Proceedings of the 2012 Winter Simulation Conference
C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, eds.

ON THE MODELING AND FORECASTING OF CALL CENTER ARRIVALS

Rouba Ibrahim

Department of Management Science and Innovation
University College London
Gower Street, London, WC1E 6BT, UK

Pierre L'Ecuyer

DIRO, Université de Montreal
C.P. 6128, Succ. Centre-Ville
Montréal (Québec), H3C 3J7, CANADA

Nazim Regnard

DIRO, Université de Montreal
C.P. 6128, Succ. Centre-Ville
Montréal (Québec), H3C 3J7, CANADA

Haipeng Shen

Department of Statistics and Operations Research
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599, USA

ABSTRACT

We review and discuss the key issues in building statistical models for the call arrival process in telephone call centers, and then we survey and compare various types of models proposed so far. These models are used both for simulation and to forecast incoming call volumes to make staffing decisions and build (or update) work schedules for agents who answer those calls. Commercial software and call center managers usually base their decisions solely on point forecasts, given in the form of mathematical expectations (conditional on current information), but distributional forecasts, which come in the form of (conditional) probability distributions, are generally more useful, in particular in the context of simulation. Building realistic models is not simple, because arrival rates are themselves stochastic, time-dependent, dependent across time periods and across call types, and are often affected by external events. As an illustration, we evaluate the forecasting accuracy of selected models in an empirical study with real-life call center data.

1 INTRODUCTION

Demand arrivals are a primary source of uncertainty in various types of service systems such as health care systems, emergency services, retail stores, transportation systems, hotels, restaurants, call centers, etc. In those systems, demand arrives according to complicated stochastic processes that are difficult to model, because arrival intensities vary with time, are themselves stochastic, are not independent across successive time periods, and often depend on external events that are more or less predictable.

In this paper, we review some models proposed for call arrivals to a telephone call center (Gans, Koole, and Mandelbaum 2003, L'Ecuyer 2006, Akşin, Armony, and Mehrotra 2007). These centers have a huge economic importance. A key aspect of their management is to try to optimize their staffing and the work schedules of agents, to minimize the operating cost while providing a sufficiently good quality of service. The latter is usually quantified by imposing constraints on certain performance measures such as the fraction of calls answered within a given number of seconds (the so-called *service level*), the fraction of calls where callers lose patience and hang up before being answered (the *abandonment ratio*), or the average waiting time before an agent answers, often called the *average speed of answer*. These constraints can be expressed as expected values of these measures, or their averages in the long run, or they could also be probabilistic constraints on their (random) values over given time periods (for example, one may impose that over any given day, with probability at least 0.9, 80% of the calls are to be answered within 20 seconds); see Gans, Koole, and Mandelbaum (2003), Avramidis et al. (2010), and Gurvich et al. (2010).

In modern call centers, incoming calls are categorized in different types, each call type requiring a specific ability or skill. The number of different call types could be anywhere from 1 to nearly 100. Some call centers may employ no more than half a dozen agents, while large ones may have over 1000 agents working simultaneously. Each agent is trained to have a given subset of these skills, and can handle only call types for which she has the required skill. We then have a *multiskill* call center. An arriving call can be served immediately if an agent with the appropriate skill is available, or may have to wait in a queue. An *abandonment* occurs whenever the waiting time of a call exceeds its (random) patience time. Then, either that call is lost, or this person will call again later (this is a *retrial*). Skill-based routing strategies are rules that specify the agent-to-call and call-to-agent assignment mechanisms.

Demand forecasts are needed for workforce planning and management, which can be categorized in three levels of decision making (Gans, Koole, and Mandelbaum 2003, Mehrotra, Ozlük, and Saltzman 2010): long-term planning, short-term scheduling, and real-time schedule adjustments. Examples of long-term decisions are determining how many agents to hire and train, with which skills, at what times. These decisions can be made 6 to 12 months ahead of time and take into account aggregate call forecasts, agent availability and productivity assumptions, and anticipated staff attrition. Short-term scheduling determines which agents are assigned to work on which shift, on which days, and at what times, over the course of a scheduling period. This is typically done one to three weeks ahead of time, with the set of available agents. In multiskill centers, call routing rules must be selected together with the scheduling. Real-time schedule adjustments can be made after agent schedules are created, when new information becomes available such as updated forecasts of arrival volumes, agent absenteeism for some reason, etc. This can occur a day or two in advance, and during the day. These decisions can be made based on simulations or based on simplified approximations. The former requires good models for the arrival processes.

For call centers with a single call type, where all agents can handle any call, it is customary to use Erlang-C or Erlang-A queueing formulas (the Erlang-A model accounts for abandonments, in contrast to Erlang-C) to determine how many agents are needed in each time period of the day (typically using 15- or 30-minute periods) to satisfy constraints on the required quality of service (Gans, Koole, and Mandelbaum 2003, Avramidis and L'Ecuyer 2005). These staffing requirements are used in turn to construct work schedules (Pot et al. 2008, Avramidis et al. 2010). Erlang formulas assume that, in each time period, we have a first-in-first-out (FIFO) queue operating in steady-state, with independent and identically distributed (i.i.d.) exponential service times and Poisson arrivals at a constant rate. Their implementation only requires knowledge of the arrival rate, the mean service time, and the number of agents, for each period. Traditional forecasting methods for call centers have been designed to provide input parameters for these formulas. They usually provide only *point forecasts*, in the form of the expected number of calls in each time period of the day, because the formulas do not require additional information. These forecasts can be updated at any time (e.g., at the end of each time period) to account for the latest available information.

Erlang formulas only provide rough approximations, based on unrealistic simplifying assumptions. Moreover, they do not apply to multiskill centers. Stochastic simulation provides a more realistic evaluation tool in this context. It only requires arrival models in the form of stochastic processes that are easy to simulate. In these models, the arrival times and counts in each time period must be well-defined random variables, whose distributions may depend on current available information. In particular, at the beginning of each time period, the "forecasts" for the coming periods are *distributional forecasts*, defined in the form of (conditional) probability distributions (for both the arrival counts and arrival rates). These conditional distributions are not necessarily available or computed explicitly; they are often implicit in the models.

Some stochastic models provide explicit distributions for stochastic time-dependent *arrival rates*, in which case a simulation would normally generate a realization of the arrival rate process, and then generate arrivals according to a Poisson process conditional on the realized arrival rate process. Other models provide distributions directly for the *arrival counts* in different time periods, and assume a Poisson process with a constant rate over each period. To simulate the process using such a model, one would generate the arrival count in each period, and then generate the arrival times independently and uniformly over that

period. Models where we generate the rates and where the rates are constant in each period are usually more convenient to simulate, because arrival times can then be generated one by one only as needed, by generating independent exponential inter-arrival times, and placing those in the future event list one at a time, a mechanism that agrees with standard simulation software (Buist and L'Ecuyer 2005). Whenever the arrival rate changes at the beginning of a period, the next arrival event is rescheduled to account for the change in the arrival rate. In the ContactCenters software (Buist and L'Ecuyer 2005), the implementation takes care of this rescheduling automatically. The arrival times do not need to be stored and sorted as when counts are generated first. Note that for the sole purpose of generating point forecasts, using counts or rates is equivalent (they have the same expectation), provided that they satisfy the same modelling assumptions.

One important fact to take into account is that arrival times of individual calls are rarely available in call center data; only aggregated arrival counts per time period are available, and one must be able to estimate the models from that. For an exception, see Brown et al. (2005).

The remainder of this paper is organized as follows. In Section 2, we summarize important stylized features that have been observed empirically in call center arrival data, and we briefly discuss how these facts can be taken into account in arrival process models. Then, we examine and compare specific models in more detail, by focusing first on modeling arrivals within one day in Section 3, and extending to arrival models over several days in Section 4. In Section 5, we compare selected models in terms of forecasting performance, in an empirical study based on real-life call center data.

2 KEY FEATURES OF ARRIVAL PROCESSES

A modeling hypothesis that may appear natural is that calls arrive according to a Poisson process with time-dependent arrival rates. For sure, the arrival rate in call centers is not constant. In typical business-related call centers, there is a peak period just before lunch and another one just after lunch, with a lower arrival rate during lunch, and even lower in the early morning and late afternoon; see, e.g., Figure 2 of Avramidis, Deslauriers, and L'Ecuyer (2004) and Channouf and L'Ecuyer (2012). But different shapes of arrival rate patterns occur as well in other types of call centers, such as those handling food orders, ambulance and police calls, etc. Moreover, different shapes of arrival rate patterns are often observed over different days of the week, in different periods of the year, or on certain particular days (e.g., the first Monday of the month), in a given call center. For example, business-related centers often have higher call volumes on Mondays than on other days, while ambulance and police calls have a higher rate late at night over the weekend, but not during the week. This leads to our first key feature (or property) of call arrival processes:

(P1) The arrival rate varies (sometimes considerably) with the time of day and exhibits daily, weekly, yearly, and other types of seasonalities.

A Poisson process assumption with a deterministic arrival rate function implies that the number of arrivals over any given time period is a Poisson random variable, whose variance is equal to its expectation. However, empirical evidence invalidates this assumption; the observed variance of arrival counts is typically much larger than the mean, sometimes by a factor of 5 or more (Jongbloed and Koole 2001, Avramidis, Deslauriers, and L'Ecuyer 2004, Steckley, Henderson, and Mehrotra 2005):

(P2) The total demand (number of incoming calls) over any given time period has overdispersion relative to the Poisson distribution (the variance is significantly greater than the mean).

To reconcile the Poisson process model with the reality in (P2), one must take the arrival rate itself as stochastic; we will return to this in the next section. Other important features observed empirically concern the dependence between successive rates or counts:

(P3) There is significant (strong) positive dependence between arrivals rates (or counts) in different time periods of the same day, and this positive dependence usually decreases when the considered time

periods are taken farther apart (Tanir and Booth 1999, Whitt 1999, Avramidis, Deslauriers, and L'Ecuyer 2004, Shen and Huang 2008b, Channouf and L'Ecuyer 2012).

Models that fail to account for the positive dependence in (P3) or the overdispersion in (P2) give an overoptimistic view of call center performance measures. Such errors can be very significant; see Avramidis, Deslauriers, and L'Ecuyer (2004), Avramidis and L'Ecuyer (2005), Steckley, Henderson, and Mehrotra (2005), and Steckley, Henderson, and Mehrotra (2009). When observing arrival data over several months or years, we have (Aldor-Noiman et al. 2009, Brown et al. 2005, Channouf et al. 2007):

(P4) After correcting for detectable seasonalities, noticeable correlations remain between arrival counts over successive days.

(P5) After accounting for the dependence between total daily volumes in two successive days, the dependence that remains between the last period(s) of the first day and the first period(s) of the second day could be significant in call centers that operate 24 hours a day (such as for emergency services, police, etc.), and negligible in centers that close during the night.

For 24-hour-a-day centers, it may then seem natural to state the model of arrival rates (counts) per period as a single univariate time series after removing seasonalities and perhaps daily random effects (Channouf, L'Ecuyer, Ingolfsson, and Avramidis 2007, Section 4.2, consider this type of model). For other centers, it would be more fitting to use a multivariate time series for the sequence of vectors of arrival rates (counts).

When incoming calls are classified into multiple types, arrival rates (and counts) of certain pairs of call types are sometimes correlated (usually positively), while other pairs are approximately independent. Positive correlations may arise, for example, in multilingual call centers where certain service requests are handled in different languages. Neglecting this positive dependence can lead to serious overloads, particularly when some agents handle calls in multiple languages.

(P6) In call centers with multiple call types, there is sometimes strong dependence between arrival rates (and counts) of certain call types during the same time period.

Auxiliary information is often available in call centers to improve point or distributional forecasts considerably. For example, when a company sends notification letters to customers, or makes advertisements, this may trigger a larger volume of calls (Landon, Ruggeri, Soyer, and Tarimcilar 2010). Also, the number of abandonments in recent periods could be used as a covariate in a forecasting model for forthcoming hours, to account for retrials.

(P7) External knowledge can often be used to improve forecasting accuracy (and reduce the variance of distributional forecasts) by introducing covariates in models.

In certain types of call centers, for example in emergency services, unpredictable bursts of high arrival rates over short periods of time do occur. In this context, an important accident or similar event may trigger several dozen different calls (or more) within a few minutes, all related to the same event, whereas the usual expected number of calls within those few minutes is, say, no more than 2 or 3.

(P8) In certain types of call centers, the arrival rate has sometimes unexpected high peaks over short periods of time.

Ideally, we want arrival models to be as realistic as possible and to account for the above-named Properties (P1) to (P8). Their number of parameters should remain reasonably small to avoid overfitting, and these parameters should be easy to estimate from available data. Moreover, these estimates should not

be too hard to update (e.g., via Bayesian methods) to obtain distributional forecasts when new information becomes available at the end of any given time period.

3 MODELING ARRIVALS OVER A SINGLE DAY

In this section, we focus on modeling arrivals over a single day. The day is divided into *p time periods*, usually of equal length (although this is not essential). For example, if the center receives calls from 8:00 to 21:00, we may have $p = 52$ quarter hours. We denote by $\mathbf{X} = (X_1, \ldots, X_p)$ the vector of arrival counts in those periods. When arrivals are from a Poisson process with a random rate function, we denote by Λ_j the *cumulative* arrival rate (its integral) over period j. Then, conditional on Λ_j, X_j has a Poisson distribution with mean Λ_j. To simplify the notation, we assume in this paper that the periods have the same length and also that the time unit is one period. Then, when the arrival rate is constant over each period, this rate is the same as the cumulative rate Λ_j, and we denote both by Λ_j.

One of the most convenient types of arrival models for simulation is a Poisson process. But in view of Property (P2), the arrival rate of the Poisson process must be taken as stochastic (Jongbloed and Koole 2001, Avramidis, Deslauriers, and L'Ecuyer 2004, Steckley, Henderson, and Mehrotra 2009, Shen 2010b). Whitt (1999) proposed to do that by starting with a deterministic arrival rate function $\{\lambda(t), t_0 \le t \le t_e\}$, where t_0 and t_e are the opening and closing times of the call center for the considered day, and to multiply this function by a random variable W with mean $\mathbb{E}[W] = 1$, called the *busyness factor* for that day. The (random) arrival rate process for that day is then $\Lambda = \{\Lambda(t) = W\lambda(t), t_0 \le t \le t_e\}$. To simulate this process, it suffices to generate W first and then generate arrivals from the Poisson process with rate function Λ. Under this model, the arrival rates at any two given times are perfectly correlated, and $\mathrm{Corr}[\Lambda_j, \Lambda_k] = 1$ for all j, k. We also expect the X_j's to be strongly correlated. More specifically, let I_j denote the time interval of period j, let $\bar{\lambda}_j = \int_{I_j} \lambda(t)\mathrm{d}t$, and let X_j be the number of arrivals in I_j. Using variance and expectation decompositions, one can find that $\mathrm{Var}[X_j] = \bar{\lambda}_j(1 + \bar{\lambda}_j \mathrm{Var}[W])$ and, for $j \neq k$,

$$\mathrm{Corr}[X_j, X_k] = \mathrm{Var}[W]\left[(\mathrm{Var}[W] + 1/\bar{\lambda}_j)(\mathrm{Var}[W] + 1/\bar{\lambda}_k)\right]^{-1/2}.$$

This correlation is zero when $\mathrm{Var}[W] = 0$ (a deterministic rate) and approaches 1 when $\mathrm{Var}[W] \to \infty$.

Avramidis, Deslauriers, and L'Ecuyer (2004) have studied this model in the special situation where W has a gamma distribution with $\mathbb{E}[W] = 1$ and $\mathrm{Var}[W] = 1/\gamma$. Then, each Λ_j has a gamma distribution, $\mathbf{X}$ has a negative multinomial distribution, the parameters of this distribution are easy to estimate, and the variance of the arrival counts can be made arbitrarily large by decreasing γ toward zero. Jongbloed and Koole (2001) examined a similar model, but with independent busyness factors, one for each period of the day. Under their model, the Λ_j's are independent, as are the X_j's, which is inconsistent with (P3).

Avramidis, Deslauriers, and L'Ecuyer (2004) assume a piecewise constant arrival rate, $\lambda(t) = \lambda_j$ when t belongs to the jth period of the day, and show how to compute the joint maximum likelihood estimator of γ and the λ_j's. Channouf (2008) considers a variant of the model where $\lambda(t)$ is defined by a cubic spline over the day, with a fixed set of knots, and also shows how to estimate model parameters. This can provide a smoother (perhaps more realistic) model of the arrival rate. On the other hand, simulating arrivals from this process is more complicated and time consuming. Moreover, in empirical experiments, call center performance measures observed with this model were not much different from those observed with the piecewise constant rate model.

The model with a single busyness factor W accounts for (P3), but its flexibility is rather limited, because given the $\bar{\lambda}_j$'s, $\mathrm{Var}[X_j]$ and $\mathrm{Corr}[X_j, X_k]$ for $j \neq k$ are all determined by a single parameter value, namely $\mathrm{Var}[W]$. A larger variance necessarily implies larger correlations, and vice-versa. In an attempt to increase the flexibility of the covariance matrix $\mathrm{Cov}[\mathbf{X}]$, and in particular to enable a reduction of the correlations, Avramidis, Deslauriers, and L'Ecuyer (2004) introduced two different models for $\mathbf{X}$, based on the multivariate Dirichlet distribution, which yield relatively smaller correlations than the model based on W. Nevertheless, they remain larger than real-life estimates. In all these models, $\mathrm{Corr}[X_j, X_k]$ depends

on $\mathbb{E}[X_j]$, $\mathbb{E}[X_k]$, and $\mathrm{Var}[W]$, but not on the spacing between periods j and k. This disagrees with the second part of (P3).

Channouf (2008) and Channouf and L'Ecuyer (2012) proposed models that account for (P1) to (P3), with much more flexibility to match the correlations between the X_j's, by using a normal copula to specify the dependence structure between these counts. The vector $\mathbf{X}$ is assumed to have a discrete multivariate distribution with arbitrary one-dimensional marginals, which are estimated separately and independently. In their implementation, the authors take these marginals as negative binomial and find that this agrees with the data they have. Conditional on X_j, the arrival times in period j are again assumed to be independent and uniformly distributed in that period. This corresponds to Poisson arrivals at a rate Λ_j which is constant and gamma distributed in each time period. The correlation matrix for the normal copula is selected so that rank correlations between the counts match those in the data as closely as possible, under the constraint that this still gives a valid correlation matrix. To reduce the number of parameters in the copula model (especially when the number of time periods in the day is large, because there are then too many correlations to estimate), the authors also restrict the correlation matrices to certain parametric subclasses. For example, one may force the entries $r_{i,j}$ of the correlation matrix to be functions of $|j-i|$ only, and even to have the special form $r_{i,j} = \rho^{|j-i|}$ where $0 < \rho < 1$. The latter corresponds to an autoregressive structure of order 1 (AR(1)) for the series $X_1, X_2, \ldots, X_p$. Channouf and L'Ecuyer (2012) test their model on three data sets taken from real-life call centers and find that, in all three cases, it matches the correlations and the coefficients of variation of the counts better than all the models examined by Avramidis, Deslauriers, and L'Ecuyer (2004). In principle, similar copula models could be developed for the vector of arrival rates, $(\Lambda_1, \ldots, \Lambda_p)$, instead of for the vector of counts.

In call centers with multiple call types, to account for (P6), one must model the multivariate distribution of the vector giving the number of arrivals of each type in any given time period. Again, the dependence here can be modeled via a copula, after fitting the marginals individually. The simplest and more practical type of copula for this is probably the normal copula, used for example by Kim, Kenkel, and Brorsen (2012) and Ibrahim and L'Ecuyer (2012). However, empirical data suggests that for certain pairs of call types, the coefficient of upper or lower tail dependence, which measures the strength of the dependence in the right or left tail of the distribution, is quite different from that implied by a normal copula. Moreover, the choice of copula can have a significant impact on performance measures in call centers, because of the strong effect of tail dependence on the quality of service (Jaoua and L'Ecuyer 2011).

4 ARRIVAL PROCESS MODELS OVER SEVERAL DAYS

We now consider arrival process models over several days or months. We introduce some additional notation. Again, we suppose that the operating time of the call center over one day is partitioned into p time periods of equal length. We take the same p for all days. If some days have shorter opening hours than others (e.g., over the weekend), then we can just assume that the arrival rate is zero in the periods where the center is not open. We consider observations over q successive days. Each day i has a type of day $d_i \in \{1, \ldots, 7\}$, where $d_i = 1$ means that day i is a Monday, $d_i = 2$ means that day i is a Tuesday, and so on. When there are special days (c.f. (P7)), we index their type by k and let $S_k = \{i : \text{day } i \text{ is a special day of type } k\}$. In the models considered here, we consider a single call type and we assume that arrivals are from a Poisson process with a (random) constant arrival rate $\Lambda_{i,j}$ over period j of day i, for $i = 1, \ldots, q$ and $j = 1, \ldots, p$. This rate is expressed in "arrivals per period." Conditional on $\Lambda_{i,j}$, the number $X_{i,j}$ of arrivals in that period has a Poisson distribution with mean $\Lambda_{i,j}$.

Early studies, for example Mabert (1985), Andrews and Cunningham (1995) and Bianchi, Jarrett, and Hanumara (1998), used standard time series methods such as ARIMA models, with covariates to account for advertising, special-day effects, etc., to forecast arrival volumes in call centers. Exponential smoothing is a popular forecasting technique, where the forecast is constructed from an exponentially weighted average of past observations. The Holt-Winters method is an extension of exponential smoothing which accommodates both a trend and a seasonal pattern. Taylor (2008) compared various time series models, including a

Holt-Winters exponential smoothing model with multiple seasonal patterns. For forecasts more than a few days in advance, he concluded that very simple methods such as the additive fixed-effect model described below are hard to beat. However, in the short term, one can take advantage of the dependence structure between rates (or counts) in successive days and within the same day. Shen (2010a) comments about Taylor's work, highlighting the difference between modeling arrivals as a single time series, and as a vector time series where each day is modeled as a component of that vector; c.f. (P5). Channouf et al. (2007) developed simple additive models for the (small) number of ambulance calls in each hour, in the city of Calgary. Their models capture daily, weekly, and yearly seasonalities, selected second-order interaction effects (e.g., between the time-of-day and day-of-the week effects), special-day effects (such as the Calgary Stampede), and autocorrelation of the residuals between successive hours. Their best model outperforms a doubly-seasonal ARIMA model for the residuals of a model that captures only special-day effects.

Some linear models proposed recently use the "root-unroot" data transformation $Y_{i,j} = (X_{i,j} + 1/4)^{1/2}$ to stabilize the variance (Brown et al. 2005, Brown et al. 2010). The unconditional distribution, with random $\Lambda_{i,j}$, is then a mixture of such normal distributions, and therefore has larger variance, but one can nevertheless "assume" (as an approximation) that the square-root transformed counts $Y_{i,j}$ are normally distributed and fit Gaussian linear models to the transformed data, if $\mathrm{Var}[\Lambda_{i,j}]$ is not too large; see Brown et al. (2005), Weinberg, Brown, and Stroud (2007), Aldor-Noiman, Feigin, and Mandelbaum (2009), and Ibrahim and L'Ecuyer (2012).

One example of a general additive *fixed-effects* (FE) model for the square-root-transformed counts $Y_{i,j}$ is

$$Y_{i,j} = \alpha_{d_i} + \beta_j + \theta_{d_i,j} + \sum_k (\gamma_k + \delta_{k,j}) \mathbb{I}[i \in S_k] + \varepsilon_{i,j},$$

where the fixed coefficients α_{d_i}, β_j, $\theta_{d_i,j}$, γ_k, and $\delta_{k,j}$ (to be estimated from the data) represent the day-of-the-week effect, the period-of-day effect, day-period interaction effect, special-day effect of type k, and special-day-period interaction effect, respectively, and $\mathbb{I}[\cdot]$ is the indicator function. The residuals $\varepsilon_{i,j}$'s are assumed to be i.i.d. normal with mean 0 and variance σ_ε^2.

As an improvement, and based on real call center data analysis, Aldor-Noiman, Feigin, and Mandelbaum (2009) propose the following linear *mixed-effects* (ME) model:

$$Y_{i,j} = \alpha_{d_i} + \beta_j + \theta_{d_i,j} + \sum_k (\gamma_k + \delta_{k,j}) \mathbb{I}[i \in S_k] + D_i + \varepsilon_{i,j},$$

where D_i is a random effect for day i, the $\varepsilon_{i,j}$'s are no longer independent, and the other terms are the same as in the FE model defined above. They assume that the D_i's have mean 0 and obey an AR(1) process: $D_i = \rho_\mathrm{D} D_{i-1} + \varepsilon_i$ where $0 < \rho_\mathrm{D} < 1$ and the ε_i's are i.i.d. normal with mean 0 and variance $\sigma_\mathrm{D}^2(1-\rho_\mathrm{D}^2)$. This implies that $(D_1, \ldots, D_q)$ is multinormal with mean zero and covariance matrix with elements $\sigma_{i,i'} = \sigma_\mathrm{D}^2 \rho_\mathrm{D}^{|i'-i|}$. The residuals $\varepsilon_{i,j}$ are also assumed to have an AR(1) structure within each day. That is, the vector $(\varepsilon_{i,1}, \ldots, \varepsilon_{i,p})$ has a multinormal distribution with mean zero and covariance matrix with elements $\sigma_{j,j'} = \sigma_\varepsilon^2 \rho_\varepsilon^{|j'-j|}$. The additive random effect D_i plays a similar role as the multiplicative random busyness factor W in the single-day model discussed earlier. Here, by playing with the variance and correlation parameters σ_D^2, ρ_D, σ_ε^2 and ρ_ε, we have four degrees of freedom to adjust the overdispersion and the dependence across days and across periods. Brown et al. (2005) proposed an earlier version of this model, also based on call-center data, without intraday correlations and without special-day effects.

Ibrahim and L'Ecuyer (2012) extend this ME model to two bivariate ME models, where $Y_{i,j}$ is replaced by the pair of transformed arrival counts for two different call types. These models account for the dependence between the two call types by assuming that the vectors of random effects or the vectors of residuals across call types are correlated multinormal. This corresponds to using a normal copula.

To reduce the dimensionality of the vectors $(Y_{i,1}, \ldots, Y_{i,p})$, Shen and Huang (2005) proposed the use of singular-value decomposition to define a small number of vectors whose linear transformations capture most of the information relevant for prediction. Based on this, Shen and Huang (2008b) then developed a

dynamic updating method for the distributional forecasts of arrival rates. Shen and Huang (2008a) proposed a method to forecast the latent rate profiles of a time series of inhomogeneous Poisson processes to enable forecasting future arrival rates based on a series of observed arrival counts. Aktekin and Soyer (2011) recently proposed a model based on a Poisson-gamma process, where $\Lambda_{i,j} = W_{i,j}\lambda_{i,j}$ for fixed $\lambda_{i,j}$'s, and where the multiplicative factors $W_{i,j}$ have a gamma distribution and obey a gamma process. Soyer and Tarimcilar (2008) analyzed the effect of advertisement campaigns on call arrivals. Theirs is a Bayesian analysis where they model the Poisson rate function using a mixed model approach. This mixed model is shown to be superior to using a fixed-effects model instead. Weinberg, Brown, and Stroud (2007) also use Bayesian techniques in their forecasts. They exploit the (normal) square-root transformed counts to include conjugate multivariate normal priors, with specific covariance structures. They use Gibbs sampling and the Metropolis Hastings algorithm to sample from the forecast distributions, which unfortunately requires long computational times. Moreover, it is unclear how to incorporate exogenous covariates in such a model.

5 A CASE STUDY

We now report partial results of an empirical study using real-life data gathered at the call center of a major Canadian company. The data were collected over $q = 275$ days (excluding holidays when the center is closed, and weekends), from October 19, 2009, to November 11, 2010. The center operates from 8:00 to 19:00 on weekdays (Monday to Friday), i.e. each weekday contains $p = 22$ half-hour periods. The data consisted of arrival counts $X_{i,j}$ in the ith day and jth half-hour period (here, we consider a single call type).

Figure 1 shows the time series of the counts $X_{i,j}$ over two weeks, from August 19, 2010, to September 1, 2010. In Figure 2, we plot the average counts over those days, as functions of the period, for each weekday. We observe two major daily peaks, one shortly before 11:00 and the other around 13:30. Exploratory analysis of our data shows reveals: (i) positive correlations between the daily counts over successive days, which decrease with the distance between those days (see Table 1); and (ii) positive correlations between the counts over successive periods of the same day. This agrees with (P3) and (P4).

Table 1: Correlations between arrival counts on successive weekdays.

Weekday	Mon.	Tues.	Wed.	Thurs.	Fri.
Mon.	1.0	0.48	0.35	0.35	0.34
Tues.		1.0	0.68	0.62	0.62
Wed.			1.0	0.72	0.67
Thurs.				1.0	0.80
Fri.					1.0

We applied the square-root transformation $Y_{i,j} = \sqrt{X_{i,j} + 1/4}$ to our data, and then adjusted the FE and ME models described in Section 4, without special days; see Ibrahim and L'Ecuyer (2012) for more details. We also considered the Holt-Winters (HW) smoothing method, with a daily seasonality. Finally, we considered a way of splitting the existing daily forecasts used at the company by applying a *top-down* (TD) approach, as explained in Gans et al. (2003) and Taylor (2008), which in our context splits forecasts of total daily arrival counts into forecasts of half-hour counts based on estimates of historical proportions of calls in successive half-hour intervals of the day. The reason for considering the company's daily forecasts is that they incorporate important external information (not in the data set) that impacts the arrival process, such as major marketing campaigns and recent price increases.

To compare the performances of these models, we generate out-of-sample forecasts for the horizon ranging from August 19, 2010, to November 11, 2010. That is, we make forecasts for a total of 85 days and generate $85 \times 22 = 1320$ predicted values. We consider two forecasting lead times to mimic real-life challenges faced by call center managers: two weeks and one day. We let the learning period include all days in the data set, up to the beginning of the forecasting lag. When we generate a forecast for all periods

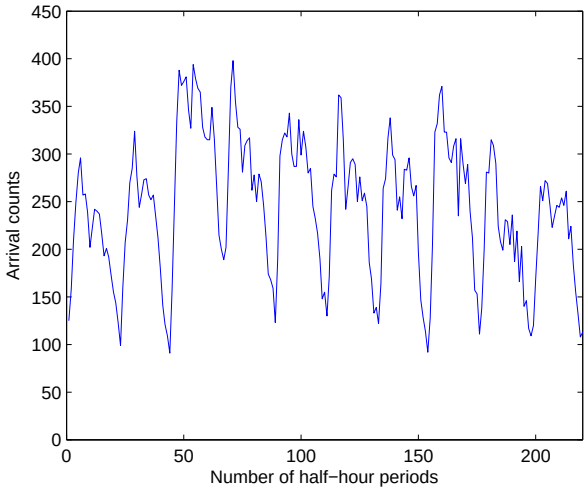

Figure 1: Arrivals for two weeks starting 8/19/10.

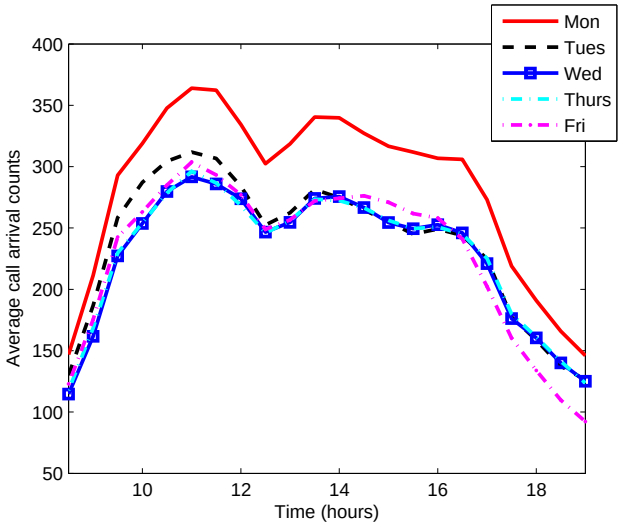

Figure 2: Average arrival counts per half-hour.

of a given day, we roll the learning period forward so as to preserve the length of the forecasting lead time. We re-estimate model parameters after each daily forecast. The results are summarized in Table 2.

We quantify the accuracy of a point prediction using the *mean squared error* (MSE), defined by:

$$\text{MSE} \equiv \frac{1}{K}\sum_{i,j}(X_{i,j} - \hat{X}_{i,j})^2 \ ,$$

where $\hat{X}_{i,j}$ is the value of $X_{i,j}$ predicted by the model, and K is the total number of predictions $\hat{X}_{i,j}$ made. To evaluate the distributional forecasts, we compute the empirical coverage probability of the 95% prediction intervals provided by the model, which is defined as:

$$\text{Cover} = \frac{1}{K}\sum_{i,j}\mathbb{I}(X_{i,j} \in (\hat{L}_{i,j}, \hat{U}_{i,j})),$$

where $\hat{L}_{i,j}$ and $\hat{U}_{i,j}$ are the lower and upper bounds of the 95% prediction interval on $X_{i,j}$. If the model is correct, we expect this coverage to be near 0.95. Note that the TD and HW methods only provide point forecasts, so we do not compute the coverage for those methods.

Table 2 shows that the FE model generates the most accurate forecasts with a lead time of two weeks, consistent with what we said earlier. The ME model performs worse than the FE model in this case: the ME model apparently overfits the data by estimating more parameters. We also find that the TD approach, based on 2-weeks-ahead forecasts made by the company, is outperformed by both the FE and ME models. Finally, the HW approach yields disappointing results. The empirical coverage probabilities for FE and ME are ≈ 0.52 (very bad) and ≈ 0.95 (as desired), respectively. The FE model largely underestimates the uncertainty in the data by not capturing the correlation structure between the arrival counts.

As expected, the superiority of the ME model becomes evident for one-day-ahead forecasts, where the short-term correlation structure can be exploited. The TD approach is also competitive here for point forecasts; it generates the second most accurate forecasts after ME (presumably because the company's forecasts contain valuable additional information). HW smoothing leads to the least accurate forecasts.

Figures 3 and 4 show normal Q-Q plots for the out-of-sample residuals of FE and ME (for the transformed counts), respectively, using a forecasting lead time of half a day. (We also include in the plots corresponding envelopes at the 95% confidence level.) For more detailed results, see Ibrahim and L'Ecuyer (2012). We see that the normal distribution is a better fit for the ME residuals than for the FE residuals. This indicates that ME provides more reliable distributional forecasts.

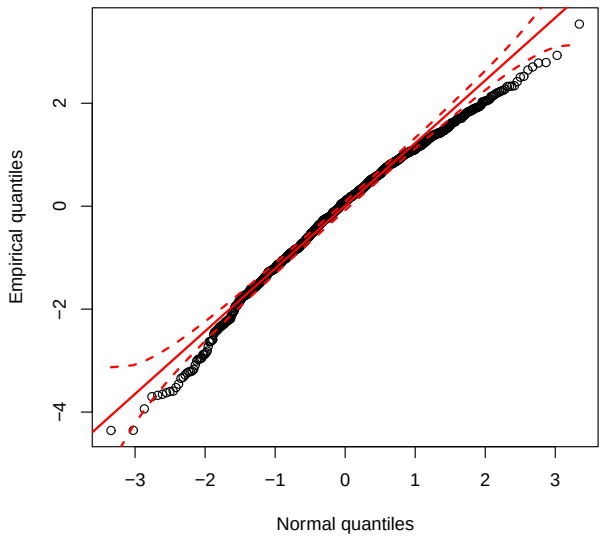

Figure 3: FE Model

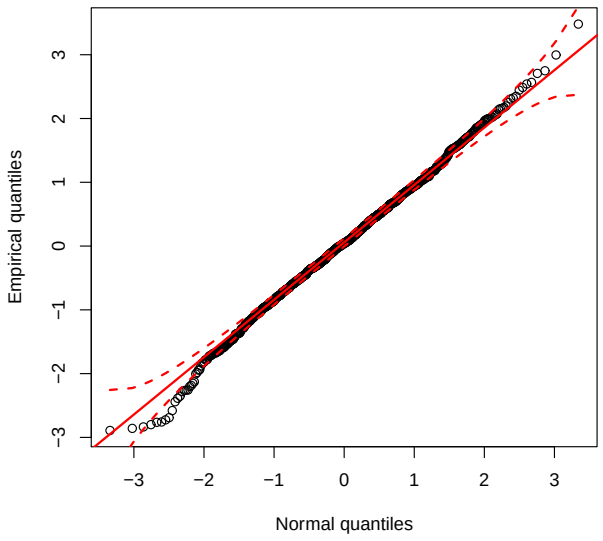

Figure 4: ME Model

Table 2: Accuracy of point and interval predictions for two forecasting lead times.

	Forecast lead time of 14 days				*Forecast lead time of 1 day*			
	ME	TD	FE	HW	ME	TD	FE	HW
MSE	41.7	45.8	40.3	67.0	30.4	33.9	35.7	60.7
Cover	0.96	-	0.52	-	0.95	-	0.50	-

ACKNOWLEDGMENTS

This work has been supported by grants from NSERC-Canada and Hydro-Québec, and a Canada Research Chair, to P. L'Ecuyer, and from US-NSF to Haipeng Shen.

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.

Aktekin, T., and R. Soyer. 2011. "Call Center Arrival Modeling: A Bayesian State Space Approach". *Naval Research Logistics* 58 (1): 28–42.

Aldor-Noiman, S., P. Feigin, and A. Mandelbaum. 2009. "Workload forecasting for a call center: Methodology and a case study". *Annals of Applied Statistics* 3:1403–1447.

Andrews, B., and S. M. Cunningham. 1995. "L.L. Bean improves call-center forecasting". *Interfaces* 25:1–13.

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.

Avramidis, A. N., A. Deslauriers, and P. L'Ecuyer. 2004. "Modeling Daily Arrivals to a Telephone Call Center". *Management Science* 50 (7): 896–908.

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.

Bianchi, L., J. Jarrett, and R. C. Hanumara. 1998. "Improving Forecasting for Telemarketing Centers by ARIMA Modeling with Intervention". *International Journal of Forecasting* 14:497–504.

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* 100:36–50.

Brown, L. D., T. Cai, R. Zhang, L. Zhao, and H. Zhou. 2010. "The root-unroot algorithm for density estimation as implemented via wavelet block thresholding". *Probability Theory and Related Fields* 146:401–433.

Buist, E., and P. L'Ecuyer. 2005. "A Java Library for Simulating Contact Centers". In *Proceedings of the 2005 Winter Simulation Conference*, edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 556–565: IEEE Press.

Channouf, N. 2008. *Modélisation et optimisation d'un centre d'appels téléphoniques: étude du processus d'arrivée*. Ph. D. thesis, Département d'Informatique et de Recherche Opérationnelle, Université de Montréal, Canada.

Channouf, N., and P. L'Ecuyer. 2012. "A Normal Copula Model for the Arrival Process in a Call Center". *International Transactions in Operational Research*. to appear.

Channouf, N., P. L'Ecuyer, A. Ingolfsson, and A. N. Avramidis. 2007. "The Application of Forecasting Techniques to Modeling Emergency Medical System Calls in Calgary, Alberta". *Health Care Management Science* 10 (1): 25–45.

Gans, N., G. Koole, and A. Mandelbaum. 2003. "Telephone Call Centers: Tutorial, Review, and Research Prospects". *Manufacturing and Service Operations Management* 5:79–141.

Gurvich, I., J. Luedtke, and T. Tezcan. 2010. "Staffing Call Centers with Uncertain Demand Forecasts: A Chance-Constrained Optimization Approach". *Management Science* 56 (7): 1093–1115.

Ibrahim, R., and P. L'Ecuyer. 2012. "Forecasting Call Center Arrivals: Fixed-Effects, Mixed-Effects, and Bivariate Models". *Manufacturing and Services Operations Management*. to appear.

Jaoua, A., and P. L'Ecuyer. 2011. "Modeling and assessing the effect of the nonlinear dependence between call types in multi-skill call centers". see https://symposia.gerad.ca/system/documents/0000/0244/jaoua-poster.pdf.

Jongbloed, G., and G. Koole. 2001. "Managing Uncertainty in Call Centers using Poisson Mixtures". *Applied Stochastic Models in Business and Industry* 17:307–318.

Kim, T., P. Kenkel, and B. W. Brorsen. 2012. "Forecasting Hourly Peak Call Volume for a Rural Electric Cooperative Call Center". *Journal of Forecasting* 31:314–329.

Landon, J., F. Ruggeri, R. Soyer, and M. M. Tarimcilar. 2010. "Modeling Latent Sources in Call Center Arrival Data". *European Journal of Operations Research* 204 (3): 597–603.

L'Ecuyer, P. 2006. "Modeling and Optimization Problems in Contact Centers". In *Proceedings of the Third International Conference on Quantitative Evaluation of Systems (QEST'2006)*, 145–154. University of California, Riversdale: IEEE Computing Society.

Mabert, V. A. 1985. "Short Interval Forecasting of Emergency Phone Call (911) Work Loads". *Journal of Operations Management* 5 (3): 259–271.

Mehrotra, V., O. Ozlük, and R. Saltzman. 2010. "Intelligent Procedures for Intra-day Updating of Call Center Agent Schedules". *Production and Operations Management* 19 (3): 353–367.

Pot, A., S. Bhulai, and G. Koole. 2008. "A simple staffing method for multi-skill call centers". *Manufacturing and Service Operations Management* 10:421–428.

Shen, H. 2010a. "Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles: Comments". *International Journal of Forecasting* 58:652–654.

Shen, H. 2010b. "Statistical Analysis of Call-Center Operational Data: Forecasting Call Arrivals, and Analyzing Customer Patience and Agent Service". In *Wiley Encyclopedia of Operations Research and Management Science*, edited by J. J. Cochran. John Wiley.

Shen, H., and J. Z. Huang. 2005. "Analysis of Call Centre Arrival Data Using Singular Value Decomposition". *Applied Stochastic Models in Business and Industry* 21:251–263.

Shen, H., and J. Z. Huang. 2008a. "Forecasting Time Series of Inhomogeneous Poisson Processes with Application to Call Center Workforce Management". *Annals of Applied Statistics* 2 (2): 601–623.

Shen, H., and J. Z. Huang. 2008b. "Interday forecasting and intraday updating of call center arrivals". *Manufacturing and Service Operations Management* 10 (3): 391–410.

Soyer, R., and M. M. Tarimcilar. 2008. "Modeling and Analysis of Call Center Arrival Data: A Bayesian Approach". *Management Science* 54 (2): 266–278.

Steckley, S. G., S. G. Henderson, and V. Mehrotra. 2005. "Performance measures for service systems with a random arrival rate". In *Proceedings of the 2005 Winter Simulation Conference*, edited by M. E. Kuhl, N. M. Steiger, F. B. Armstrong, and J. A. Joines, 566–575: IEEE Press.

Steckley, S. G., S. G. Henderson, and V. Mehrotra. 2009. "Forecast errors in service systems". *Probability in the Engineering and Informational Sciences* 23 (2): 305–332.

Tanir, O., and R. J. Booth. 1999. "Call center simulation in Bell Canada". In *Proceedings of the 1999 Winter Simulation Conference*, 1640–1647. Piscataway, New Jersey: IEEE Press.

Taylor, J. W. 2008. "A Comparison of Univariate Time Series Methods for Forecasting Intraday Arrivals at a Call Center". *Management Science* 54 (2): 253–265.

Weinberg, J., L. D. Brown, and J. R. Stroud. 2007. "Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data". *Journal of the American Statistical Association* 102 (480): 1185–1198.

Whitt, W. 1999. "Dynamic Staffing in a Telephone Call Center Aiming to Immediately Answer All Calls". *Operations Research Letters* 24:205–212.

AUTHOR BIOGRAPHIES

ROUBA IBRAHIM is an Assistant Professor at the Department of Management Science and Innovation of University College London. Her doctorate degree, supervised by Ward Whitt, was completed in 2010 at the Department of Industrial Engineering and Operations Research of Columbia University. Her research interests include stochastic modeling applications in call centers and healthcare systems, and simulation.

PIERRE L'ECUYER is Professor in the DIRO, at the Université de Montréal, Canada. He holds the Canada Research Chair in Stochastic Simulation and Optimization. He is a member of the CIRRELT and GERAD research centers. His main research interests are random number generation, quasi-Monte Carlo methods, efficiency improvement via variance reduction, sensitivity analysis and optimization of discrete-event stochastic systems, and discrete-event simulation in general. He is currently Editor-in-Chief for *ACM Transactions on Modeling and Computer Simulation*, and Associate/Area Editor for *ACM Transactions on Mathematical Software*, *Statistics and Computing*, *International Transactions in Operational Research*, and *Cryptography and Communications*.

NAZIM RÉGNARD is a postdoc in the DIRO, at the Université de Montréal, Canada. His doctorate degree, supervised by Jean-Michel Zakoïan, was completed in 2011 at the EQUIPPE Laboratory of Lille 3 University, France. His research interests are stochastic modeling applications in call centers, financial time series modeling with applications to energy markets and econometric theory of financial time-series.

HAIPENG SHEN is an Associate Professor at the Department of Statistics and Operations Research, University of North Carolina at Chapel Hill. His doctorate degree, supervised by Lawrence D. Brown, was completed in 2003 at the Department of Statistics, the Wharton School of Business, University of Pennsylvania. His research interests include data-driven workforce management problems in labor-intensive service systems, and high-dimensional inference in statistics.