

The mathematics of call centers

Ger Koole

Department of Mathematics, Vrije Universiteit Amsterdam

“Research highlight”, Annual report Stieltjes Institute 2000/2001

1 Introduction

In 1999 a new research group is created at the Vrije Universiteit under the name “optimization of business processes”. This group does research on the theory and applications of stochastic operations research. Main research axes are queueing theory, Markov decision chains, and their applications in logistics, telecommunication, and call centers. In this research highlight I will focus on the last application area. While doing so I will not only show the crucial role of probability theory in call center management, I will also show connections with other fields of mathematics.

2 Some background

Call centers are systems that deliver services by telephone. People are of course the prime resource, and they account for 60 to 70% of the operational costs. Another important resource is communication and computer equipment. Without computers call centers would have been impossible: they replaced the paper files allowing for immediate access to customer data. The successors of call centers are often called contact centers, because of the shift of communication towards fax, email, and internet. However, in most contact centers inbound calls still form the majority of contacts with customers. Inbound calls are also the most demanding when it comes to waiting times or response times: nobody expects an email to be answered within minutes, but waiting on the telephone for several minutes is usually considered unacceptable.

Most companies nowadays have call centers, or hire specialized firms to handle their communications with customers through call centers. As such there is an enormous financial interest in call centers. Current trends are towards an increase in size and in complexity.

The management of call centers can be split in two parts. Issues related to the *effectiveness* of the services offered by the call center require ICT and product-related skills. The *efficiency* is for a considerable part the domain of mathematics. Because the costs are dominated by personnel costs, this amounts to an efficient use of the workforce. The basis of workforce management in call

centers is the well-known Erlang formula [3]: it allows one to calculate the waiting time distribution under given load and number of servers. It is used in call centers to calculate the number of servers needed to meet the required waiting time objectives. From this daily requirements are calculated which form the basis of the agent schedules. The research projects that we discuss in the next section can all be seen as extensions of the Erlang formula.

For an extensive overview of call center research, see Gans et al. [4].

3 Call center research

We will discuss a number of papers related to call centers. In the first statistics and queueing theory are used together to solve a problem related to variations in the arrival rate.

An interesting property of the waiting time in a call center, given by the Erlang formula, is the steepness of its curve as a function of the load or the number of servers for values corresponding to a high productivity. This is exactly the region where call centers operate: managers naturally want a high productivity. This means that one employee more or less has a big impact on the waiting times; the same holds for a small variation in offered load. Thus it is very important that the arrival rate (the load equals arrival rate times average service time) is well estimated. Although sophisticated workforce management tools are used for this job, we see that in practice call centers seem to be unable to predict offered traffic with the necessary precision. Using more sophisticated models is not the solution, for the simple reason that certain effects that influence the offered load cannot be predicted by the time the forecast is needed to make employee schedules. An example is an insurance company, where the number of claims increases drastically after a storm. Of course a storm cannot be predicted several weeks in advance, when the workforce schedule is made. The "solution" is to take the randomness into account when using the Erlang formula, by assuming that the arrival rate itself has a certain probability distribution. This gives upper and lower bounds to the arrival rate, and using the Erlang formula, upper and lower bounds to the number of employees needed. A call center manager can use this to schedule personnel, for example by hiring employees that can be scheduled on a very short notice. The mathematics behind this idea is worked out in Jongbloed & Koole [6].

Another way to deal with the fluctuations in load caused by variations in arrival rate is *call blending*. Assume that a call center has next to inbound calls also outbound calls that has to be done at some point in time. Call blending consists of dynamically assigning employees either to inbound calls, to outbound calls, or to let them idle. The objective is to maximize the number of outbound calls handled, while meeting waiting time requirements for the inbound calls. It is optimal to assign free employees to inbound calls if any is waiting; however it is usually not a good idea to use any free employee for outbound calls if no inbound calls are waiting! This is because this would mean that all inbound calls have to wait in queue for a free agent, which often makes the waiting time

unacceptably high. A control rule has to decide how many employees have to be kept free for incoming calls. This threshold level is computed in Bhulai & Koole [2] for various situations.

After having decided how many agents are needed in every interval using the Erlang formula (or one of its extensions) the employee schedules have to be determined. Usually this is done with a mathematical programming approach, where during each time interval the service level constraint must be met. Over the whole planning period this leads to overcapacity: because the minimum number of servers is the smallest integer number satisfying the service constraint and because the scheduled number of agents is not equal to this minimum at all periods. See Figure 1 (from [7]) for an example of the effect in a small sized call center. In the figure the minimal numbers of employees for each interval, the best schedule satisfying these minima, and the corresponding service levels are plotted for a typical small sized call center. The service level constraint is 95% for each interval; we see that the schedule satisfying these constraints for each interval, and that uses the minimal number of agents, has an average service level of 98.6%. This calls for a new objective that has a single service level constraint for the entire planning period (in Figure 1 a day). It was shown in Koole & van der Sluis [7] that a local search methods exist to find the best schedule. In the situation of Figure 1 we were able to reduce with this method the number of scheduled agents from 28 to 24.

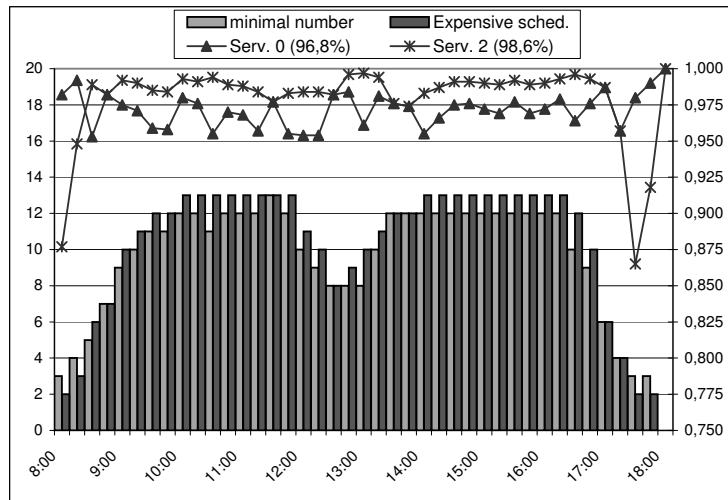


Figure 1: Numbers of employees and service levels for a call center

In the model above we assumed for each interval a constant arrival rate, and we approximated the service level in each interval with the Erlang formula. This is standard practice in call centers, and called the pointwise stationary approximation in Green & Kolesar [5]. In certain situations however this approx-

imation is not a good one, especially not if there is undercapacity during some interval. In these situation fluid approximation (as introduced in Newell [8]) perform better. In Altman et al. [1] it was shown that for queueing models such as call centers fluid approximations show a better performance than the original queueing systems for a considerable class of performance measures. Testing the usefulness of fluid approximations in the context of call center modeling is part of ongoing research.

References

- [1] E. Altman, T. Jiménez, and G.M. Koole. On the comparison of queueing systems with their fluid limits. *Probability in the Engineering and Information Sciences*, 15:165–178, 2001.
- [2] S. Bhulai and G.M. Koole. A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control*, 2003. To appear.
- [3] A.K. Erlang. Solutions of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Electroteknikeren*, 13:5–13, 1917. In Danish.
- [4] N. Gans, G.M. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 2003. To appear.
- [5] L. Green and P. Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37:84–97, 1991.
- [6] G. Jongbloed and G.M. Koole. Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17:307–318, 2001.
- [7] G.M. Koole and H.J. van der Sluis. Optimal shift scheduling with a global service level constraint. *IIE Transactions*, 2003. To appear.
- [8] G.F. Newell. *Applications of Queueing Theory*. Chapman and Hall, 1971.