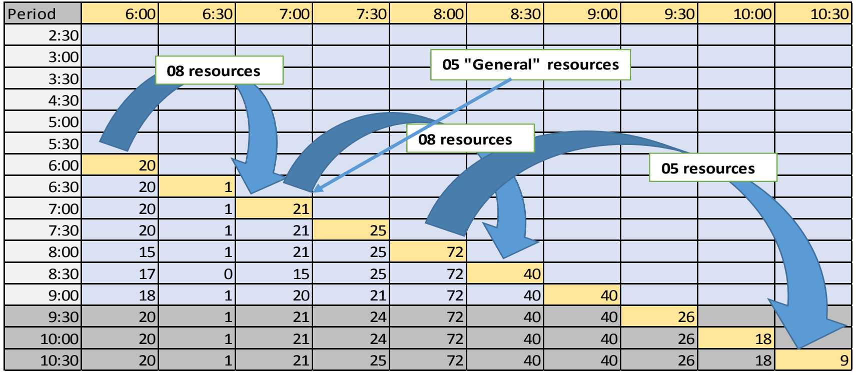
$$E_C\,(m,u) = \frac{\frac{u^m}{m!}}{\frac{u^m}{m!}+(1-\rho)\sum_{k=0}^{m-1}\frac{u^k}{k!}}$$

$\rho = \frac{u}{m}$ = Agents occupancy

It is predicted 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 basic motivation of the analyzer is to maximize the day’s average service level. The service level during eac0068 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 adaptation amongst the requested and offered capacity within the day. A case of the changes made in order to obtain a better dimensioning can be seen on Figure 3. During the last part of the dimensioning and staffing processes, the analyzer attempts to estimate how the operation service level will be, 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, using the Erlang Methodology, the service levels to be achieved for each day and hour, within the specific month.

Table 1. Changes made during the dimensioning process, Sept/15


08 resources
05 "General" resources
08 resources
05 resources

Period	6:00	6:30	7:00	7:30	8:00	8:30	9:00	9:30	10:00	10:30
2:30										
3:00										
3:30										
4:30										
5:00										
5:30										
6:00	20									
6:30	20	1								
7:00	20	1	21							
7:30	20	1	21	25						
8:00	15	1	21	25	72					
8:30	17	0	15	25	72	40				
9:00	18	1	20	21	72	40	40			
9:30	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

For the planning and dimensioning team, the Erlang formula used to calculate the service level is not really precise, but it is not totally inaccurate to the point of creating situations where the actual service level is far from the one calculated. Actually, some internal questions came out regarding this formula, directed by some empiric considerations, like in the situation where the service level calculated for a time band giving a deficit of 3 agents was 77% and precipitated to 0% when the deficit was 12 agents. Consequently, there is not a general agreement about this methodology being the best tool to compute the service level, but the staffing team did not find another approach more accurate during researches.

5. Suggested method for the dimensioning and Analysis of the Results

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 an elasticity on the definition of distribution of the handling time.

The notion 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).

The use of Simulation makes it possible to calculate experimentally (rather than estimate analytically) − within the use of some historical premises and the quantity of agents selected – some relevant performance indicators, like the service level, the abandonment rate, the average waiting time, the number of agents busy and the ineffective times.

For the dimensioning and staffing of the agents to handle the extra clients of the Lottery, in September 2015 it was utilized the assumption that 586 calls would come to the phone workstation (according to the clients' forecast sent before the months' start) with an AHT of 29 seconds in the first 30 minutes of the day (from 00:00 a.m. to 00:30 a.m.), based on the historical data of this campaign. It would be necessary, based on the analytical formulas, to assign 12 agents for that part of the day, so that it is possible to obtain a service level of 85%. The staffing team requested then 12 agents and the analytical formulas forecasted a service level of 88.04% during this period. With the intention of examining these numbers, it was created a model in Arena Contact Center software 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. As it is known, the standard deviation of the distribution is equivalent to its mean divided by the $\sqrt{k}$. To be capable to consider a moderate variance of the data, the model takes the Erlang distribution with k = 4, being equal to 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 attitude of the abandoned calls. In order to do this, it was used the premise that 80% of the abandoned calls are recalled between 1 and 9 minutes after the disconnection (a uniform distribution).

In Tables 2 and 3, three periods were considered to compare the analytical results, the Simulation ones with the real results of the database.

Table 2. Queue Theory Results

Period	00:00-00:30 (less than 10sec)	02:00-02:30 (less than 10sec)	05:30-06:00 (less than 10sec)
Agent	12	4	11
Calls	586 (516)	196 (86.34)	253 (252.44)
AHT	29 sec	28 sec	31 sec
SL	88.04%	44.05%	99.78%

Table 3. Simulation Results

Period	00:00-00:30 (less than 10sec)	02:00-02:30 (less than 10sec)	05:30-06:00 (less than 10sec)
Agent	12	4	11
Calls	579 (541)	193 (144)	253.27 (253.22)
AHT	29.34 sec	27.53 sec	31.41 sec
SL	93.31%	74.44%	99.98%

Table 4 Teleperformance’s database

Period	00:00-00:30 (less than 10sec)	02:00-02:30 (less than 10sec)	05:30-06:00 (less than 10sec)
Agent	12	4	7
Calls	592 (549)	192 (136)	249 (248)
AHT	29.4 sec	27.6 sec	31,5
SL	92.74%	70.83%	99.60%

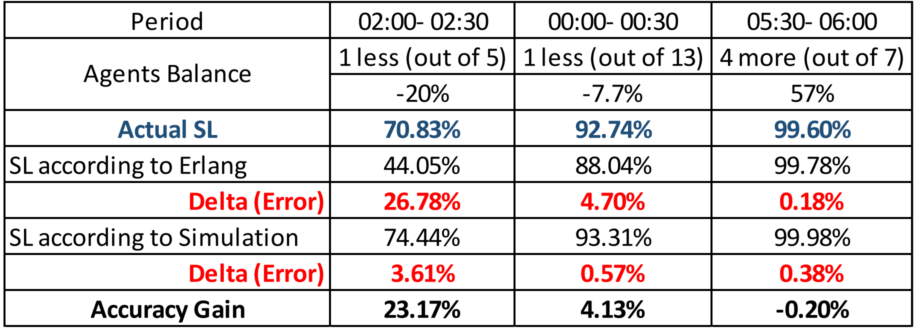
In the first period, the real service level was of 92.74%; as a result much closer to the value experimentally calculated by the simulation (93.31%) than the value analytically estimated by Erlang formulas (88.04%). Based on what was said, the underestimation of the service level sponsored by the Queue Theory is principally due to the non-reflection of the call abandonment. From the 595 calls generated in each replication, 14.5 were abandoned by the clients, producing an abandonment rate equal to 2.44%.

The second and the third scenarios are made, in order to notice the empirical approach accuracy during other time periods, with different service level platforms (very low and very high values). In the second scenario the number of agents is 4 (-1 of the required). The service level generated by the analytical formulas is 44.05%, against 74.44% generated by the Simulation. Consulting the database the real service level for that time period was 70.83%, so again the Simulation gave us a better accurate result.

In the third time period, the number of agents was higher than the sufficient ones (+4 agents), in this case both approaches (analytical and empirical) were very close to the real value. It looks like there is no accuracy benefit for the service level forecasted – in scenarios with very large values for this variable – obtained by the usage of the empirical approach.

Finally, it is possible to introduce a more complete comparison. The accuracy benefit behavior −obtained by the usage of the Simulation approach − in relation to the service level platform can be recapped in Table 5.

Table 5 Actual and estimated Service Level by Erlang formulas and Simulation, estimation deltas and accuracy benefit for different service level platforms



Period	02:00- 02:30	00:00- 00:30	05:30- 06:00
Agents Balance	1 less (out of 5)	1 less (out of 13)	4 more (out of 7)
	-20%	-7.7%	57%
Actual SL	**70.83%**	**92.74%**	**99.60%**
SL according to Erlang	44.05%	88.04%	99.78%
Delta (Error)	**26.78%**	**4.70%**	**0.18%**
SL according to Simulation	74.44%	93.31%	99.98%
Delta (Error)	**3.61%**	**0.57%**	**0.38%**
Accuracy Gain	**23.17%**	**4.13%**	**-0.20%**

6. Conclusions

It was possible to validate, during this research and via the utilization of simulation models to manage the handling capacity dimensioning problem experienced by the studied call center, some advantages of the empirical approach when compared to the analytical ones, principally in more complex operations: (a) it is possible to involve more details of the operation, to utilize statistical distributions that is more appropriate with the input data and to have the model near to reality, persuading the collection of more accurate results; (b) the service level calculated by Erlang formulas is generally underestimated, basically because these formulas don't consider the calls abandonment; (c) other performance indicators can be considered, presented and analyzed; (d) minimum and maximum values of every important indicator can be achieved, the analyst not being limited to the average values as when using the Queuing Theory; (e) a better understanding of the operation is obtained with the adoption of the empirical approach, which gives the possibility to dynamically check out the system attitude and its performance indicators attitude and, this way, comprehend why the queues are created and what causes a high waiting time, for example, while within the Erlang method it is possible to see only the generated outputs regarding the provided inputs, making more difficult the complete understanding of the operation. This paper also permitted the conclusion that the accuracy benefit for the service level, promoted by Simulation, tends to be higher when this variable demonstrates lower values, based on Table 1, once shown.

6.1. Suggestions and recommendations

A lot has been done since the complexity regarding the management of modern call centers was recognized. Surely there is a lot to be answered yet, and in a more precise way, like the questions regarding the call centers optimization, amongst more discussions and new researches, and the improving of a really important aspect for the modeling systems: the adjacency to the real world. An interesting research object could be to design the clients' post-abandonment behavior and to involve it in the model. In addition, it should also be considered the fact that after disconnecting the first call, a client trying to contact again could be more impatient and decide to stay less time in the queue, before disconnecting again. In the end, future Simulation works to be applied to the call centers industry could explore other particularities present on this type of operation that are not utilized to be well approached via analytical methods, like, for example: (a) the call transmission process during a client attending operation before being handled by the right agent; (b) conferences between the client and more than one agent at the same time; (c) conditional call deviates towards specialized services; and (d) other queue disciplines than the traditional FIFO. Since most of the Simulation software address such operational characteristics it would not be difficult − for a researcher interested in these suggestions and that wish to expand their knowledge about the used tool functional details − to model these characteristics and reach interesting conclusions for the industry being focused here.

References

[1] Araujo, M., Araujo, F. and Adissi, P. (2004) "Model for targeting the demand of a call center on multiple priorities: Implementation of the study in one telecommunications call center", Production Online Magazine

[2] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyin, S. and Zhao, L. (2002), "Statistical analysis of a telephone call center: a queuing-science perspective", Wharton Financial Institutions Center.

[3] Chassioti, E. and Worthington, D. (2004), "A new model for call center queue management", Journal of the Operational Research
Society, Vol. 55

[4] Gulati, S. and Malcolm, S. (2001), “Call center scheduling technology evaluation using simulation”, Winter Simulation Conference.

[5] Hall, B. and Anton, J. (1998), “Optimizing your call center through simulation”, Call Center Solutions Magazine

[6] Mapo online, Call-center: Biznesi që vijon lulëzimin.

[7] Mehrotra, V. and Fama, J. (2003), “Call Center Simulation Modeling: Methods, Challenges and Opportunities”, Winter Simulation Conference

[8] Mehrotra, V., Profozich, D. and Bapat, V. (1997), “Simulation: the best way to design your call center”, Telemarketing & Call Center Solutions

[9] Paragon (2005), Simulation of the Call Center with Arena Contact Center, www.paragon.com

[10] Riley, D. (2005), “Simulating a Virtual Customer Service Center”, Winter Simulation Conference

[11] Steckley, S., Henderson, S. and Mehrotra, V. (2005), “Performance Measures for Service Systems with a Random Arrival Rate”, Winter Simulation Conference

[12] Tanir, O. and Booth, R. (1999), “Call center simulation in Bell Canada”, Winter Simulation Conference

[13]www.teleperformance.com