Proceedings of the 2009 Winter Simulation Conference
M. D. Rossetti, R. R. Hill, B. Johansson, A. Dunkin and R. G. Ingalls, eds.

COMPARISON OF CALL CENTER MODELS

Luiz Augusto G. Franzese
Marcelo Moretti Fioroni
PARAGON Consultoria
Rua Clodomiro Amazonas 1435, 5th Floor – ZIP 04537-012
Sao Paulo BRAZIL

Rui Carlos Botter
Escola Politécnica da Universidade de São Paulo
Av. Prof. Mello Moraes, 2231
ZIP 05508-900
Sao Paulo BRAZIL

Paulo José de Freitas Filho
Universidade Federal de Santa Catarina UFSC
PerformanceLab Office 507 - Trindade
88040-900 PO Box 476
Florianopolis BRAZIL

ABSTRACT

Call Centers are important channels of communication within the consumer relationship and a point of integration between suppliers and their customers. Correctly sizing the capacity of a given Call Center can bring benefits not only in terms of improved customer service (efficacy), but also in terms of reduced operating costs (efficiency). However, specifying the capacity of a Call Center is not a trivial task, but one that demands a significant knowledge of mathematics, in particular of analytical models. This paper presents the Erlang B, Erlang C and Simulation models followed by a comparison based on a case study, in order to identify the advantages of using simulation.

1 INTRODUCTION

Telephone Call Centers account for a considerable proportion of the contact channels between suppliers and their customers, and are a channel particularly suited to conducting consumer relations. According to Mehrotra (1997), a Call Center can be defined as, "any group whose principal business is talking on the telephone to customers or prospects."

According to Brizola (2002), a Call Center is a system that offers complete management of all communication channels between a business and its customers, optimizing processes, eliminating duplicated work and making better use of time. The strategic vision of customer care was described by Cusak (1998) as being to satisfy the customer as cheaply as possible while acquiring data for market and business research.

Sizing is the appropriate distribution of resources (telephony, people, hardware, software, etc.) in order to guarantee that customers are dealt with within predefined service quality levels and in line with their expectations.

2 CALL CENTER SIZING

One of the major challenges involved in sizing service capacity is to achieve a balance between efficacy of service (which is very often measured purely in terms of call times and in terms of content or outcome of the contact) and the efficiency of the system (generally measured in terms of the cost or financial results of the operation).

According to Brizola (2002), optimum sizing is directly dependent on making project projections of call volumes with the greatest possible precision. According to Cleveland (1997), capacity sizing can be defined as follows:

Sizing is the art of guaranteeing the necessary resources at the right time in order to deal with the predicted volume while guaranteeing the quality of service level required. Franzese (2002) proposed a modification to this definition in order to include efficiency and efficacy: Sizing is the technique and the art of continuously guaranteeing the necessary resources and trained personnel at the right time in order to deal with the predicted volume while guaranteeing the quality of service level required at the lowest possible cost.

Correct sizing when setting up a Call Center can mean significant savings, bearing in mind that rolling out updated technology and expanding capacity in existing centers generally involve large additional costs when compared with the initial investment. A well-sized Call Center, in addition to reducing operating and start-up costs, offers control over a series of factors such as:

1. Waiting Time: when the number of Service Positions (SP) is appropriate for the call demand, the time waiting in the queue is acceptable to the customer. Telephone costs are also reduced as a result of shorter waiting times (in the case of 0800 numbers or similar).
2. Idle Time or Work Overload of Agents : one of the results of correct sizing is preventing those agents who are available at a given time from spending large periods of time without handling calls, or, the opposite, that is service agents have rates of occupation that are too high, increasing customer waiting times and agent tiredness.
3. Busy Tone : this factor is directly influenced by the number of trunks available, i.e., if the proportion of calls being barred is high, the number of lines is too low and a good proportion of customers are unable to get through, which could result in losing contacts and, possibly, income from these customers;
4. Abandonment: if there are not enough service agents to meet demand then a queue will form and therefore the possibility of customers abandoning the queue is introduced.

3 ANALYTICAL MODELS

The theory of queues (derived from the Latin *cauda*) was initially the idea of Erlang, who published his first article on the subject in 1909 and is considered the founder of Telephone Traffic Theory and of the Theory of Queues. After the Second World War, interest was solidified with formal operational research and since then much work has been published on the subject (Cooper 1997).

Call Centers can be seen as Stochastic Systems and can be treated mathematically as queue models (Mandelbaum 2001). In one queue model, illustrated in Figure 1, the customers are calls received, the servers are the telephone trunks, agents or service positions, and the queues are populated by customers waiting to be served.

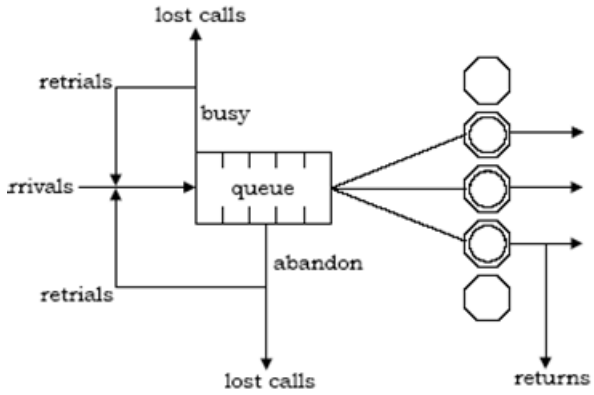

Figure 1: Operational Schematic of a Simple Call Center (Mandelbaum 2001)

In 1953 Kendall introduced the A/B/C notation to define stochastic models, which was extended to 1/2/3/(4/5/6), where the numbers are substituted for the following:

- Codes to describe incoming calls. The following are used:
  - M for "Markovian", implying that call handling times or time between incoming calls have an exponential distribution.
  - D for “deterministic” distribution, for call handling times.
  - Ek for Erlang distribution where the “k” is the factor of the model.
  - G for "General Distribution ".
- Codes to describe the service process. The following are used:
  - M for "Markovian", implying that call handling times or time between incoming calls have an exponential distribution.
  - D for “deterministic” distribution, for call handling times.
  - Ek for Erlang distribution where the “k” is the factor of the model.
  - G for "General Distribution ".

- The number of service resources or channels
- The order of priority with which activities on the queue will be performed:
  - First Come First Served (FCFS) or First In First Out (FIFO)
  - Last Come First Served (LCFS) or Last In First Out (LIFO),
  - Service In Random Order (SIRO)
- Maximum system size. The maximum number of customers allowed into the system, including those being served. When this limit is reached further incoming calls will be ignored.
- The size of the origin of calls. This is the size of the population from which the customers arrive. This is a limit on the rate of arrivals. The more jobs there are on the queue, the less are available to come into the system.

Erlang developed 2 formulae to calculate system congestion (E), E1=E1,n(A) and E2=E2,n(A), where n is the number of servers and A is the traffic demand. These are known as the Erlang B and Erlang C formulae respectively.

3.1 Erlang B Formula or Erlang Loss Formula

The Erlang B formula is a model that is applicable to calculating lost calls due to blockages, i.e., when servers are not available the request for service is refused and the customer must try again. This is the situation where all of the resources (trunks) of a telephone system are occupied and the customer receives the busy tone and therefore hangs up and starts to dial repeatedly until a server (in this case trunk) is available.

Cooper (1997) points out that for there not to be any violations of this system, the formula is only valid when the time between incoming calls follows Poisson distribution, although, surprisingly, it is valid for any type of call handling time distribution.

The Erlang B formula calculates the likelihood that calls will be blocked (the probability of losses) for a given level of traffic and a given number of servers.

3.2 Erlang C Formula or Erlang Delay Formula (M/M/s)

In contrast with the Erlang B model, in the Erlang C model described below, when calls cannot be taken immediately they wait in a queue until a server is available. This model, therefore, defines the probability C(s,a) that all agents will be occupied or the probability that a customer will have to wait in the queue, where s agents have been assigned to serve the traffic of a erlang and where a < s (obligatory to maintain stability. Koole 2004).

The assumptions behind this model are:

- Arrivals: exponentially equivalent Poisson distribution, or with Markovian (M) time between arrivals
- Call handling time: exponential distribution, or Markovian (M)
- There is no abandonment, customers remain waiting in the queue indefinitely
- Service priority is FIFO, following order of arrival
- Expected call handling time is identical for all types of customers and all agents in group s
- No traffic is lost from trunks, i.e., all traffic received is accepted and there are no retrials
- The system is in equilibrium

It is also possible to calculate the probability that waiting time P (q < T) will be less than a given T, i.e. the probability that clients will be served within Waiting Time < T, given s agents, and Ws predicted call handling time, which is termed the Service Level (SL). For example, how many agents are necessary for 90% of customers to be served after a waiting time of 20 seconds or less?

3.3 Simulation of discrete events

According to Anton (1999), Bapat and Pruitte (1999) analytical tools and models are widely used in Call Centers and are primarily focused on calculating the number of trunks and agents. However, some of the assumptions involved mean that certain analyses are limited in the current environment. In general these assumptions are:

- All calls received of the same type.

- Calls on the queue never abandon.
- Agents process calls by order of arrival (FIFO: First In First Out).
- Each agent deals with every call with the same adroitness.

Many more recent analytical models and tools have been developed in order to try to improve certain aspects of Erlang's formulae, whether by the addition of certain random factors or by taking account of abandonment, however so far none of them offer as robust a solution as that obtained by discrete event simulation.

At customer service centers that have more than one type of contact channel and also deal with e-mail, faxes or Internet contact, which are known as Contact Centers, the level of complexity is even higher. Bapat & Pruitte (1999) and also Koole (2004), indicate that models based on the Erlang formula can be used to generate the initial parameters for constructing simulation scenarios.

Among the difficulties involved in constructing Simulation models, Pichitlmken (2003), describing experiments carried out in partnership with Bell Canada and Bell Labs, highlights the difficulty involved in obtaining data, since much data produced at Call Centers is already aggregated, usually over 30 minute periods, and data on individual calls is unavailable. This being so, the variations in the arrival process and call handling times cannot always be modeled by mathematical distributions based on analysis of sample data, but must be selected arbitrarily.

Furthermore, Mehrotra & Fama (2003) state that the assumption of exponential variation in arrival times is not contested by the Call Center industry as a matter of convenience, since equipment and software only records average figures. This has also been pointed out by Koole (2004) and by Anton (1999) who point to problems with traditional assumptions of variation adopted in queue models:

- Calls do not always arrive in Poisson distribution, indeed, many experiments carried out by these authors suggested uniform, lognormal and other distributions.
- Call handling times cannot always be generalized with exponential variation and in some cases may even be bimodal.

Among the countless reasons for using simulation, Krungle (1998) lists certain situations in which it is applicable:

- There are no adequate analytical models or those that exist are highly complex;
- Static results using analytical models have proved inadequate;
- Analytical models cannot identify bottlenecks or recommend modifications to Call Center planning;
- Analytical models provide averages, but not variability or extremes;
- Analytical models do not generally provide sufficient details or identify interactions;
- Animation may be better for demonstrating results to management/decision-makers

Mehrotra & Fama (2003) illustrate the three principal applications for simulation in Call Centers:

- Traditional analysis: models created to analyze operations, using data obtained from a variety of sources.
- Applications embedded in routers: many router / DAC software packages include flow simulation to allow the impact of different routing rules to be estimated;
- Applications incorporated in timetabling software: some shift timetabling software packages include simulation as part of their calculation resources.

Simulation models and related studies must be focused on achieving predefined objectives (for example: an analysis of alternative flow routing set ups), however the intention is not to imitate the real system precisely (Chokshi 1998).

The application of simulation in Call Centers has generally been aimed at strategic and tactical analyses, due not only to the large scope of the models and their ability to correlate many different variables and events, but also because of the complexity of creating simulation models and the prerequisites for doing so.

It is obligatory to mention, among pioneering simulation tools that have been applied to Call Centers, the Call Processing Simulator software developed by AT&T in the United States in 1979 (Brigandi 1994) using the GPSS language and which was later ported out to SLAMSYSTEM/PC. Using this tool AT&T analyzed their customers' call centers in order to identify bottlenecks and demonstrate the viability of solutions and by 1993 they had carried out approximately 2,000 case studies.

Simulation models can be developed in many different languages and graphic modeling environments, but from the mid-1990s some software companies began to create specialized Call Center simulations tools.

3.4 Comparison of different models

There is consensus with relation to the criticisms of the typical assumptions of models (the use of waiting times with exponential variation or Poisson arrivals). This could be avoided in the future were the manufacturers of call center equipment to record events individually rather than averaging over periods of time.

With respect to sizing, quantitative models based on the Erlang C model were criticized by all of the authors researched for this paper, with individual preferences for simulation models or variant queue models, with alternative customer arrival distributions and/or call handling time distributions.

Both Koole (2004) and Mandelbaum (2001) criticize the use of Erlang B for calculating trunk requirements since they may result in overcapacity or even undercapacity as a result of abandonment, providing evidence from a study that found good results with just 10% more trunks than agents.

Wolff (2003) analyzed the results published by Mandelbaum (2001) and observed that the average waiting time during peak traffic times is overestimated when calculated using Erlang C, in comparison with an equivalent simulation model developed on ARENA.

Anton (1999) is more emphatic in preferring simulation modeling using commercial tools and discrete events. Koole (2004) supports combining models with abandonment metric and simulation models arrived at by fine adjustment. The quantitative models and their principal characteristics are listed in Table 1.

Table 1: Comparison between Quantitative Models

Characteristics	**Erlang B**	**Erlang C M/M/C**	**Simulation**
Arrivals	Poisson	Poisson	Distribution defined by modeler
Traffic queued or refused	Refused	Queued	Complex
Call flow and routing	Single queue	Single queue	Complex
Call overflow	No	No	Yes
Abandonment	No	No	Distribution defined by modeler
Retrials	No	No	Distribution defined by modeler
Call handling time	Exponential	Exponential	Distribution defined by modeler
Prioritization between different types of call	No -all calls are equal	No -all calls are equal	Yes
Agent ability (performance by type of call)	No -all agents are equal	No -all agents are equal	Yes
Interaction between events	No	No	Yes
Queue priority	FIFO	FIFO	User-defined

4 APPLICATION OF MODELS TO A CASE STUDY

The models presented were applied to a case study, an IT company Call Center, with focus on its primary operation, which is a specialized service using exclusive resources. The principal characteristics of the case studied are provided in Table 2.

Table 2: Characteristics of Call Center Studied

Business	Specialized Call Center.
Service studied	One type of specialist technical support, among many provided.
Size of operation	50 dedicated trunks, up to 30 service positions, specialized to this service.
Service provided	Incoming, Monday to Saturday, from 6 a.m. to 10 p.m.
Horizon	Tactical, weekly
Routing	Several different types of calls and service teams, but the focus is on the most representative service.
Call handling times	Obtained from statistical analysis of Call Center reports.
Time to abandonment	Obtained from analysis of Call Center reports.
Demand forecast	Provided by the Call Center.

4.1 Simulation Model

The approach taken to the simulation model, its input variables and its results was based on the terminology employed by Rockwell Software's ARENA CONTACT CENTER, and includes animation to facilitate analysis and understanding of the model (Figure 2).

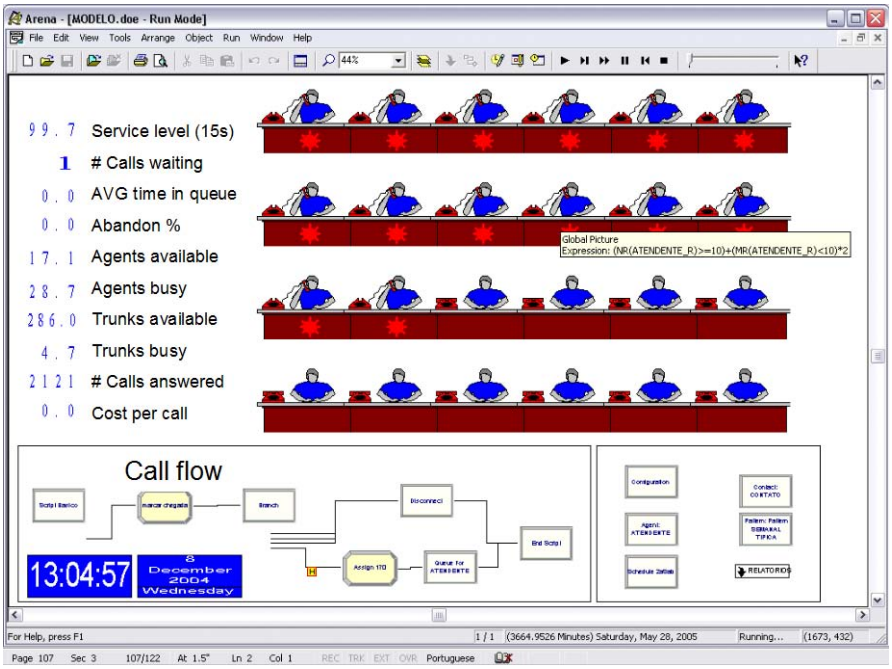

Figure 2: Simulation Model on ARENA

The period of execution for the model was defined as a typical week (from Monday to Saturday) and it was run for a total of 100 replications. The service level metric is calculated every 15 seconds while all other system times are in minutes.

4.2 Erlang Models

In order to allow for comparison of the simulation model results with results from the analytical models, the traffic generated by the simulator were used as data for the following models and assumptions:

- ERLANG B: 2% probability of lost traffic
- ERLANG C: 5% probability of queuing
- ERLANG C Service Level: 85% of calls answered in 15 seconds or less

The analytical models Erlang B, Erlang C and Erlang C Service Level were developed using Microsoft Excel. Erlang tables data was downloaded from <www.erlang.com.br> (in Portuguese Language).

5 RESULTS

Even though this was a single service flow, with just one type of call and one service team, the impact of abandonment on agent sizing by simulation was significant (21 Agents). Observing Figure 3, which illustrates the day of heaviest traffic, Monday, it can be concluded that:

- Erlang C and Erlang C Service Level, which do not include abandonment, calculated the number of agents needed as 27.4% 10.6% higher respectively;

- The difference between the two simulation scenarios, the basic high capacity set up (50 Agents) and the chosen configuration (21 Agents), was just 2%, which indicates that one could proceed directly to excess capacity scenarios for purposes of sizing.

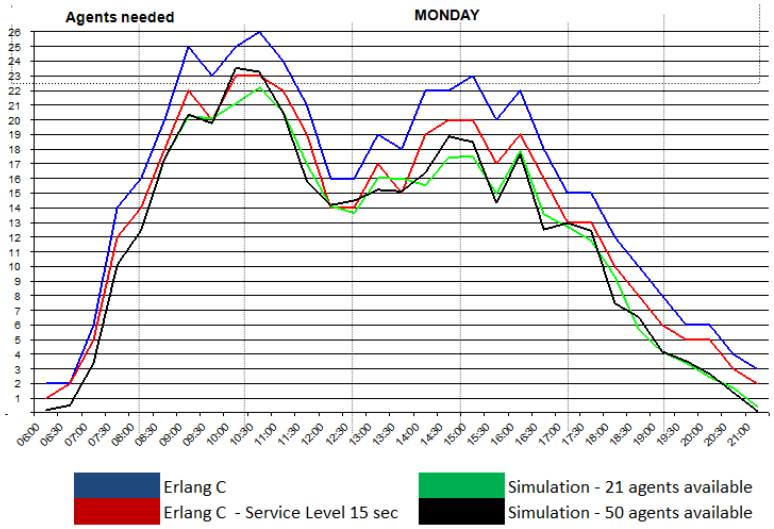

Figure 3: Differences between Erlang C, Erlang C Service Level and the Simulation

The results of trunk sizing also exhibited significant differences between the Erlang B model and the simulation:

- Simulation: a peak of 23 trunks busy when 50 are available. When remodeled with 23 trunks, just 5 calls in 6,000 were blocked, i.e. 0.081%.
- Erlang B: a peak of 32 trunks busy to achieve a probability of blockage loss of 0.01%

6 CONCLUSIONS

Because of its underlying assumptions, the Erlang C formula has been extensively criticized in the literature by many authors including Cleveland (1997), Mehrotra (1997), Anton (1999), Klunge (1999), Brown (2002), Mandelbaum (2001), Koole (2004) among other researchers.

This practical application of modeling has demonstrated the numerical advantages of using simulation models when the abandonment metric, which is indispensable to sizing resources, is included.

Another of the advantages of using simulation rather than the Erlang models is the fact that the numbers of resources (trunks and agents) needed are fractioned, which avoids the propagation of rounding errors due to continual adoption of correction factors followed by rounding.

Furthermore, whereas with the Erlang models trunks and contact channels are sized independently, with simulation it is possible to define them together. In other words a lack of capacity in the contact channel restricts flow and has an impact on sizing of resources.

While the simulation results are more accurate (assuming one has the "right" real-world situation reflected in the model), the closed-form methods are very quick and easy to obtain - not requiring owning a simulation software package. This positive fact (with regard to Erlang B and C) may explain why these tools continue to be taught to our students, until affordable simulators can be widely available.

REFERENCES

Anton, J., Bapat and B. Hall. 1999. Call Center Performance Enhancement Using Simulation and Modeling. 1 ed. Indiana, ICHOR Business Books.

Bapat, V. and E. B. Jr., Pruitte. 1999. Using Simulation in Call Centers. In: Proceedings of the 1999 Winter Simulation Conference, ed. P.A. Farrington, H.B. Nembhard, D.T. Sturrock , G.W. Evans, 1395-1400. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Brizola, N. , S. W. Costa, T. A. Pazeto, and P. J. F. Freitas. 2001. Planejamento de Capacidade de Call Center. In : ICIE, Florianópolis.

Chokshi R. 1999. Simulation: A Key to Call Center Management, AT&T Laboratories. In : Arenasphere 98. Nemacolin. EUA

Cleveland, B. and J. Mayben. 1997. Call Center Management On Fast Forward: Succeeding In Today's Dynamic Inbound Environment. 1 ed. Annapolis, Maryland, Call Center Press.

Cooper R. B. 1997. Introduction to Queueing Theory. 2 ed. North Holland, New York.

Cusak M. 1998. Online Customer Care: Strategies for Call Center Excellence. 1 ed. Wisconsin, ASQ.

Franzese, L. A. G. 2002. Apostila de Engenharia de Tráfego. Paragon Tecnologia, São Paulo.

Gulati, S. and S. A. Malcolm. 2001. Call Center Scheduling Technology Evaluation Using Simulation. In Proceedings of the 2001 Winter Simulation Conference, ed. B. A. Peters, J. S. Smith, D. J. Medeiros, M. W. Rohrer, 1438-1442. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Klunge R. 1998. The role of Simulation in Call Center Management. In : Arenasphere 98. Nemacolin.

Koole G. 2004. The Calculus of Call Center: Server Level Definitions and computations. In: MSON Conference. Endihoven. Vrije Universiteit Amsterdam.

Mandelbaun A., A. Sakov and S. Zeltyn. 2001. Empirical Analysis of a Call Center. Technion Israel Institute of Technology, Israel.

Mehrotra V. and J. Fama. 2003. Call Center Simulation modeling: Methods, Challenges, and opportunities, In: Proceedings of the 2003 Winter Simulation Conference, USA.

Mehrotra V. 1997. Ringing up Big Business, Avaialable at: <http://www. CallCenter Staffing.htm>

Pichitlmken, J., A. Deslauries, P. L. Ecuyer, and A. N. Avramidis. 2003. Modeling and Simulation of a Telephone Call Center. In Proceedings of 2003 Winter Simulation Conference, eds. S. E. Chick, P. J. Sanchez, D. M. Ferrin, D. J. Morrice, 144-152. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Wolff J. F. 2003. Simulação de uma Central de Atendimento: Uma Aplicação – Universidade Federal de Santa Catarina, Florianópolis , Santa Catarina, Dissertação de Mestrado.

AUTHOR BIOGRAPHIES

LUIZ AUGUSTO G. FRANZESE is a simulation consultant with a degree in Production Engineering and an MSc. in Logistics who has completed 250 successful projects with simulation. He founded PARAGON Tecnologia in 1992, the pioneer and leading consulting company in simulation in South America. He has trained more than 1,200 professionals in simulation. He can be contacted by email at <augusto@paragon.com.br>

MARCELO MORETTI FIORONI is a simulation consultant with an Electrical Engineering degree, an MSc. in Manufacturing and a PhD in Logistics from the University of Sao Paulo (USP). He has participated in 250 successful projects involving simulation. Consultant with PARAGON Tecnologia, leading consulting company in simulation in South America. Teaches Simulation at Faculdades Metropolitanas Unidas (FMU) in Sao Paulo, Brazil. He has trained more than 1,200 professionals in simulation. He can be contacted by email at <marcelo@paragon.com.br>

RUI C. BOTTER is an Associate Professor at the University of Sao Paulo in the Naval and Ocean Engineering Department. He has been using simulation as a research tool since 1993 and his activities are focused mainly on logistics and transportation. He can be contacted by email at <rcbotter@usp.br>

PAULO FREITAS is an associate professor at the Department of Computer Science at the Federal University of Santa Catarina (UFSC), Brazil. He received a doctoral degree in production engineering from UFSC in 1994. His research interests include simulation of computer systems for performance improvement, Monte Carlo methods, risk modeling and simulation, analysis for input modeling, and output analysis. He is a member of the Society for Computer Simulation (SCS) and the Brazilian Computer Society (SBC). His e-mail address is <freitas@inf.ufsc.br> and his web address is <www.inf.ufsc.br/~freitas>