Forecasting and Staffing Call Centers with Multiple Interdependent Uncertain Arrival Streams

Han Ye
Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599, hanye@email.unc.edu

James Luedtke
Department of Industrial and Systems Engineering, University of Wisconsin-Madison, Madison, WI 53706, jrluedt1@wisc.edu

Haipeng Shen
Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599, haipeng@email.unc.edu

We consider forecasting and staffing call centers with multiple interdependent uncertain arrival streams. We first develop general statistical models that can simultaneously forecast multiple-stream arrival rates that exhibit inter-stream dependence. The models take into account several types of inter-stream dependence. With distributional forecasts, we then implement a chance-constraint staffing algorithm to generate staffing vectors and further assess the operational effects of incorporating such inter-stream dependence, considering several system designs. Experiments using real call center data demonstrate practical applicability of our proposed approach under different staffing designs. An extensive set of simulations is performed to further investigate how the forecasting and operational benefits of the multiple-stream approach vary by the type, direction, and strength of inter-stream dependence, as well as system design. Managerial insights are discussed regarding how and when to take advantage of the inter-stream dependence operationally.

Key words: cross training, vector time series, quality of service, chance constraint, stochastic programming

History:

1. Introduction
Call centers are modern service networks in which agents provide services to customers via telephones. They have become a primary contact channel between service providers and customers. There are several good reviews about research related to call center operations, including Gans et al. (2003) and Aksin et al. (2007). Inbound call centers take calls that are initiated by customers, and they use a hierarchical staffing and scheduling system. The process begins with forecasting call arrival rates over a planning horizon, ranging from a day to several weeks.

The current paper focuses on call centers with multiple uncertain call arrival streams, such as from multiple customer classes, that are dependent. For example, the numbers of English-speaking and French-speaking customers that call a major Canadian telecommunication company during the same time intervals are positively dependent (Ibrahim and L’Ecuyer 2013). We aim at answering
the following questions: what statistical models lead to better forecasts? do the better forecasts translate into improved operational performance? if they do, what kind of systems can benefit the most? To answer the questions, we first develop forecasting models that take into account the inter-stream dependence to give more accurate distributional forecasts. In addition, we use a chance-constraint formulation (Gurvich et al. 2010) to investigate the operational benefits of incorporating such dependence on quality of service (QoS) and staffing costs, and offer insights into how the benefits interact with the flexibility of the service system in terms of system design and cross-training.

Traditionally, call centers assume that arrival rate forecasts are accurate and use point forecasts of the arrival rates to drive the staffing and scheduling plans, which very often results in mismatch between the point forecasts and realized arrival volumes and hence improper staffing levels. A lot of progress has been made recently in the statistics and operations management (OM) literatures to better cope with arrival rate uncertainty. However, almost all the statistical work is about forecasting a single arrival stream, such as Avramidis et al. (2004), Brown et al. (2005), Weinberg et al. (2007), Taylor (2008), Shen and Huang (2008a), Shen and Huang (2008b), Aldor-Noiman et al. (2009), Matteson et al. (2011), Taylor (2012), and references therein. The one exception is the recent work of Ibrahim and L’Ecuyer (2013) that extends the one-stream model of Aldor-Noiman et al. (2009) to develop two mixed-effects models for forecasting bivariate arrival streams.

On the other hand, OM papers account for uncertainty when making workforce management decisions. Several papers use stochastic programming (SP) (Birge and Louveaux 1997) to account for arrival rate uncertainty when making staffing and call-routing decisions, including Harrison and Zeevi (2005), Bassamboo et al. (2005, 2006), Bassamboo and Zeevi (2009), Bertsimas and Doan (2010), Gurvich et al. (2010). More recent papers extend the SP formulation to scheduling, such as Robbins and Harrison (2010), Robbins et al. (2010), Liao et al. (2012). To cope with biased initial arrival-rate forecasts, Mehrotra et al. (2010) uses mid-day recourse actions to adjust pre-scheduled staffing levels. Some of the above papers deal with staffing/scheduling when there are multiple arrival streams. For example, Gurvich et al. (2010) proposes a chance-constraint formulation to staff multiple-stream call centers.

The above papers have made important progress addressing the problems caused by arrival-rate uncertainty, although only partially. Statistical forecasting papers have evaluated their methods using traditional forecasting accuracy measures based on realized arrival counts, while ignoring the operational effects the forecasting errors might have on cost and QoS measures. On the other hand, OM papers have carefully demonstrated the cost and QoS implications of their procedures,
assuming that the arrival rate distributions are given, although in practice they have to be estimated from data. Gans et al. (2012) is the only paper that aims at solving the whole problem, integrating arrival rate forecasting with stochastic programming to illustrate the operational effects of SP with or without recourse using arrival-rate distributions forecasted and updated from real data. As in almost all the forecasting papers, Gans et al. (2012) only considers a single arrival stream.

Our paper focuses on call centers with multiple uncertain arrival streams, and for the first time, integrates arrival rate forecasting and staffing together for such systems. We extend the univariate forecasting model of Gans et al. (2012) to the multivariate cases, to simultaneously provide distributional forecasts for multiple arrival streams. Given the distributional forecasts, we then implement the chance-constraint formulation proposed by Gurvich et al. (2010) to evaluate the downstream effects of the improved forecasts on staffing costs and QoS measures. We use a real call center application and extensive simulation studies to demonstrate the effects of incorporating inter-stream dependence on both forecasting accuracy and operational efficiency, and how the advantages of our multiple-stream approach vary by the type and strength of inter-stream dependence, as well as staffing designs. Hence our paper nicely complements the earlier works of Ibrahim and L’Ecuyer (2013), Gans et al. (2012), Gurvich et al. (2010).

More specifically, we make the following contributions. We introduce our statistical forecasting models in Section 2. Our general models take into account three types of inter-stream dependence. The models impose a multiplicative structure on the arrival rates, and treat them as products of daily-total rate and within-day proportion profile. This formulation reduces the dimensionality of the problem. We theoretically and numerically evaluate the forecasting benefits of incorporating inter-stream dependence of different type and strength. Furthermore, to test the operational effects of incorporating inter-stream dependence, we consider the chance-constraint staffing formulation with sample based approximation by Gurvich et al. (2010), which is briefly reviewed in Section 3. We integrate our forecasting method and the chance-constraint staffing formulation as an entire solution to staff call centers with multiple uncertain inter-dependent arrival streams. We test the forecasting and operational benefits of our forecasting approach using a real call center application in Section 4 and extensive simulation studies in Section 5. In particular, we consider 225 simulated scenarios of different types and strengths of inter-stream dependence as well as 9 staffing designs, to understand how the forecasting and operational benefits vary by direction and strength of dependence among the streams, as well as the infrastructure of the staffing system.

Our results suggest that the multiple-stream approach provides more accurate distributional forecasts; the stronger the dependence on the other streams’ past information, the better point forecasting improvement the multiple-stream approach has. Operationally, our forecasting approach
that incorporates inter-stream dependence leads to stable QoS performance regardless of staffing
design and type/strength of inter-stream dependence, while the approach that ignores such de-
dpendence causes mismatch between demand and staffing level, either understaffing or overstaffing,
and the severity of the mismatch increases when the staffing system becomes more flexible. This
suggests that it is even more crucial for flexible systems to use our multiple-stream forecasting
approach, in order to fully utilize the flexibility of the system. Finally, our study suggests that
cross-training improves cost-efficiency of the staffing system, and the magnitude of improvement
depends on the direction and strength of inter-stream dependence.

2. Forecasting Multiple Arrival Streams
We present our general statistical model for forecasting multiple arrival streams, and discuss various
types of inter-stream dependence that our model covers. We rigorously illustrate the forecasting
benefits of incorporating inter-stream dependence. We then develop a sequential iterative model
estimation algorithm, and obtain point and distributional forecasts for arrival rates and counts.

2.1. General Statistical Model
Denote the number of customer types (or arrival streams) as \( I \). For the \( i \)th arrival stream, \( i = 1, \ldots, I \), we observe the number of call arrivals during time period \( t \) on day \( d \), and denote it as \( N^{(i)}_{d,t} \) for \( t = 1, \ldots, T, d = 1, \ldots, D \). In practice, each time period can be either quarter hour or half hour.

We model \( N^{(i)}_{d,t} \) as a Poisson random variable with a random arrival rate \( \lambda^{(i)}_{d,t} \) that depends
on customer type \( i \), time period \( t \), and day \( d \) (most likely through day of the week, denoted as \( w_d \)) (Brown et al. 2005, Kim and Whitt 2013). Following common practice in call center arrival
forecasting (see Ibrahim and L’Ecuyer (2013) and references therein), we first apply the following
square-root transformation to normalize the arrival counts and stabilize the variance as \( \sigma^2_{(i)} \):

\[
X^{(i)}_{d,t} = \sqrt{N^{(i)}_{d,t} + \frac{1}{4}} \sim \mathcal{N} \left( \sqrt{\lambda^{(i)}_{d,t}}, \sigma^2_{(i)} \right).
\]

We then consider the following statistical model on the square-root-transformed counts \( X^{(i)}_{d,t} \):

\[
\begin{align*}
X^{(i)}_{d,t} &= \sqrt{\lambda^{(i)}_{d,t} + \epsilon^{(i)}_{d,t}}, \quad i = 1, \ldots, I, \\
\theta^{(i)}_{d,t} &= \sqrt{\lambda^{(i)}_{d,t} = u^{(i)}_d f^{(i)}_{w_d,t}}, \quad f^{(i)}_{w_d,t} \geq 0, \quad \sum_{t=1}^{T} f^{(i)}_{w_d,t} = 1, \\
u_d - \alpha_{w_d} &= A \left( u_{d-1} - \alpha_{w_{d-1}} \right) + z_d, \quad z_d = \left( z^{(1)}_d, z^{(2)}_d, \ldots, z^{(I)}_d \right) \sim \mathcal{N}(0, \Omega),
\end{align*}
\]

where \( w_d \) is the day-of-week of day \( d \), \( u_d = (u^{(1)}_d, u^{(2)}_d, \ldots, u^{(I)}_d)' \) is the vector of daily total arrival
rates of all customer streams (on the square-root scale), \( \alpha_{w_d} = (\alpha^{(1)}_{w_d}, \alpha^{(2)}_{w_d}, \ldots, \alpha^{(I)}_{w_d})' \) models the
day-of-week effect on daily total arrival rate (on the square-root scale), and \( f_{w_d,t}^{(i)} \) is the intraday rate proportion of the \( t \)th time interval for customer type \( i \) that also depends on day of week.

Model (1) can be viewed as the multi-arrival-stream extension of the forecasting model considered by Gans et al. (2012). It captures inter-day, intra-day and inter-stream dependences that have been identified in previous call center work and the current study (Section 4). We can understand the model in the following way. The square-rooted count data approximately follow a multivariate Gaussian distribution (Brown et al. 2010). On the square-root transformed scale, the arrival rate profile for any customer type \( i \) on day \( d \), i.e. \( \theta_d^{(i)} \equiv (\theta_{d,1}^{(i)}, \theta_{d,2}^{(i)}, \ldots, \theta_{d,T}^{(i)})' \), is assumed to have a multiplicative structure, where it can be modelled as the product of the daily total rate \( u_d^{(i)} \) and the intraday proportion profile, \( f_{w_d}^{(i)} \equiv (f_{w_d,1}^{(i)}, f_{w_d,2}^{(i)}, \ldots, f_{w_d,T}^{(i)})' \), that depends on the corresponding day-of-week. Such multiplicative structure has been proposed for call centers by as early as Whitt (1999), and shown to work well empirically by Brown et al. (2005), Shen and Huang (2008a,b), Gans et al. (2012). In addition, the vector of daily total rates of all customer types \( u_d \) follows a first-order vector autoregressive (VAR(1)) time series model (Reinsel 2003), after being adjusted for the day-of-week effect \( \alpha_{w_d} \). Diagnostics suggest that the model works well for our data.

Furthermore, Model (1) incorporates three types of inter-stream dependence as we now elaborate:

- Type (a) dependence: \( \Sigma = (\Sigma_{dt})_{I \times I} \) is the covariance matrix that accounts for inter-stream dependence among the (transformed) arrival counts, conditional on knowing the arrival rates;
- Type (b) dependence: \( A = (a_{st})_{I \times I} \) is the autoregressive coefficient matrix that models the inter-stream dependence among the daily total arrival rates;
- Type (c) dependence: \( \Omega = (\Omega_{st})_{I \times I} \) is the covariance matrix that captures the inter-stream dependence among the innovations \( z_d \).

### 2.1.1. An Illustrative Example: Bivariate Arrival Streams

To better understand the three types of inter-stream dependence, we consider a special case of Model (1) for bivariate arrival streams, i.e. \( I = 2 \). We first illustrate how the inter-stream dependence is modeled, and then compare our model with the two bivariate mixed models proposed by Ibrahim and L’Ecuyer (2013).

The two-dimensional model is as follows:

\[
\begin{align*}
\lambda_{d,t}^{(i)} &= \sqrt{\lambda_d^{(i)}} + \epsilon_{d,t}^{(i)}, \quad i = 1, 2, \\
\theta_d^{(i)} &= \frac{\lambda_d^{(i)}}{u_d^{(i)}} f_{w_d}^{(i)}, \quad i = 1, 2, \\
\theta_{d,t}^{(i)} &= \sqrt{\Lambda_d^{(i)}} + \epsilon_{d,t}^{(i)}, \quad i = 1, 2, \\
\left(\begin{array}{c}
u_d^{(1)} - \alpha_{w_d}^{(1)} \\ u_d^{(2)} - \alpha_{w_d}^{(2)} \end{array}\right) &= \left(\begin{array}{cc}
a_{11} & a_{12} \\ a_{21} & a_{22} \end{array}\right) \left(\begin{array}{c}
u_{d-1}^{(1)} - \alpha_{w_d-1}^{(1)} \\ u_{d-1}^{(2)} - \alpha_{w_d-1}^{(2)} \end{array}\right) + \left(\begin{array}{c}z_d^{(1)} \\ z_d^{(2)} \end{array}\right), \\
\left(\begin{array}{c}
u_d^{(1)} - \alpha_{w_d}^{(1)} \\ u_d^{(2)} - \alpha_{w_d}^{(2)} \end{array}\right) &\sim \mathcal{N}\left(\left(0, 0\right), \left(\begin{array}{cc}\Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{array}\right)\right)
\end{align*}
\]

According to Model (2), the two arrival streams are dependent in the following three ways:
Type (a) dependence: conditional on the arrival rates \((\theta^{(1)}_{d,t}, \theta^{(2)}_{d,t})'\), the two (square-rooted) counts, \(X^{(1)}_{d,t}\) and \(X^{(2)}_{d,t}\), are dependent as measured by the correlation \(r \equiv \Sigma_{12} / \sqrt{\Sigma_{11} \Sigma_{22}}\);

Type (b) dependence: there is a bivariate lag-1 dependence among the daily total arrival rates, \(u^{(1)}_d\) and \(u^{(2)}_d\), and the lag-1 cross dependence is measured by \(a_{12}\) and \(a_{21}\);

Type (c) dependence: the innovations in the bivariate lag-1 equations of the daily total arrival rates are dependent, measured by the correlation \(\rho \equiv \Omega_{12} / \sqrt{\Omega_{11} \Omega_{22}}\).

Ibrahim and L’Ecuyer (2013) proposed two additive mixed models specifically for bivariate arrival streams. Different from our formulation, they imposed an additive structure on the arrival rates, taking into account day-of-week and time-of-day effects. In addition, each of their two models respectively considered one type of inter-stream dependence: dependence of the counts between the two streams of the same day and same time interval, as in our Type (a) dependence, and dependence of the two daily rates of the same day, similar to our Type (c) dependence. However, they did not consider the Type (b) dependence, which we later argue is crucial in reducing forecasting errors. Hence, our model is more general and flexible in that it works for any number of arrival streams and allows more general inter-stream dependence. Our numerical studies (Section 4.4) also suggest that our approach reaches the QoS target better at the downstream staffing stage, and our estimation algorithm is much faster than the fitting approach used for their mixed models.

### 2.2. Forecasting Error

We show theoretically that our Model (1) will not increase the forecasting error over alternative models that are based on only a subset of the arrival streams. As a result, fitting Model (1) simultaneously on all arrival streams (in theory) will perform at least as well as the alternative approach that models each arrival stream separately. Later in Sections 5 we empirically demonstrate such improvement via simulation studies when the models have to be estimated from data.

To fix ideas, let \(y\) denote the random variable that we want to forecast, with mean \(\mu_y\) and variance \(\Gamma_y\), and let \(\xi_n, n = 1, 2, \ldots\), denote a series of random variables that we use to build the forecasting model. Define \(\Gamma_n \equiv \text{Cov}(y, \xi_n), \Gamma_{(n)} = (\Gamma_1, \Gamma_2, \ldots, \Gamma_n)', \xi_{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)', \mu_{(n)} = \text{E}(\xi_{(n)}), \text{m}_{s,l} = \text{Cov}(\xi_s, \xi_l), s, l = 1, 2, \ldots, \text{M}_{(n)} = \text{Cov}(\xi_{(n)}) = (m_{s,l})_{n \times n}, \text{m}_{(n+1)} = (m_{1,n+1}, m_{2,n+1}, \ldots, m_{n,n+1})'\). We assume that \(\text{M}_{(n)}\) is non-singular, \(n = 1, 2, \ldots\).

Consider the forecasting variances of \(y\) when using either \(\xi_{(n)}\) or \(\xi_{(n+1)}\) as the forecaster, denoted as \(\tilde{\Gamma}_n\) and \(\tilde{\Gamma}_{n+1}\), respectively. Proposition 1 suggests that using \(\xi_{(n+1)}\) never increases the forecasting variance, which makes intuitive sense as there is no less information in \(\xi_{(n+1)}\). In the current context, the results suggest that it won’t hurt to use all available arrival streams (rather than a
subsets) to build the forecasting model, barring any estimation errors. The proof of the proposition follows naturally from multivariate normal theory, see Section A of the online supplement.

**Proposition 1.** Assuming that $\xi_{n+1}$ is non-deterministic given $\xi(n)$, then

$$\tilde{\Gamma}_n - \tilde{\Gamma}_{n+1} = \frac{(\Gamma_{n+1} - \Gamma^T_{(n)} M^{-1}_{(n)} m_{n(n+1)})^2}{m_{n+1,n+1} - m^T_{(n)} M^{-1}_{(n)} m_{n(n+1)}} \geq 0. \quad (3)$$

Considering Proposition 1 in the context of Model (1), we can understand the forecasting variance reduction obtained by introducing more arrival streams in our forecasting model. Suppose we forecast $u_{d}^{(1)}$ using $u_{d-1}^{(1)}, u_{d-1}^{(2)}, \ldots, u_{d-1}^{(I)}$, Proposition 1 suggests that including more streams in the forecasting model will not increase the forecasting variance. The variance reduction, under the autoregressive structure in Model (1), is a function of $A$ and $\Omega$, but has a complicated form. However, in the two-dimension example, when there is no Type (b) dependence, i.e. $a_{12} = a_{21} = 0$, we can show that the variance reduction (3) is zero, suggesting there is no need to consider a bivariate forecasting approach when the model does not incorporate Type (b) dependence. When $a_{21} \neq 0$ and $a_{12} \neq 0$, the variance reduction (3) is non-zero, and we will understand the effects of Type (b) dependence in the simulation studies of Section 5. The above theoretical study suggests that considering Type (b) dependence is essential for improving point forecast accuracy. This finding provides one plausible explanation for why the two bivariate models of Ibrahim and L’Ecuyer (2013) did not improve point forecast accuracy over the univariate mixed model, because their bivariate models didn’t consider the Type (b) dependence between two arrival streams.

### 2.3. Model Estimation

Model (1) is in fact a Hidden Markov Model (HMM) (Cappé et al. 2005), since the autoregressive daily total rates $u_{d}^{(i)}$ are unobserved. HMMs can be estimated using maximum likelihood. However, the very general dependence structures assumed in Model (1) makes direct maximum likelihood computationally very expensive, hence less practical. Here we propose to use a sequential algorithm to estimate all the parameters involved in Model (1). We first describe how to obtain estimates for $u_{d}^{(i)}, f_{w_d,t}^{(i)}$, and $\Sigma$ for $i = 1, \ldots, I$, $t = 1, \ldots, T$, $d = 1, \ldots, D$, and then discuss estimation of the vector time series parameters $\alpha_{w_d}, A$ and $\Omega$. For univariate arrival forecasting, Shen and Huang (2008a) have shown that the sequential approach performs similarly to maximum likelihood.

**Estimation of $u_{d}^{(i)}, f_{w_d,t}^{(i)}$ and $\Sigma$.** Consider the following model equation:

$$X_{d,t}^{(i)} = u_{d}^{(i)} f_{w_d,t}^{(i)} + \epsilon_{d,t}^{(i)}, \quad i = 1, \ldots, I, \quad \epsilon_{d,t} = \left(\epsilon_{d,t}^{(1)}, \epsilon_{d,t}^{(2)}, \ldots, \epsilon_{d,t}^{(I)}\right) \overset{iid}{\sim} \mathcal{N}(0, \Sigma),$$

with the constraints $f_{w_d,t}^{(i)} \geq 0$ and $\sum_{t=1}^{T} f_{w_d,t}^{(i)} = 1$. This is a multiplicative linear model that can be estimated using an iterative General Least Squares (GLS) algorithm as we discuss below.
1. Initialize:

   (a) Estimate \( f_{\text{old}}^{(i),d,t} \) with the proportion of the transformed counts of time interval \( t \) out of all transformed counts in the same day-of-week:
   \[
   \hat{f}_{\text{old}}^{(i),d,t} = \frac{\sum_{d':w_{d'}=w_d} X_{d',t}^{(i)}}{\sum_{d':w_{d'}=w_d} \sum_{t'} X_{t',t'}^{(i)}}.
   \]

   (b) Fit Ordinary Least Squares (OLS) to the following model to obtain \( \hat{u}_{\text{old}}^{(i),d} \):
   \[
   X_{d,t}^{(i)} = \hat{u}_{\text{old}}^{(i),d} + \epsilon_{d,t}^{(i)}, \quad \epsilon_{d,t}^{(i)} \sim \mathcal{N}(0, \sigma^2), \quad i = 1, 2, \ldots, I, \quad d = 1, 2, \ldots, D, \quad t = 1, 2, \ldots, T.
   \]

   (c) Use the empirical covariances among the model residuals to get \( \hat{\Sigma}^{\text{old}} \).

2. Update:

   (a) Fit GLS to the following model to get the updated estimates \( \hat{f}_{\text{new}}^{(i),d,t} \):
   \[
   X_{d,t}^{(i)} = \hat{u}_{\text{old}}^{(i),d} + \epsilon_{d,t}^{(i)}, \quad \epsilon_{d,t}^{(i)} \sim \mathcal{N}(0, \hat{\Sigma}^{\text{old}}),
   \]

   (b) Obtain the updated estimate \( \hat{\Sigma}^{\text{new}} \) using the model residual covariance matrix.

   (c) Fit GLS to the following model to get the updated \( \hat{u}_{\text{new}}^{(i),d} \):
   \[
   X_{d,t}^{(i)} = \hat{f}_{\text{new}}^{(i),d,t} \quad \hat{u}_{\text{new}}^{(i),d} + \epsilon_{d,t}^{(i)}, \quad \epsilon_{d,t}^{(i)} \sim \mathcal{N}(0, \hat{\Sigma}^{\text{new}}),
   \]

   (d) Update \( \hat{\Sigma}^{\text{new}} \) with the model residual covariance matrix.

3. Repeat Step 2 setting \( \hat{u}_{\text{old}}^{(i),d} = \hat{u}_{\text{new}}^{(i),d} \) and \( \Sigma^{\text{old}} = \Sigma^{\text{new}} \) until convergence such that
   \[
   \sqrt{\frac{T}{T}} \sum_{t=1}^{T} \left( \hat{f}_{\text{old}}^{(i),d,t} - \hat{f}_{\text{new}}^{(i),d,t} \right)^2 / T < 10^{-8}, \quad \forall w_d, i.
   \]

At convergence, \( \hat{u}_{\text{new}}^{(i),d} \) and \( \hat{f}_{\text{new}}^{(i),d,t} \) are the final estimates for \( u_{\text{old}}^{(i),d} \) and \( f_{\text{old}}^{(i),d,t} \), respectively. Then we use the residual covariance matrix to estimate \( \hat{\Sigma} \) with d.f. = \( DT - D - w^* \cdot (T - 1) \), where \( w^* \) is the number of working days in a week. For example, \( w^* = 5 \) for centers that only open on weekdays.

Note that fitting GLS regression can be time-consuming. Alternatively, we can replace GLS with OLS in the above algorithm, which results in great reduction on computing time. This alternative provides estimates that are quite close to those from the GLS method. In our call center application (Section 4), the relative distance of \( \{f_{\text{old}}^{(i),d,t}\}_{t=1,\ldots,T} \) between the two estimates is around 0.02%.
Estimation of \( \alpha_{wd}, \lambda \) and \( \Omega \). Given the estimates \( \hat{u}_d, \hat{f}_{wd,t}^{(i)} \) and \( \hat{\Sigma} \) obtained from the first stage estimation, we have the following vector time series model:

\[
\hat{u}_d - \alpha_{wd} = A \cdot (\hat{u}_{d-1} - \alpha_{wd-1}) + z_d, \quad z_d \sim \mathcal{N}(0, \Omega), \quad d = 2, \ldots, D.
\]

Under the Gaussian assumption, the above model can be fitted using maximum likelihood to obtain estimates \( \hat{\alpha}_{wd}, \hat{A} \) and \( \hat{\Omega} \).

### 2.4. Point and Distributional Forecast

Given the above parameter estimates, we can obtain the point and distributional forecasts for future arrival rates and counts as follows. For a positive integer \( h \), the \( h \)-step-ahead point forecast for the daily total rate on day \( D + h \), \( u_{D+h} \) is given by

\[
\hat{u}_{D+h} = \hat{\alpha}_{wd} + \hat{A}^h \cdot (\hat{u}_D - \hat{\alpha}_{wd}),
\]

with the forecasting error as \( \sum_{h=0}^{h-1} \hat{A}^h z_{D+h} \), where \( z_{D+h} \sim \mathcal{N}(0, \hat{\Omega}) \). In particular, the covariance matrix of the forecast error of \( \hat{u}_{D+h} \) is

\[
\hat{\Omega}^{(D+h)} = \sum_{h=0}^{h-1} \hat{A}^h \hat{\Omega} (\hat{A}^h)'\]

Given the mean in (6) and variance in (7), and under the Gaussian assumption, the distributional forecast for the daily arrival rate \( u_{D+h} \) is

\[
u_{D+h} \sim \mathcal{N}\left(\hat{u}_{D+h}, \hat{\Omega}^{(D+h)}\right).
\]

Define \( \hat{F}_{D+h,t,} = \text{diag}\{\hat{f}_{wd+h,t}^{(i)}, \ldots, \hat{f}_{wd+h,t}^{(I)}\} \). Then the distributional forecast for the arrival rate vector \( \theta_{D+h,t} \equiv (\theta_{D+h,t}^{(1)}, \ldots, \theta_{D+h,t}^{(I)})' = \hat{F}_{D+h,t} u_{D+h} \) is as follows:

\[
\theta_{D+h,t} \sim \mathcal{N}\left(\hat{F}_{D+h,t} \hat{u}_{D+h}, \hat{F}_{D+h,t} \hat{\Omega}^{(D+h)} \hat{F}_{D+h,t}'\right).
\]

In particular, the point forecast for \( \theta_{D+h,t}^{(i)} \) is \( \hat{f}_{wd+h,t}^{(i)} \cdot \hat{u}_{D+h} \), \( i = 1, \ldots, I \), and the forecasting covariance between \( \theta_{D+h,t}^{(i)} \) and \( \theta_{D+h,t}^{(i')} \) is \( \hat{f}_{wd+h,t}^{(i)} \cdot \hat{\Omega}^{(D+h)} \cdot \hat{f}_{wd+h,t}^{(i')}' \), where \( \hat{\Omega}^{(D+h)} \) is the \((i,i')\)th entry of \( \hat{\Omega}^{(D+h)} \). The point and distributional forecasts for the arrival rate \( \lambda_{D+h,t}^{(i)} \) naturally follow.

In terms of the arrival counts, define \( X_{D+h,t} \equiv (X_{D+h,t}^{(1)}, \ldots, X_{D+h,t}^{(I)})' \) and \( \epsilon_{D+h,t} \equiv (\epsilon_{D+h,t}^{(1)}, \ldots, \epsilon_{D+h,t}^{(I)})' \sim \mathcal{N}(0, \hat{\Sigma}) \). Notice that

\[
X_{D+h,t} = \theta_{D+h,t} + \epsilon_{D+h,t}.
\]

Then the distributional forecast for the arrival count vector \( X_{D+h,t} \) is

\[
X_{D+h,t} \sim \mathcal{N}(\hat{F}_{D+h,t} \hat{u}_{D+h}, \hat{F}_{D+h,t} \hat{\Omega}^{(D+h)} \hat{F}_{D+h,t}' + \hat{\Sigma}).
\]

In particular, the point forecast of \( X_{D+h,t}^{(i)} \) is \( \hat{f}_{wd+h,t}^{(i)} \cdot \hat{u}_{D+h} \), and the forecasting covariance between \( X_{D+h,t}^{(i)} \) and \( X_{D+h,t}^{(i')} \) is \( \hat{f}_{wd+h,t}^{(i)} \cdot \hat{\Omega}^{(D+h)} \cdot \hat{f}_{wd+h,t}^{(i')}' + \hat{\Sigma}_{ii'} \).
2.5. Forecasting Performance Measures

We consider two measures to evaluate the accuracy of point forecasts. For any single arrival stream, let $\lambda_{d,t}$ denote the arrival rates in day $d$ and time interval $t$, and let $\hat{\lambda}_{d,t}$ denote the corresponding point forecast. Suppose we are interested in forecasting the arrival rates for one day. Then the Root Mean Squared Error (RMSE) and Mean Relative Error (MRE) for Day $d$ are defined as follows:

\[
\text{RMSE}_d = \sqrt{\frac{1}{T} \sum_{t=1}^{T} (\hat{\lambda}_{d,t} - \lambda_{d,t})^2}, \quad \text{MRE}_d = \frac{100}{T} \sum_{t=1}^{T} \frac{|\hat{\lambda}_{d,t} - \lambda_{d,t}|}{\lambda_{d,t}}.
\]

Similar measures can be defined for the arrival counts.

3. Operational Performance

Recently, operations management papers attempt to account for arrival rate uncertainty while making staffing and scheduling policies. Among those, Gurvich et al. (2010) proposed a multiple stream chance-constrained approach providing staffing levels that meet the uncertain demand at a pre-chosen probability level. In their approach, they assume the existence of a forecasting distribution for the arrival rates before the staffing planning process can be taken. In this paper we consider a simplified version of their model to explore the operational effects of simultaneously modeling multiple arrival streams instead of independently modeling each of them. We remark, however, that the improved forecasts obtained from modeling multiple arrival streams simultaneously have potential to improve performance of any staffing procedure that accepts as input a multivariate distribution of arrival rates.

3.1. A Brief Review of the Chance Constraint Formulation

Consider a call center with $I$ customer classes and $J$ server pools. Set $I = \{1, \ldots, I\}$ and $J = \{1, \ldots, J\}$. Servers in the same pool have the same skills in terms of the set of customer classes they are capable of serving and the corresponding service rates. Denote $J(i)$ as the set of server pools with skill $i$, and $I(j)$ as the set of skills that server pool $j$ has. The staffing vector is denoted by $N = (N_1, N_2, \ldots, N_J)'$ where $N_j$ denotes the number of agents from server pool $j$ to have staffed.

We consider the call center as a parallel server system, where customers go through a single stage of service before departing from the system.

The staffing process is performed for one time interval and all the discussion following is focused on an arbitrary time interval. Gurvich et al. (2010) considered a dynamic queueing model that requires determining staffing levels for the server pools and a control policy for dynamically assigning customers to servers. In the first step of their solution procedure, they approximate the dynamic queueing process with a static model that only considers the forecast uncertainty to obtain an
initial staffing solution, which they then modify using a simulation-based procedure to meet QoS constraints. To simplify our experiments, we directly use the static model as the system we study.

In this static model, the number of class-$i$ customers that arrive in a period is $\Lambda_i$, where $\Lambda = (\Lambda_1, \ldots, \Lambda_I)$ is a multivariate random variable following a known (forecasted) distribution. The staffing vector $N$ must be chosen before knowing the true arrival counts. After $N$ is chosen, the arrival counts are revealed, and then customers are (fractionally) allocated to servers. The service rate of a class-$i$ customer when served by a pool-$j$ server is $\mu_{ij}$, for $j \in J(i)$. It is assumed that $\psi_i$ proportion of class-$i$ customers are allowed to abandon, and hence these customers do not need to be served. Thus, in this model the QoS targets will all be met if the system of inequalities,

$$\sum_{j \in J(i)} \mu_{ij} v_{ij} \geq \Lambda_i (1 - \psi_i), \quad i \in I,$$

$$\sum_{i \in I(j)} v_{ij} \leq N_j, \quad j \in J,$$

has a feasible solution, where $v_{ij}$ is the number of pool-$j$ servers allocated to serve class-$i$ customers.

Let $\delta > 0$ be a given risk-level, representing the proportion of periods in which we are willing to allow our QoS targets to be violated. Then, the Random Static-Planning Problem (RSPP) of Gurvich et al. (2010) is as follows:

$$\min \quad c \cdot N$$

subject to

$$\mathbb{P}(\Lambda \in \mathcal{B}(N)) \geq 1 - \delta,$$

$$N \in \mathbb{R}_+^J,$$  \hspace{1cm} (8)

where

$$\mathcal{B}(N) = \{ \Lambda \in \mathbb{R}_+^I : \exists v \in \mathbb{R}_+^{I \times J}, \text{ with } \sum_{j \in J(i)} \mu_{ij} v_{ij} \geq \Lambda_i (1 - \psi_i), i \in I, \sum_{i \in I(j)} v_{ij} \leq N_j, j \in J \}.$$

A vector $\lambda \in \mathbb{R}_+^I$ is in $\mathcal{B}(N)$ if there exists an allocation of the servers specified by the staffing solution $N$ such that the QoS target can be met. Therefore, (8) finds the minimum cost staffing solution that meets the quality-of-service target with probability $1 - \delta$.

To solve the RSPP, Gurvich et al. (2010) used a discrete approximation of the random arrival rate $\Lambda$ and formulated the RSPP as a mixed-integer program. They considered two discretization methods (fix grid approximation and Monte Carlo sampling), among which we use the Monte Carlo sampling approximation method. In particular, independent samples of size $K$ are generated from the distribution of $\Lambda$ and each sample point is assigned the same probability $1/K$. Denote the $k^{th}$ sample by $\Lambda(k) = (\Lambda_1(k), \ldots, \Lambda_I(k))'$. Hence the sample-based RSPP is given by:
\[ \begin{align*}
\text{min} & \quad c \cdot N \\
\text{subject to} & \quad \sum_{j \in J(i)} \mu_{ij} v_{ij}^{(k)} \geq y_k \Lambda_i(k)(1 - \psi_i), \quad i \in I, \ k = 1, \ldots, K, \\
& \quad \sum_{i \in I(j)} v_{ij}^{(k)} \leq N_j, \quad j \in J, k = 1, \ldots, K, \\
& \quad \sum_{k} y_k \geq K(1 - \delta), \\
& \quad N \in \mathbb{R}_+^J, \ y_k \in \{0, 1\}, v^{(k)} \in \mathbb{R}^I \times J, k = 1, \ldots, K. 
\end{align*} \] (9)

For each sample, a binary variable \( y_k \) is introduced, where \( y_k = 1 \) implies the QoS targets are met for that sample, and a set of continuous variables \( v_{ij}^{(k)}, i \in I, j \in J \) are introduced, representing the allocation of server type \( i \) to customer type \( j \) in sample \( k \). Then the constraint \( \sum_k y_k \geq K(1 - \delta) \) ensures that the QoS targets are met in a high proportion of the samples. Luedtke and Ahmed (2008) show how the sample-based RSPP can be used to construct feasible solutions and statistical optimality bounds for the original, unsampled, RSPP. In particular, it is shown that, with high probability, solving a sample-based RSPP with max violation target \( \delta \) yields a solution feasible to the RSPP with target \( \delta + \gamma(K) \), where \( \gamma(K) \approx O(\sqrt{I/K}) \). In addition, errors in the forecast distribution will further lead to a mismatch between the target \( \delta \) used in the sample-based RSPP and the actual violation probability of the resulting solution. Thus, while we cannot expect the violation probabilities will exactly match the target, the extent to which a solution obtained using a forecasting method matches the target is an important metric for comparing forecasting methods.

### 3.2. Operational Performance Measures

Given a distributional forecast of \( \Lambda = (\Lambda_1, \ldots, \Lambda_I)' \), a staffing vector \( N \) can be obtained with the above algorithm. Given the realization of \( \Lambda \), we are able to evaluate the performance of the staffing vector and further the distributional forecast.

Given realized counts \( C \) and a staffing vector \( N \), we solve the following linear program, which minimizes the amount by which the QoS constraint is violated:

\[ \begin{align*}
\text{min} & \quad \sum_{i \in I} y_i \\
\text{subject to} & \quad \sum_{j \in J(i)} \mu_{ij} v_{ij} + y_i \geq C_i(1 - \psi_i), \quad i \in I, \\
& \quad \sum_{i \in I(j)} v_{ij} \leq N_j, \quad j \in J, \\
& \quad v \in \mathbb{R}^I \times J, \ y \in \mathbb{R}_+^I.
\end{align*} \]

Let \( s(C, N) \) be the optimal value of this linear program, so that \( s(C, N) \) measures how much the QoS constraint is violated. Denote \( v(C, N) \) as the violation indicator function for the QoS constraint: \( v(C, N) = 1 \) indicates that the QoS constraint is violated (\( s(C, N) > 0 \)), and \( v(C, N) = 0 \).
indicates that the QoS constraint is satisfied \( s(C, N) = 0 \). Denote \( c(N) := c \cdot N \) as the staffing cost under staffing vector \( N \). Suppose we’ve tested the performance of the staffing vector for \( D' \) days. Let \( C^{(d, t)} \) and \( N^{(d, t)} \) denote the observed counts and staffing vector for the \( t \)th interval in the \( d \)th day, respectively. The violation probability is defined as

\[
V.\text{PROB} := \frac{1}{D'T} \sum_{d=1}^{D'} \sum_{t=1}^{T} v(C^{(d, t)}, N^{(d, t)}).
\]

Denote the daily staffing cost on day \( d \) as

\[
\text{COST}_d := \sum_{t=1}^{T} c(N^{(d, t)}).
\]

4. Real Call Center Application

We present a real call center application to demonstrate the practical applicability of our multiple-stream forecasting method, illustrating both the statistical and operational benefits of incorporating inter-stream dependence in forecasting multiple-stream arrivals as well as staffing call centers with multiple customer classes. We first provide the relevant background information for the data in Section 4.1, and then report the empirical results with some discussion about managerial insights in Sections 4.2 and 4.3, and lastly compare our methods with the existing ones in Ibrahim and L’Ecuyer (2013) in Section 4.4.

4.1. Background of the Call Center Data

The data we analyze were collected at an Israel telecom call center, provided to us as a courtesy from the Technion Service Enterprise Engineering Lab. There are several customer types, with Private and Business being the two major ones, which account for about 30% and 18% of the overall incoming calls respectively. Aldor-Noiman et al. (2009) presented an earlier workload forecasting study using a subset of the data, focusing on the single arrival stream of Private customers. More details of the data can be found in Aldor-Noiman et al. (2009).

We focus on the two main arrival streams: Private and Business, and demonstrate the effects of incorporating inter-stream dependence on forecasting accuracy and staffing performance measures. Note that our method in general can be applied to more than two arrival streams.

The data cover 300 days between 06/19/2004 and 04/14/2005. For both customer types, the call center operates from 7:00 am to midnight everyday, while the call volumes are very low on Fridays and Saturdays compared with the other weekdays, so we focus on the weekdays from Sunday to Thursday, which result in 215 days. For each day we divide the 17 working hours into 34 half-hour
intervals, and record the arrival count during each interval. The square-root transformation is then applied to the arrival counts to stabilize the variance and normalize the data.

Figures 1 and 2 plot the square root transformed arrival counts of Private and Business customers, respectively. Within each figure, the left panel displays the call volume profiles for each one of the 215 weekdays, where different colors are used for different days of the week; the middle panel shows the average arrival volume profile for each day-of-the-week, highlighting the day-of-the-week effect on the within-day arrival pattern; finally, the right panel depicts the time series plot of the daily-total arrival volumes against the day index, with different days of the week indicated by different symbols and colors. The two arrival streams exhibit similar patterns in both the within-day arrival profiles and the daily total volumes: the within-day profiles have two peaks around 1:00pm and 6:00pm, Sundays have the highest total volume relative to the other weekdays, and there are significant day-of-the-week effects on both the within-day arrival profile and the daily total volume. Model (2) is designed to capture the above empirical features in the data. Section E
of the supplement contains some diagnostic results that suggest Model (2) fits our data well.

Figure 3 suggests that the two arrival streams are dependent on each other. The left panel shows the scatter plot of the daily total arrival volumes on the transformed scale with the day-of-the-week effect removed. The correlation between the two time series is around 0.72 which is fairly strong. The right panel is the scatter plot for the square-rooted interval call volumes between the two customer types after we remove both the day-of-the-week and the interval effects. Fourteen outliers are excluded from the plot. We observe a moderately strong correlation of 0.38.

We refer to the Private arrivals as Type 1 and the Business arrivals as Type 2. The two-dimensional Model (2) is fitted to the data. Table 1 reports the corresponding estimates for the inter-stream dependence parameters. We observe that both the daily-total dependence (Type (c)) and the interval dependence (Type (a)) between the two streams are strong, with the correlation being 0.6671 and 0.6875 respectively, while the cross-lag dependence (Type (b)) is relatively weak in this case, being 0.0922 and -0.1018.

### 4.2. Forecasting Performance Comparison

#### 4.2.1. Competing Methods
To examine whether accounting for inter-stream dependence helps improve forecasting performance, we first consider the following two forecasting methods:

- MU1: fit Model (2) separately on data from each arrival stream, which ignores the dependence between the two arrival streams and essentially considers them as independent;
Table 2 Rolling point forecast comparison based on RMSE and MRE.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE</th>
<th>MRE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Private</td>
<td>Business</td>
</tr>
<tr>
<td>HA</td>
<td>Median</td>
<td>Mean</td>
</tr>
<tr>
<td>MU1</td>
<td>36.80</td>
<td>38.53</td>
</tr>
<tr>
<td>MU1-AR</td>
<td>36.83</td>
<td>38.51</td>
</tr>
<tr>
<td>MU2</td>
<td>37.10</td>
<td>38.50</td>
</tr>
<tr>
<td>MU2G</td>
<td>37.14</td>
<td>38.47</td>
</tr>
</tbody>
</table>

- MU2: fit Model (2) simultaneously on the two arrival streams.

Since the only difference between the two methods is whether the inter-stream dependence is taken into account, any observed difference between the resulting forecasting distributions can then be attributed to the inter-stream dependence.

In addition, we also consider the following three forecasting methods:

- HA: use the historical average of the same day-of-week as the forecast. This method has been repeatedly used as the benchmark in, for example, Weinberg et al. (2007), Shen and Huang (2008a), Aldor-Noiman et al. (2009), Ibrahim and L’Ecuyer (2013).

- MU1-AR: fit Model (2) separately on the two streams, and at the same time impose a within-day AR(1) correlation structure to \( \{\epsilon_{d,t}\}_{t=1,2,...,T} \), similar to Aldor-Noiman et al. (2009) and Ibrahim and L’Ecuyer (2013).

- MU2G: fit Model (2) simultaneously on the two streams, and use the GLS estimation.

### 4.2.2. Point Forecast Comparison

The following rolling forecasting exercise is performed: for each day between Day 101 and Day 215, we use its previous 100 consecutive days to generate the distributional forecast of the arrival profiles for that day using the five methods, and calculate the RMSE and MRE as defined in Section 2.5. As a result, for each forecasting method, we have 115 values of RMSE and MRE, the median and mean of which are summarized in Table 2.

In Table 2, the best performer under each measure is highlighted in boldface. The comparison suggests that, in terms of point forecasting accuracy, the historical average method performs the worst, and there is no strong evidence that a single forecasting method gives the “best” point forecast since the forecasting accuracy depends on the performance measure and customer type. Paired t-tests based on RMSE and MRE further confirm that

- HA performs significantly worse than any one of the other four methods, which again supports the incorporation of dependence;

- MU1 and MU1-AR perform similarly, which offers no evidence that imposing the within-day AR(1) structure helps improve the point forecasts;
There is no solid evidence to conclude that MU2 gives more accurate point forecast than MU1, which is consistent with the theoretical discussion in Section 2.2 since the particular real data possesses very weak Type (b) dependence.

- MU2G slightly improves over MU2, even though MU2G takes much longer to compute.

**4.2.3. Distributional Forecast Comparison** Although MU1 and MU2 perform comparably in terms of point forecast accuracy, the distributional forecasts between MU1 and MU2 are significantly different. In the 115 rolling forecasting experiments, MU2 estimates the Type (a) correlation parameter $r$ to be between 0.6 and 0.7, and MU2G gives similar correlation estimates. As an illustration, Figure 4 shows the mesh plots of the forecasting distributions offered by MU1 and MU2 for one randomly-chosen time interval in the forecasting exercise, with the parameter estimates appropriately rounded. Note that the shape of the contours is affected by the different correlation estimates. Such difference between the forecasting distributions helps to explain the operational benefits of MU2, reported below in Section 4.3.

**4.3. Operational Performance Comparison**

We now compare the operational effects of using the forecasts generated by MU1 and MU2, via extensive staffing experiments based on the 115-day rolling forecasting exercise in Section 4.2. For illustration purposes, we implement the chance-constraint staffing formulation of Gurvich et al. (2010) as reviewed in Section 3, and feed the forecasting distributions into it to obtain the staffing costs and the associated quality of service (QoS) measures. More specifically, we consider three system structures with a range of system parameters, and obtain managerial insights about how system design and cross-training affect staffing cost and QoS performance. The comparisons clearly demonstrate the operational benefits of accounting for inter-stream dependence in the forecasting stage, and reveal an important message for the managers - the more flexible a staffing design is, the more crucial it is to appropriately incorporate inter-stream dependence.
4.3.1. Effects of System Designs Note that the staffing decision given by the chance-constraint program depends on the specific structure of the staffing system and the corresponding parameters. With two arrival streams, we consider three interesting system designs as shown in Figure 5, which are referred as the I-design, (or the II-design in our case), the M-design, and the X-design (Gans et al. 2003). These staffing designs cover a wide range of system complexity and flexibility, as we now discuss:

- the II-design: there are two dedicated server pools, each one serving one customer class.
- the M-design: there are two dedicated server pools, one for each customer class. There is also a flexible server pool, that serves both arrival streams. In comparison, the M-design adds a flexible server pool to the II-design.
- the X-design: there are two separate pools of cross-trained servers: the servers in each pool primarily serve one particular customer class, although they can serve the other customer class if needed. Compared with the II-design, the X-design allows resource sharing between the two classes, for example, when there are overloads in one or both classes (Perry and Whitt 2009).

We expect that the flexibility level of a staffing design interacts with the comparison between MU1 and MU2. A more flexible system shall be more capable of taking advantage of the benefits from incorporating inter-stream dependence; hence MU2 shall lead to more operational benefits in a more flexible system. In addition to the various design structures in Figure 5, we consider two more factors that affect system flexibility level:

- salary of the servers who serve more than one customer classes, i.e. $c_2$ of the M-design in Figure 5;
- service rate of the cross-trained servers when they handle their non-primary customer class, i.e. $\mu_{12}$ and $\mu_{21}$ of the X-design in Figure 5.
It is natural to consider these three factors. First, managers in large-scale call centers often cross train servers to increase system flexibility, which results in a more complex staffing structure. Second, cross-trained servers with more skills usually get higher pay, which on the other hand makes the system less cost-efficient. Third, cross-trained servers may be slower in serving their non-primary customers compared with the dedicated servers, which to some degree reduces the system operational efficiency. In summary, the system becomes more flexible, when one increases the number of server pools, or decreases the cost for cross-trained servers, or increases the non-primary service rate of cross-trained servers. Below we perform two numerical comparisons to present the effects of these three factors on forecasting and staffing performance. We then generalize some managerial insights from the two comparisons.

**Comparison 1: “II” vs. “M”**. We consider five M-designs, with different flexible server costs. The parameter settings are presented in Table 3, where the M-designs are arranged in the order of decreasing system flexibility. The violation probability target $\delta$ is 0.05, the allowed abandonment proportion $\psi_i$ is 0.04 for both customer classes, and the service rate $\mu_{ij}$ is 1 across all server pools and customer classes. Each dedicated server costs 1, while the flexible server cost $c_2$ ranges between 1.1 and 2.

The various settings are chosen in a way so that the II-design can be viewed as the limit of the various M-designs. More specifically, when a single flexible server costs as much as two dedicated servers, the M2.0-design is basically the II-design.

Figure 6 displays the results of the staffing experiment under the various settings. The left panel compares the daily mean of the realized violation probability, obtained from averaging over the 115 out-of-sample forecasting days, while the right panel compares the corresponding daily mean of the staffing cost between MU1 and MU2. The following observations can be made:

- MU2 is more stable than MU1 in violation probability across all the settings; MU1’s violation probability increases (more severe understaffing) when the system becomes more flexible (that is,
when the flexible server cost decreases). An intuitive explanation is that MU1 pays more penalty for ignoring inter-stream dependence in a more flexible system which is better at exploiting the benefits of inter-stream dependence. More detailed comparison between MU2’s and MU1’s violation probabilities is given in the bullet point below.

- Under the most flexible settings - M1.1 and M1.3, the violation probabilities of MU2 are smaller and closer to the target value 0.05 than MU1’s. Under the inflexible II-design, MU1 has a smaller violation probability than MU2, and the reason is as follows. When the design is inflexible with two separate queues, the staffing decision is driven only by the marginal forecasting distributions instead of the joint distribution of the two arrival streams, and this fact will later be demonstrated in the simulation studies of Section 5. MU1 and MU2 perform similarly regarding point forecast accuracy as shown in Table 2. However, we observe that MU1 produces wider marginal confidence intervals than MU2 and hence makes the staffing program account for a larger region of arrival rates. Therefore, MU1 has a smaller violation probability under the II-design since it generates similar point forecasts but wider marginal confidence intervals, compared to MU2.

- For each forecasting method, the M2.0-design and the II-design have very similar staffing performance, because these two designs are basically the same.

- When the system becomes more flexible (i.e. from II to M1.7 to M1.5, etc.), the staffing cost of MU2 decreases with the violation probability staying stable, indicating improved cost-efficiency of the service system after cross-training with stable QoS performance.

**Comparison 2: “II” vs. “X”**. We consider three X-designs where the rate at which an agent serves a non-primary customer class varies, and study how the varying cross-service rate affects the operational benefits of incorporating inter-stream dependence using MU2. We use the II-design as the benchmark, because when the cross-service rates are 0, the X-design reduces to the II-design. We choose to compare the X-design with the II-design in this staffing experiment, since Perry and Whitt (2009) have carefully studied the X-design as a potential remedy to unexpected overload under the II-design. Furthermore, note that Perry and Whitt (2009) consider two independent arrival streams, while we are interested in the effects of inter-stream dependence.

The parameter settings for the various X-designs are listed in Table 3, in the order of decreasing system flexibility. The values of $\delta, \psi_i, \mu_{ii}, c_j$ are fixed across all the X-designs, while only the cross-service rates $\mu_{12} = \mu_{21}$ change from 0.8 to 0.6 to 0.4. It makes sense that the cross-service rates are less than the dedicated service rates, which satisfy the *strong inefficient-sharing condition* (Perry and Whitt 2009).
Figure 6  Real data: daily staffing comparison among the II-design and various M-designs with different flexible server costs.

Figure 7  Real data: daily staffing comparison among the II-design and various X-designs with different cross service rates.

Figure 7 presents the daily staffing comparison between MU2 and MU1 under the X-design with varying cross-service rates. We can understand Figure 7 similarly as Figure 6: they present very analogous messages. Therefore, detailed explanations are omitted and instead we state the following two major messages:

- MU2’s violation probabilities are closer to the target $\delta = 0.05$ across all the X-designs. MU1 understaffs with violation probabilities greater than 0.05 in all the X-designs.
- When the system becomes more flexible (i.e. from II-design to X0.4 to X0.6, etc.), the staffing cost decreases and the violation probability also decreases for MU2.

In light of the two comparisons reported above, the following managerial insights are observed based on the particular call center data:

- When the service system is flexible; incorporating inter-stream dependence ensures stable performance in QoS; ignoring the dependence results in understaffing due to the positive dependence between the two classes, and the severity increases when the system becomes more flexible. When
the service system is inflexible with separate service queues: there is no benefit to account for inter-stream dependence in regards to staffing performance.

- When the system becomes more flexible, the staffing cost associated with using MU2 forecasts decreases while maintaining stable QoS performance, implying potential benefits of cross-training. Based on the amount of staffing cost reduction, call center managers can make cross-training decisions that balance between enhanced system cost-efficiency and training expense. In the simulation study (Section 5), we will show that the cross-training decision shall depend on the direction and strength of inter-stream dependence.

4.4. Comparison with Existing Bivariate Forecasting Models
To our best knowledge, Ibrahim and L’Ecuyer (2013) is the only other paper that develops forecasting models for call centers with two arrival streams. In contrast, our models are applicable for call centers with any number of arrival streams; in addition, we also study the downstream impact of incorporating inter-stream dependence on staffing under various system designs.

Considering only two arrival streams, one major difference between their models and ours is that they propose an additive structure to decompose the daily effect and the interval effect while ours consider a multiplicative structure. Under the additive structure, the forecasting variances of the interval counts within the same day are identical, while under the multiplicative structure the forecasting variance for one interval depends on its arrival volume. Since we observe heteroscedasticity of the interval counts as shown in the left panels in Figures 1 and 2, we think a multiplicative structure is more realistic. A similar preference has been stated in Shen and Huang (2008a) based on their numerical studies. Moreover, Ibrahim and L’Ecuyer (2013) proposed two bivariate mixed effect models (BME1 and BME2), where BME1 takes into account the Type (c) dependence and BME2 takes into account the Type (a) dependence respectively, while our model simultaneously accommodate the three types of inter-stream dependence.

We now compare MU2, BME1 and BME2 using the real data, in terms of forecasting accuracy, operational performance, and computation time. (We obtained codes from the authors to estimate BME1 and BME2.) For those two models, it can take as long as 4.45 hours to forecast one day with a learning period of 100 days, so we choose a shorter learning period of 30 days, when performing the rolling forecast experiment. We encounter convergence problems to forecast Day 198 and Day 40 using the BME1 method; hence we finally focus on the results from forecasting Day 41 to Day 197. Similar challenges have been noted by the authors as well.

Our forecasting comparison (see Section B of the online supplement for details) shows that BME1 is the least accurate one among the three methods, while MU2 and BME2 are comparable in terms
of point forecast accuracy: MU2 tends to forecast the Business arrivals better while BME2 is better at forecasting the Private arrivals. Hence we exclude BME1 from the follow-up staffing and QoS comparison. In Section 4.3, we have shown that the multivariate methods have more benefits in more flexible staffing systems, so we only consider the two most flexible staffing settings: M1.1 and X0.8 in the current comparison. Table 4 shows the achieved violation probabilities of MU2 and BME2 under the two staffing designs. Under the M1.1 setting, MU2 and BME2 have similar violation probabilities, with a p-value of 0.53 from the associated pairwise two-sample test. Under the X0.8 setting, MU2’s violation probability is significantly smaller than BME2’s, with a p-value of 0.0008 in the pairwise test.

Finally, we compare the computation time of the three methods. Table 5 gives a summary of the time it takes to forecast one day using each of the three methods, in the rolling forecast experiment with a learning lag of 30 days. We can see that MU2 is clearly the fastest with computing time always shorter than 0.15 second. The average computing time for BME1 is 13.43 minutes with a maximum of 37.5 minutes, and the average computing time for BME2 is 10.81 minutes with a maximum of 38.43 minutes. As the learning period increases, the computing times of BME1 and BME2 increase dramatically. For example, it takes BME1 and BME2 more than 50 minutes to forecast Day 51 based on the data from Day 1 to Day 50, and it takes them more than 4.3 hours to forecast Day 101 using the data from Day 1 to Day 100.

5. Simulation Studies

We only have access to one real data set with multiple customer streams, while there may exist other patterns of inter-stream dependence in service systems. Hence, we conduct simulation studies in this section to investigate the performance of our multiple-stream forecasting method under a range of scenarios with different inter-stream dependence in terms of type, direction and strength. We consider both positive and negative inter-stream dependence at low, medium, and high levels. Managerial insights are obtained and discussed based on the simulation results in terms of staffing.
setting and agent cross-training. For illustration purposes, we first focus on scenarios with two arrival streams to identify the main messages in Sections 5.1 - 5.3, and then present in Section 5.4 an example of three streams to demonstrate the general applicability of our methods to scenarios with more than two streams.

5.1. Effects of the Existence of Inter-stream Dependence

In this section we study the effects of the existence of the three types of inter-stream dependence. We consider eight scenarios with or without a particular type of dependence, and perform simulation studies to compare the performance between MU2 and MU1.

The parameter setting for the simulation is as follows. Based on the real data estimates and consider Model (2), we set $\Sigma_{11} = 0.8, \Sigma_{22} = 0.6, \Omega_{11} = 300, \Omega_{22} = 120, a_{11} = 0.6, a_{22} = 0.4$. The within-day proportion profiles $f_{wd,t}^{(i)}$ and the average daily total rates $\alpha_{wd}$ are set to be the real data estimates. We consider a Type (a) dependence scenario with $r = 0.7$, a Type (b) dependence scenario with $(a_{12}, a_{21}) = (0.5, 0.3)$, and a Type (c) dependence scenario with $\rho = 0.7$. Considering all the combinations of existence/nonexistence of the three types of inter-stream dependence, we come up with eight dependence scenarios as indicated in the first three columns of Table 6. Under each scenario, we generate data of 200 days and perform 100 rolling forecast and staffing experiments using MU2 and MU1, and compare their performance. In the staffing experiments, we choose the flexible staffing setting M1.1 for illustration.

Table 6 shows the comparison under each of the eight scenarios. The table lists the point forecast accuracy criteria, $RMSE$ and $MRE$, along with the staffing performance measures: $V.PROB - \delta$ is the difference between the achieved violation probability and the target value of 0.05, and $V.PROB.DIFF$ shows the difference in violation probability between MU1 and MU2. The best performer in each category is highlighted in boldface.

We can clearly observe the following messages from the results:

- Although V.PROB is always above 0.05, MU2 is consistently much closer to the target.
- When Type (b) dependence exists, i.e. the first four rows of the table, there are solid improvements in point forecast accuracy of MU2 over MU1. This message is consistent with the theoretical results in Proposition 1, and further complements it in the sense that the reported improvements take into account the estimation errors. Note that the models of Ibrahim and L’Ecuyer (2013) ignore this type of dependence.
- With Type (a) dependence, the operational improvement of MU2 over MU1 is larger. This observation is consistent with the findings from the real call center data which exhibits strong Type (a) dependence.
<table>
<thead>
<tr>
<th>Dependence Scenario</th>
<th>Method</th>
<th>Stream1</th>
<th>Stream2</th>
<th>V.PROB-δ</th>
<th>V.P.DIFF.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>MRE</td>
<td>RMSE</td>
<td>MRE</td>
</tr>
<tr>
<td>Y</td>
<td>Y</td>
<td>43.69</td>
<td>9.354</td>
<td>23.64</td>
<td>11.756</td>
</tr>
<tr>
<td></td>
<td></td>
<td>44.06</td>
<td>9.410</td>
<td>24.24</td>
<td>12.005</td>
</tr>
<tr>
<td>N</td>
<td>Y</td>
<td>43.26</td>
<td>9.018</td>
<td>22.01</td>
<td>12.551</td>
</tr>
<tr>
<td></td>
<td></td>
<td>43.92</td>
<td>9.117</td>
<td>23.59</td>
<td>11.508</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>40.13</td>
<td>9.579</td>
<td>22.73</td>
<td>12.678</td>
</tr>
<tr>
<td></td>
<td></td>
<td>41.33</td>
<td>9.939</td>
<td>22.32</td>
<td>12.678</td>
</tr>
<tr>
<td>Y</td>
<td>N</td>
<td>41.53</td>
<td>9.323</td>
<td>22.40</td>
<td>11.720</td>
</tr>
<tr>
<td></td>
<td></td>
<td>42.32</td>
<td>9.446</td>
<td>23.02</td>
<td>12.015</td>
</tr>
<tr>
<td>Y</td>
<td>N</td>
<td>42.63</td>
<td>9.459</td>
<td>23.46</td>
<td>11.823</td>
</tr>
<tr>
<td></td>
<td></td>
<td>42.47</td>
<td>9.415</td>
<td>23.51</td>
<td>11.819</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>40.89</td>
<td>9.182</td>
<td>22.85</td>
<td>11.816</td>
</tr>
<tr>
<td></td>
<td></td>
<td>41.81</td>
<td>9.163</td>
<td>22.84</td>
<td>11.806</td>
</tr>
<tr>
<td>N</td>
<td>N</td>
<td>41.51</td>
<td>9.129</td>
<td>21.97</td>
<td>11.405</td>
</tr>
<tr>
<td></td>
<td></td>
<td>41.71</td>
<td>9.144</td>
<td>21.93</td>
<td>11.388</td>
</tr>
</tbody>
</table>

Table 6: Simulation of two stream data and comparison between MU1 and MU2. Staffing experiments are under the M1.1 design.

The above messages motivate to expand our simulation studies in Sections 5.2 and 5.3 to study in more details about how Type (b) and Type (a) dependences affect forecasting and staffing performances, respectively.

5.2. Effects of Type (b) Dependence

Having observed that Type (b) dependence is essential for improving point forecast accuracy, we further study the effects of the direction and the magnitude of such dependence.

We consider 225 scenarios that cover almost all patterns of inter-stream dependence. Under each scenario we simulate 300 days of data and perform 200 rolling forecasts using MU1 and MU2 with a learning lag of 100 days. We compare the forecasting accuracy through RMSE and MRE. Detailed simulation settings and results are presented in Section C of the online supplement. The main messages from the numerical comparison are as follows:

- The stronger dependence of Stream 1’s arrival volume on Stream 2’s past information, reflected by a larger value of $a_{12}$, the bigger the improvement of MU2 over MU1 in forecasting Stream 1.
- Similarly, the stronger dependence of Stream 2’s volume on Stream 1’s past information, as in a larger value of $a_{21}$, more improvement of MU2 over MU1 in forecasting Stream 2.
- In terms of the point forecast accuracy, there is no clear relationship between the magnitude/sign of $\rho$ (or $\rho$) and the advantage of using MU2. However, MU2 is always more accurate in the distributional forecast.

The above findings are consistent with the theoretical implications of Proposition 1. Furthermore, they suggest that, even taking into account estimation errors, Type (b) dependence is the main driver for the improvement in point forecasting accuracy of the multivariate method.
5.3. Effects of Type (a) Dependence

Section 5.1 implies that, with Type (a) dependence, MU2 can lead to a more significant improvement in terms of violation probability than MU1. We now study in more details how the strength and direction of Type (a) dependence affect the operational advantages of MU2. We learned from Section 4.3 that the benefits of MU2 depend on the system flexibility level, in particular the following three factors: (1) complexity of the staffing structure, (2) salary of the flexible server, (3) cross-service rate for the non-primary customer class. We consider the same factors in the current study, and investigate how they interact with the strength/direction of Type (a) dependence.

5.3.1. Experiment 1: Different Structures

In this experiment, we consider three staffing settings: II, M1.1, and X0.8. The reason for choosing these settings has been discussed in Section 4.3.1. We also consider three scenarios of inter-stream dependence under Model (2). We fix Type(b) and Type(c) dependence with $\rho = 0$, $a_{12} = -0.5$, $a_{21} = -0.3$ since they have less impact on staffing (Table 6). We then vary the direction and strength of Type (a) dependence and consider $r = -0.8, 0, 0.8$. Under each of the three dependence scenarios, we simulate 200 days of two-stream data, and use MU1 and MU2 to generate 100 days of rolling forecasts. Then with each of the three staffing designs, we run rolling staffing experiments that feed on the distributional forecasts generated by MU1 and MU2 and compare their operational performance.

Figure 8 shows the interaction between Type (a) dependence and the system structure. We observe the following messages:

- When the system is inflexible (i.e. the II setting): MU2 and MU1 achieve stable and similar violation probabilities across all Type (a) dependence scenarios; the same holds for the staffing cost. The stability of MU2 under the II-design across all dependence scenarios implies that an inflexible system only accounts for the marginal distributions of the two arrival streams, and thus staffing performance merely depends on the marginal forecasting distributions.

- When the system is flexible (i.e. the M1.1 and X0.8 settings): MU2 is more stable in violation probability; MU1 understaffs (i.e. V.PROB>0.05) for positively correlated arrival streams and overstaffs (i.e. V.PROB<0.05) for negatively correlated arrival streams. An intuitive explanation is that a flexible system has the potential to take advantage of inter-stream dependence, which will be lost if using MU1 forecast; ignoring such dependence (as in MU1) will result in the mismatch between the staffing level and the demand, either inadequate staffing (with positive $r$) or excessive capacity (with negative $r$).

- When the system structure becomes more complex/flexible (i.e. from II to M1.1, or from II to X0.8), the staffing cost decreases for both MU2 and MU1. For MU2, the cost reduction is larger with
negatively correlated arrivals than with positively correlated arrivals. Since staffing cost reduction varies across different arrival-stream dependence for MU2, cross-training decisions shall take into account inter-stream dependence. For example, managers should be more conservative towards cross-training when there exists positive inter-stream dependence and more aggressive otherwise. For MU1, the staffing cost does not change much as the inter-stream dependence changes, since it ignores such dependence and produces very similar distributional forecasts under each scenario.

5.3.2. Experiment 2: Varying Parameters The previous staffing experiment examines the system structure, while in this section we look at the effects of two system parameters: (1) flexible server cost; (2) cross service rate of multi-skill servers to handle their non-primary customer class. We consider the staffing settings shown in Table 3 with different choices for the two parameters. We fix the Type (b) and Type (c) dependences as in the previous section and consider two Type (a) dependence scenarios: \( r = -0.8 \) and \( r = 0.8 \).

Figures 9 and 10 show the staffing comparison under various M-designs and X-designs, respectively. We obtain the following messages:

- MU2 is more stable in violation probability than MU1 across all staffing parameter sets.
- MU1’s violation probability varies as the system flexibility level changes. MU1 pays more penalty under more flexible systems for ignoring the inter-stream dependence. We consider MU1’s performance under the II-design as the limiting benchmark while varying system flexibility levels. The more flexible the system, the more deviation of MU1’s V.PROB from its limiting performance under the II-design. Furthermore, the direction of deviation depends on the direction of inter-stream dependence. Specifically, when inter-stream dependence is positive, MU1 tends to understaff more as the system becomes more flexible; when inter-stream dependence is negative, MU1 tends to overstaff more as the system becomes more flexible.
MU2 achieves better violation probability with positive inter-stream dependence than with negative inter-stream dependence. We observe from another simulation study (not shown here) that the chance-constraint program is more accurate in approaching the target violation probability when facing positively correlated distributions than negatively correlated distributions with the same marginal distributions. One reason is that the boundary of the $\left(1 - \delta\right)$ support region (where the QoS level is met) is shorter and clearer when the inter-stream correlation is positive.

As the system becomes more flexible (i.e., flexible server salary decreases in the M design, or cross service rate increases in the X design), the staffing cost of MU2 decreases and the amount of decrement depends on inter-stream dependence. More specifically, cost reduction induced by enhanced system flexibility is higher with negative inter-stream dependence than with positive inter-stream dependence. This phenomenon again suggests that cross-training has more benefits for negative inter-stream dependence than for positive inter-stream dependence.

5.4. Simulation Study with Three Arrival Streams
To our best knowledge, we are the first one to propose a statistical forecasting model for call centers with more than two streams of arrivals. In this section, we conduct a simulation experiment similar to Section 5.1, but with three arrival streams instead of two, and study how the existence of various inter-stream dependence affects the forecasting and operational performance.

We consider eight dependent scenarios and simulate three-stream call arrival data under each scenario. We then fit two forecasting models MU1 and MU3 (i.e. fit Model (1) simultaneously on all three streams), and feed the obtained distributional forecasts into the chance-constraint staffing program to generate the staffing vectors and the associated costs. We consider the staffing structure shown in Figure 11, which is a subgraph of the design studied in Gurvich et al. (2010). (See Section D of the online supplement for detailed settings of the data simulation and the staffing design.)

Analogous to Table 6, Table 7 compares point forecast and staffing performance between MU1 and MU3 under the eight scenarios of inter-stream dependence. The messages delivered are similar to those from Table 6. The main observations from Table 7 can be summarized as follows:

- The violation probability from using the MU3 forecast is smaller than that from MU1, and almost always closer to the QoS target.
- When Type (b) dependence exists, MU3 is more accurate in point forecasting.
- Without Type (b) dependence, the two methods give similar point forecasts; but when Type (a) dependence exists, MU3 significantly improves the violation probability over MU1.

![Figure 11](image-url)  
Staffing design under 3 arrival streams.
6. Conclusion

Our paper has provided interesting insights on forecasting and staffing call centers with multiple uncertain arrival streams. We developed statistical models to simultaneously forecast multiple stream arrivals considering three types of inter-stream dependence. We theoretically and numerically demonstrated the benefits of considering such inter-stream dependence. Our models gave more accurate forecasts and more stable QoS performance than the univariate alternative that ignores such dependence. In the real data application, our models better approached the QoS target, and were more stable and faster in computation than the alternative bivariate models. We used extensive numerical studies to investigate how the forecasting and operational benefits of the multiple stream approach varied by type, direction, and strength of inter-stream dependence, as well as across multiple staffing designs. We obtained the following insights:

- When multi-stream data exhibits strong cross-lag dependence, i.e. Type (b) dependence, incorporating inter-stream dependence improves point forecast accuracy.

- When the staffing system is inflexible with multiple queues operating separately, independent forecasting method that ignores dependence performs well enough.

- When the staffing system is flexible, our multivariate method leads to stable QoS performance regardless of inter-stream dependence and staffing design; ignoring dependence causes understaffing with positively correlated interval arrivals and overstaffing with negatively correlated interval arrivals, and the severity of understaffing/overstaffing increases as the system becomes more flexible. Therefore, it is beneficial to incorporate inter-stream dependence in flexible staffing systems.

- With our multiple-stream forecasting method, cross-training reduces operational staffing cost while maintaining stable QoS performance, and the cost-reduction is more with negative inter-stream dependence than with positive inter-stream dependence. Cross training decision may be made based on the improved system cost-efficiency and the initial training expense. Such decisions shall depend on the type of inter-stream dependence.

There are a couple of further research directions that are worth pursuing. We observed that the realized violation probability always deviated from the target. This deviation could be caused by the approximation error of using simple random sampling to approximate the forecasting distribution. A natural question is to come up with more efficient sampling schemes that can better approximate the forecasting distribution. Gans et al. (2012) demonstrated the benefits of using parametric Gaussian quadrature discretization (Miller and Rice 1983) over random sampling; however, their stochastic programming formulations imposed QoS expectation constraint instead of the chance-constraint. Whether Gaussian quadrature can be used in a chance-constraint framework is
worth investigating. Another related question is to combine the multiple-stream forecasting with staffing/scheduling formulations that impose expectation constraints, such as expected abandonment rate below a target, similar to what Gans et al. (2012) did for a single arrival stream.

References


Supplement

A Proof of Proposition 1

To forecast $y$ based on $\xi_{(n)}$, we consider the joint distribution of $(y, \xi_{(n)}^T)^T$ and assume it's multivariate Gaussian as follows,

$$
\begin{pmatrix}
y \\
\xi_{(n)}
\end{pmatrix}
\sim
\mathcal{N}
\left(
\begin{pmatrix}
\mu_y \\
\mu_{(n)}
\end{pmatrix},
\begin{pmatrix}
\Gamma_y & \Gamma_{(n)}^T \\
\Gamma_{(n)} & M_{(n)}
\end{pmatrix}
\right).
$$

Then given the vector $\xi_{(n)}$, $y$ has the following distribution

$$
y|\xi_{(n)} \sim \mathcal{N}(\tilde{\mu}_n, \tilde{\Gamma}_n), \quad n = 1, 2, \ldots
$$

where

$$
\tilde{\mu}_n = \mu_y + \Gamma_{(n)}^T M_{(n)}^{-1} (\xi_{(n)} - \mu_{(n)}), \quad \tilde{\Gamma}_n = \Gamma_y - \Gamma_{(n)}^T M_{(n)}^{-1} \Gamma_{(n)}.
$$

Notice that

$$
\text{Var}(\xi_{n+1}|\xi_{(n)}) = m_{n+1,n+1} - m_{(n+1)}^T M_{(n)}^{-1} m_{(n+1)} \geq 0.
$$

Assuming that $\xi_{n+1}$ is non-deterministic conditional on $\xi_{(n)}$, then we have

$$
\text{Var}(\xi_{n+1}|\xi_{(n)}) = m_{n+1,n+1} - m_{(n+1)}^T M_{(n)}^{-1} m_{(n+1)} > 0.
$$

Thus, the forecasting variance reduced by introducing one more variable $\xi_{n+1}$ is given by

$$
\tilde{\Gamma}_n - \tilde{\Gamma}_{n+1} = -\Gamma_{(n)}^T M_{(n)}^{-1} \Gamma_{(n)} + \Gamma_{(n+1)}^T M_{(n+1)}^{-1} \Gamma_{(n+1)}
= -\Gamma_{(n)}^T M_{(n)}^{-1} \Gamma_{(n)} + (\Gamma_{(n)}, \Gamma_{n+1}) \begin{pmatrix}
M_{(n)} & m_{n+1} \\
m_{n+1}^T & m_{n+1,n+1}
\end{pmatrix}^{-1}
\begin{pmatrix}
\Gamma_{(n)} \\
\Gamma_{n+1}
\end{pmatrix}
= \frac{(\Gamma_{n+1} - \Gamma_{(n)}^T M_{(n)}^{-1} m_{(n+1)})^2}{m_{n+1,n+1} - m_{(n+1)}^T M_{(n)}^{-1} m_{(n+1)}} \geq 0.
$$

1
Table 1: Rolling forecasts comparison on real data among MU2, BME1 and BME2.

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE Private Queue</th>
<th>RMSE Business Queue</th>
<th>MRE Private Queue</th>
<th>MRE Business Queue</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Median  Mean</td>
<td>Median  Mean</td>
<td>Median  Mean</td>
<td>Median  Mean</td>
</tr>
<tr>
<td>MU2</td>
<td>38.00  41.48</td>
<td>20.89  22.03</td>
<td>8.34  8.84</td>
<td>11.31  11.67</td>
</tr>
<tr>
<td>BME1</td>
<td>38.63  43.75</td>
<td>21.95  23.26</td>
<td>8.77  9.45</td>
<td>11.71  12.30</td>
</tr>
<tr>
<td>BME2</td>
<td><strong>37.18</strong>  <strong>41.21</strong></td>
<td>21.57  22.29</td>
<td>8.38  <strong>8.83</strong></td>
<td>11.32  11.83</td>
</tr>
</tbody>
</table>

Table 1 displays the results of the rolling forecasts comparison among the three forecasting methods: MU2, BME1 and BME2, which are discussed in Section 4.4 of the main article. In the rolling forecast experiment, we forecast Day 41 through Day 197 with a learning period of 30 days, and record the accuracy measures RMSE and MRE when forecasting each day. The means and medians of RMSE and MRE are reported in Table 1 with the best performers highlighted in boldface.

C Simulation: Effects of Type (b) Dependence

In this section, we present the detailed settings and results of the simulation study in Section 5.2 of the main article. To study the effects of Type (b) dependence, the setting of the simulation is as follows.

We first consider multiple scenarios of Type (a) and Type (c) dependence, by varying $(r, \rho)$ in the set $(-0.8, 0, 0.8) \times (-0.8, 0, 0.8)$, and come up with 9 scenarios for $(r, \rho)$. For each chosen pair of $(r, \rho)$, we consider 25 scenarios of Type (b) dependence, by varying $(a_{12}, a_{21})$ in $(-0.5, -0.25, 0, 0.25, 0.5) \times (-0.3, -0.15, 0, 0.15, 0.3)$. These Type (b) dependence scenarios ensure that the vector time series of daily total arrival rates generated by the matrix $A$ are stationary. Hence, in total there are $3 \times 3 \times 5 \times 5 = 225$ scenarios of inter-stream dependence.

Under each of the 225 scenarios, we simulate 300 days of data and perform rolling forecast experiments to forecast Day 101 through Day 300 using a learning period of 100 days. The rolling forecast is repeated 200 times under each scenario, thus we have for each method 200 records of the performance measures. To compare the point forecasting performance, we denote the mean RMSE of the 200 rolling forecasts for each method as $\text{RMSE}_{MU1}$ and $\text{RMSE}_{MU2}$. We then calculate the relative mean RMSE
reduction of MU2 over MU1 as

\[
\Delta_{\text{RMSE}} = \frac{\text{RMSE}_{MU1} - \text{RMSE}_{MU2}}{\text{RMSE}_{MU1}}
\]

Figures 1 and 2 display the RMSE reduction for forecasting Stream 1 and Stream 2, respectively. Each figure has \(3 \times 3 = 9\) plots, where each plot corresponds to a specific pair of \((r, \rho)\). For each plot in the two figures, there are \(5 \times 5 = 25\) lattices referring to different pairs of \((a_{12}, a_{21})\), where \(a_{12}\) varies in the horizontal direction and \(a_{21}\) varies in the vertical direction. The RMSE reduction is shown in colors according to the color bar besides each panel, where magenta indicates that MU2 is better, i.e. gives a smaller RMSE, and cyan indicates otherwise. Similar plots are obtained for MRE, which are not included here as the patterns look similar to the plots for RMSE.

![Figure 1: \(\Delta_{\text{RMSE}}\) for forecasting Stream 1.](image)

\(3\)
D Simulation Settings for the Three Stream Example

The settings of the inter-stream dependence scenarios in Section 5.4 of the main article are as follows:

- Use real data estimates for the intra-day proportion profiles of Stream 1 and Stream 2. Set the intra-day proportion profiles of Stream 3 the same as Stream 2.

- Use real data estimates for the average daily total rates $\alpha_{wd}$ of Stream 1 and Stream 2. Since Stream 1 and Stream 2 are two major streams, we assume that Stream 3’s volumes are a little smaller than Stream 2’s. Therefore, we set the average daily total rates of Stream 3 to be 90% of Stream 2’s. In addition, the standard deviations of Stream 3’s interval counts and daily total innovations are proportional to those of Stream 2’s at the ratio 0.9.

- Set Type (a) dependence as $\Sigma = \begin{pmatrix} 0.81 & 0.5 & 0.45 \\ 0.50 & 0.63 & 0.39 \\ 0.45 & 0.39 & 0.50 \end{pmatrix}$, where the correlation between any
two streams is 0.7. Set Type (b) dependence as $A = \begin{pmatrix} 0.6 & 0.30 & 0.3 \\ 0.3 & 0.41 & 0.2 \\ 0.2 & 0.10 & 0.4 \end{pmatrix}$. Matrix $A$ is chosen such that the vector time series generated by it is stationary. Set Type (c) dependence as 

$$
\Omega = \begin{pmatrix} 310.00 & 136.13 & 123.25 \\ 136.13 & 122.00 & 77.32 \\ 123.25 & 77.32 & 100.00 \end{pmatrix},
$$

where the correlation between any two streams is 0.7.

- Consider the existence/nonexistence of each type of inter-stream dependence, and we come up with eight scenarios.

The settings for the staffing experiment are as follows:

- The staffing design contains three agent pools. The first agent pool only serves Stream 1; the second agent pool serves Stream 1 and Stream 2; the third agent pool serves all the three streams. The service rate is 1.0 regardless of the agent pool and customer class.

- Allowed abandonment proportion is 0.04 for each of the three arrival streams, i.e. $\psi_1 = \psi_2 = \psi_3 = 0.04$.

- The cost vector for the three agent pools is $(c_1, c_2, c_3) = (1, 1.1, 1.2)$, as agents with more skills are paid more.

### E Model Diagnostics

In this section, we provide some diagnostics results that suggest that Model (2) of the main paper fits our call center data well.

To begin with, we show that the VAR(1) model is reasonable for $u_d$ across days. Table 2 reports the lag-1 correlations of the estimates $u_d$ (after removing the data-of-week effects), and all the correlations are significant at $10^{-8}$ level. Table 3 then shows the lag-1 correlations based on the estimated $z_d$, which are all insignificant at 0.1 level. Furthermore, Figure 3 presents the normal quantile-quantile envelope plots for the VAR(1) residuals $\hat{z}_d$, which suggest that they are approximately normally distributed.

<table>
<thead>
<tr>
<th>Correlations</th>
<th>$u_d^{(1)}$</th>
<th>$u_d^{(2)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_d^{(1)}$</td>
<td>0.5581</td>
<td>0.3743</td>
</tr>
<tr>
<td>$u_{d+1}^{(1)}$</td>
<td>0.4592</td>
<td>0.5237</td>
</tr>
</tbody>
</table>

*Table 2: Lag-1 correlations of the daily total rates after removing day-of-week effects.*
Table 3: Lag-1 correlations of the daily total rate innovations after fitting the VAR(1) model.

<table>
<thead>
<tr>
<th>Correlations</th>
<th>$\hat{z}_{d}^{(1)}$</th>
<th>$\hat{z}_{d}^{(2)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{z}_{d+1}^{(1)}$</td>
<td>0.0450</td>
<td>0.1083</td>
</tr>
<tr>
<td>$\hat{z}_{d+1}^{(2)}$</td>
<td>0.0481</td>
<td>0.0520</td>
</tr>
</tbody>
</table>

Figure 3: Q-Q plots for daily total rate innovations.

Finally, Figure 4 plots the normal quantile-quantile envelopes for the residuals $\epsilon_{d,t}^{(i)}$ for each service type. Again, the residuals are reasonably normally distributed.

Figure 4: Q-Q plots for interval residuals.