Modeling bursts in the arrival process to an emergency call center

K. Gustavsson, P. L'Ecuyer, L. Olsson G-2018-29 April 2018

La collection <i>Les Cahiers du GERAD</i> est constituée des travaux de recherche menés par nos membres. La plupart de ces documents de travail a été soumis à des revues avec comité de révision. Lorsqu'un document est accepté et publié, le pdf original est retiré si c'est nécessaire et un lien vers l'article publié est ajouté.	The series <i>Les Cahiers du GERAD</i> consists of working papers carried out by our members. Most of these pre-prints have been submitted to peer-reviewed journals. When accepted and published, if necessary, the original pdf is removed and a link to the published article is added.
Citation suggérée: K. Gustavsson, P. L'Ecuyer, L. Olsson (Mai 2018). Modeling bursts in the arrival process to an emergency call center, document de travail, Les Cahiers du GERAD G-2018-29, GERAD, HEC Montréal, Canada.	Suggested citation: K. Gustavsson, P. L'Ecuyer, L. Olsson (May 2018). Modeling bursts in the arrival process to an emergency call center, Working paper, Les Cahiers du GERAD G-2018-29, GERAD, HEC Montréal, Canada.
Avant de citer ce rapport technique, veuillez visiter notre site Web (https://www.gerad.ca/fr/papers/G-2018-29) afin de mettre à jour vos données de référence, s'il a été publié dans une revue scientifique.	Before citing this technical report, please visit our website (https:// www.gerad.ca/en/papers/G-2018-29) to update your reference data, if it has been published in a scientific journal.
La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.	The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.
Dépôt légal – Bibliothèque et Archives nationales du Québec, 2018 – Bibliothèque et Archives Canada, 2018	Legal deposit – Bibliothèque et Archives nationales du Québec, 2018 – Library and Archives Canada, 2018
GERAD HEC Montréal 3000, chemin de la Côte-Sainte-Catherine Montréal (Québec) Canada H3T 2A7	Tél.: 514 340-6053 Téléc.: 514 340-5665 info@gerad.ca www.gerad.ca

Modeling bursts in the arrival process to an emergency call center

Klas Gustavsson^{*a*} Pierre L'Ecuyer^{*b*} Leif Olsson^{*a*}

^a Department of Information Systems and Technology, Faculty of Science, Technology and Media, Mid Sweden University, Sweden

^b GERAD & DIRO & CIRRELT, Pavillon André-Aisenstadt, Montréal (Québec), Canada

klas.gustavsson@miun.se
lecuyer@iro.umontreal.ca
leif.olsson@miun.se

April 2018 Les Cahiers du GERAD G-2018-29 Copyright © 2018 GERAD, Gustavsson, L'Ecuyer, Olsson

Les textes publiés dans la série des rapports de recherche *Les Cahiers du GERAD* n'engagent que la responsabilité de leurs auteurs. Les auteurs conservent leur droit d'auteur et leurs droits moraux sur leurs publications et les utilisateurs s'engagent à reconnaître et respecter les exigences légales associées à ces droits. Ainsi, les utilisateurs:

- Peuvent télécharger et imprimer une copie de toute publication du portail public aux fins d'étude ou de recherche privée;
- Ne peuvent pas distribuer le matériel ou l'utiliser pour une activité à but lucratif ou pour un gain commercial;

• Peuvent distribuer gratuitement l'URL identifiant la publication. Si vous pensez que ce document enfreint le droit d'auteur, contacteznous en fournissant des détails. Nous supprimerons immédiatement l'accès au travail et enquêterons sur votre demande. The authors are exclusively responsible for the content of their research papers published in the series *Les Cahiers du GERAD*. Copyright and moral rights for the publications are retained by the authors and the users must commit themselves to recognize and abide the legal requirements associated with these rights. Thus, users:

- May download and print one copy of any publication from the public portal for the purpose of private study or research;
- May not further distribute the material or use it for any profitmaking activity or commercial gain;
- May freely distribute the URL identifying the publication.

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim. **Abstract:** In emergency call centers (for police, firemen, ambulances, rescue teams) a single event can sometimes trigger many incoming calls to the center in a short period of time. Several people may call to report the same fire or the same accident, for example. Such a sudden *burst* of incoming traffic can have a significant impact on the responsiveness of the call center for other events in the same period of time. We examine data from the SOS Alarm center in Sweden, related to this type of situation. We also build a stochastic model for the bursts. We show how to estimate the model parameters for each burst by maximum likelihood, how to model the multivariate distribution of those parameters using copulas, and how to simulate the burst process from this model. In our model, certain events trigger an arrival process of incoming calls with a random time-varying rate over a finite period of time of random length. The time period can be short and the arrival rate over that period can be quite large.

Acknowledgments: This research project has been funded by the Swedish emergency call center *SOS Alarm Sverige Ab*, who also provided the data. The work of P. L'Ecuyer was also supported by a discovery grant from NSERC-Canada, a Canada Research Chair, and an Inria International Chair.

1 Introduction

Emergency call centers receive phone calls for various types of urgent situations such as medical emergencies, fires, accidents, rescue, terrorist acts, etc. In North America (Canada, USA, and recently Mexico) calling 911 will connect you to an *emergency dispatch office*, also called a *public-safety answering point*, in which operators can organize and dispatch the appropriate responses such as ambulances, firefighters, police, rescue resources, etc. In Europe and many other countries, 112 is the corresponding calling number. Our analysis in this paper is based on data from SOS-Alarm, which handles the 112 number in Sweden.

Managing an emergency call center involves deciding (among other things) what goals or constraints we want to impose on the quality of service, how to route calls and assign priorities, how many operators (general or specialized) to have in the center in each period (e.g., half-hour) of each day (this is called *staffing*), and what would be the work schedule of each available operator, e.g., over a given week (this is *scheduling*). The staffing and scheduling decisions must be made under various constraints on the work schedules of operators, based on union agreements for example, on the number of operators that can be available, the tasks for which they have been trained, etc. The staffing and scheduling problems are usually formulated as stochastic optimization problems in which the objective is to minimize the operating costs, under constraints on the quality of service which are defined as probabilities or mathematical expectations. For example, one can impose the constraint that the average (expected) waiting time of calls must be less than s_0 seconds, that the fraction of calls answered within less than s_1 seconds must be at least 95%, etc. These constraints are often imposed separately within given time periods (e.g., each day, each hour, etc.) and sometimes separately for different call types. For further details, see for example Aksin et al. (2007), Cezik and L'Ecuyer (2008), Avramidis et al. (2010), Koole (2013), Ta et al. (2016).. At the SOS Alarm Emergency call center, it is requested that 99% of the calls are answered within 30 seconds and that the average waiting time is less that 8 seconds.

To solve such a problem, one needs a reasonably realistic model of how things happen in the call center. Erlang formulas have been used for a long time for staffing call centers. These formulas are based on the simplifying assumptions that all calls have the same exponential service-time (or duration) distribution, and that calls arrive according to a Poisson process with a known constant arrival rate. But these assumptions are unrealistic. In particular, in typical emergency call centers (and other call centers as well), the arrival rate is time-dependent and is itself random (Avramidis et al. 2004; Channouf and L'Ecuyer 2012; Ibrahim et al. 2016; Oreshkin et al. 2016). The service times are also not exponential and their distribution may depend on the server, on time, and on other factors (Avramidis and L'Ecuyer 2005; Brown et al. 2005; Ibrahim et al. 2016). For realistic call center models, there is no reliable approximation formula for the measures of performance or quality of service, and one must rely on simulation (Pichitlamken et al. 2003). Simulation-based stochastic optimization algorithms have been proposed and experimented for call centers; see for example Atlason et al. (2004), Cezik and L'Ecuyer (2008), Avramidis et al. (2009), Avramidis et al. (2010), Chan et al. (2016).

In this paper, we focus on the modeling of one particular aspect of the arrival process in emergency call centers: the presence of arrival *bursts* triggered by a single event. For example suppose that a large fire or accident occurs in a city or along a highway. Within a few minutes, several people may call the 112 number to report the same incident. In some cases, a single event may trigger over 40 calls in less than 2 minutes, for instance. During a burst, the arrival rate of calls increases momentarily, possibly by a very large factor. This can overload the call center capacity and, as a result, urgent calls for unrelated events could be lost or may have to wait too long, potentially with serious consequences. It is important to understand how these bursts occur and to develop realistic models of the arrival rate process within a burst. We do this based on data from the SOS Alarm call centers in Sweden. One important difficulty in modeling a time-dependent and stochastic arrival-rate process like this one is that the arrival rate itself cannot be observed, only the arrival times can be observed. This complicates significantly the estimation of model parameters (Ibrahim et al. 2012; Oreshkin et al. 2016). We explain how to handle this for our models.

The rest of the paper is organized as follows. In Section 2, we describe the data we have on the bursts, and we provide some examples and summary information. In Section 3, we define a model we came up with, based

on the observed data, and we show how the bursts can be simulated once the model parameters have been estimated. In our model, a burst has a random length, during which the arrival rate is an exponential function with random initial value and whose exponent is also random and can be either negative (the rate decreases), positive (the rate increases exponentially), or zero (the rate is constant). With this type of exponential rate function, the arrivals can be simulated by inversion, via a change of variable that transforms a standard Poisson process (with rate 1) to a Poisson process with the desired exponential rate. The starting point, length, initial rate, and exponent, are random variables and our goal is to estimate their joint distribution from a given parameterized class of distributions. In Section 4, we explain how these parameters can be estimated by maximizing the log-likelihood of the data for our model. We illustrate this numerically in Section 5. Our proposed model is not perfect. We mention possible improvements and extensions at the end of the paper.



Figure 1: Impact of a large burst on the cumulative arrival count process.



Figure 2: The cumulative arrivals for two bursts; one with a decreasing arrival rate (left) and one with an increasing arrival rate (right).

2 The available data

Our study is based on detailed data from the Swedish 112 emergency call center, which is managed by a semiprivate company named SOS Alarm Sverige AB. The SOS Alarm call center works as a single virtual center which serves all of Sweden, although the operators (or agents) are physically in several different locations. The main one is an underground bunker in Stockholm. There are some locations with a small number of operators, e.g., in Northern Sweden. Calls are handled preferably by operators at the closest location, but if no operator is available at there, the call can be taken at another location.

The center handles about 60,000 calls per week (i.e., 6 calls per minute) on average. The available data contains call-by-call information that includes (among other things) the arrival time of each call, its waiting time, its area of origin, its duration, and, very importantly for us, the event number to which this call is associated. With this information, we can identify all the calls related to the same event, i.e., the calls that belong to a given burst. This last information is sometimes unavailable in emergency call centers; then it is much more difficult to identify the bursts in the data and to estimate model parameters.

For our data analysis in this paper, we consider only the calls related to "rescue" operations in a wide sense. i.e, calls that request for some emergency response and action by the police, fire department, or an ambulance, for example. This covers event types such as accidents, aggressions or attacks, fires, etc., that are likely to produce significant bursts.

We used data collected from January 1 to June 30, 2016. There was approximately three million calls overall during that period. Out of that, we extracted all the rescue events that generated at least 5 calls. There was 984 such events. The average number of calls per burst was 6.7 and the average duration of bursts was 591.6 seconds. Among those, we found 155 bursts of 15 calls or more, with an average burst size of 23.7 calls.

Figure 1 illustrates the impact of a burst of about 10 minutes on the cumulative number of arrival as a function of time. The burst causes the larger slope from about 3700 to 4300 seconds (a period of 10 minutes). It has a visible impact on the arrival process. Figure 2 shows the cumulative rate for two bursts, one with a decreasing arrival rate (on the left) and one with an increasing arrival rate (on the right).

3 Modeling and simulating a burst

3.1 The model

Based on what we have observed in the data, we designed the following model. When an event occurs that triggers a burst of calls, we assume that the calls related to the event, and which constitute the burst, arrive according to a non-homogeneous Poisson process with a certain arrival rate, after the time of the event. Note that the time of the triggering event is not observed, only the arrival times of the calls are observed. For this reason, we find it convenient to start our time clock when the first call of the burst arrives. This is time $t = T_1 = 0$ in our model of a burst. One alternative way of modeling could be to assume that the burst starts at the time when the event occurs, say time T_0 , and try to estimate T_0 for each burst. We do not take this more complicated path here.

After the first call which arrives at time $T_0 = 0$, additional calls related to the same event arrive at rate $\lambda(t)$ at time $t, t \ge 0$, and the (random) rate function λ is assumed to have the form

$$\lambda(t) = \begin{cases} Ae^{-tB} & \text{for } 0 \le t \le C, \\ 0 & \text{elsewhere,} \end{cases}$$
(1)

where A > 0, $B \in \mathbb{R}$, C > 0, and the vector (A, B, C) has some joint continuous distribution over $\Omega = [0, \infty) \times \mathbb{R} \times [0, \infty) \subset \mathbb{R}^3$. The burst has *intensity* parameter A, exponential rate with exponent B, and duration C. Its arrival rate $\lambda(t)$ for $t \in [0, C]$ is constant if B = 0, decreasing if B > 0, and increasing if B < 0. Note that *none* of the parameters A, B, and C is observed in the data.

In a simulation model, once (A, B, C) are known, the arrival times of calls from the burst can be generated using inversion and an appropriate transformation from a standard Poisson process, as we will explain. The cumulative rate of the Poisson process from time 0 to time t is

$$a(t) = \int_0^t \lambda(s) \mathrm{d}s = \frac{A}{B} \left(1 - e^{-tB} \right), \qquad 0 \le t \le C.$$

Its inverse can be found by writing a(t) = x and expressing t as a function of x, using the above expression. This gives

$$t = a^{-1}(x) = -\frac{\log(1 - Bx/A)}{B}$$
 for $B \neq 0$

For B = 0, these expressions for a(t) and $a^{-1}(x)$ are indeterminate, and using them for B near 0 will lead to numerical instabilities, but we can compute a stable approximation around 0 by expanding the exponential and the log in Taylor series and dividing each term by B. For a(t), using $1 - e^{-\epsilon} = \epsilon - \epsilon^2/2 + \epsilon^3/6 - \cdots$, we get

$$a(t) = At\left(1 - \frac{tB}{2} + \frac{(tB)^2}{6} - \ldots\right)$$

when B is close to 0. For $a^{-1}(x)$, using $-\log(1-\epsilon) = \epsilon + \epsilon^2/2 + \epsilon^3/3 + \cdots$, so when B is very close to 0,

$$a^{-1}(x) = \frac{x}{A} \left(1 + \frac{Bx}{2A} + \frac{B^2 x^2}{3A^2} + \dots \right).$$

In each case, we can truncate the series to a finite number of terms to obtain an accurate approximation, and the first term gives the exact value when B = 0.

3.2 Simulating the arrivals

It is known that if we simulate the arrival times X_1, X_2, X_3, \ldots of a standard Poisson process, with constant rate equal to 1, and we set $T_j = a^{-1}(X_j)$ for $j \ge 1$, then the T_j are the arrival times for a Poisson process with cumulative rate function a. See for example Çinlar (1975), Chapter 4, Section 7. Generating the X_j is easy: We put $X_0 = 0$ and the interarrival times $X_j - X_{j-1}$ are independent exponential random variables with mean 1, for $j \ge 1$. This gives Algorithm 1 to generate the arrival times T_j and their number N. In this algorithm, Expon(1) denotes an exponential random variable with mean 1. When $|B| < \epsilon_B$, we use the series to approximate $a^{-1}(X_j)$ instead of the direct formula. We add terms of this series until the last term is smaller than ϵ_S . At the end, we return the arrival times that are smaller than C.

Algorithm 1 : Generating the arrivals of a burst with exponential rate

4 Parameter estimation by maximum likelihood

4.1 Parameter estimation for a single burst

We start by writing the loglikelihood function for a single burst, as a function of (A, B, C), given that the arrival times for that burst are T_1, \ldots, T_N , and N is the number of arrivals. Note that $T_1 = 0$ does not contribute to the likelihood. The loglikelihood of these observations is then as follows; see, e.g., Daley and Vere-Jones (2003) for how to derive such a formula:

$$\log L = \sum_{j=2}^{N} \log \lambda(T_j) - \int_0^C \lambda(t) dt$$
$$= \sum_{j=2}^{N} (\log A - BT_i) - \int_0^C A e^{-tB} dt$$
$$= (N-1) \log A - B \sum_{j=2}^{N} T_i - AH$$
(2)

where

$$H = \begin{cases} \left(1 - e^{-CB}\right)/B & \text{if } B \neq 0; \\ C - C^2 B/2 + C^3 B^2/3! - C^4 B^3/4! + \cdots & \text{if } B \text{ is near } 0; \\ C & \text{if } B = 0. \end{cases}$$

Clearly, one must have $C \geq T_N$.

To estimate the parameters (A, B, C) for a single burst, for given N and T_1, \ldots, T_N , we can maximize this loglikelihood with respect to (A, B, C), under these constraints. We now look at how to do this by first deriving a set of necessary optimality conditions that should be satisfied when log L is maximized. At the optimum, for each of the parameters A, B, C, either the derivative of log L with respect to this parameter is zero, or this parameter cannot move further in the direction of the positive derivative because it is blocked by a constraint. The derivatives of log L with respect to the different parameters are:

$$\begin{split} \frac{\partial \log L}{\partial A} &= (N-1)/A - H, \\ \frac{\partial \log L}{\partial B} &= -\sum_{j=2}^{N} T_i - A \frac{\partial H}{\partial B} = -\sum_{j=2}^{N} T_i + \frac{A}{B^2} \left(1 - (1+CB)e^{CB} \right), \\ \frac{\partial \log L}{\partial C} &= -Ae^{-CB}, \end{split}$$

We are therefore looking for (A, B, C) for which each of these partial derivatives is zero or the parameter is blocked by a constraint such as $C \ge T_N$. Let us examine these conditions more closely.

The partial derivative with respect to C is always negative, so C should be taken as small as possible, which means $C = T_N$. Zeroing the derivative with respect to A tells us that we must take A = (N - 1)/H. Replacing A by (N - 1)/H in the partial derivative with respect to B yields

$$\frac{\partial \log L}{\partial B} = -\sum_{j=2}^{N} T_i - \frac{(N-1)}{H} \frac{\partial H}{\partial B} = -\sum_{j=2}^{N} T_i - (N-1) \frac{\partial \log H}{\partial B} = -\sum_{j=2}^{N} T_i - (N-1) \left(\frac{Ce^{-CB}}{1 - e^{-CB}} - \frac{1}{B} \right)$$

To equal this to zero, we need to find B for which

$$\frac{1}{B} - \frac{Ce^{-CB}}{1 - e^{-CB}} = \frac{1}{N-1} \sum_{j=2}^{N} T_j =: S.$$

Note that when $B \to 0$, the left side converges to C/2. This can be verified by replacing e^{-CB} by its Taylor expansion around B = 0, then putting the two terms on the same denominator, simplifying, and taking the limit. Therefore, if S = C/2, then B = 0 is the solution. If S < C/2, then the solution B is positive. This makes sense, because S < C/2 means that the arrivals tend to occur earlier than C/2 on average, which suggests that the arrival rate should be decreasing. If S > C/2, we have the opposite. Once we know the sign of B, we can find it using a standard root-finding technique.

4.2 Meta-parameter estimation

Suppose now that the vector Y = (A, B, C) has density $h_{\theta}(y)$ which depends on some unknown parameter (vector) $\theta \in \Theta$. Our goal is to estimate θ from the available data. A standard strategy for this, at least conceptually, is to maximize the loglikelihood of the data with respect to θ . See Munger et al. (2012) and the references given there. This loglikelihood is the log of the expectation with respect to the density h_{θ} of the product of likelihoods of all the bursts:

$$\log L(\theta) = \log \prod_{k=1}^{m} \mathbb{E}_{\theta} L_k(Y) = \sum_{k=1}^{m} \log \int_{\Omega} L_k(y) h_{\theta}(y) dy$$

where *m* is the number of bursts in the data and $L_k(y)$ is the likelihood function for the *k*th burst as a function of *y*, which is given by the exponential of the expression (2) in which (A, B, C) is replaced by *y*, and N, T_2, \ldots, T_N depend on *k*. Maximizing this integral with respect to θ is not easy. Even evaluating the integral for a single θ is usually too hard to be done exactly. What can be done is to approximate the integral by an average obtained by Monte Carlo. For any given θ and each *k*, we sample *n* independent realizations

of Y, say $y_{k,1}(\theta), \ldots, y_{k,n}(\theta)$, from the density h_{θ} . We can then replace the integral $\int_{\Omega} L_k(y) h_{\theta}(y) dy$ by the average

$$\frac{1}{n}\sum_{i=1}^{n}L_k(y_{k,i}(\theta))$$

in the loglikelihood expression. This gives the overall loglikelihood estimator

$$\log \hat{L}_n(\theta) = \sum_{k=1}^m \log\left(\frac{1}{n} \sum_{i=1}^n L_k(y_{k,i}(\theta))\right).$$
(3)

Conceptually, we can assume that the Monte Carlo samples are defined for all $\theta \in \Theta$, with common random numbers across all values of θ . After "fixing" the common random numbers, the vector $(y_{k,1}(\theta), \ldots, y_{k,n}(\theta))$ and the estimator $\log \hat{L}_n(\theta)$ become deterministic functions of θ . The idea is then to maximize the deterministic function $\log \hat{L}_n(\theta)$ with respect to θ . This function can be computed at any desired value of θ by reusing the common random numbers. Under appropriate assumptions on h_{θ} and on the sampling method, this is usually a smooth function of θ , although it is typically not concave and it may have multiple local maxima, so it is generally not easy to maximize. Note that (3) is a biased estimator of $\log L$, because the expectation of the log is not equal to the log of the expectation, but the bias vanishes when $n \to \infty$. This bias can also be reduced by using the Delta method with one additional term in the Taylor expansion.

A key ingredient for applying this methodology is that one must first select a parameterized density family $\{h_{\theta}, \theta \in \Theta\}$ for Y. This is also not trivial, mostly because the three components of Y are usually not independent and it is generally not easy to model this dependence. We will look at it in our numerical examples in Section 5.

A simpler (perhaps more naive) approach to estimate the density h_{θ} is to first estimate the vector Y = (A, B, C) separately for each burst, by maximizing its own loglikelihood function as explained in Section 4.1, then look at the distribution of the realizations of Y thus obtained, and fit some three-dimensional density h_{θ} to these data. This is what we will do in the next section.

5 Numerical examples

For each of the 984 bursts of size 5 or more collected in our data, we estimated the three parameters A, B, C by MLE as explained earlier. We took only the bursts of size 5 or more because for the smaller bursts we can hardly estimate the three parameters. Figure 3 shows the cumulative number of calls and the estimated cumulative arrival rate with our model for two examples of bursts, one with approximately constant rate and the other with decreasing rate. The rate model does not fit perfectly for those two bursts, but it provides a reasonable approximation, better than just assuming a constant rate. In the left picture, there is a significant delay between the first and second call. We have observed these types of gaps in other bursts as well, perhaps in around 10% of them. In some cases this delay was pretty long relative to the length of the burst, so the cumulative rate had a hockey stick shape. In a few (rare) cases we observed a significant gap in between two intervals of high-frequency arrivals, and in one case there were gaps between four groups of arrivals. These types of bursts have explanations (e.g., a fire first notices only from inside a building, and later on seen from outside, etc.). We did not try to model these occasional delays in the bursts for now; we leave this for future work.

Figure 4 shows scatter plots of the pairs (B, A), the pairs (B, C), and the pairs (A, C). We observe a strong dependence in the first two pairs, but not much for the (A, C) pair. In the left plot, we also see that there is no point (B, A) below the read line, i.e., with $0 \le A < 4B$. We will model the dependence using copulas. The usual way to do this is to fit a univariate distribution to each marginal, then transform the three variables of each point to uniforms by applying the probability integral transformation (i.e., take the cdf of the estimated marginal), and fit a three-dimensional copula to these uniform points. We did this and it did not work well because the dependence behaves differently when B > 0 than when B < 0, and it was hard to capture this difference by a standard copula.



Figure 3: The cumulative arrivals (step function) and the estimated cumulative rate function (smooth function in red) for two bursts.



Figure 4: Scatter plot of points (B, A) (left), (B, C) (middle), and (A, C) (right).

For this reason, we decided to separate the two cases, B > 0 and B < 0, and construct separate models for the two. For each case, we have a marginal distribution for each variable, A, B, and C. This gives six marginal distributions. We estimated each marginal distribution in two ways. The first approach was to select and fit parameterized distributions and the second was to estimate each density by a kernel density estimator (KDE) with a Gaussian kernel. The reason for using these two different methods is the following. We found that the KDE provides a better fit than the parameterized distributions, so we used it to transform the data to uniform to obtain an empirical copula. On the other hand, when generating triples (A, B, C)using the copula, we need to apply the inverse (estimated) cdf to a uniform to generate each coordinate of this vector, and the inverse cdf is much easier to compute for a parameterized distribution than for a KDE. Therefore for that purpose, we used the parameterized versions of the marginals. We now describe the process in more details.

Let F_A^+ , F_B^+ , F_C^+ denote the cdf's of the marginal distributions of A, B, C obtained by KDE when B > 0, and let F_A^- , F_B^- , F_C^- be the marginal distributions from KDE when B < 0. Each of these KDEs was constructed using a Gaussian kernel with a bandwidth selected by a heuristic formula of Silverman (1986). After computing these cdf's, we applied the probability integral transformation to transform each parameter vector (A, B, C) in the data to a vector $\mathbf{U} = (U_A, U_B, U_C) = (F_A^+(A), F_B^+(B), F_C^+(C))$ if B > 0, and similarly using the other marginals if B < 0, For the case B < 0, we actually modeled the density and cdf of -Binstead of B. The resulting vector \mathbf{U} has marginals that are (approximately) uniform over (0, 1), so its distribution is (approximately) a copula. Scatter plots of the two-dimensional projections of the resulting vectors \mathbf{U} are shown in Figure 5. This figure reveals negative dependence for all pairs, except for (B, A)when B > 0 for which the dependence is positive. Some corners are totally empty. For example, when B > 0and U_A is small, U_B is never large and U_C is never small.

It is hard to fit a three-dimensional copula model that matches all this dependance. What we did is model the copulas for the pairs (U_B, U_A) and (U_B, U_C) for each sign of B, using two-dimensional Archimedean copulas. To generate a triple (A, B, C), we first select the sign of B, which is positive with some probability p, then we generate U_B from the uniform distribution over (0, 1), then U_A conditional on U_B and also U_C



Figure 5: Scatter plot of pairs (U_B, U_A) , (U_B, U_C) , and (U_A, U_C) for B > 0 (above) and for B < 0 (below).

conditional on U_B , each from the appropriate copula, and finally we apply the appropriate inverse cdf to each uniform to obtain the final triple. For this last step, the inverse cdf's of the marginals must be easily computable, which is not the case for the KDE's. For this reason, for this step we uses parametric distributions for the marginals. The selected parametric distributions were lognormal for A, gamma for \sqrt{B} , and Weibull for C, for the case B > 0. For the case B < 0, we took the generalized extreme value (GEV) distribution for \sqrt{A} , gamma for $\sqrt{-B}$, and GEV for C. To obtain (A, B, C) from **U**, if B > 0 we put $(A, \sqrt{B}, C) = ((G_A^+)^{-1}(U_A), (G_B^+)^{-1}(U_B), (G_C^+)^{-1}(U_C))$ where G_A^+, G_B^+, G_C^+ denote the cdf's of the estimated parametric marginals, and similarly for the case where B < 0.



Figure 6: Scatter plot of the two-dimensional projections of a sample of 984 realizations of (A, B, C) simulated with our model.

After estimating all the parameters, we generated a sample of 984 realizations of (A, B, C) from our model, using the method just described. Figure 6 shows scatter plots of the two-dimensional projections of these points. These plots can be compared with the plots of the raw data in Figure 4. We find that the model is reasonably representative.

9

Conclusions

We developed a stochastic model for bursts of call arrivals in emergency call centers, based on data from the SOS Alarm call center in Sweden. The probabilistic behavior of each burst is determined by a vector of three parameters. We modeled the three-dimensional distribution of this vector using a copula construction and found that this distribution matches very well the empirical distribution of the parameter vectors estimated directly from the data. Further work that we intend to do includes trying to model the delays that sometimes occur in the bursts, trying a KDE of the three-dimensional copula (instead of parametric two-dimensional ones), and implementing the methodology described in Section 4.2. The latter would permit one to consider all the bursts from the data, and not only those of size 5 or more (say) to estimate the model. On the other hand, maximizing the likelihood is likely to be much more difficult.

References

- Akşin, O. Z., M. Armony, and V. Mehrotra. 2007. The Modern Call Center: A Multi-Disciplinary Perspective on Operations Management Research. Production and Operations Management 16(6):665–688.
- [2] Atlason, J., M. A. Epelman, and S. G. Henderson. 2004. Call center staffing with simulation and cutting plane methods. Annals of Operations Research 127:333–358.
- [3] Avramidis, A. N., W. Chan, M. Gendreau, P. L'Ecuyer, and O. Pisacane. 2010. Optimizing Daily Agent Scheduling in a Multiskill Call Centers. European Journal of Operational Research 200(3):822–832.
- [4] Avramidis, A. N., W. Chan, and P. L'Ecuyer. 2009. Staffing multi-skill call centers via search methods and a performance approximation. IIE Transactions 41(6):483–497.
- [5] Avramidis, A. N., A. Deslauriers, and P. L'Ecuyer. 2004. Modeling Daily Arrivals to a Telephone Call Center. Management Science 50(7):896–908.
- [6] Avramidis, A. N., and P. L'Ecuyer. 2005. Modeling and Simulation of Call Centers. In Proceedings of the 2005 Winter Simulation Conference, edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 144–152: IEEE Press.
- [7] Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. 2005. Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. Journal of the American Statistical Association 100(469):36–50.
- [8] Cezik, M. T., and P. L'Ecuyer. 2008. Staffing Multiskill Call Centers via Linear Programming and Simulation. Management Science 54(2):310–323.
- [9] Chan, W., T. A. Ta, P. L'Ecuyer, and F. Bastin. 2016. Two-stage Chance-constrained Staffing with Agent Recourse for Multi-skill Call Centers. In Proceedings of the 2016 Winter Simulation Conference, 3189–3200. Piscataway, NJ, USA: IEEE Press.
- [10] Channouf, N., and P. L'Ecuyer. 2012. A Normal Copula Model for the Arrival Process in a Call Center. International Transactions in Operational Research 19(6):771–787.
- [11] Çinlar, E. 1975. Introduction to Stochastic Processes. Englewood Cliffs, N. J.: Prentice-Hall.
- [12] Daley, D. J., and D. Vere-Jones. 2003. An Introduction to the Theory of Point Processes. Second ed. New-York: Springer-Verlag.
- [13] Ibrahim, R., P. L'Ecuyer, N. Régnard, and H. Shen. 2012. On the Modeling and Forecasting of Call Center Arrivals. In Proceedings of the 2012 Winter Simulation Conference, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 256–267: IEEE Press.
- [14] Ibrahim, R., P. L'Ecuyer, H. Shen, and M. Thiongane. 2016. Inter-Dependent, Heterogeneous, and Time-Varying Service-Time Distributions in Call Centers. European Journal of Operational Research 250:480–492.
- [15] Ibrahim, R., H. Ye, P. L'Ecuyer, and H. Shen. 2016. Modeling and Forecasting Call Center Arrivals: A Literature Study and a Case Study. International Journal of Forecasting 32(3):865–874.
- [16] Koole, G. 2013. Call Center Optimization. MG books, Amsterdam.
- [17] Munger, D., P. L'Ecuyer, F. Bastin, C. Cirillo, and B. Tuffin. 2012. Estimation of Mixed Logit Likelihood Function by Randomized Quasi-Monte Carlo. Transportation Research Part B: Methodological 4(2):305–320.
- [18] Oreshkin, B., N. Régnard, and P. L'Ecuyer. 2016. Rate-Based Daily Arrival Process Models with Application to Call Centers. Operations Research 64(2):510–527.
- [19] Pichitlamken, J., A. Deslauriers, P. L'Ecuyer, and A. N. Avramidis. 2003. Modeling and Simulation of a Telephone Call Center. In Proceedings of the 2003 Winter Simulation Conference, 1805–1812: IEEE Press.

- [20] Silverman, B. 1986. Density Estimation for Statistics and Data Analysis. London: Chapman and Hall.
- [21] Ta, T. A., P. L'Ecuyer, and F. Bastin. 2016. Staffing optimization with chance constraints for emergency call centers. In MOSIM 2016-11th International Conference on Modeling, Optimization and Simulation. See http://www.iro.umontreal.ca/~lecuyer/myftp/papers/mosim16emergency.pdf.