OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0030-364X | EISSN 1526-5463 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Rate-Based Daily Arrival Process Models with Application to Call Centers

Boris N. Oreshkin, Nazim Régnard, Pierre L'Ecuyer
Département d'Informatique et de Recherche Opérationnelle, Pavillon Aisenstadt, Université de Montréal, C.P. 6128, Succ. Centre-Ville, Montréal, Québec, Canada H3C 3J7, oreshkin@iro.umontreal.ca, lecuyer@iro.umontreal.ca

We propose, develop, and compare new stochastic models for the daily arrival rate in a call center. Following standard practice, the day is divided in time periods of equal length (e.g., 15 or 30 minutes), the arrival rate is assumed random but constant in time in each period, and the arrivals are from a Poisson process, conditional on the rate. The random rate for each period is taken as a deterministic base rate (or expected rate) multiplied by a random busyness factor having mean 1. Models in which the busyness factors are independent across periods, or in which a common busyness factor applies to all periods, have been studied previously. But they are not sufficiently realistic. We examine alternative models for which the busyness factors have some form of dependence across periods. Maximum likelihood parameter estimation for these models is not easy, mainly because the arrival rates themselves are never observed. We develop specialized techniques to perform this estimation. We compare the goodness-of-fit of these models on arrival data from three call centers, both in-sample and out-of-sample. Our models can represent arrivals in many other types of systems as well. Estimating a model for the vector of counts (the number of arrivals in each period) is generally easier than for the vector of rates, because the counts can be observed, but a model for the rates is often more convenient and natural, e.g., for simulation. We examine and provide insight on the relationship between these two types of modeling. In particular, we give explicit formulas for the relationship between the correlation between rates and that between counts in two given periods, and for the variance and dispersion index in a given period. These formulas imply that for a given correlation between the rates, the correlation between the counts is much smaller in low traffic than in high traffic.

Key words: arrival process; arrival rate; doubly stochastic Poisson process; input modeling; copula; correlation; call center

1. Introduction

The randomness of customer arrivals is a prime source of uncertainty in service systems such as restaurants, retail stores, emergency services, and call centers, to name a few. In those systems,

customers (or demands) arrive according to stochastic processes whose intensity (or rate) varies with time in a stochastic way, often influenced by external events that are not always predictable, and are generally difficult to model in a realistic way. This modeling is nevertheless essential to study the performance of these systems and manage them effectively.

In this paper, we are concerned with modeling the arrival process in a call center for one day of operation. Call centers (or contact centers) are a key component of many organizations. They employ several million people in North America alone, and much of their operating costs is to pay the agents who answer the calls (Gans et al. 2003, Akşin et al. 2007, Koole 2013). To optimize the staffing and work schedules of these agents, good models are required to forecast the call arrival volumes (the demand) and also to simulate the detailed operations of the call centers. Most large call centers are indeed complicated stochastic systems whose realistic models can only be handled via stochastic discrete-event simulation (Mehrotra 1997, Avramidis and L'Ecuyer 2005, Buist and L'Ecuyer 2005, 2012, Ibrahim et al. 2015a,b).

Our discussion here targets call centers, but our models could apply to many other settings, such as customer arrivals in retail stores, basket arrivals to cashiers in grocery stores, emergency arrivals in healthcare services, demands for ambulances, for taxis, for pizza deliveries, demands for a specific product in a store or online, party arrivals in restaurants, and many more. We leave it to the readers to test how well our proposed models can fit data sets from these other areas.

Call arrivals can usually be assumed independent over a short time scale, because they are initiated by individuals who make decisions (approximately) independently in the short term. For a given expected number of arrivals within a selected minute, say, the calls typically arrive (approximately) independently of each other. It is then natural and quite standard to model arrivals by a Poisson process, which is equivalent to assume that the arrivals occur one by one, independently of each other, conditional on the arrival rate. There are many systems where this modeling choice makes sense, at least to a good approximation, including call centers. It is supported mathematically by the Poisson superposition theorem and is ubiquitous in all the work on the modeling of call centers and other similar service systems. For detailed justifications and some examples, we refer the reader to Whitt (2001), Brown et al. (2005), Koole (2013), Kim and Whitt (2014b), Ibrahim et al. (2015b) and the references given there. There are a few situations were call center arrivals can hardly be seen as Poisson, e.g., when dozens of people call the police or ambulance at almost the same time for the same accident, or if many people call to order an item immediately after seeing a tv commercial for that item (Soyer and Tarimcilar 2008). These arrivals can be modeled by a Poisson process whose arrival rate has large narrow peaks once in a while, when such events occur. We do not model these special types of bursts in this paper.

Empirical evidence shows that if arrivals are from a Poisson process, the arrival rate must change with time and also be stochastic. Such evidence is given later in this paper and also in Tanir and Booth (1999), Deslauriers (2003), Avramidis et al. (2004), Brown et al. (2005), Steckley et al. (2009), Channouf and L'Ecuyer (2012), Ibrahim et al. (2015b) and references therein. If the arrival rate is taken as a deterministic function of time, the Poisson process model implies that the variance and the mean of the number of arrivals in any given time period are equal. This disagrees with what is observed in call center data (and for many other systems): the variance of daily arrival counts is typically larger than the mean, and often much larger. This is because the arrival rate changes and depends on several factors that are too hard to predict.

A simple way of defining a stochastic arrival rate over a given time period is to assume that a deterministic base rate over that period is multiplied by a single random factor with mean one; see Whitt (1999b) and Avramidis et al. (2004). When the time period is one day, the base rate as a function of time is called the *daily profile* and the random factor is the *busyness factor* for the day. With a single factor for the day, however, the arrival rates over any two disjoint periods of the day are perfectly correlated, which is more correlation than what is implied by observed data. At the other extreme, Jongbloed and Koole (2001) proposed a model in which the day is divided into periods, each period has its own random busyness factor, and these factors are independent. Call center data strongly disagrees with this independence assumption: the arrival rates over disjoint periods are typically positively correlated. The explanation is that factors that affect the arrival rates (e.g., weather conditions, etc.) typically span over several periods of the day, so larger call volumes in the morning are often associated with larger call volumes in the afternoon, for example. Neglecting this dependence leads to an underestimation of the queue build up process and of waiting times. Our aim is to develop more realistic arrival rate models that are between these two extremes and for which the means, variances, and correlations of arrival counts better match those observed in data. In the models we consider, the arrivals are Poisson with a stochastic rate. Under such models, the variance of the arrival counts in any given period cannot be smaller than the mean; i.e., underdispersion is not possible. We have never observed underdispersion, or negative correlations between periods, in call center arrival data.

The daily profile can be taken in principle as any fixed function of time. Although a continuous function may appear more realistic, the most popular choice by far is a piecewise-constant function, for which the day is divided into time periods of equal length (usually 30 or 15 minutes) and the arrival rate is assumed constant over each period (Gans et al. 2003, Avramidis et al. 2004, Brown et al. 2005, Akşin et al. 2007, Channouf and L'Ecuyer 2012, Koole 2013, Kim and Whitt 2014a, Ibrahim et al. 2015b). There are many reasons for this. For one, most call center managers determine their required staffing by using approximations via Erlang formulas. For each time period, they

would compute how many agents they need to achieve a given target performance measure (e.g., 80% of the calls answered within 20 seconds, or less than 3% abandonment, etc.) by assuming a steady-state model over that period and using the Erlang formula to approximate the performance. This can be done even if the arrival rate over the period is random with a known distribution. The random arrival rates do not have to be independent across periods, there can be global performance targets for the day, and also multiple call types (Gans et al. 2003, Harrison and Zeevi 2005, Atlason et al. 2008, Ceẓik and L'Ecuyer 2008, Avramidis et al. 2009a, Gurvich et al. 2010, Koole 2013). If one wishes to estimate the performance by a more detailed simulation instead of queuing formulas, it is easier and faster to simulate a Poisson process with piecewise-constant rate than one whose rate changes continuously. For the former it suffices to generate independent exponential inter-arrival times in each period, and eventually reschedule the next arrival time when the arrival rate changes at a period boundary (Buist and L'Ecuyer 2005, Ibrahim et al. 2012). A third justification for using piecewise-constant rates is that exact call arrival times are rarely available in call center data. Typically, the available data is in the form of number of arrivals (the arrival *counts*) in each time period, and the call by call arrival process model must be constructed based on this data only. Moreover, agents can typically be added or removed only at period boundaries.

It is of course possible to construct models in which the arrival rate changes continuously. For example, Kim and Whitt (2014a) consider a piecewise-linear approximation and compare with a piecewise-constant rate. Channouf (2008) developed a methodology that uses smoothing splines to model the arrival rate function, together with a single random busyness factor for the day. The process is simulated via a thinning technique. His methodology can estimate the parameters from aggregated data (arrival counts per period) and is implemented in the ContactCenters software (Buist and L'Ecuyer 2005, 2012). In numerical experiments with real data from three different call centers, Channouf (2008) observed a small difference in simulated performance measures (a small improvement) when switching from a piecewise-constant rate to a spline rate, both with a single random busyness factor for the day. Note that if the (smooth) spline base rate is multiplied by busyness factors that differ across periods, the resulting rate function will no longer be continuous. One could think of a model in which some base rate is multiplied by different random busyness factors in different periods and the smoothing spline is fitted to the resulting random rates afterward, or models in which both the base rate and the random busyness factor are continuous functions of time, probably parameterized, and whose parameters would be fitted to call-by-call arrival times. These types of models are beyond our scope; we leave them for future work.

Assuming a piecewise-constant rate is reasonable if the arrival rate does not change too rapidly and the time periods are taken small enough so the arrival rate is approximately constant in each period. Kim and Whitt (2014b,a) have studied this issue and developed tests for the Poisson

process assumption (with either deterministic or random rates). When the individual arrival times are available, their tests can be used to select an appropriate period length in which the rate can be taken as approximately constant. They also provide practical guidelines for selecting this length and discuss various other related issues. In their tests with real-life data, they found that the Poisson assumption was quite reasonable with 30-minute periods for call center data, and with 60-minute periods for arrivals to a hospital emergency system. The choice of period length may also depend on how much data is available to estimate the rate for each period: if the periods are taken too short, the estimates can be too noisy and we can end up with an overfitting problem. The methodology developed in this paper works for an arbitrary period length.

We focus on modeling the call center arrival process over one day of operation. This is useful for simulating the call center one day at a time, more than a few days before the days that are simulated, so the dependence between the currently available data and the days that are simulated can be seen as negligible. This type of setting occurs when managers construct work schedules one or more weeks in advance. The predictive power of known time-series forecasting models in this scenario is weak, because the target day is too far ahead (Avramidis et al. 2004, Ibrahim and L'Ecuyer 2013). These models can be useful not only for simulation, but also to obtain distributional forecasts that can be used in other algorithms or formulas.

For the day-to-day management of real-life call centers, one would also need to model the dependence across successive days, the seasonalities at weekly and yearly levels, special-day effects, and in many cases the dependence between different call types (Brown et al. 2005, Jaoua et al. 2013, Ibrahim et al. 2015b). This is beyond the scope of the present paper, but could be done in combination with the models proposed here. Shen (2010) and Ibrahim et al. (2015b) give overviews of existing models, which typically have a regression or time series flavor. For example, Weinberg et al. (2007) used a Bayesian approach to sample from the forecast distributions based on a linear regression data model, while Ibrahim et al. (2012), Aldor-Noiman et al. (2009), Brown et al. (2005) used different variants of linear regression models to produce point forecasts of daily call volumes. One common characteristic of these papers is that they all use different variants of the root-unroot variance stabilizing transformation proposed by Brown et al. (2001), which approximates the square root of the sum of 1/4 and a Poisson random variable with large mean, minus the square root of the mean, by a normal random variable with mean 0 and variance 1/4. However, this approximation is often questionable because (i) the expected count in a given time period is not always large and (ii) the arrival counts are typically not Poisson, but have over-dispersion compared with the Poisson distribution (the variance is larger than the mean). Most of these works focus on the *point forecasting* of call volumes, i.e., estimating the expectation, to plug it eventually into an Erlang formula to determine the required number of agents, rather than distributional forecasts.

Models for the distribution of the vector of arrival counts over a given day, with the day divided into equal-length periods, have also been proposed. Avramidis et al. (2004) introduced two such models. In the first one, the vector of counts has a negative multinomial distribution whose parameters have a (multivariate) Dirichlet distribution. In the second model, the total number of arrivals during the day has an arbitrary distribution (they take the gamma distribution in their experiments) and the vector of proportions of arrivals in each period has an independent Dirichlet distribution. The coordinates of the resulting vector of "counts" must then be rounded to obtain integer counts. These models are more flexible and improve the matching of correlations compared with the model with a single busyness factor that multiplies the daily profile, but the correlations still remain too strong. Channouf and L'Ecuyer (2012) illustrate this with real call center data and propose a general multivariate distribution model for the vector of counts in which the marginal distributions are specified individually to match the distribution of counts in each period, and the dependence between the counts is modeled separately via a Gaussian copula, that matches (approximately) the pairwise rank correlations. They also have versions in which the correlation matrix of the copula is parameterized, to reduce the number of model parameters. Their model provides a much better match to data even after accounting for the fact that it has more parameters (via the Akaike information criterion). These authors model the vector of arrival *counts*, whereas in the present work we want to model the vector of arrival *rates*.

With a distributional model for the *counts*, assuming a piecewise constant rate, one can simulate the arrivals by first generating the number of arrivals (the count) in each period, and then generating the arrival times uniformly and independently over the given time period. This is consistent with the assumption that arrivals are from a Poisson process with unknown (perhaps random) constant rate over the period. But this is much less convenient and efficient than generating the arrivals one by one directly from the constant rate when this rate is known (Ibrahim et al. 2012). The former requires generating, storing, and sorting all arrival times in the period before doing the discrete-event simulation, whereas with the latter we only need to store the next arrival time. Our preference is therefore for distributional models of the rates. In a simulation, we can then first generate the (random) arrival rates for all periods of the day and then run the simulation with those rates. Another important motivation for the rate-based models is that many staffing models need the distribution of the rate as one of their inputs (Gans et al. 2003).

In this paper, the day is divided into periods of equal length and the arrivals are assumed to be from a Poisson process with constant rate in each period. All our models are developed for the setting in which data are aggregated in the form of arrival counts per period. For recommendations on how to select the period length when detailed call by call arrival times are available, we refer the reader to Kim and Whitt (2014b,a). We start from two simple models mentioned previously,

in which a deterministic daily profile for the rate is multiplied either by a single busyness factor for the day, or by independent busyness factors, one for each period. Our first idea is to combine these two models: we take one random local busyness factor for each period plus a global one for the day. This provides more freedom to control the correlations. As in Jongbloed and Koole (2001) and Avramidis et al. (2004), our busyness factors have a gamma distribution. Then we generalize this to a model in which the global gamma factor is raised to a different power (and normalized) in each time period, which gives further flexibility in matching the correlations. Finally, we propose a model in which the multivariate gamma random vector of busyness factors, rather than the vector of counts as in Channouf and L'Ecuyer (2012), is determined by a Gaussian (normal) copula. This gives even more flexibility to match both the correlations and the variance within each period, at the expense of having many more parameters to estimate. To reduce (and control) the number of parameters, we consider variants of this model for which the correlation matrices are restricted to classes having a special structure. We compare the goodness of fit of these different models on real data sets.

For the models of Jongbloed and Koole (2001), Avramidis et al. (2004), and Channouf and L'Ecuyer (2012), parameter estimation was relatively easy, via maximum likelihood and correlation matching. For our new models, it is much harder, mainly because we do not observe the arrival rates themselves, but only the counts, which give only indirect information on the rates. An important part of our contribution is to develop viable methods to estimate the parameters for all the proposed models, via maximum likelihood.

The rest of the paper is organized as follows. In Section 2, we define our general setting with busyness factors and piecewise constant arrival rates, and we examine some of its properties. In Section 3 we introduce a two-level busyness factor model, which combines a single busyness factor for the day and a busyness factor for each period. Section 4 generalizes this model by including period-wise exponentiation of the daily busyness factors. In Section 5, we model the dependence structure in the vector of stochastic arrival rates via a normal copula. Estimation methods for these models are developed in the Online Supplement. They constitute an important part of our contribution. Section 6 reports the results of our experiments. Additional results and plots are given in the Online Supplement. All sections whose "numbers" start by a letter from A to E are in the Online Supplement.

2. General Setting, Notation, and Relationships Between Rates and Counts

We consider one day of operation of a call center. The opening hours are divided into p *time periods* of equal length. For example, if the center receives calls from 8:00 to 18:00 and the periods are 30 minutes long, we have $p = 20$. Let $\mathbf{X} = (X_1, \ldots, X_p)$ be the vector of arrival counts in those p periods.

We assume that the arrivals are from a Poisson process with a random rate Λ_j, constant over period j. To simplify the notation, the time unit for this rate is assumed to be one period, i.e., the rate is expressed in (expected) number of arrivals per period. The vector $\mathbf{\Lambda} = (\Lambda_1, \ldots, \Lambda_p)$ can have an arbitrary multivariate distribution over $[0, \infty)^p$. Taking its mean $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ as a scaling factor (or *base rate*), we shall assume that $\Lambda_j = B_j \lambda_j$ where B_j is a non-negative random variable with $\mathbb{E}[B_j] = 1$ for each j. We call B_j the *busyness factor* for period j and we denote $\mathbf{B} = (B_1, \ldots, B_p)$. Conditional on $\mathbf{\Lambda}$, the X_j's are independent and each X_j has a Poisson distribution with mean Λ_j. To summarize, we have

$$\Lambda_j = B_j \lambda_j \quad \text{and} \quad X_j \sim \text{Poisson}(\Lambda_j), \tag{1}$$

where Poisson(λ) denotes the Poisson distribution with mean λ. By taking the logarithm on each side of the equation in (1), we obtain a linear model with mixed effects. However, this model is non-standard because the rates Λ_j are hidden and can be inferred only indirectly through the counts X_j, and its parameters are often hard to estimate for this reason. Note that this setting does not capture trends or changes that might occur over time frames longer than a day.

The mean and covariance matrix of the counts $\mathbf{X}$ can be expressed in terms of those of the rates $\mathbf{\Lambda}$ as follows. For each j, we have $\mathbb{E}[X_j] = \mathbb{E}[\Lambda_j] = \lambda_j$ and

$$\text{Var}(X_j) = \mathbb{E}[\text{Var}[X_j|B_j]] + \text{Var}[\mathbb{E}[X_j|B_j]] = \mathbb{E}[B_j \lambda_j] + \text{Var}(B_j \lambda_j) = \lambda_j (1 + \lambda_j \, \text{Var}(B_j)). \tag{2}$$

The *coefficient of dispersion*, or *dispersion index* (DI), defined as the ratio of variance to the mean (Cox and Lewis 1966), is then

$$\text{DI}(X_j) = \frac{\text{Var}(X_j)}{\lambda_j} = 1 + \lambda_j \, \text{Var}(B_j) \geq 1. \tag{3}$$

We use the DI rather than the coefficient of variation to measure the relative variability because it better indicates the overdispersion with respect to the Poisson distribution. Eq. (3) shows in particular that the variance can never be smaller than the mean under a Poisson process model. When $\text{Var}(B_j) = 0$, X_j has a Poisson distribution with mean λ_j, so $\text{DI}(X_j) = 1$. More interestingly, for a fixed value of $\text{Var}(B_j)$, $\text{DI}(X_j)$ increases linearly with λ_j, and $\text{DI}(X_j) \to 1$ when $\lambda_j \to 0$. This means that under this general model in which the rate λ_j is multiplied by a single factor over the time interval, the arrival process behaves pretty much like a Poisson process over very short time intervals, for which λ_j is small, and the overdispersion increases with the length of the time interval, because λ_j is then larger. In particular, if the base rate is multiplied by a single factor over the entire day, the count for the entire day under this model will typically have a much larger DI than that for one hour or one half-hour.

We also introduce a standardized version of the DI (SDI), defined as

$$\mathrm{SDI}(X_j) = \frac{\mathrm{DI}(X_j) - 1}{\lambda_j}, \tag{4}$$

which under our model is equal to $\mathrm{Var}(B_j)$, so it measures the variability of the rate independently of λ_j. Note that if we merge two periods j and k having a common busyness factor $B = B_j = B_k$, the SDI of the merged periods is the same as that of the original ones: $\mathrm{SDI}(X_j + X_k) = \mathrm{SDI}(X_j) = \mathrm{SDI}(X_k) = \mathrm{Var}(B)$, and this applies as well to more than two periods. Thus, looking at how the SDI behaves when we merge periods permits one to test the dependence between their busyness factors. In particular, for the model with a single common busyness factor $B = B_1 = \cdots = B_p$, the SDI always remains the same when we merge periods. In general, we have

$$\begin{aligned}
\mathrm{SDI}(X_j + X_k) &= \frac{\lambda_j^2 \mathrm{Var}(B_j) + \lambda_k^2 \mathrm{Var}(B_k) + 2\lambda_j\lambda_k \mathrm{Cov}(B_j, B_k)}{(\lambda_j + \lambda_k)^2} \\
&\le \frac{(\lambda_j^2 + \lambda_k^2 + 2\lambda_j\lambda_k)\max(\mathrm{Var}(B_j), \mathrm{Var}(B_k))}{(\lambda_j + \lambda_k)^2} \\
&= \max(\mathrm{Var}(B_j), \mathrm{Var}(B_k)) \;=\; \max(\mathrm{SDI}(X_j), \mathrm{SDI}(X_k)),
\end{aligned} \tag{5}$$

with equality holding when $\mathrm{Corr}(B_j, B_k) = 1$. This generalizes to more than two periods. On the other hand, one can have $\mathrm{SDI}(X_j + X_k) < \min(\mathrm{Var}(B_j), \mathrm{Var}(B_k)) = \min(\mathrm{SDI}(X_j), \mathrm{SDI}(X_k))$, so the SDI of merged periods can be smaller than the smallest SDI of those periods. For example, if $\mathrm{Var}(B_j) = \mathrm{Var}(B_k) > 0$ and $\mathrm{Corr}(B_j, B_k) < 1$, then

$$\mathrm{SDI}(X_j + X_k) = \frac{\mathrm{Var}(B_j)[\lambda_j^2 + \lambda_k^2 + 2\lambda_j\lambda_k \mathrm{Corr}(B_j, B_k)]}{(\lambda_j + \lambda_k)^2} < \mathrm{Var}(B_j) = \min(\mathrm{SDI}(X_j), \mathrm{SDI}(X_k)).$$

For $j \neq k$, we also have

$$\mathrm{Cov}(X_j, X_k) = \mathbb{E}[(B_j\lambda_j)(B_k\lambda_k)] - \lambda_j\lambda_k = \lambda_j\lambda_k \mathrm{Cov}(B_j, B_k) \tag{6}$$

and

$$\begin{aligned}
\mathrm{Corr}(X_j, X_k) &= \frac{\mathrm{Cov}(B_j, B_k)}{[(\mathrm{Var}(B_j) + 1/\lambda_j)(\mathrm{Var}(B_k) + 1/\lambda_k)]^{1/2}} \\
&= \frac{\mathrm{Corr}(B_j, B_k)}{[((1 + 1/(\mathrm{Var}(B_j)\lambda_j))(1 + 1/(\mathrm{Var}(B_k)\lambda_k))]^{1/2}}.
\end{aligned} \tag{7}$$

For a fixed distribution of $\mathbf{B} = (B_1, \ldots, B_p)$, we have $\mathrm{Corr}(X_j, X_k) \to 0$ when $\lambda_j \to 0$ or $\lambda_k \to 0$, whereas $\mathrm{Corr}(X_j, X_k) \to \mathrm{Corr}(B_j, B_k)$ when both $\lambda_j \to \infty$ and $\lambda_k \to \infty$. That is, the process behaves like a Poisson process over short time periods (the correlation between the counts over short disjoint periods is near zero), while the correlation is higher over larger disjoint time periods. For fixed values of λ_j and λ_k, $\mathrm{Corr}(X_j, X_k)$ is close to $\mathrm{Corr}(B_j, B_k)$ when B_j and B_k have large variance and is smaller otherwise.

Three special cases of this model, studied earlier, are defined as follows. We will use them for comparison with our new models.

The first (simplest) case is the degenerate case where $B_j = 1$ for all j, so $\mathrm{Var}(B_j) = 0$ and $\mathrm{Corr}(X_j, X_k) = 0$. It gives an ordinary nonhomogenous Poisson arrival process with piecewise constant rate, as used in Brown et al. (2005) for example. We will refer to this case simply as *Poisson*. It is well known that this type of model is unacceptable because it typically underestimates the variability of the counts in a significant way; see, e.g., (Jongbloed and Koole 2001, Avramidis et al. 2004, Steckley et al. 2009, Shen 2010).

In the second special case, introduced by Whitt (1999a), one takes $B_j = B$ for all j, where $\mathbb{E}[B] = 1$, so the base rate is multiplied by the same random busyness factor for the entire day. Then the SDI over any union of periods is $\mathrm{Var}(B)$. Avramidis et al. (2004) further studied this model in the case where B has a gamma distribution with $\mathrm{Var}(B) = 1/\gamma$, which can take any value in the range $(0, \infty)$, and $\mathrm{Corr}(B_j, B_k) = 1$. In this case, the vector $\mathbf{X}$ has a negative multinomial distribution whose parameters are easy to estimate from the counts. We call this model *Poisson-gamma-single-factor*, or simply *PGsingle*. Two important drawbacks of this model are that (i) it tends to overestimate the positive correlation between the counts in different periods and (ii) it does not fit the variance equally well for all the periods of the day (see Avramidis et al. 2004 and Channouf and L'Ecuyer 2012). These problems are due to the fact that there is a single busyness factor common to all periods of the day, and (given the λ_j's) a single parameter γ to be chosen that determines all the variances and correlations.

The third simplified setting, from Jongbloed and Koole (2001), uses independent busyness factors B_j for the different periods of the day. This gives $\mathrm{Corr}(X_j, X_k) = 0$. These authors use the gamma distribution for the B_j and show that the parameters are easy to estimate by maximum likelihood. We refer to this model as *Poisson-gamma-independent*, or *PGindep*. It has the important limitation of neglecting the dependence (often strong) between counts across different periods.

In the remainder, we propose new instances of the general arrival process model outlined earlier, which allow more flexibility for matching the variances within periods and correlations between periods. First, we combine the PGsingle and PGindep models into a two-level busyness factor model that includes both a daily busyness factor and a busyness factor per period. We then further extend this model by introducing an exponentiation of the daily busyness factor in every period. These new models are more flexible than the previous ones and they contain only a few extra parameters compared to the PGindep model. Then we propose a normal copula model for $\mathbf{B}$ which is even more flexible to fit the variances and correlations of the counts, at the expense of having more parameters.

3. Two-Level Busyness Factor Model

We consider the following two-level arrival process model, named PG2, based on the multiplicative combination of independent period busyness factors $\tilde{B}_j$ and the busyness factor for the day, $\bar{B}$. Let Gamma(a, b) denote a gamma distribution with mean a/b and variance a/b^2. We assume that $\bar{B}, \tilde{B}_1, \ldots, \tilde{B}_p$ are independent with

$$\bar{B} \sim \text{Gamma}(\beta, \beta) \quad \text{and} \quad \tilde{B}_j \sim \text{Gamma}(\alpha_j, \alpha_j) \text{ for each } j, \tag{8}$$

for some positive parameters $\beta, \alpha_1, \ldots, \alpha_p$, and we take

$$B_j = \tilde{B}_j \bar{B} \tag{9}$$

as the busyness factor of period j. This combination permits one to better control the correlation between the B_j's, in comparison with the previous special cases where it was either 0 or 1.

Simple formulas are available for the moments in this model:

$$\text{Var}(B_j) = \frac{(1+\beta+\alpha_j)}{\beta\alpha_j} \quad \text{and} \quad \text{Cov}(B_j, B_k) = \frac{1}{\beta}.$$

Then, $\text{Var}(X_j)$ and $\text{Cov}(X_j, X_k)$ are easily obtained from (2) and (6). Table 1 summarizes some statistical properties of this model and compares with the simpler special cases. We see how the additional terms for this PG2 model provide more flexibility to match the variances and correlations. Since we have simple formulas for the moments, moment-matching estimators (MMEs) for this model are easy to compute (see Section A.1). However, in experiments where we generated data sets by simulation from the model with known parameters, and then estimated these parameters by moment matching, we found that these MMEs often returned values far off the correct ones, which indicates a lack of accuracy and robustness.

Table 1 Some statistical properties (moments) for the PGsingle, PGindep, and PG2 models.

Model	$\mathbb{E}[X_j]$	$\text{Var}(X_j)$	$\text{Corr}(B_j, B_k)$	$\text{Corr}(X_j, X_k)$
PGsingle	λ_j	$\lambda_j + \lambda_j^2/\beta$	1	$[(1+\beta/\lambda_j)(1+\beta/\lambda_k)]^{-1/2}$
PGindep	λ_j	$\lambda_j + \lambda_j^2/\alpha_j$	0	0
PG2	λ_j	$\lambda_j + \frac{(1+\beta+\alpha_j)\lambda_j^2}{\beta\alpha_j}$	$[(1+\frac{1+\beta}{\alpha_j})(1+\frac{1+\beta}{\alpha_k})]^{-1/2}$	$[(1+\frac{\beta}{\lambda_j}+\frac{1+\beta}{\alpha_j})(1+\frac{\beta}{\lambda_k}+\frac{1+\beta}{\alpha_k})]^{-1/2}$

Maximum likelihood estimators (MLEs) are generally more robust and accurate. However, while these estimators were readily available for the previous special cases, here they are much harder to compute. More specifically, the available expression for the density of X_j, which appears in the likelihood function for each j, involves an integral with respect to the realization of the vector $\mathbf{B}$ of unobserved daily busyness factors (see (3) in Section A.2) and we do not know how to evaluate this

integral in closed form. We opted to develop parameter estimation methods for this model based on Monte Carlo estimation of the log-likelihood function. The stochastic optimization algorithm that we use to approximate the MLEs for the PG2 model is described and studied in Section A.2. When comparing the MLEs with the MMEs using simulation experiments as described earlier, using mean square error in the parameter estimation, for various sets of parameter values, MLEs were always clear winners.

In this PG2 model, the parameters $\alpha_1, \ldots, \alpha_p$ can be specified independently from each other, without any functional relationship between them. Alternatively, one may impose as an additional constraint that α_j as a function of j belongs to some class of smooth functions, e.g., a cubic spline. This can provide a more realistic model in the situation where α_j does not vary with j too wildly on a given day. Forcing α_j to obey a spline as a function of j is a way to reduce the overfitting of the model. We will examine the smoothing spline variant of the model named *PG2sp* in our case studies. The additional ingredients required to compute the MLE for the PG2sp model are described in Section A.6.

4. Extended Two-Level Busyness Factor Model

The PG2 model of Section 3 is more flexible than PGindep and PGsingle, but on close examination we see that in comparison with PGindep, it has only one additional parameter β, so the additional flexibility in matching all the variances and correlations is still limited. This is because the busyness factor $\bar{B}$ for the day affects all the periods in exactly the same way. To remove this restriction, and to add flexibility in matching the correlations, here we raise the factor $\bar{B}$ to some power p_j in each period j, where the exponents p_j's may differ across periods, and we normalize so that the expectation of $\bar{B}^{p_j}$ remains equal to 1 in each period. The exponent permits one to modulate the impact of $\bar{B}$ differently across periods. For example, it could cover a situation where the busyness factor for the day affects the arrival rates much more strongly in the middle of the day than in the evening. According to our data, such types of situations do occur.

This yields the following model, which we call *PG2pow*:

$$B_j = \tilde{B}_j \bar{B}^{p_j} / \gamma(p_j) \tag{10}$$

where $\gamma(p_j) = \mathbb{E}[\bar{B}^{p_j}] = \beta^{-p_j}\Gamma(p_j + \beta)/\Gamma(\beta)$ is the appropriate normalization constant. This model remains in the class of gamma-Poisson processes as the distribution of $\bar{B}^{p_j}$ belongs to the class of generalized gamma. For any fixed value of $\beta > 0$, when $p_j \to 0$ we have $\mathrm{Var}(\bar{B}^{p_j}) \to 0$ and $\bar{B}^{p_j}$ becomes degenerate at 1, while when $p_j \to \infty$, $\mathrm{Var}(B^{p_j}) \to \infty$. Therefore, the impact of the daily busyness factor $\bar{B}$ on period j can be made arbitrarily small by decreasing p_j, eventually completely

decorrelating this period j from the rest of the day, and arbitrarily large by increasing p_j. The variances and covariances of the counts are given by

$$\mathrm{Var}(X_j) = \lambda_j + \lambda_j^2 \left[\frac{(1+\alpha_j)\Gamma(\beta)\Gamma(2p_j+\beta)}{\alpha_j \Gamma(p_j+\beta)^2} - 1 \right], \tag{11}$$

$$\mathrm{Cov}(X_j, X_k) = \lambda_j \lambda_k \left[\frac{\Gamma(\beta)\Gamma(p_j+p_k+\beta)}{\Gamma(p_j+\beta)\Gamma(p_k+\beta)} - 1 \right]. \tag{12}$$

We see that if either $p_j = 0$ or $p_k = 0$, then $\mathrm{Cov}(X_j, X_k) = \mathrm{Corr}(X_j, X_k) = 0$. Thus, in this extended model, the correlations and the variances can be further disentangled compared to the PG2 model.

Parameter estimation for this model can be performed using techniques similar to those used for the PG2 model. Further details are given in Section B.

5. A normal copula for the vector of rates

The most general way of modeling the distribution of $\mathbf{B} = (B_1, \ldots, B_p)$ is to select an arbitrary marginal distribution for each B_j, and to model the dependence by a copula (Nelsen 1999). Here we propose a model that captures the dependence by a normal copula. The advantages of using a normal copula are that it can match (approximately) all the correlations between the X_j's, the parameters are not too hard to estimate even when the dimension p is large, and it is not difficult to generate the vector $\mathbf{B}$ from this model. The resulting model has much more flexibility to match the variances and correlations than the previous ones, at the expense of having many more parameters. This type of model based on a normal copula is also known as a NORTA (NORmal To Anything) model (see, e.g., Hörmann et al. 2004, Avramidis et al. 2009b). We call it *PGnorta*. Channouf and L'Ecuyer (2012) also proposed a normal copula model, but it was to model directly the vector $\mathbf{X}$ of arrival counts instead of the vector $\mathbf{B}$ as we do here. Estimating the copula parameters is more difficult in our case because $\mathbf{B}$ is unobserved. The marginal distributions of each B_j can be arbitrary over $[0, \infty)$. In our development and experiments, we use a gamma distribution with mean 1, as in the previous models.

5.1. Normal copula with arbitrary correlation matrix and gamma marginals

We assume that each B_j has a Gamma(α_j, α_j) distribution, with cumulative distribution function (cdf) G_j. Each α_j's is estimated individually by MLE as in the PGindep model. The dependence between the B_j's is modeled by a normal copula, defined as follows. Recall that a *copula* is a multivariate distribution whose marginals are all uniform over the interval $(0, 1)$. A *normal copula* in p dimensions is a special type of copula that can be specified by selecting an arbitrary (valid) $p \times p$ correlation matrix $\mathbf{R}^{\mathrm{Z}}$. To generate a random vector $\mathbf{U}$ from that copula, we generate $\mathbf{Z} = (Z_1, \ldots, Z_p)$ from the multinormal distribution with mean zero and covariance matrix $\mathbf{R}^{\mathrm{Z}}$, then we

put $\mathbf{U} = (U_1, \ldots, U_p) = (\Phi(Z_1)), \ldots, \Phi(Z_p))$ where Φ is the standard normal cdf. Then to generate $\mathbf{B}$ from this copula, we simply put $B_j = G_j^{-1}(U_j) \stackrel{\text{def}}{=} \inf\{x \in \mathbb{R} : G_j(x) \geq U_j\}$ for all j. The choice of $\mathbf{R}^{\text{Z}}$ defines implicitly the covariance matrix of $\mathbf{B}$, which in turn determines the covariance matrix of $\mathbf{X}$. We want to choose $\mathbf{R}^{\text{Z}}$ to match the empirical correlations for $\mathbf{X}$ observed in the data. As generally recommended because it is more robust (Hörmann et al. 2004, Avramidis et al. 2009b), we want to match the Spearman (or rank) correlations between the X_j's. With this approach, the modeling of marginal distributions and correlations is highly decoupled, since the correlations are estimated separately from the marginals. The variances and correlations of the X_j's can be matched very closely, as with the model of Channouf and L'Ecuyer (2012). Our results with real data will confirm this.

For each j, let F_j be the cdf of X_j and $\sigma_{F_j}^2 = \text{Var}(F_j(X_j))$. The Spearman correlation between X_j and X_k is

$$r_{j,k}^X = \frac{\mathbb{E}\{F_j(X_j)F_k(X_k)\} - \mathbb{E}\{F_j(X_j)\}\mathbb{E}\{F_k(X_k)\}}{\sigma_{F_j}\sigma_{F_k}}.$$

Here, $F_j(X_j)$ is not a uniform random variable over $(0, 1)$, because X_j is a discrete random variable. After the parameters α_j have been estimated by MLE, we have estimates $\hat{F}_j$ of the marginal distributions F_j (obtained by replacing each α_j by its estimate in the function F_j), and the Spearman correlation for the model that uses these estimates is

$$\tilde{r}_{j,k}^X = \frac{\mathbb{E}\{\widehat{F}_j(X_j)\widehat{F}_k(X_k)\} - \mathbb{E}\{\hat{F}_j(X_j)\}\mathbb{E}\{\hat{F}_k(X_k)\}}{\sigma_{\widehat{F}_j}\sigma_{\widehat{F}_k}}, \tag{13}$$

Let $\widehat{r}_{j,k}^X$ be the (empirical) Spearman correlation coefficient in the data. For each pair (j, k), we want to find $\rho_{j,k}^{\text{Z}}$ for which

$$\tilde{r}_{j,k}^X = \widehat{r}_{j,k}^X. \tag{14}$$

This is a root finding problem in which the left side, given in (13), contains an expectation defined as a double integral inside a double sum: we must integrate with respect to the joint distribution of (B_j, B_k), then sum with respect to the (conditional) Poisson distributions of X_j and X_k. This is more complicated than in Channouf and L'Ecuyer (2012), who could use the root-finding methodology of Avramidis et al. (2009b), which does not apply to our case. To approximate the root, we use a stochastic approximation (SA) root-finding method in which we estimate the multivariate expectation by Monte Carlo for each value of $\rho_{j,k}^{\text{Z}}$ that is considered. The algorithm is given in Section C. For a recent coverage of stochastic root finding methods, see Pasupathy and Kim (2011).

As is typically the case when a large correlation matrix is estimated from data, the matrix whose entries are the $\rho_{j,k}^{\text{Z}}$ just obtained may not be a valid (positive definite) correlation matrix.

In our experiments, these matrices were always either positive definite or only slightly negative definite. Those that were not positive definite were modified slightly into positive definite ones by applying a small perturbation using the heuristic of Davenport and Iman (1982), which finds a valid correlation matrix which is as close as possible to the matrix with entries $\rho^{\mathrm{Z}}_{j,k}$. This method was also used successfully by Channouf and L'Ecuyer (2012).

5.2. Parametric models for the correlation matrix

In the PGnorta model presented so far, there are $p(p-1)/2$ correlations to specify in $\mathbf{R}^{\mathrm{Z}}$. This can be too many when p is large, and may open the door to overfitting. To prevent this, we can parameterize the matrix $\mathbf{R}^{\mathrm{Z}}$ by a small number of parameters, then estimate those parameters. Here we consider two such parametric models, namely *PGnortaAR1*, where entries of the correlation matrix $\mathbf{R}^{\mathrm{Z}}$ are assumed to obey an AR(1) process, $\rho^{\mathrm{Z}}_{j,k} = \rho^{|j-k|}$, and *PGnortaARM*, where the AR(1) process is extended to $\rho^{\mathrm{Z}}_{j,k} = a\rho^{|j-k|} + c$ for $j \neq k$, and $\rho^{\mathrm{Z}}_{j,j} = 1$. Channouf and L'Ecuyer (2012) have used successfully similar parameterizations in their model for $\mathbf{X}$ and pointed our their usefulness for the situation where the correlation between the counts may drop sharply between lag 0 and lag 1, then decays slowly as a function of the lag $|j-k|$, and does not approach 0. We have observed this type of behavior in some of our data sets. It can happen when some factor has a strong impact over the entire day. The parameters of these two new models are estimated by matching the models to the correlation matrix estimated by the algorithm of Section C, using least-squares fitting.

6. Case Studies

In this section we report the results of fitting the different models discussed previously to real data sets obtained from three call centers located in Canada. The first one is a 24-hour emergency call center, the second one is the commercial call center considered in Ibrahim and L'Ecuyer (2013) and the third one is the call center of the Quebec electricity provider, Hydro-Québec. In all cases, our data comprises only a subset of the call types. We distinguish the different days of the week, but assume otherwise that the arrival rates have the same distributions across successive weeks. We will see that notwithstanding the different nature of these call centers and the fact that the data they generate have different statistical patterns, the results of the fit are qualitatively consistent across different datasets. We compare the statistical performance of the following nine models defined earlier: Poisson, PGindep, PGsingle, PG2, PG2sp, PG2pow, PGnorta, PGnortaAR1, and PGnortaARM.

6.1. An Emergency Call Center

The emergency call center operates 24 hours a day for 7 days a week. We had access to the call-by-call data (exact arrival time of each call) over 616 successive days. The center receives calls categorized in several dozen types. For the results reported here, we selected a subset of those types for which the daily patterns were similar and we consider the aggregated arrival process for those types. These results are representative of a larger set of statistical analyzes performed over different subsets and over individual call types having sufficiently large volume. Confidentiality agreements prevent us from providing further details.

The days are divided into $p = 48$ half-hour periods. At first, our days started and ended at midnight, but we found in the data that there is often significant positive correlation between the call volume in the evening (say between 8 p.m. and midnight) and the call volume during the following night (say between midnight and 5 a.m.), which could not be captured well by our models when starting the days at midnight. This type of dependence across successive days was in fact causing estimation artifacts such as spurious humps in the correlation curves between past and future arrival volumes during the day, due to the correlations between parts of different days in the training dataset and not by the within-day effects. This problem was resolved by starting the days at 5 a.m. instead of midnight. Around 5 a.m., the traffic is usually very low and there is very little correlation between the arrival volumes before and after 5 a.m. Holidays are an exception to this rule: the call volumes are usually larger during the night before a holiday, and smaller in the morning of the holiday. For this reason, in a preprocessing phase before fitting our models, we removed the data corresponding to special days (Quebec statutory holidays), for which the arrival volumes and patterns differ significantly from the ordinary days. The arrival process for those days would have to be modeled separately.

Some descriptive statistics. Preliminary analysis of data revealed that Friday, Saturday and Sunday have particular statistical patterns different from the rest of the dataset. On the other hand, statistical characteristics of weekdays from Monday to Thursday are very similar. The results reported here are for the data from Monday to Thursday regrouped in a single dataset. This data represents normal weekdays of the call center. Figure 1 shows the average number of calls received per half-hour period.

To see how the DI and SDI behave when we aggregate periods, we define

$$Y_{j,d} = X_j + \cdots + X_{j+d-1}$$

for $j \geq 1$ and $d \leq p - j + 1$. This represents the count for an aggregation of d successive periods starting at period j. Fig. 2 shows $\mathrm{DI}(Y_{j,d})$ and $\mathrm{SDI}(Y_{j,d})$ as a function of j, for $d = 1, 2, 4, 8$ (i.e., time

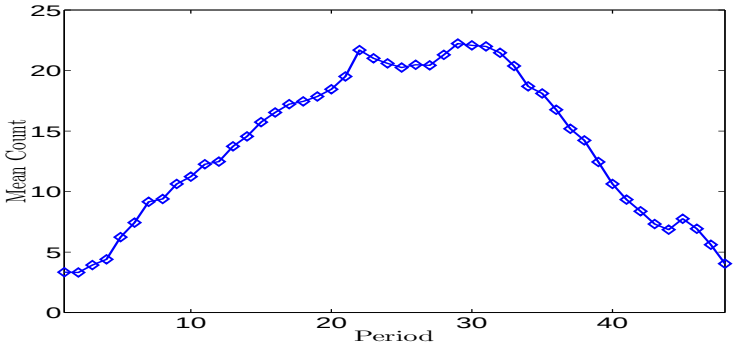

Figure 1 Mean count per period for the emergency center.

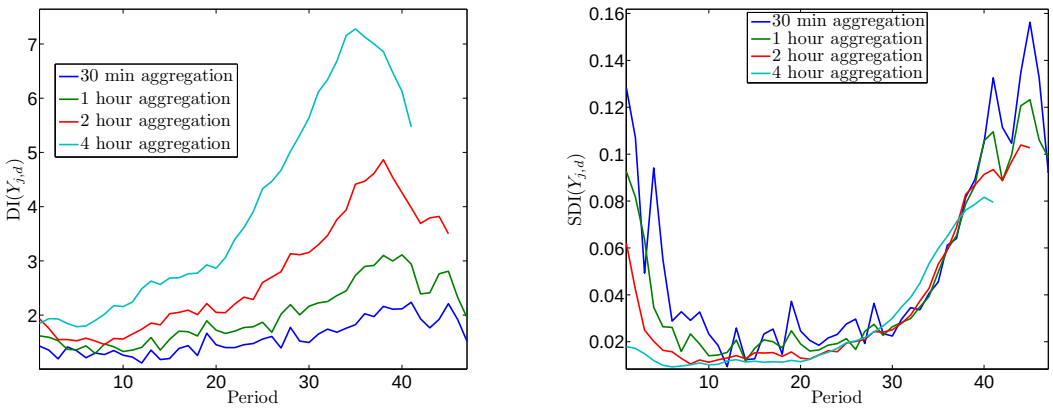

Figure 2 DI($Y_{j,d}$) (left panel) and SDI($Y_{j,d}$) (right panel) as a function of j, for $d = 1$, 2, 4, 8, for the emergency center.

slots of 30 minutes to 4 hours). When we increase d for a fixed j, we find that DI($Y_{j,d}$) increases as expected from (3), and that SDI($Y_{j,d}$) remains pretty stable, which suggests that the B_j's are strongly correlated (we are close to a single busyness factor model; see (5)) at least within 4-hour time slots. The SDI also depends much on j, which shows that there is much more overdispersion (for the selected call types) in early morning, the evening, and during the night, than in the middle of the day. We also see in the figure that the curve for $d = 8$ is higher than the other three for j between about 30 and 37. This may appear to contradict (5). The explanation is that SDI($Y_{j,8}$) is for a time slot of 8 periods that extends on the right of period j, and the maximum of each of the other three curves over this time slot is indeed larger than SDI($Y_{j,8}$).

Fig. 1 in Section E provides a picture of all the correlations between pairs of periods, and also between aggregated periods in blocks of 1, 2, and 4 hours. The correlations are quite small in this case compared with other (typical) call centers. They are larger in the evening than in the rest of the day.

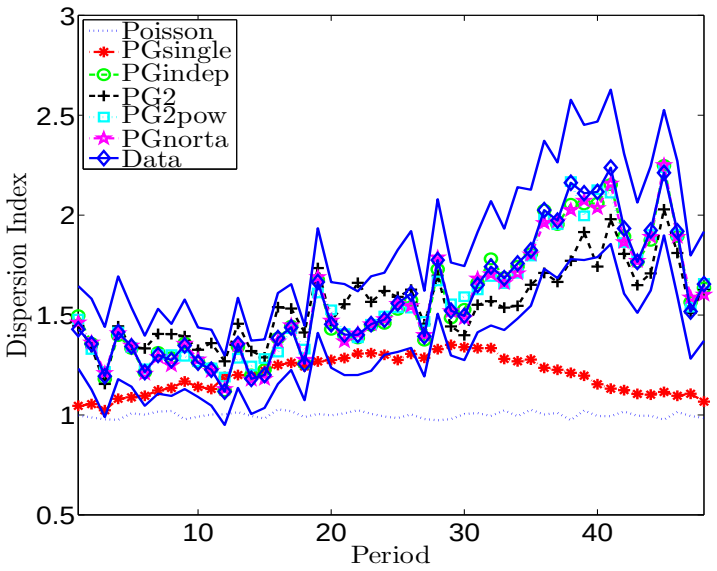

Figure 3 Comparison of $\mathrm{DI}(X_j)$ as a function of j for different models and for the data, for the emergency call center.

How the models fit the data. Fig. 3 compares the DI obtained for the six models with the sample DI calculated from the data (also given as the lower curve in Figure 2(a)), in each 30-minute period j. For the latter, we also provide a 95% confidence interval for the DI (indicated by the solid lines) calculated using bootstrap from a *kernel density estimator* (KDE) of the data sample, with a Gaussian (normal) kernel and a bandwidth chosen so that the variance associated with the estimated density is equal to the sample variance in the data; see Section D for the details.

The Poisson and PGsingle models are far from matching the empirical DI; they both give an SDI that is always too low. For PGsingle, this is due to the fact that a single parameter is available, namely the variance of the single busyness factor, which is the SDI common to all periods in the model. This parameter also affects the correlation between the counts across periods. If it was set higher, so that the average SDIs would match the average SDI in the data for example, then these correlations would be much too large compared with those in the data. The MLEs make a compomise. The PG2 model improves the DI, but the values differ from those in the data and wander around the boundaries of the confidence interval. The PG2pow and PGnorta provide a much better fit.

Fig. 4 shows the correlation between the demand $Y_{1,j}$ in the first j periods and the demand $Y_{j+1,p-j}$ in the remaining $p-j$ periods, as a function of j, for the data and for the models where this correlation is nonzero. The solid lines indicate 95% confidence intervals calculated from the data using the same KDE bootstrap methodology as in the previous plot, but with a KDE based on a two-dimensional Gaussian kernel (see Section D). We see that only PGnorta has sufficient

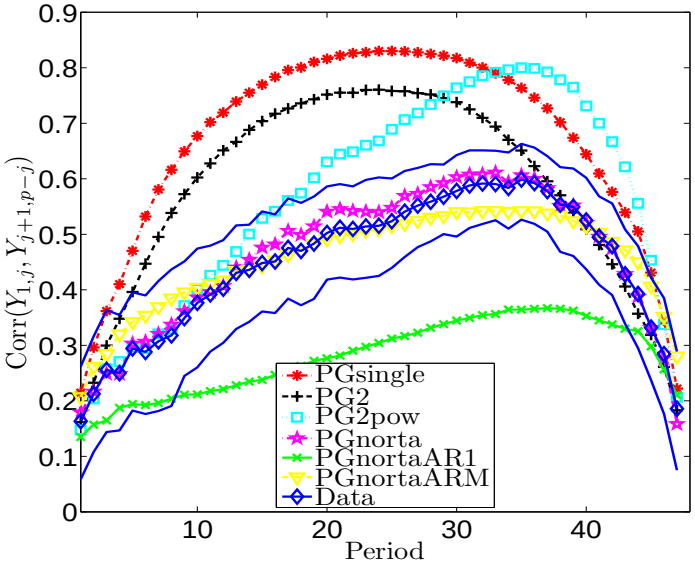

Figure 4 Comparison of the sample coefficient of correlation of past and future demand for the models with correlation, for the emergency call center.

flexibility to fit the correlations. The simplified model PGnortaARM also does well, with much fewer parameters, while PGnortaAR1 significantly underestimates the correlations. PGsingle and PG2 are significantly off. PG2pow does better and captures the shape of the correlation curve, but it does not fit the correlations accurately.

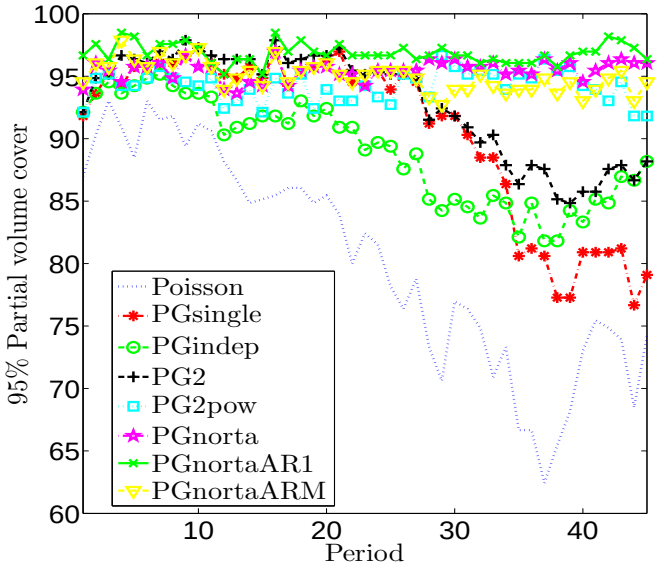

Figure 5 Comparison of the empirical coverage probability of a 95% PI of the partial demand $Y_{j,4}$, for different models, for the emergency call center.

To test the quality of fit of both the dispersion and correlation in a combined way, we compare some quantiles of the distribution of the partial demand over the two-hour interval (four periods) starting at j, $Y_{j,4}$, for the empirical distribution of the data and the distributions implied by the models. For each model, we computed a *prediction interval (PI)* (Geisser 1993) whose boundaries are the 0.025 and 0.975 quantiles of the implied distribution, and then computed the empirical coverage probability of this PI, defined as the fraction of observations of $Y_{j,4}$ in the data that fall in the interval. This coverage should be close to 0.95. A coverage smaller than 0.95 indicates that the model underestimates the dispersion, and vice-versa. Larger positive correlations across periods, or larger variances of the counts within periods, tend to increase the variance of $Y_{j,4}$, so if a model fails to properly capture any of these two effects, the coverage should deviate from 0.95, signalling an incorrect distribution for $Y_{j,4}$ in the model, which in turn would lead to an incorrect distribution of the waiting time distribution and of performance measure estimates when we simulate the model.

Fig. 5 shows the empirical coverage probability for each model. We see that PG2pow and PGnorta best capture the distribution of the partial demand, and PGnorta is the best performer. For the other four models, the PI coverage is too small in intervals 30 to 45, which correspond to the counts in periods 30 to 48. These are the periods with the largest DIs (see Fig. 3). Recall that a larger DI is associated with a stronger departure from the Poisson distribution and larger correlations between the X_j's for given correlations between the B_j's, as we have seen earlier. As expected, PGindep performs poorly here even though it models well the DI for one period at a time (see Fig. 3), because it totally neglects the correlation across periods. PGsingle fails because it largely underestimates the DI in this region. PG2 provides only a small improvement. The PG2pow model, even though it does not exactly capture the correlations of the partial demand (see Fig. 4), is reasonably close to the 95% percent target in all segments. This suggests that this model, whose number of parameters remains small, could be a reasonable choice in some situations such as the one illustrated here.

An out-of-sample goodness-of-fit analysis. In what we have seen so far, PGnorta appears to be the best performer overall. It is the only model that matches both the variances and correlations in the data, and provides 0.025 and 0.975 quantiles of the partial demand $Y_{j,4}$ that match those of the data, in all cases. On the other hand, this model has many more parameters (the entire correlation matrix $\mathbf{R}^{\mathrm{Z}}$) than the other ones, so it could perhaps overfit the data. For a fairer comparison, we now examine how the models perform for out-of-sample distributional forecasts, using a *leave-one-out* technique as follows. For each i, we remove day i from the data set, re-estimate the model without that day, and compute a PI $[L_{i,j,d}, U_{i,j,d}]$ with integer bounds, in which $Y_{j,d}$ will fall with probability $P_{i,j,d} \approx 1-\alpha$ under this model, where $1-\alpha$ is a fixed number selected in advance.

We do this for each i, j, and selected values of d. The PI boundaries are integers because $Y_{j,d}$ is always an integer, and for this reason it is generally not possible to select the bounds so that the probability of falling in the interval is exactly equal to $1-\alpha$. The difference can be significant when $Y_{j,d}$ has a small mean. Then we compute the proportion of days i for which the realization $Y_{i,j,d}$ of $Y_{j,d}$ for the removed day falls in the interval, and we compare this proportion with our best estimate of $P_{i,j,d}$, using a sum of squares criterion, as we now explain.

To estimate the quantiles $L_{i,j,d}$ and $U_{i,j,d}$, we simulate $n=10^5$ replicates of the vector of counts $(X_{i,1},\ldots,X_{i,p})$ for day i from the model estimated with day i removed, compute $Y_{i,j,d}$ for each (j,d) of interest and each replicate, say $y_{i,j,d,k}$ for replicate k, then we compute the integers $L_{i,j,d}$ and $U_{i,j,d}$ so that the proportion of the n realizations $y_{i,j,d,k}$ that fall below $L_{i,j,d}$ is approximately $\alpha/2$, and the proportion that fall above $U_{i,j,d}$ is approximately $\alpha/2$. After that, we estimate the average over i of the true probability $P_{i,j,d}$ under the estimated model by the sample average

$$\mathrm{CVM}_j = \frac{1}{I}\frac{1}{n}\sum_{i=1}^{I}\sum_{k=1}^{n}\mathbb{I}(\hat{Y}_{i,j,d,k}\in[L_{i,j,d},U_{i,j,d}]), \tag{15}$$

where I is the number of days. On the other hand, the proportion of observations in the data that fall in the interval is

$$\mathrm{CVD}_j = \frac{1}{I}\sum_{i=1}^{I}\mathbb{I}(X_{i,j}\in[L_{i,j},U_{i,j}]). \tag{16}$$

To measure the overall difference between the values of CVM_j and CVD_j, we compute the following root mean squares deviation (RMSD) between them, averaged over the periods, scaled by a factor of 100 (just for better readability):

$$\mathrm{RMSD} = 100\sqrt{\frac{1}{p}\sum_{j=1}^{p}(\mathrm{CVD}_j-\mathrm{CVM}_j)^2}. \tag{17}$$

Table 2 reports this scaled RMSD for the various models, for $\alpha=0.50$, 0.25, and 0.10, and for $d=1$, 2, 4, 8. The RMSD for one period at a time ($d=1$) measures the out-of-sample fit of the marginal distributions, while those for larger d also measure the adequacy of the correlation structure, which affects the distributions of the sums of counts over successive periods, for given marginals. The ability to match the distribution of $Y_{j,d}$ for $d>1$ is then related to the ability of modeling the dependence.

Table 2 reveals that the models in which the busyness factor in each period has a gamma distribution (PGindep, PGnorta, PGnortaAR1, PGnortaARM) have the best ability to fit the marginal distributions ($d=1$). For these models, the marginal distribution of the count in each period is negative binomial. The PG2pow model is almost as good in fitting the marginals. The

Table 2 RMSD for various models and pairs (j, d), for the emergency call center.

	$1-\alpha=50\%$					$1-\alpha=75\%$					$1-\alpha=90\%$				
	0.5 h	1 h	2 h	4 h	8 h	0.5 h	1 h	2 h	4 h	8 h	0.5 h	1 h	2 h	4 h	8 h
Poisson	10.4	14.2	19.5	24.2	27.8	10.7	16.6	23.5	31.3	37.5	8.5	13.8	21.0	30.1	38.7
PGsingle	7.1	8.9	11.3	12.4	9.9	7.2	10.0	12.5	13.5	12.0	5.3	7.8	10.1	11.4	9.0
PGindep	1.8	5.3	11.6	17.8	22.5	1.3	5.3	12.7	21.4	29.9	0.8	4.1	10.2	18.7	29.1
PG2	2.6	5.2	8.6	11.2	9.7	2.1	4.9	8.7	11.4	11.6	1.6	3.8	6.8	9.4	8.8
PG2sp	2.7	5.1	8.6	11.2	9.8	2.2	4.8	8.7	11.4	11.6	1.6	3.8	6.8	9.4	8.8
PG2pow	2.0	3.2	4.6	5.0	5.4	1.5	2.9	4.4	5.1	5.0	1.0	2.0	3.1	3.4	3.0
PGnorta	1.8	2.3	2.2	2.8	3.0	1.3	1.7	1.7	1.7	1.3	0.8	1.1	1.2	1.3	0.8
PGnortaAR1	1.8	3.3	4.6	6.2	5.5	1.3	2.4	4.5	6.1	5.9	0.9	1.4	2.4	3.5	3.6
PGnortaARM	1.8	2.8	3.8	5.3	5.3	1.3	2.4	3.4	4.3	4.5	0.9	1.5	2.2	2.7	2.8

models with strong daily busyness factor (PGsingle, PG2) do not fit well the marginal distributions, because they are unable to track the changes in the variance of the stochastic rate over the whole day. For $d > 1$, PGnorta is the best model to capture the behavior. Its simplified parametric versions PGnortaAR1 and PGnortaARM give significantly more error in the distributional forecast. We attribute this to the fact that the correlation between counts within the day for this particular center varies from period to period in a more complicated way than what is assumed by these models. In particular, there is much more correlation between nearby periods in the evening than between nearby periods during the day; this can be seen clearly in Fig. 1 of the Online Supplement, which shows the correlations between counts in all pairs of periods. PG2pow has a quality of fit comparable to that of the latter two parametric models. The other models, Poisson, PGindep, PGsingle, and PG2, do not model the correlations well and provide poor joint distributional forecasts.

6.2. A Commercial Call Center

Our second data set is from the call center of a major Canadian company, and is taken from Ibrahim et al. (2012) and Ibrahim and L'Ecuyer (2013). The center operates from 8 a.m. to 7 p.m. on weekdays and 8 a.m. to 6 p.m. on Saturdays. We have arrival counts data over half-hour periods ($p = 22$ periods per day on weekdays) for 394 days from October 13, 2009 to November 11, 2010. In a preprocessing step, we removed the special days (Quebec and Ontario statutory holidays). The call center handles several call types; for our study, we have selected one of these call types having a large volume. Preliminary data analysis revealed that Monday and Saturday have particular statistical patterns that differ from the other days, whereas Tuesday to Friday have a very similar pattern. Thus, we have retained the data from Tuesday to Friday, in a single dataset. It represents a normal weekday in the call center. We will perform the same empirical analysis, model fitting, and goodness-of-fit assessment as for the previous data set.

Figure 6 shows the average number of arrivals per period. Note that the vertical scale starts at 100. The DI and SDI as functions of the period index j for different aggregation levels d are shown in Fig. 7. They behave similarly to those of Fig. 2, although the DI here is roughly 10 to 20 times

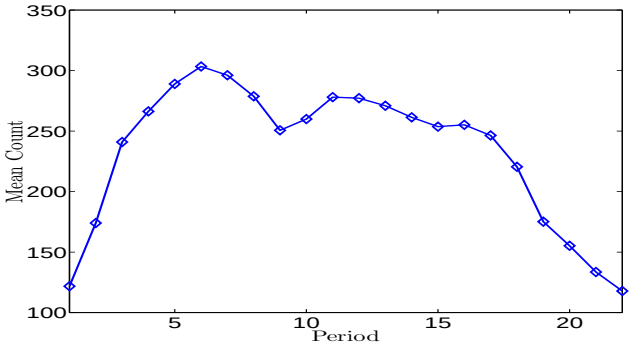

Figure 6 Mean arrival count per period for the commercial call center.

larger, depending on the j and d, whereas the SDI is slightly smaller than in Fig. 2(b) on average (about half of it in the evening). The call volumes here are about 12 to 20 times larger, and we can conclude from (3) that this is the main reason for the larger DI. The SDI appears larger for $d=8$ than for the lower aggregations from $j=4$ to $j=14$, which seems to contradict (5). but this is because the height of the $d=8$ curve at j is for the aggregation of periods j to $j+7$, and this height remains smaller than the maximum height of the 30-minute curve over those 8 periods. The correlations between counts over all pairs of periods, and for aggregated periods in blocks of 1, 2, and 4 hours, are pictured in Fig. 2 of Section E. These correlations are much larger than for the emergency center. There are somewhat stronger between the periods in the middle of the day, and weaker between the early morning periods and the rest of the day.

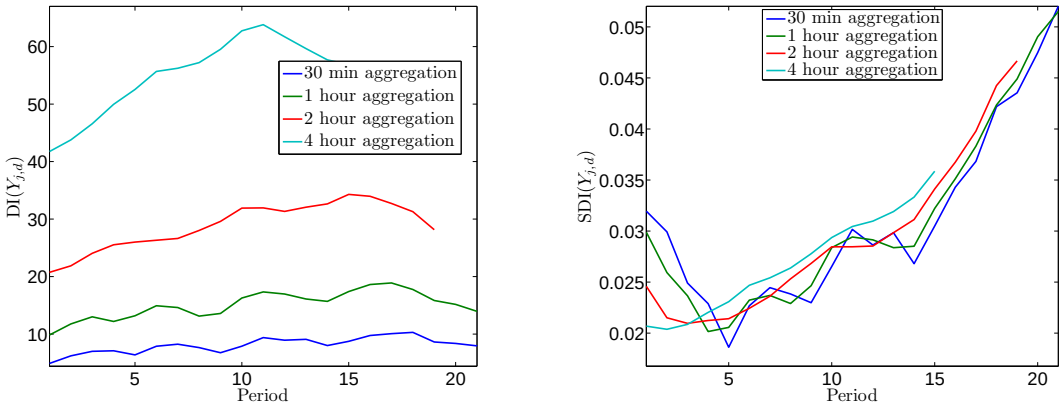

Figure 7 DI($Y_{j,d}$) (left panel) and SDI($Y_{j,d}$) (right panel) as functions of j, for $d=1,2,4,8$, for the commercial call center.

Fig. 8 shows the DIs per period for different models and for the data, with 95% confidence intervals computed as in Fig. 3. Despite the larger DI here, the qualitative ranking of models remains the same: PGindep, PG2pow, and PGnorta provide the best fit for the DI. Fig. 9 shows

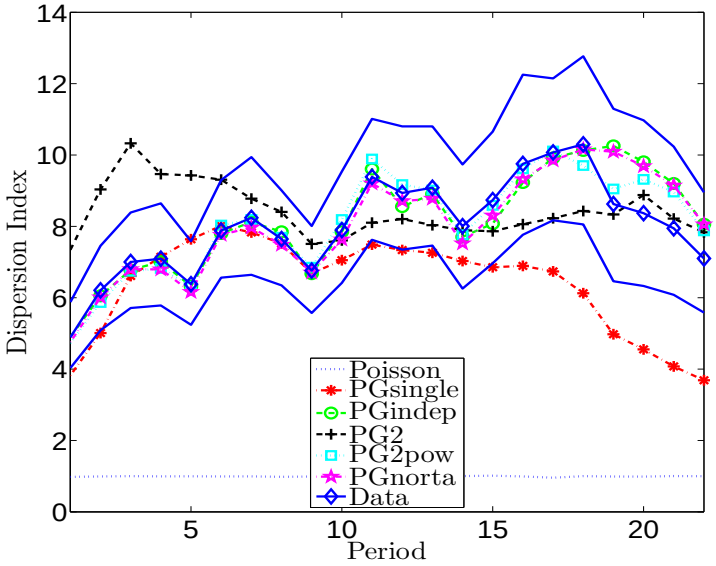

Figure 8 Comparison of $\mathrm{DI}(X_j)$ as a function of j for different models and for the data, for the commercial call center.

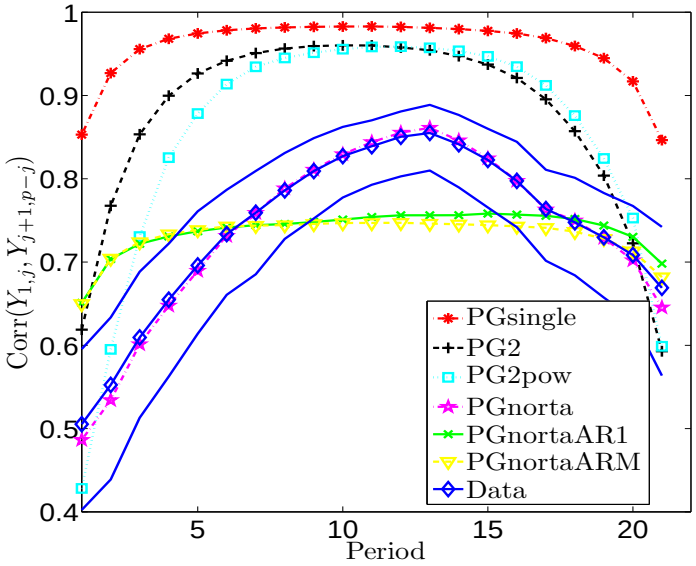

Figure 9 Comparison of the sample coefficient of correlation of past and future demand, for the models with correlation, for the commercial call center.

the quality of fit for the correlation between past and future demand (the confidence interval for this quantity is calculated as for Fig. 4). Here we observe more correlation than for the emergency center. This is not only due to the larger volumes; the correlations are also larger between the underlying busyness factors in our NORTA model. Only PGnorta gives a good fit here.

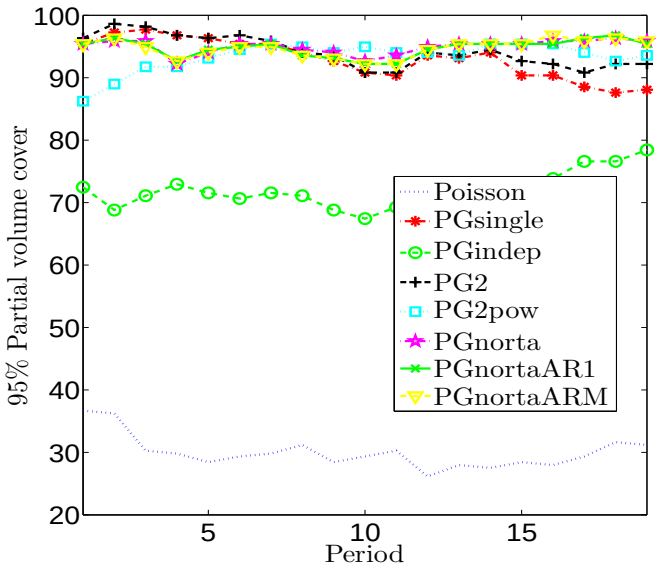

Figure 10 Comparison of the empirical coverage probability of a 95% PI of the partial demand $Y_{j,4}$, for different models, for the commercial call center.

Fig. 10 presents the PI coverage for the partial demand, calculated as for the emergency center. The Poisson and PGindep are doing very poorly, because they neglect the positive dependence across periods. All other models perform somewhat reasonably, but PGnorta is clearly the best. The other models fail to properly cover the data either near the beginning or near the end of the day, or both.

Table 3 RMSD for various models and pairs (j, d), for the emergency call center.

	$1-\alpha=50\%$					$1-\alpha=75\%$					$1-\alpha=90\%$				
	0.5 h	1 h	2 h	4 h	8 h	0.5 h	1 h	2 h	4 h	8 h	0.5 h	1 h	2 h	4 h	8 h
Poisson	31.7	36.1	40.2	42.5	44.8	43.1	50.9	57.5	61.9	66.7	44.7	55.8	64.4	71.3	77.7
PGsingle	6.3	5.4	4.6	2.8	1.6	7.6	7.1	6.1	4.0	2.3	5.8	6.1	5.4	4.0	3.4
PGindep	3.4	11.2	20.6	28.1	35.0	3.1	13.2	27.3	39.3	48.7	2.0	12.1	26.4	41.2	51.9
PG2	5.0	4.1	3.4	2.4	1.5	4.8	4.1	5.1	4.3	2.6	3.0	2.9	3.3	2.9	2.3
PG2sp	4.7	4.2	3.4	2.4	1.3	4.5	3.9	5.2	4.3	2.7	2.8	2.7	3.5	3.1	2.8
PG2pow	3.0	3.1	2.8	2.9	1.4	2.5	3.3	4.1	2.0	0.8	1.7	3.3	3.7	2.8	2.7
PGnorta	3.3	3.5	3.1	2.0	2.3	3.2	3.0	2.7	1.2	1.3	2.0	2.4	2.2	1.7	2.0
PGnortaAR1	3.3	3.6	3.4	1.7	0.9	3.2	3.1	2.9	1.8	0.5	2.0	2.4	2.2	2.3	2.8
PGnortaARM	3.4	3.6	3.4	1.8	1.1	3.2	3.1	2.8	1.9	0.5	2.0	2.4	2.3	2.2	2.8

The results of the out-of-sample test of fit, in Table 3, are similar to those of the emergency center. The main difference is that here, the parametric models PGnortaAR1 and PGnortaARM perform almost as well as the PGnorta, due to the fact that the correlations between periods in the data vary much less and are in better agreement with those parametric models. The single busyness factor model also provides a more adequate model here than for the emergency center,

especially when periods are aggregated, due to the fact that the correlations between periods are significantly larger here. Finally, the PG2pow model provides a better fit for the marginals than the models in which the marginal busyness factors are gamma distributed.

6.3. A call center from a Quebec utility society

Our third example is based on data from the call center of Hydro-Québec, a governmental society that produces and provides all the electricity for the province of Quebec. The center operates from 8 a.m. to 6 p.m. on weekdays. We have arrival counts data over 15 minute periods ($p = 40$ periods per day) for 247 days from January 1, 2011 to December 31, 2011. In a preprocessing step, we removed the special days (Quebec statutory holidays and 29 August 2011, corresponding to the Monday immediately following August 28, 2011 when hurricane Irene hit province of Quebec). The call center handles several call types; for our study, we have selected one of these call types having a large volume. Preliminary data analysis revealed that weekdays from Monday to Friday have similar statistical patterns. Thus, we have retained the data from Monday to Friday, which represent a normal weekday in the center, in a single dataset. We do the same analysis as for the previous two cases.

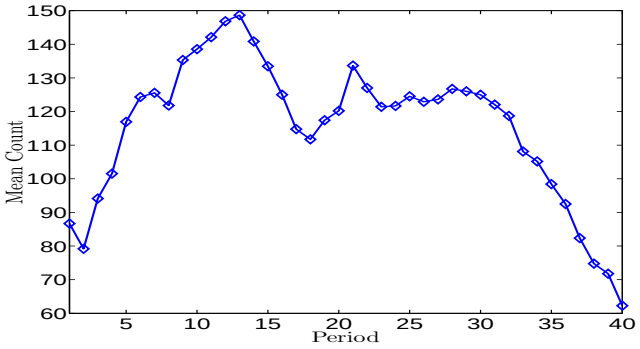

Figure 11 Mean arrival count per period for the utility call center.

Fig. 11 shows the average number of calls per period, while Fig. 12(a) gives the DI and SDI. The arrival volumes here, as well as the DI and SDI, all lie between those of the two previous call centers, except for the SDI in the middle of the day, which is slightly higher here. Again, the numbers agree with (3). The call volumes are about half of those of the commercial center The correlations between counts are pictured in Fig. 3 of the Online Supplement. They are slightly smaller overall than for the commercial center.

Fig. 13 shows $\mathrm{DI}(X_j)$ as a function of j, for different models and for the data, with a confidence interval computed for the data as in Fig. 3. PG2 provides a better fit for the DI here than for the other datasets. Overall, all models have more difficulty fitting the DI here than in the previous

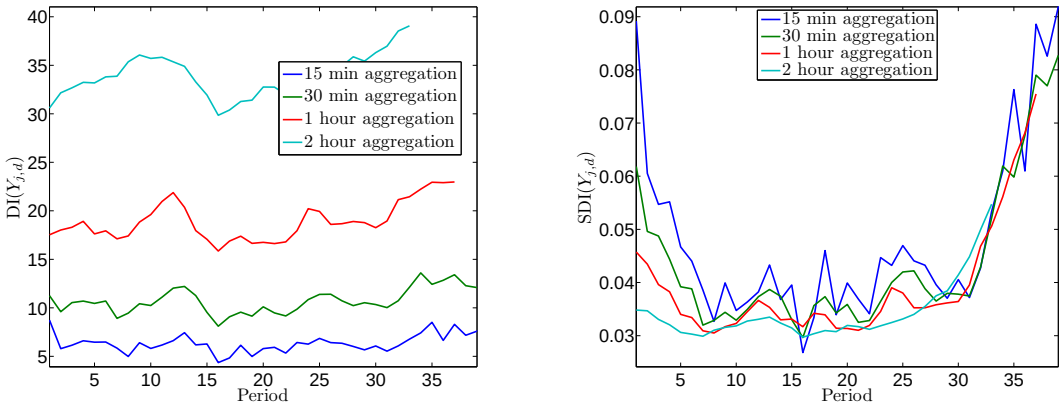

Figure 12 DI($Y_{j,d}$) (left panel) and SDI($Y_{j,d}$) (right panel) as functions of j, for $d = 1, 2, 4, 8$, for the utility call center.

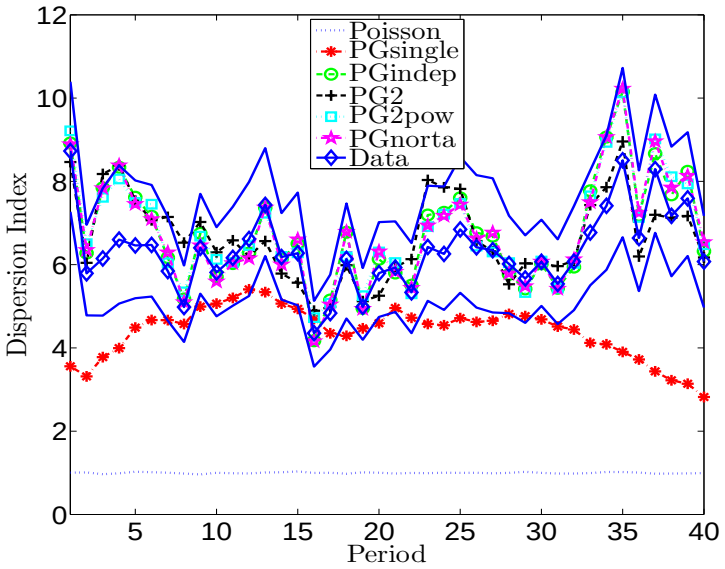

Figure 13 Comparison of DI(X_j) as a function of j for different models and for the data, for the utility call center.

cases, but they stay within the confidence interval most of the time. Fig. 14 shows the quality of fit for the correlation between past and future demand, similar to Fig. 4. Only PGnorta gives a good fit. Fig. 15 gives the PI coverage for the partial demand $Y_{j,4}$, calculated as in Fig. 5. The Poisson and PGindep models are doing very poorly as usual. All other models perform somewhat reasonably, but PGnorta is the best among them. The other models miss the proper coverage near the beginning or the end of the day, or both.

Table 4 presents the results of the out of sample analysis. The PG2pow model again provides a bit better fit for the marginals than the models based on gamma marginals for the busyness factors. This suggests that one could improve the modeling by selecting alternative marginal distribution

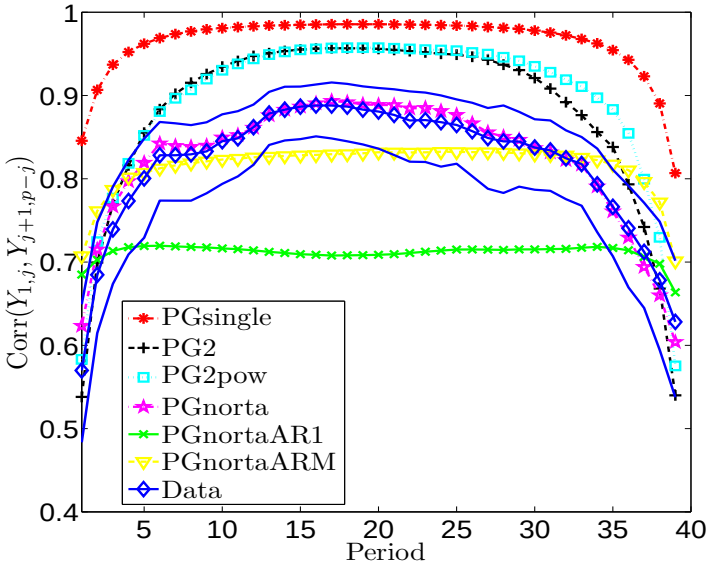

Figure 14 Comparison of the sample coefficient of correlation of past and future demand for the models with correlation, for the utility call center.

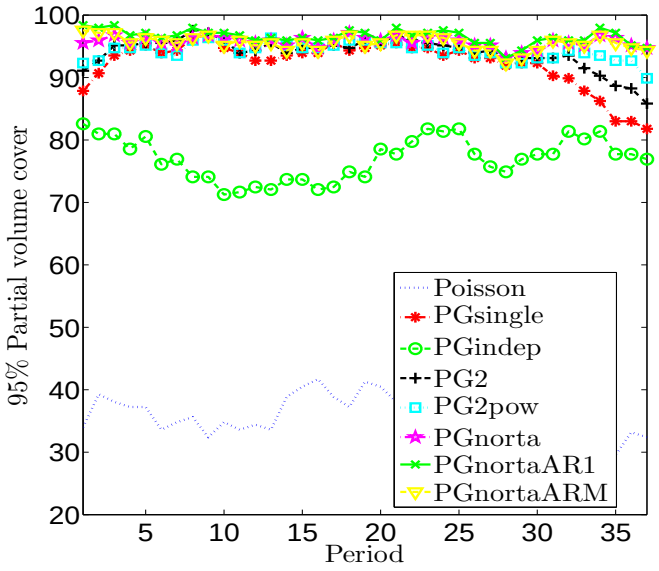

Figure 15 Comparison of the empirical coverage probability of a 95% PI of the partial demand $Y_{j,4}$, for different models, for the utility call center.

of the counts (i.e., explore alternatives to the gamma distribution for the rates). For the tails of the distribution (75% and 90% target cover), PG2pow provides the best coverage (smallest RMSD) for all aggregation levels. This can be explained by the fact that in this dataset we have a strong influence of the daily busyness factor, as could be seen in Fig. 14 and in Fig. 3(a) of the Online Supplement. which show that the correlations between periods rarely falls below 0.4 or 0.5, even

Table 4 RMSD from the model cover, for the utility call center.

	$1-\alpha=50\%$					$1-\alpha=75\%$					$1-\alpha=90\%$				
	0.25 h	0.5 h	1 h	2 h	4 h	0.25 h	0.5 h	1 h	2 h	4 h	0.25 h	0.5 h	1 h	2 h	4 h
Poisson	29.5	34.2	38.3	41.4	44.2	38.9	47.1	53.9	59.6	64.6	39.2	50.9	59.7	67.5	74.4
PGsingle	8.1	7.6	6.1	4.4	2.1	8.6	8.0	6.9	4.0	1.7	7.3	7.0	5.4	3.1	1.4
PGindep	5.4	9.4	18.6	26.5	33.6	4.5	10.5	24.6	36.5	46.3	1.8	8.4	22.3	37.8	51.2
PG2	5.1	4.1	4.6	4.1	3.2	4.4	3.4	3.8	3.3	2.2	2.0	3.0	3.5	2.5	1.7
PG2sp	5.1	4.0	4.6	4.1	3.1	4.4	3.3	3.8	3.2	2.1	2.1	3.0	3.5	2.4	1.7
PG2pow	4.8	3.4	3.1	3.1	2.5	4.0	2.3	2.4	2.7	2.0	1.5	1.7	1.6	1.1	1.1
PGnorta	5.3	4.7	3.5	2.6	1.6	4.4	4.1	3.9	3.4	2.7	1.8	2.2	2.4	2.3	2.4
PGnortaAR1	5.2	5.9	6.1	5.6	4.7	4.4	5.4	6.0	6.6	5.7	1.8	2.8	3.7	4.0	3.6
PGnortaARM	5.3	4.8	4.3	3.5	2.2	4.4	4.0	4.2	4.0	3.2	1.8	2.3	2.5	2.7	2.2

for periods that are far away in the day. At the same time, $\mathrm{Var}(B_j)$ varies significantly over the day (see Fig 12(b)); this cannot be captured by PG2 or PG2sp, but can be captured by PG2pow, which then provides a better cover. The parametric model PGnortaARM is close to PGnorta here, but it still fails to capture some of the variability in the intraday correlations. It performs better than PGnortaAR1 because it better captures the slowly decreasing correlation profile. Overall, the data here seems to agree quite well with PG2pow.

7. Impact of the Model on Performance Measures: An Illustration

In this section, we illustrate the impact of the choice of the arrival process model on the service level (SL) and the average waiting time (AWT) of calls, in each period of the day. The SL in a period is defined in our example as the percentage of calls with waiting time less than 120 seconds, among those that arrived in that period and did not abandon before 120 seconds. The SL and AWT are averages per call in the long-run (i.e., over an infinite number of days). We took exactly the same data as for the Quebec utility society of Section 6.3, and the same estimated arrival process models. There is a single call type and the service times have a lognormal distribution with mean 206.44 and variance 23667 (in seconds), as estimated from the data. Each waiting call abandons (if not yet served) after an exponential time with mean 2443 seconds. We computed a reasonable staffing vector (the number of agents answering the calls in each period) using the optimization tools in ContactCenters (Buist and L'Ecuyer 2012), with the Poisson arrival-rate model. The retained staffing vector for the 40 time periods is (16, 24, 31, 36, 43, 48, 51, 52, 56, 60, 62, 65, 67, 67, 66, 65, 62, 61, 60, 61, 64, 64, 63, 63, 64, 64, 64, 64, 65, 65, 64, 64, 62, 60, 58, 56, 53, 49, 48, 44).

We performed 10,000 simulation runs to estimate the SL and AWT for each period of the day, for each of the following five arrival-process models (independently across models): Poisson, PGsingle, PGindep, PG2, PG2pow, and PGnorta. The results are in Figure 16. We see that the choice of model makes a significant difference on both the SL and AWT. For example, the average AWT for the day is below 100 seconds with the PG2 and PG2pow models, and over 180 seconds with

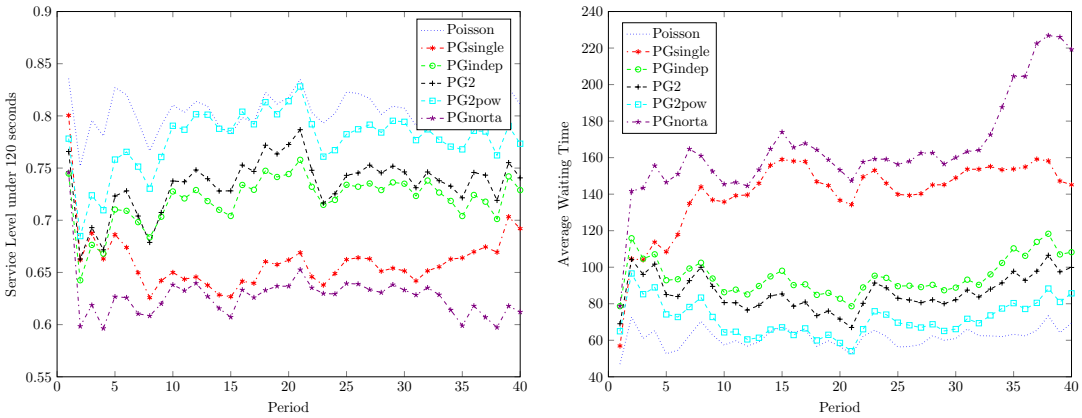

Figure 16 Evolution of the SL (left) and AWT in seconds (right) during the day for the Quebec utility society.

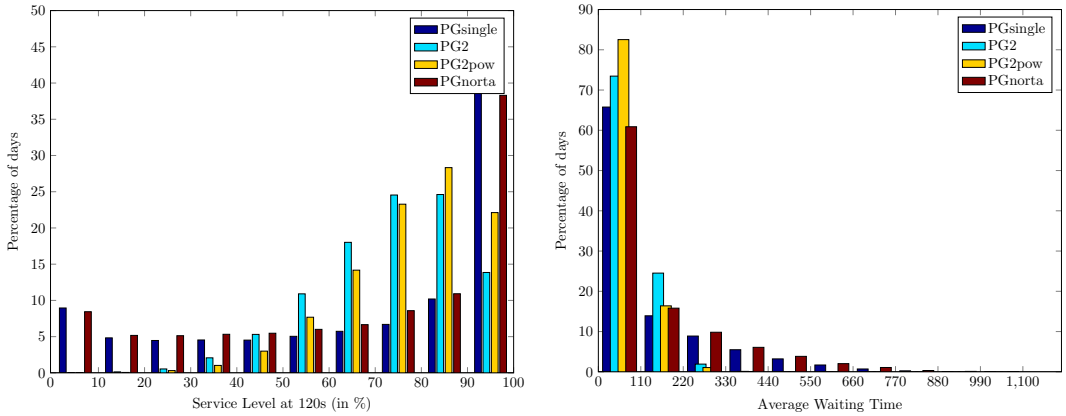

Figure 17 Histogram of the distribution of the daily SL (left) and daily AWT (right), with different models, for the Quebec utility society.

the PGnorta model. The quality of service is the worst with PGnorta, and this is particularly true around the end of the day. Note that the correlations are much higher with PGsingle than PGnorta (Figure 14), so the worst quality of service with PGnorta could seem surprising. It can be explained by fact that the variance of the busyness factor is smaller with PGsingle than with PGnorta (Figure 13). In view of Figures 14 and 13, PGnorta seems the most appropriate model.

We also computed the SL and the AWT for each day separately (they are now random variables), over the 10,000 days. Histograms of the corresponding empirical distributions are given in Figure 17, for PGsingle, PG2, PG2pow, and PGnorta. We find that the percentage of days in which the SL is very low (e.g., less than 30%) or the AWT is very small (e.g., more than 300 seconds) is much larger with PGsingle and PGnorta than for the other models.

8. Conclusions

In this paper we have proposed several new models for the daily arrival process in a call center. These models are based on a doubly stochastic process for which the daily rates are assumed to

have an arbitrary joint distribution and the arrival counts are Poisson given the rates. They include the two-level busyness factor model, its extended version, and a normal copula model for the vector of rates. The proposed models generalize existing rate-based models for daily arrival processes and offer different degrees of complexity and detail in modeling daily arrivals. We have developed statistical parameter estimation approaches for these models and tested the quality of fit of the proposed models on three data sets from real call centers. Our study reveals that the proposed models are capable of better fitting the data and capturing their important statistical characteristics such as overdispersion and strong correlations across the day than the existing models. According to our out-of-sample analysis, among the proposed models the extended two-level busyness factor model and the normal copula for the vector of rates model are the best models overall over our three datasets. An important message coming out from our study is that none of the models is universally good and model selection should be made based on the available data for a specific call center. For example, in our study the extended two-level busyness factor model outperformed the more complex copula model for the vector of rates on the utility call center dataset, which seems to exhibit the statistical patterns closely mimicking those of the extended two-level busyness factor model.

Acknowledgments

This work has been supported by an NSERC postdoctoral scholarship to B. Oreshkin, as well as grants from NSERC-Canada and Hydro-Québec, a Canada Research Chair, and an Inria International Chair, to P. L'Ecuyer.

References

Akşin, O. Z., M. Armony, V. Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* **16**(6) 665–688.

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

Atlason, J., M. A. Epelman, S. G. Henderson. 2008. Optimizing call center staffing using simulation and analytic center cutting plane methods. *Management Science* **54**(2) 295–309.

Avramidis, A. N., W. Chan, P. L'Ecuyer. 2009a. Staffing multi-skill call centers via search methods and a performance approximation. *IIE Transactions* **41** 483–497.

Avramidis, A. N., N. Channouf, P. L'Ecuyer. 2009b. Efficient correlation matching for fitting discrete multivariate distributions with arbitrary marginals and normal copula dependence. *INFORMS Journal of Computing* **21** 88–106.

Avramidis, A. N., A. Deslauriers, P. L'Ecuyer. 2004. Modeling daily arrivals to a telephone call center. *Management Science* **50**(7) 896–908.

Avramidis, A. N., P. L'Ecuyer. 2005. Modeling and simulation of call centers. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, J. A. Joines, eds., *Proceedings of the 2005 Winter Simulation Conference*. IEEE Press, 144–152.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, 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., R. Zhang, L. Zhao. 2001. Root-unroot methodology for non-parametric density estimation. Tech. rep. Technical report, University of Pennsylvania, Dept. of Statistics.

Buist, E., P. L'Ecuyer. 2005. A Java library for simulating contact centers. M. E. Kuhl, N. M. Steiger, F. B. Armstrong, J. A. Joines, eds., *Proceedings of the 2005 Winter Simulation Conference*. IEEE Press, 556–565.

Buist, E., P. L'Ecuyer. 2012. *ContactCenters: A Java Library for Simulating Contact Centers*. Software user's guide, available at `http://www.simul.umontreal.ca/contactcenters`.

Ceşik, M. T., P. L'Ecuyer. 2008. Staffing multiskill call centers via linear programming and simulation. *Management Science* **54**(2) 310–323.

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., P. L'Ecuyer. 2012. A normal copula model for the arrival process in a call center. *International Transactions in Operational Research* **19** 771–787.

Cox, D. R., P. A. W. Lewis. 1966. *The Statistical Analysis of Series of Events*. Methuen, London.

Davenport, J. M., R. L. Iman. 1982. An iterative algorithm to produce a positive definite correlation matrix from an approximate correlation matrix. Tech. rep., Sandia National Laboratories, Albuquerque, New Mexico.

Deslauriers, A. 2003. Modélisation et simulation d'un centre d'appels téléphoniques dans un environnement mixte. Master's thesis, Department of Computer Science and Operations Research, University of Montreal, Montreal, Canada.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management* **5** 79–141.

Geisser, S. 1993. *Predictive Inference: An Introduction*. Chapman and Hall, New York, NY.

Gurvich, I., J. Luedtke, T. Tezcan. 2010. Staffing call centers with uncertain demand forecasts: A chance-constrained optimization approach. *Management Science* **56**(7) 1093–1115.

Harrison, J. M., A. Zeevi. 2005. A method for staffing large call centers based on stochastic fluid models. *Manufacturing and Service Operations Management* **7**(1) 20–36.

Hörmann, W., J. Leydold, G. Derflinger. 2004. *Automatic Nonuniform Random Variate Generation*. Springer-Verlag, Berlin.

Ibrahim, R., P. L'Ecuyer. 2013. Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing and Services Operations Management* **15**(1) 72–85.

Ibrahim, R., P. L'Ecuyer, N. Régnard, H. Shen. 2012. On the modeling and forecasting of call center arrivals. C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, A. M. Uhrmacher, eds., *Proceedings of the 2012 Winter Simulation Conference*. IEEE Press, 256–267.

Ibrahim, R., P. L'Ecuyer, H. Shen, M. Thiongane. 2015a. Inter-dependent, heterogeneous, and time-varying service-time distributions in call centers. *European Journal of Operational Research* To appear.

Ibrahim, R., P. L'Ecuyer, H. Ye, H. Shen. 2015b. Modeling and forecasting call center arrivals: A literature study and a case study. *International Journal of Forecasting* To appear.

Jaoua, A., P. L'Ecuyer, L. Delorme. 2013. Call type dependence in multiskill call centers. *Simulation* See `http://sim.sagepub.com/content/early/2013/04/01/0037549713479405`.

Jongbloed, G., G. Koole. 2001. Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry* **17** 307–318.

Kim, S.-H., W. Whitt. 2014a. Are call center and hospital arrivals well modeled by nonhomogeneous poisson processes? *Manufacturing and Service Operations Management* **16**(3) 464–480.

Kim, S.-H., W. Whitt. 2014b. Choosing arrival process models for service systems: Tests of a nonhomogeneous poisson process. *Naval Research Logistics* **61**(1) 66–90.

Koole, G. 2013. *Call Center Optimization*. MG books, Amsterdam.

Mehrotra, V. 1997. Ringing up big business. *ORMS Today* **24**(4) 18–24.

Nelsen, R. B. 1999. *An Introduction to Copulas*, *Lecture Notes in Statistics*, vol. 139. Springer-Verlag, New York, NY.

Pasupathy, R., S. Kim. 2011. The stochastic root finding problem: Overview, solutions, and open questions. *ACM Transactions on Modeling and Computer Simulation* **21**(3) Article 19.

Shen, H. 2010. Statistical analysis of call-center operational data: Forecasting call arrivals, and analyzing customer patience and agent service. J. J. Cochran, ed., *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley.

Soyer, R., 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, V. Mehrotra. 2009. Forecast errors in service systems. *Probability in the Engineering and Informational Sciences* **23**(2) 305–332.

Tanir, O., R. J. Booth. 1999. Call center simulation in Bell Canada. P. A. Farrington, H. B. Nemhard, D. T. Sturrock, G. W. Evans, eds., *Proceedings of the 1999 Winter Simulation Conference*. IEEE Press, Piscataway, New Jersey, 1640–1647.

Weinberg, J., L. D. Brown, 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. 1999a. Dynamic staffing in a telephone call center aiming to immediately answer all calls. *Operations Research Letters* **24** 205–212.

Whitt, W. 1999b. Improving service by informing customers about anticipated delays. *Management Science* **45**(2) 192–207.

Whitt, W. 2001. *Stochastic-Process Limits*. Springer-Verlag, New York, NY.