DATA-DRIVEN SERVICE OPERATIONS MANAGEMENT

Han Ye

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2014

Approved by:

Haipeng Shen

Nilay Argon

Chuanshu Ji

Shu Lu

J. S. Marron

ABSTRACT

Han Ye: DATA-DRIVEN SERVICE OPERATIONS MANAGEMENT
(Under the direction of Haipeng Shen)

This dissertation concerns data driven service operations management and includes three projects. An important aim of this work is to integrate the use of rigorous and robust statistical methods into the development and analysis of service operations management problems. We develop methods that take into account demand arrival rate uncertainty and workforce operational heterogeneity. We consider the particular application of call centers, which have become a major communication channel between modern commerce and its customers. The developed tools and lessons learned have general appeal to other labor-intensive services such as healthcare.

The first project concerns forecasting and scheduling with a single uncertain arrival customer stream, which can be handled by parametric stochastic programming models. Theoretical properties of parametric stochastic programming models with and without recourse actions are proved, that optimal solutions to the relaxed programs are stable under perturbations of the stochastic model parameters. We prove that the parametric stochastic programming approach meets the quality of service constraints and minimizes staffing costs in the long-run.

The second project considers 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 operational advantage of the inter-stream dependence.

The third project of this dissertation studies operational heterogeneity of call center agents with regard to service efficiency and service quality. The proxies considered for agent service efficiency and service quality are agents' service times and issue resolution probabilities, respectively. Detailed analysis of agents' learning curves of service times are provided. We develop a new method to rank agents' first call resolution probabilities based on customer call-back rates. The ranking accuracy is studied and the comparison with traditional survey-driven methods is discussed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

In recent years, call centers have shared a vast industry and are experiencing dramatic growth. It was estimated in 1999 that the U.S. had 1.55 million call center agents (Gans et al. (2003)). In 2008, the number of call centers in the U.S. was estimated to be 47,000 with 2.7 million employees (Aksin et al. (2007)). As a primary customer-facing communication channel, call centers have become an integral part of many businesses and are playing an increasingly significant role in bridging service providers to their customers. It's estimated that more than 70% of all customer-business interactions are handled through call centers (Brown et al. (2005)).

The essential challenge for call center managers is to develop efficient staffing and scheduling strategies to achieve both desired levels of service quality and operating expenses. The staffing and scheduling process usually begins with forecasting arrival demand volumes over a planning horizon, which ranges from a day to several weeks. Call centers also need to evaluate the service quality and efficiency of their agents during the planning horizon. With the demand forecasts and service evaluation, call centers then determine the staffing and scheduling plan for short intervals (varying from 15-min to 1-hour) within the horizon, which minimizes the operational costs subject to a pre-specified Quality of Service (QoS) level. The final step is rostering where agents are assigned to the planned schedules.

Traditionally, call centers assume that the arrival rate forecasts are accurate, and use point forecasts of the arrival rates to derive the staffing and scheduling plans. However, very often the rate forecasts and the realized arrival rates do not match perfectly, and the use of inaccurate rate forecasts would result in improper staffing levels, and further causes the system performance to diverge from operational target. Recently, people in both statistics and operations management/research have become aware of the arrival rate uncertainty and have begun to deal with the problem of the discrepancy between forecasts and realizations. Statistical papers try to develop more accurate point forecasts and at the same time carefully characterize the arrival rate forecasting distribution. On the other hand, operations management researchers incorporate arrival rate uncertainty into the staffing and scheduling methodologies. Papers are rare that integrate statistical forecasting process and operational staffing/scheduling process to jointly cope with the arrival rate uncertainty problem. The first two projects of the thesis aim to bridge the gap between existing statistical and operational research. The first project concerns call centers with a single arrival stream while the second one concerns call centers with multiple arrival streams and investigates the benefits of incorporating inter-stream dependence.

For a single stream and a single pool of agents, Gans et al. (2012) developed and tested a combined forecasting and parametric stochastic programming approach which takes into account arrival rate uncertainty with inter-day and intra-day dependence. Chapter 2 of this thesis extends their work, and performs theoretical stability analysis of the stochastic programming models with and without recourse actions. In particular, we prove that there exist optimal solutions for the parametric stochastic model relaxation, and the optimal solutions are continuous with respect to small perturbations of the model parameters.

Under the context of multiple-stream arrivals, Ibrahim and L'Ecuyer (2012) built linear mixed models to jointly forecast the arrival counts for two different call types handled at a single call center. Operations management papers utilize skill-based routing strategies to deal with multiple-stream staffing/scheduling problem. A recent work by Gurvich et al. (2010) proposed a chance-constraint optimization approach to staff multiple-stream call centers in the short run.

In Chapter 3, we consider both forecasting and staffing together to solve a complete multiple-stream call center staffing problem with uncertain arrival demands. We evaluate our approach on a real call center data set. We also provide theoretical assessment and simulation tests of our approach under various scenarios. Our study demonstrates the importance of incorporating dependence structure among arrival streams in both forecasting and staffing stages. We also show how the performance of our multiple-stream approach varies by type and strength of dependence among the streams. Our efforts naturally extend the work of Gans et al. (2012)., Ibrahim and L'Ecuyer (2012) and Gurvich et al. (2010).

More specifically, we conduct the following analyses.

- We develop statistical methods to generate simultaneous distributional forecasts of multiple arrival streams. In particular, we decompose within-day arrival volumes of each stream into the product of daily-total rate and within-day proportion profile. We then apply vector time series models to jointly forecast multiple-stream daily-total arrival rates. Compared with the linear mixed effect models in Ibrahim and L'Ecuyer (2012), our method is more attractive in two ways: it models the inter-stream and within-stream dependence in a more general form; it's more practicable and faster in computation.
- We theoretically evaluate the forecasting benefits of incorporating dependence among the streams under different type and strength of dependence. In particular, we derive the forecasting variance reduction of the multiple-stream forecasting method over the single-stream forecasting method, as a function of inter-stream and within-stream correlations.
- To evaluate the operational effects of incorporating dependence among the

streams, we implement the chance-constraint staffing approach with the sample based approximation of Gurvich et al. (2010). We demonstrate how the performance of this algorithm varies by the direction and strength of dependence among the streams.

- We integrate our forecasting method and the chance-constraint staffing approach as an entire solution to staffing call centers with multiple uncertain demand streams. We test our approach on a real call center data set. Results suggest that the multiple-stream approach provides more accurate distributional forecasts and the follow-up staffing algorithm is closer to meet the quality-of-service constraint, compared with the single-stream approach which ignores the inter-stream dependence. We also test our approach under 125 simulated scenarios of different type and strength of inter-stream dependence. Our results show that: the stronger the dependence on the other streams' past information, the better the forecasting performance of the multiple-stream approach; for negatively correlated streams, the multiple-stream approach saves money while at the same time provides the same service quality.

In the third project of the thesis (Chapter 4), we consider agent heterogeneity in terms of service efficiency and service quality. Service time is a basic measurement of service efficiency. In classical queuing models such as Erlang-C (M/M/N) and Erlang-A (Garnett et al. (2002), which allows the customer to abandon), agent service times are assumed independent and identically distributed (iid) according to an exponential distribution. Thus agents are assumed to be homogeneous in 2 ways:

- exponentiality in service times,
- time stationarity in service time attributes.

Such assumptions are imposed mainly for mathematical tractability. However they rarely prevail in practice. The empirical analysis of Brown et al. (2005) reveals that service-times are log-normally distributed (as opposed to being exponential). We then perform a detailed analysis of agents' learning-curves, and show various learning patterns of agents.

Regarding agent heterogeneity in service quality, we consider issue resolution probability as the proxy, since it is directly related to customers' demands and perception of the call service. Issue resolution probability, by definition, is the probability that a customer's issue is resolved by the end of the call. Traditionally, most staffing/scheduling methods assume the issue resolution probability to be one, i.e., all the problems given to the agents are solved by the end of the call. However, agents' capability of solving customers' problems has been empirically observed to be noticeably different, and customers with unsolved problems may call back, which increases system load and wastes the recourse (de Vericourt and Zhou (2005)). Conventionally, issue resolution probability is estimated via customer surveys, which may require extra agent endeavor to call back and suffer from extremely high non-response rate. Thus the evaluations obtained are limited, unreliable and very likely biased. We propose an innovative estimate for issue resolution probability, which requires no extra agent efforts other than historical operational data. We also discuss factors that affect issue resolution probability such as agents intentionally hanging up on customers.

2 Stability Analysis on Stochastic Programming Models

In this chapter we concern forecasting and staffing call centers with a single uncertain arrival stream, regarding which Gans et al. (2012) developed and tested a combined forecasting and parametric stochastic programming approach which takes into account arrival rate uncertainty with inter-day and intra-day dependence. Our work is an extension of Gans et al. (2012). More specifically, we conduct theoretical analyses on their parametric stochastic programming models.

2.1 Background and Motivation

In this section, we first briefly review the forecasting and stochastic programming scheduling approach by Gans et al. (2012) and then highlight our motivation.

In their paper, they first derived parametric forecasts for call centers, then demonstrated that the parametric forecasts can be used to drive stochastic programming models whose results are stable with a relatively small number of scenarios. They then developed a Bayesian procedure to update the forecast distribution during the later stage and extended their stochastic models to be suitable for recourse action given the forecast updates.

Regarding the model performance: on one hand, they conducted a numerical study which shows that the inclusion of multiple arrival-rate scenarios allows the call centers to meet long-run average QoS targets, while the use of recourse actions help them to lower long-run average costs; on the other hand, theoretical properties of the integrated forecasting and scheduling models have not been discussed yet. In particular, the main question of interest is: do their models minimize scheduling cost and satisfy the QoS constraint in the long-run, theoretically? This big question can be decomposed into three sub-problems:

- P1) to show the consistency of the statistical parametric forecasts with or without later stage Bayesian updates.
- P2) to show the stability of the parametric stochastic programming models in terms of small perturbations of the model parameters (statistical forecasts). Since the parametric stochastic programming models are formulated as Integer Programs (IP), the problem P2) is further decomposed to the following two problems:
  - P2.1) to show the stability of the model IP's in terms of relaxation to Linear Programs (LP).
  - P2.2) to show the stability of the relaxed LP's in terms of small perturbations of the LP parameters (statistical forecasts).

This chapter provides detailed analyses to address problem P2.2).

In the following is a summary of the three parametric stochastic models proposed in Gans et al. (2012). Regarding problem P2.2), we consider the LP relaxations to the following three IP's.

- Integer Program (IP) (6) in Gans et al. (2012), that solves to get optimal schedules for the planning horizon given the parametric forecasts for the horizon, subject to the QoS constraint that expected abandonment rate in the planning horizon less than a threshold.
- IP (10) in Gans et al. (2012), that solves to get optimal recourse actions for the later stage in the planning horizon given the early stage schedule and forecast

updates, subject to the QoS constraint that expected later stage abandonment rate less than a threshold.

- IP (12) in Gans et al. (2012), that solves to get optimal schedules for the planning horizon with recourse given the parametric forecasts for the horizon, subject to the QoS constraint that expected abandonment rate in the planning horizon less than a threshold. This model provides optimal schedule before the planning horizon, consolidating all possible recourse actions for the later stage before the early stage is observed, compared to (6) and (10) in Gans et al. (2012).

All the LP relaxations to the above IP's are driven by the arrival forecasts, in particular the discretized forecasting distribution. It is non-trivial to substantiate the existence of optimal solutions to the LP relaxations and that the optimal solutions are stable with respect to small perturbations of the discretized forecast distribution, as well as the way it's discretized. We then address this problem in next section by providing theoretical stability analysis for the LP relaxations of the above IP's.

2.2 Model Stability Analysis

Our analysis is based on the findings of Williams (1963) and Robinson (1977). In particular, Robinson (1977) proves that a necessary and sufficient condition for the primal and dual optimal solution sets of a solvable, finite-dimensional linear programming problem to be stable under small but arbitrary perturbations in the parameters of the problem is that both of these sets are bounded.

With a slight abuse of notation, denote the primal LP as (P): $\max\{cx \mid Ax \leq b; x \geq 0\}$, and its corresponding dual as (D): $\min\{\pi b \mid \pi A \geq c\}$. Then we would like to recall the theorem:

Theorem 1 (Robinson (1977), p.440) *The following are equivalent:*

(a) The sets S_P and S_D of optimal solutions of (P) and (D) respectively, are nonempty and bounded. Or equivalently, the conditions R1) and R2) are satisfied:

R1) for every vector $y \neq 0$, $y \geq 0$, $Ay \leq 0 \implies cy < 0$, and

R2) for every vector $\rho \neq 0$, $\rho \geq 0$, $\rho A \geq 0 \implies \rho b > 0$.

(b) There exists an $\epsilon_0 > 0$ such that for any A', b' and c' with

$$\epsilon' \equiv max\{||A' - A||, ||b' - b||, ||c' - c||\} < \epsilon_0,$$

the two dual problems (P′): $\max\{c'x \mid A'x \leq b'; x \geq 0\}$ and (D′): $\min\{\pi b' \mid \pi A' \geq c'\}$ are solvable.

If these conditions are satisfied, then there exist constants $\epsilon_1 \in (0, \epsilon_0]$ and γ such that for any A', b' and c' with $\epsilon' < \epsilon_1$, any optimal solutions x' solving (P') and u' solving (D'), one has $d[(x', u'), S_P \times S_D] \leq \gamma\epsilon'$.

2.2.1 Simple Stochastic Program

Here's a version of the IP (6) from Gans et al. (2012). It is a bit different from (6) in that it introduces an extra set of constraints and an intermediate set of variables,

α_k's, that will be useful in our analysis.

$$
\begin{array}{lcll}
\min \sum_{j\in\mathcal{J}} c_j x_j & & & \\
\text{subject to} & & & \\
(\sum_{j\in\mathcal{J}} a_{ij}x_j)m_{ikn} + b_{ikn} & \leq & \alpha_{ik} & i\in\mathcal{I}, k\in\mathcal{K}, n\in\mathcal{N}_i \\
\sum_{i\in\mathcal{I}} \alpha_{ik} & \leq & \alpha_k & k\in\mathcal{K} \\
\sum_{k\in\mathcal{K}} p_k\alpha_k & \leq & \alpha^*\bar{\lambda} & \\
x_j & \in & \mathbb{Z}^+ & j\in\mathcal{J} \\
\alpha_{ik} & \geq & 0 & i\in\mathcal{I}, k\in\mathcal{K} \\
\alpha_k & \geq & 0 & k\in\mathcal{K}.
\end{array}
\tag{2.1}
$$

Note that, by construction, $m_{ikn} < 0$ and $b_{ikn} > 0$ for all i, k, and n. To avoid technical distractions, we'll assume that $c_j > 0$ for all $j \in \mathcal{J}$, $\lambda_{ik} > 0$ for all $i \in \mathcal{I}, k \in \mathcal{K}$ and that $p_k > 0$ for all $k \in \mathcal{K}$. That is, the cost of people working on any schedule is strictly positive, as is the expected number of arrivals under any of the problem's scenarios and their probabilities.

We would like to prove the following proposition so that that the objective value and α_k's obtained by the optimal solution to the above IP are continuous with respect to perturbations of the parametric forecasts p_k's and λ_{ik}'s.

Proposition 2.1 *There exist optimal solutions for the LP relaxation of 2.1 and these optimal solutions are continuous with respect to small but arbitrary perturbations of the LP relaxation of 2.1.*

Proof

We apply Theorem 1 in our proof. To do so, we first massage the LP relaxation into a standard form:

$$\begin{array}{llll}
\min \sum_{j\in\mathcal{J}} c_j x_j & & & \\
\text{subject to} & & & \\
(\sum_{j\in\mathcal{J}} a_{ij}x_j)m_{ikn} \; - \; \alpha_{ik} & \leq & -b_{ikn} & i\in\mathcal{I}, k\in\mathcal{K}, n\in\mathcal{N}_i \\
\sum_{i\in\mathcal{I}} \alpha_{ik} \; - \; \alpha_k & \leq & 0 & k\in\mathcal{K} \\
\sum_{k\in\mathcal{K}} p_k\alpha_k & \leq & \alpha^*\bar{\lambda} & \\
x_j & \geq & 0 & j\in\mathcal{J} \\
\alpha_{ik} & \geq & 0 & i\in\mathcal{I}, k\in\mathcal{K} \\
\alpha_k & \geq & 0 & k\in\mathcal{K}.
\end{array} \tag{2.2}$$

Here's the vector-matrix form of the above LP

$$\begin{array}{lll}
\max -cx \; + \; 0\alpha^1 \; + \; 0\alpha^2 & & \\
\text{subject to} & & \\
\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} x \\ \alpha^1 \\ \alpha^2 \end{bmatrix} & \leq & \begin{bmatrix} -b \\ 0 \\ \alpha^*\bar{\lambda} \end{bmatrix} \\
x, \quad \alpha^1, \quad \alpha^2 & \geq & 0.
\end{array} \tag{2.3}$$

The decision variables are as follows:

- x, a $|\mathcal{J}|$-vector;
- α^1, a $|\mathcal{I}|\cdot|\mathcal{K}|$-vector of the α_{ik}'s; and
- α^2, a $|\mathcal{K}|$-vector of the α_k's.

The right-hand-side has three parts:

- $-b$, a $(\sum_{i\in\mathcal{I}}|\mathcal{N}_i|)\cdot|\mathcal{K}|$-vector, and
- 0, a $|\mathcal{K}|$-vector; and

- $\alpha^* \bar{\lambda}$, a scalar

and the constraint matrix is made up of the following submatrices, each with dimensions that match the corresponding segments of the right-hand side and decision variables:

1) A_{11}, a matrix of 0's and m_{ikn}'s, where each slope $m_{ikn} < 0$;
2) A_{12}, a matrix of -1's and 0's;
3) A_{13}, a matrix of 0's;
4) A_{21}, a matrix of 0's;
5) A_{22}, a matrix of 1's and 0's;
6) $A_{23} = -I$, the negative of the identity matrix;
7) A_{31}, a row vector of 0's;
8) A_{32}, a row vector of 0's; and
9) A_{33}, a row vector of p_k's.

For condition R1, we let $y = (x, \alpha^1, \alpha^2)$ be such that $y \neq 0$ and $y \geq 0$. We note that only $\alpha^2 = 0$ ensures that the left-hand side of the constraint $\sum_{k \in \mathcal{K}} p_k \alpha_k \leq \alpha^* \bar{\lambda}$ is not positive. In turn, given $\alpha^2 = 0$ only $\alpha^1 = 0$ ensures that the left-hand sides of the constraints $\sum_{i \in \mathcal{I}} \alpha_{ik} - \alpha_k \leq 0$ (for all $k \in \mathcal{K}$) are not positive. Thus if $y \neq 0$, there must be an $x_j > 0$ so that $-c_j x_j < 0$. Thus $cy < 0$, and y satisfies the conditions of R1.

For condition R2, let the sub-vectors of the (row vector) dual variable $\rho = (\rho^1, \rho^2, \rho^3)$ have the dimensions of the right-hand side $b = (-b, 0, \alpha^* \bar{\lambda})'$ and assume $\rho \neq 0$, $\rho \geq 0$. First we note that, if there is an element $\rho^1_{ikn} > 0$, then

$$(\rho^1, \rho^2, \rho^3) \begin{bmatrix} A_{11} \\ A_{21} \\ A_{31} \end{bmatrix} < 0$$

since there will be an element $m_{ikn} < 0$ of A_{11} that is multiplied with ρ^1_{ikn} and both A_{21} and A_{31} have all zeros. In this case, R2 is trivially satisfied. If $\rho^1 = 0$ and $\rho^3 > 0$, then $\rho^3 \alpha^* \bar{\lambda} > 0$ implies that $(\rho^1, \rho^2, \rho^3)(-b, 0, \alpha^* \bar{\lambda})' > 0$ and again R2 is satisfied. If $\rho^1 = 0$ and $\rho^3 = 0$, then there must be a $\rho^2_k > 0$, and in this case

$$(\rho^1, \rho^2, \rho^3) \begin{bmatrix} A_{13} \\ A_{23} \\ A_{33} \end{bmatrix} < 0$$

since the kth column of A_{13} is all 0's, and the kth column of A_{23} has a -1's in the kth row and 0's elsewhere. Again, R2 is trivially satisfied in this case.

Thus the LP relaxation (2.2) satisfies R1 and R2 and, and the optimal solutions are continuous with small but arbitrary perturbations of the LP 2.2. □

2.2.2 Later Stage Recourse Program

Here is a version of the IP (10) in Gans et al. (2012). It is different from (12) in that it keeps all scenarios in the formulation instead of using the certainty equivalent formulation and that it introduces an extra set of constraints and intermediate set of variables, α_k's.

$$
\begin{array}{llll}
\max & -\sum\limits_{j\in\mathcal{J}}\sum\limits_{h\in\mathcal{H}_j} d_{jh}z_{jh} & & \\
\text{subject to} & & & \\
\sum\limits_{j\in\mathcal{J}}\sum\limits_{h\in\mathcal{H}_j} r_{ijh}m_{ikn}z_{jh}-\alpha_{ik} & \leq & -b_{ikn}-(\sum\limits_{j\in\mathcal{J}} a_{ij}x_j)m_{ikn} & i\in\mathcal{I}_l, k\in\mathcal{K}', n\in\mathcal{N}_i \\
\sum\limits_{i\in\mathcal{I}_l}\alpha_{ik}-\alpha_k & \leq & 0 & k\in\mathcal{K}' \\
\sum\limits_{k\in\mathcal{K}'} p'_k\alpha_k & \leq & \tilde{\alpha} & \\
\sum\limits_{h\in\mathcal{H}_j} z_{jh} & \leq & x_j & j\in\mathcal{J} \\
z_{jh} & \in & \mathbb{Z}^+ & j\in\mathcal{J}, h\in\mathcal{H}_j \\
\alpha_{ik} & \geq & 0 & i\in\mathcal{I}_l, k\in\mathcal{K}' \\
\alpha_k & \geq & 0 & k\in\mathcal{K}'
\end{array}
\tag{2.4}
$$

Where $\tilde{\alpha}=\sum\limits_{k\in\mathcal{K}'} p'_k\sum\limits_{i\in\mathcal{I}_l}(s_i\cdot m_{iks_i}+b_{iks_i})$ is the expected abandonment rate of the later stage that would have been achieved by the original scheduling policy, and $s_i\equiv\sum\limits_{j\in\mathcal{J}} a_{ij}x_j$ denotes the early stage staffing levels. Without loss of generality, we assume that $x_j>0$ for all $j\in\mathcal{J}$, otherwise, we could reorganize the schedule set $\mathcal{J}$ to exclude those j's with $x_j=0$. Notice that $f(\lambda'_{ik},\mu,\theta,n)$ is non-increasing and convex in n and positive. Then we have

$$n^*\cdot m_{ikn}+b_{ikn}\leq n^*\cdot m_{ikn^*}+b_{ikn^*},\quad \text{for all } i\in\mathcal{I}_l, k\in\mathcal{K}', n, n^*\in\mathcal{N}_i.$$

$$n\cdot m_{ikn}+b_{ikn}\geq 0,\quad \text{for all } i\in\mathcal{I}_l, k\in\mathcal{K}', n\in\mathcal{N}_i.$$

The LP relaxation of (2.4) is as follows,

$$
\begin{array}{llll}
\max & -\sum\limits_{j\in\mathcal{J}}\sum\limits_{h\in\mathcal{H}_j} d_{jh}z_{jh} & & \\
\text{subject to} & & & \\
\sum\limits_{j\in\mathcal{J}}\sum\limits_{h\in\mathcal{H}_j} r_{ijh}m_{ikn}z_{jh}-\alpha_{ik} & \leq & -b_{ikn}-(\sum\limits_{j\in\mathcal{J}} a_{ij}x_j)m_{ikn} & i\in\mathcal{I}_l, k\in\mathcal{K}', n\in\mathcal{N}_i \\
\sum\limits_{i\in\mathcal{I}_l}\alpha_{ik}-\alpha_k & \leq & 0 & k\in\mathcal{K}' \\
\sum\limits_{k\in\mathcal{K}'} p'_k\alpha_k & \leq & \tilde{\alpha} & \\
\sum\limits_{h\in\mathcal{H}_j} z_{jh} & \leq & x_j & j\in\mathcal{J} \\
z_{jh} & \geq & 0 & j\in\mathcal{J}, h\in\mathcal{H}_j \\
\alpha_{ik} & \geq & 0 & i\in\mathcal{I}_l, k\in\mathcal{K}' \\
\alpha_k & \geq & 0 & k\in\mathcal{K}'
\end{array}
\tag{2.5}
$$

And we make an adjustment of (2.5) by increasing $\tilde{\alpha}$ by an arbitrarily small positive number δ, that is, to replace $\tilde{\alpha}$ by $\alpha^* = \tilde{\alpha}+\delta$. By making such adjustment, we allow the QoS constraint to be violated for a little bit. Then we consider the modified LP of (2.5):

$$\begin{array}{llll}
\max & -\sum_{j\in\mathcal{J}}\sum_{h\in\mathcal{H}_j} d_{jh}z_{jh} & & \\
\text{subject to} & & & \\
\sum_{j\in\mathcal{J}}\sum_{h\in\mathcal{H}_j} r_{ijh}m_{ikn}z_{jh}-\alpha_{ik} & \leq & -b_{ikn}-(\sum_{j\in\mathcal{J}} a_{ij}x_j)m_{ikn} & i\in\mathcal{I}_l, k\in\mathcal{K}', n\in\mathcal{N}_i \\
\sum_{i\in\mathcal{I}_l}\alpha_{ik}-\alpha_k & \leq & 0 & k\in\mathcal{K} \\
\sum_{k\in\mathcal{K}'} p'_k\alpha_k & \leq & \alpha^* & \\
\sum_{h\in\mathcal{H}_j} z_{jh} & \leq & x_j & j\in\mathcal{J} \\
z_{jh} & \geq & 0 & j\in\mathcal{J}, h\in\mathcal{H}_j \\
\alpha_{ik} & \geq & 0 & i\in\mathcal{I}_l, k\in\mathcal{K}' \\
\alpha_k & \geq & 0 & k\in\mathcal{K}',
\end{array} \tag{2.6}$$

and prove the stability of the optimal solution to the LP (2.6). We would like to make the following proposition:

Proposition 2.2 *There exist optimal solutions to the LP (2.6), and the optimal solutions are stable under small perturbations of the LP (2.6).*

Proof

We apply Theorem 1 in our proof. Specifically, we will show that R1) and R2) hold for LP (2.6).

Denote the matrix form of (2.6) by

$$\max \quad -\mathbf{dz}$$

subject to

$$
\begin{array}{c|c|c|c|}
 & \sum_{j\in\mathcal{J}}|\mathcal{H}_j| & |\mathcal{I}_l|\times|\mathcal{K}'| & |\mathcal{K}'| \\
\sum_{i\in\mathcal{I}_l}|\mathcal{K}'||\mathcal{N}_i| & (m_{ikn}r_{ijh}) & \begin{matrix} -\mathbf{I}_{\mathcal{K}}\otimes\mathbf{1}_{\mathcal{N}_1} & & \\ & \ddots & \\ & & -\mathbf{I}_{\mathcal{K}}\otimes\mathbf{1}_{\mathcal{N}_{\mathcal{I}_l}}\end{matrix} & \\ \hline
|\mathcal{K}'| & & \mathbf{I}_{\mathcal{K}} \quad \dots \quad \mathbf{I}_{\mathcal{K}} & -\mathbf{I}_{\mathcal{K}} \\ \hline
1 & & & p'_1 \quad \dots \quad p'_{\mathcal{K}} \\ \hline
|\mathcal{J}| & \begin{matrix}\mathbf{1}^T_{\mathcal{H}_1} & & \\ & \ddots & \\ & & \mathbf{1}^T_{\mathcal{H}_{\mathcal{J}}}\end{matrix} & &
\end{array}
\begin{bmatrix}\mathbf{z}\\ \boldsymbol{\alpha}^{(1)}\\ \boldsymbol{\alpha}^{(2)}\end{bmatrix} \le \begin{bmatrix}\mathbf{g}\\ 0\\ \alpha^*\\ \mathbf{x}\end{bmatrix}
$$

$$\mathbf{z}, \boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)} \geq 0 \tag{2.7}$$

Where $\mathbf{d} = (d_{11}, \dots, d_{1\mathcal{H}_1}, \dots, d_{\mathcal{J}1}, \dots, d_{\mathcal{J}\mathcal{H}_\mathcal{J}})$, $\mathbf{z} = (z_{11}, \dots, z_{1\mathcal{H}_1}, \dots, z_{\mathcal{J}1}, \dots, z_{\mathcal{J}\mathcal{H}_\mathcal{J}})^T$, $\boldsymbol{\alpha}^{(1)} = (\alpha_{11}, \dots, \alpha_{1\mathcal{K}}, \dots, \alpha_{\mathcal{I}_l 1}, \dots, \alpha_{\mathcal{I}_l\mathcal{K}})^T$, $\boldsymbol{\alpha}^{(2)} = (\alpha_1, \dots, \alpha_\mathcal{K})^T$. Further denote the matrix form as

$$
\begin{array}{ll}
\max & -\mathbf{dz} \\
\text{subject to} & \\
\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \\ A_{41} & A_{42} & A_{43} \end{bmatrix} \begin{bmatrix} \mathbf{z} \\ \boldsymbol{\alpha}^{(1)} \\ \boldsymbol{\alpha}^{(2)} \end{bmatrix} & \leq \ \mathbf{b} \\
\mathbf{z}, \boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)} & \geq \ 0.
\end{array}
\tag{2.8}
$$

And we will show that R1) and R2) holds for the above LP (2.8).

For condition R1), let $\mathbf{y} = \begin{bmatrix} \mathbf{z} \\ \boldsymbol{\alpha}^{(1)} \\ \boldsymbol{\alpha}^{(2)} \end{bmatrix}$, and $\mathbf{y} \geq 0$.

$$(A_{41}, A_{42}, A_{43})\mathbf{y} \leq 0 \Rightarrow \mathbf{z} = 0.$$

$$(A_{31}, A_{32}, A_{33})\mathbf{y} \leq 0 \Rightarrow \boldsymbol{\alpha}^{(2)} = 0,$$

$$(A_{21}, A_{22}, A_{23})\mathbf{y} \leq 0 \Rightarrow \boldsymbol{\alpha}^{(1)} = 0,$$

Then $\mathbf{y} = 0$ and R1) holds.

For condition R2), the proof is more complex. Firstly LP (2.7) is feasible. In particular, $z_{jh} = 0$, $\alpha_{ik} = s_i \cdot m_{iks_i} + b_{iks_i}$, $\alpha_k = \sum_{i \in \mathcal{I}_l} \alpha_{ik}$ form a feasible solution to (2.7). Let z_{jh}, α_{ik}, and α_k denote any feasible solutions to (2.7). Let $\boldsymbol{\rho} = (\boldsymbol{\rho}^{(1)}, \boldsymbol{\rho}^{(2)}, \rho^{(3)}, \boldsymbol{\rho}^{(4)})$ such that $\boldsymbol{\rho} \geq 0$, $\boldsymbol{\rho} \neq 0$ and $\boldsymbol{\rho} A \geq 0$. Notice that $\boldsymbol{\rho} A \geq 0$ is equivalent to the following three inequalities.

$$\begin{aligned} \sum_{ikn} \rho_{ikn}^{(1)} m_{ikn} r_{ijh} + \rho_j^{(4)} &\geq 0, \quad j \in \mathcal{J} && (2.9) \\ -\sum_{n} \rho_{ikn}^{(1)} + \rho_k^{(2)} &\geq 0, \quad i \in \mathcal{I}_l, k \in \mathcal{K}' && (2.10) \\ -\rho_k^{(2)} + p_k' \rho^{(3)} &\geq 0, \quad k \in \mathcal{K}'. && (2.11) \end{aligned}$$

To show that R2) holds, we consider four situations:

1) $\boldsymbol{\rho}^{(1)} = 0$ and $\boldsymbol{\rho}^{(2)} = 0$.

   There is an element in $(\rho^{(3)}, \boldsymbol{\rho}^{(4)})$ positive, then $\boldsymbol{\rho}\mathbf{b} = \rho^{(3)}\alpha^* + \boldsymbol{\rho}^{(4)}\mathbf{x} > 0$.

2) $\boldsymbol{\rho}^{(1)} = 0$ and $\boldsymbol{\rho}^{(2)} \neq 0$.

   By (2.11) we have $\rho^{(3)} > 0$. Then $\boldsymbol{\rho}\mathbf{b} \geq \rho^{(3)}\alpha^* > 0$.

3) $\boldsymbol{\rho}^{(1)} \neq 0$ and $\boldsymbol{\rho}^{(2)} = 0$.

   Then (2.10) does not hold.

4) $\boldsymbol{\rho}^{(1)} \neq 0$ and $\boldsymbol{\rho}^{(2)} \neq 0$.

Firstly, by (2.11) we have $\rho^{(3)} > 0$.

Notice in the primal that $x_j \geq \sum\limits_h z_{jh}$, then

$$\begin{aligned}
\boldsymbol{\rho}\mathbf{b} &= \sum_i \sum_k \sum_n \rho_{ikn}^{(1)} \left[-b_{ikn} - (\sum_j a_{ij} x_j) m_{ikn} \right] + \rho^{(3)} \alpha^* + \sum_j \rho_j^{(4)} x_j \\
&\geq \sum_i \sum_k \sum_n \rho_{ikn}^{(1)} \left[-b_{ikn} - (\sum_j a_{ij} x_j) m_{ikn} \right] + \rho^{(3)} \alpha^* + \sum_j \sum_h \rho_j^{(4)} z_{jh}, \qquad (2.12)
\end{aligned}$$

where equality holds if and only if $\rho_j^{(4)} x_j = \rho_j^{(4)} \sum\limits_h z_{jh}, j \in \mathcal{J}$.

In the primal, $\left[\sum\limits_j a_{ij} x_j + \sum\limits_j \sum\limits_h r_{ijh} z_{jh} \right] m_{ikn} + b_{ikn} \leq \alpha_{ik}$, then

$$(2.12) \;\; \geq \;\; \sum_{i,k,n} \rho_{ikn}^{(1)} \left[\sum_{j,h} r_{ijh} m_{ikn} z_{jh} - \alpha_{ik} \right] + \rho^{(3)} \alpha^* + \sum_{j,h} \rho_j^{(4)} z_{jh}, \quad (2.13)$$

where equality holds if and only if

$$\rho_{ikn}^{(1)} \left(\left[\sum_j a_{ij} x_j + \sum_{j,h} r_{ijh} z_{jh} \right] m_{ikn} + b_{ikn} \right) = \rho_{ikn}^{(1)} \alpha_{ik}, \quad i \in \mathcal{I}_l, k \in \mathcal{K}', n \in \mathcal{N}_i.$$

By (2.9), we have

$$\begin{aligned}
(2.13) \;\; &= \;\; \sum_{j,h} z_{jh} \left[\rho_j^{(4)} + \sum_{i,k,n} \rho_{ikn}^{(1)} m_{ikn} r_{ijh} \right] + \rho^{(3)} \alpha^* - \sum_{i,k,n} \rho_{ikn}^{(1)} \alpha_{ik} \\
&\geq \;\; \rho^{(3)} \alpha^* - \sum_{i,k,n} \rho_{ikn}^{(1)} \alpha_{ik}, \qquad (2.14)
\end{aligned}$$

where equality holds if and only if $z_{jh} \left[\rho_j^{(4)} + \sum\limits_{i,k,n} \rho_{ikn}^{(1)} m_{ikn} r_{ijh} \right] = 0, \; j \in \mathcal{J}, h \in \mathcal{H}_j$.

By the second and third constraints in the primal, we have

$$\begin{aligned}
(2.14) \;\; &\geq \;\; \rho^{(3)} \sum_k p_k' \alpha_k - \sum_{i,k,n} \rho_{ikn}^{(1)} \alpha_{ik} \\
&\geq \;\; \rho^{(3)} \sum_{k,i} p_k' \alpha_{ik} - \sum_{i,k,n} \rho_{ikn}^{(1)} \alpha_{ik} \qquad (2.15)
\end{aligned}$$

where equality holds if and only if $\alpha^* = \sum_k p'_k \alpha_k$ and $\sum_i \alpha_{ik} = \alpha_k, k \in \mathcal{K}'$.

By (2.10) and (2.11), we have

$$(2.15) \quad \geq \quad 0,$$

where equality holds if and only if $\alpha_k \rho_k^{(2)} = \alpha_k p'_k \rho^{(3)}, k \in \mathcal{K}'$ and $\alpha_{ik}\rho_k^{(2)} = \alpha_{ik} \sum_n \rho_{ikn}^{(1)}, i \in \mathcal{I}_l, k \in \mathcal{K}'$.

Hence, $\boldsymbol{\rho}\mathbf{b} \geq 0$ and equality holds if and only if the following equalities hold at the same time.

$$\rho_j^{(4)} x_j = \rho_j^{(4)} \sum_h z_{jh}, \qquad j \in \mathcal{J}. \tag{2.16}$$

$$\rho_{ikn}^{(1)} \left(\left[\sum_j a_{ij} x_{ij} + \sum_{j,h} r_{ijh} z_{jh} \right] m_{ikn} + b_{ikn} \right) = \rho_{ikn}^{(1)} \alpha_{ik}, \quad i \in \mathcal{I}_l, k \in \mathcal{K}', n \in \mathcal{N}_i. \tag{2.17}$$

$$z_{jh} \left[\rho_j^{(4)} + \sum_{ikn} \rho_{ikn}^{(1)} m_{ikn} r_{ijh} \right] = 0, \qquad j \in \mathcal{J}, h \in \mathcal{H}_j \tag{2.18}$$

$$\sum_k p'_k \alpha_k = \alpha^*. \tag{2.19}$$

$$\alpha_k = \sum_i \alpha_{ik}, \qquad k \in \mathcal{K}'. \tag{2.20}$$

$$\alpha_{ik} \rho_k^{(2)} = \alpha_{ik} \sum_n \rho_{ikn}^{(1)}, \qquad i \in \mathcal{I}_l, k \in \mathcal{K}. \tag{2.21}$$

$$\alpha_k \rho_k^{(2)} = \alpha_k p'_k \rho^{(3)}, \qquad k \in \mathcal{K}'. \tag{2.22}$$

Multiply both sides of (2.19) by $\rho^{(3)}$ which is positive:

$$\begin{aligned}
\rho^{(3)}\alpha^* &\overset{(2.19)}{=} \rho^{(3)}\sum_k p'_k\alpha_k \overset{(2.22)}{=} \sum_k \rho_k^{(2)}\alpha_k \overset{(2.20)}{=} \sum_{i,k}\rho_k^{(2)}\alpha_{ik} \\
&\overset{(2.21)}{=} \sum_{i,k,n}\rho_{ikn}^{(1)}\alpha_{ik} \\
&\overset{(2.17)}{=} \sum_{i,k,n}\rho_{ikn}^{(1)}\left[(\sum_j a_{ij}x_j)m_{ikn} + b_{ikn}\right] + \sum_{i,k,n}\rho_{ikn}^{(1)}\sum_{j,h} r_{ijh}z_{jh}m_{ikn} \\
&\overset{(2.18)}{=} \sum_{i,k,n}\rho_{ikn}^{(1)}\left[(\sum_j a_{ij}x_j)m_{ikn} + b_{ikn}\right] - \sum_{j,h} z_{jh}\rho_j^{(4)} \\
&\overset{(2.16)}{=} \sum_{i,k,n}\rho_{ikn}^{(1)}\left[(\sum_j a_{ij}x_j)m_{ikn} + b_{ikn}\right] - \sum_j \rho_j^{(4)}x_j \\
&\leq \sum_{ikn}\rho_{ikn}^{(1)}(s_i m_{iks_i} + b_{iks_i}) - \sum_j \rho_j^{(4)}x_j \\
&\leq \sum_{ikn}\rho_{ikn}^{(1)}(s_i m_{iks_i} + b_{iks_i}) \\
&\overset{(2.10)}{\leq} \sum_{i,k}(s_i m_{iks_i} + b_{iks_i})\rho_k^{(2)} \\
&\overset{(2.11)}{\leq} \rho^{(3)}\sum_{i,k}(s_i m_{iks_i} + b_{iks_i})p'_k. \\
&= \rho^{(3)}\tilde{\alpha}
\end{aligned}$$

which contradicts the definition of α^*. Hence there must be

$$\boldsymbol{\rho}\mathbf{b} > 0.$$

□

Remark 1 *By the proof of Proposition (2.2), we have the following properties for the LP (2.5) under perturbations of the arrival rate forecasts.*

- *For any perturbation of the arrival forecast distribution, the primal for LP (2.5) is always feasible.*

- *For any perturbation of the arrival forecast distribution such that the p'_k's remain positive:*

  R1) holds. Then by Theorem 1 in Williams (1963), the dual is feasible. Applying Theorem 2 in Williams (1963), there exist optimal solutions to both the primal and dual. The optimal solution set of the primal is bounded and the optimal solution set of the dual is unbounded. In particular, the $\boldsymbol{\rho} = (\boldsymbol{\rho}^{(1)}, \boldsymbol{\rho}^{(2)}, \rho^{(3)}, \boldsymbol{\rho}^{(4)})$ defined as follows satisfies $\boldsymbol{\rho} \geq 0$, $\boldsymbol{\rho} \neq 0$, $\boldsymbol{\rho} A \geq 0$ and $\boldsymbol{\rho}\mathbf{b} = 0$:

$$\begin{cases} \rho^{(1)}_{ikn} = \begin{cases} p'_k & n = s_i, \\ 0 & n \neq s_i, \end{cases} & i \in \mathcal{I}_l, k \in \mathcal{K}', \\ \rho^{(2)}_k = p'_k, & k \in \mathcal{K}', \\ \rho^{(3)} = 1, & \\ \rho^{(4)}_j = 0, & j \in \mathcal{J}. \end{cases}$$

2.2.3 Two-Stage Recourse Program

Here is a version of the IP (12) in the paper Gans et al. (2012). It keeps all scenarios in the formulation instead of using the certainty equivalent formulation and it introduces an extra set of constraints and intermediate set of variables, α_k's, which is useful in our analysis.

$$
\begin{array}{lll}
\max & -\sum_{j} c_j x_j - \sum_{k} p_k \sum_{j,h} d_{jh} z_{jhk} & \\
\text{subject to} & & \\
(\sum_{j} a_{ij} x_j) m_{ikn} - \alpha_{ik} & \leq \; -b_{ikn}, & i \in \mathcal{I}_e, n \in \mathcal{N}_i \\
(\sum_{j} a_{ij} x_j + \sum_{j,h} r_{ijh} z_{jkh}) m_{ikln} - \alpha_{ikl} & \leq \; -b_{ikln}, & i \in \mathcal{I}_l, k \in \mathcal{K}, l \in \mathcal{L}_k, n \in \mathcal{N}_i \\
\sum_{i \in \mathcal{I}_e} \alpha_{ik} + \sum_{i \in \mathcal{I}_l} \sum_{l \in \mathcal{L}_k} p_{kl} \alpha_{ikl} - \alpha_k & \leq \; 0 & k \in \mathcal{K} \\
\sum_{k} p_k \alpha_k & \leq \; \alpha^* \bar{\lambda} & \\
\sum_{h} z_{jkh} - x_j & \leq \; 0 & j \in \mathcal{J}, k \in \mathcal{K} \\
x_j & \in \; \mathbb{Z}^+ & j \in \mathcal{J} \\
z_{jkh} & \in \; \mathbb{Z}^+ & j \in \mathcal{J}, k \in \mathcal{K}, h \in \mathcal{H}_j \\
\alpha_{ik} & \geq \; 0 & i \in \mathcal{I}_e, k \in \mathcal{K} \\
\alpha_{ikl} & \geq \; 0 & i \in \mathcal{I}_l, k \in \mathcal{K}, l \in \mathcal{L}_k \\
\alpha_k & \geq \; 0 & k \in \mathcal{K}.
\end{array}
\tag{2.23}
$$

Notice that $c_j + d_{jh} > 0$, $j \in \mathcal{J}, h \in \mathcal{H}_j$, which denotes the final cost for schedule j when recourse action $h \in \mathcal{H}_j$ is taken. For any i, there is at least one j such that $a_{ij} = 1$, which enables the program to staff interval i. Also notice that $m_{ikn} < 0$, $i \in \mathcal{I}_e, k \in \mathcal{K}, n \in \mathcal{N}_i$ and $m_{ikln} < 0, i \in \mathcal{I}_l, k \in \mathcal{K}, l \in \mathcal{L}_k, n \in \mathcal{N}_i$.

Then consider the LP relaxation of (2.23)

$$
\begin{array}{llll}
\max & -\sum_j c_j x_j - \sum_k p_k \sum_{j,h} d_{jh} z_{jhk} & & \\
\text{subject to} & & & \\
(\sum_j a_{ij} x_j) m_{ikn} - \alpha_{ik} & \leq & -b_{ikn}, & i \in \mathcal{I}_e, n \in \mathcal{N}_i \\
(\sum_j a_{ij} x_j + \sum_{j,h} r_{ijh} z_{jkh}) m_{ikln} - \alpha_{ikl} & \leq & -b_{ikln}, & i \in \mathcal{I}_l, k \in \mathcal{K}, l \in \mathcal{L}_k, n \in \mathcal{N}_i \\
\sum_{i \in \mathcal{I}_e} \alpha_{ik} + \sum_{i \in \mathcal{I}_l} \sum_{l \in \mathcal{L}_k} p_{kl} \alpha_{ikl} - \alpha_k & \leq & 0 & k \in \mathcal{K} \\
\sum_k p_k \alpha_k & \leq & \alpha^* \bar{\lambda} & \\
\sum_h z_{jkh} - x_j & \leq & 0 & j \in \mathcal{J}, k \in \mathcal{K} \\
x_j & \geq & 0 & j \in \mathcal{J} \\
z_{jkh} & \geq & 0 & j \in \mathcal{J}, k \in \mathcal{K}, h \in \mathcal{H}_j \\
\alpha_{ik} & \geq & 0 & i \in \mathcal{I}_e, k \in \mathcal{K} \\
\alpha_{ikl} & \geq & 0 & i \in \mathcal{I}_l, k \in \mathcal{K}, l \in \mathcal{L}_k \\
\alpha_k & \geq & 0 & k \in \mathcal{K}.
\end{array}
\tag{2.24}
$$

We would like to make the following proposition for LP (2.24).

Proposition 2.3 *The LP relaxation (2.24) is solvable. And the optimal solutions of the primal and dual are stable under perturbations of this LP.*

Proof

The matrix form of (2.24)

$$\begin{array}{ll} \max & -\sum_{j} c_j x_j - \sum_{k} p_k \sum_{jh} d_{jh} z_{jkh} \\ \text{subject to} & \end{array}$$

$$\begin{array}{c|ccccc|} & |\mathcal{J}| & \sum\limits_{j\in\mathcal{J}} |\mathcal{K}|\cdot|\mathcal{H}_j| & |\mathcal{I}_e|\cdot|\mathcal{K}| & |\mathcal{I}_l| \sum\limits_{k\in\mathcal{K}} |\mathcal{L}_k| & |\mathcal{K}| \\ \sum\limits_{i\in\mathcal{I}_e} |\mathcal{K}|\cdot|\mathcal{N}_i| & A_{11} & A_{12} & A_{13} & A_{14} & A_{15} \\ \sum\limits_{i\in\mathcal{I}_l} (\sum\limits_{k\in\mathcal{K}} |\mathcal{L}_k|)|\mathcal{N}_i| & A_{21} & A_{22} & A_{23} & A_{24} & A_{25} \\ |\mathcal{K}| & A_{31} & A_{32} & A_{33} & A_{34} & A_{35} \\ 1 & A_{41} & A_{42} & A_{43} & A_{44} & A_{45} \\ |\mathcal{J}|\cdot|\mathcal{K}| & A_{51} & A_{52} & A_{53} & A_{54} & A_{55} \end{array} \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \\ \boldsymbol{\alpha}^{(1)} \\ \boldsymbol{\alpha}^{(2)} \\ \boldsymbol{\alpha}^{(3)} \end{bmatrix} \leq \begin{bmatrix} \mathbf{b}^{(1)} \\ \mathbf{b}^{(2)} \\ 0 \\ \alpha^* \bar{\lambda} \\ 0 \end{bmatrix}$$

$$\mathbf{x}, \mathbf{z}, \boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}, \boldsymbol{\alpha}^{(3)} \geq 0. \qquad (2.25)$$

where

- $A_{11} = (a_{ij} m_{ikn})$.
- $A_{13} = \begin{bmatrix} -\mathbf{I}_{\mathcal{K}} \otimes \mathbf{1}_{\mathcal{N}_1} & & & \\ & -\mathbf{I}_{\mathcal{K}} \otimes \mathbf{1}_{\mathcal{N}_2} & & \\ & & \ddots & \\ & & & -\mathbf{I}_{\mathcal{K}} \otimes \mathbf{1}_{\mathcal{N}_{\mathcal{I}_e}} \end{bmatrix}$
- $A_{21} = (a_{ij} m_{ikln})$.
- $A_{22} = (r_{ijh} m_{ikln})$.
- $A_{24} = \begin{bmatrix} -\mathbf{I}_{\sum_{k=1}^{\mathcal{K}} \mathcal{L}_k} \otimes \mathbf{1}_{\mathcal{N}_1} & & & \\ & -\mathbf{I}_{\sum_{k=1}^{\mathcal{K}} \mathcal{L}_k} \otimes \mathbf{1}_{\mathcal{N}_2} & & \\ & & \ddots & \\ & & & -\mathbf{I}_{\sum_{k=1}^{\mathcal{K}} \mathcal{L}_k} \otimes \mathbf{1}_{\mathcal{N}_{\mathcal{I}_l}} \end{bmatrix}$

- $A_{33} = \mathbf{1}_{\mathcal{I}_e}^T \otimes \mathbf{I}_{\mathcal{K}}$.
- $A_{34} = \mathbf{1}_{\mathcal{I}_l}^T \otimes \mathbf{P}$, where $\mathbf{P} = \begin{bmatrix} p_{11} & \cdots & p_{1\mathcal{L}_1} & & & & & & \\ & & & p_{21} & \cdots & p_{2\mathcal{L}_2} & & & \\ & & & & & & \ddots & & \\ & & & & & & p_{\mathcal{K}1} & \cdots & p_{\mathcal{K}\mathcal{L}_k} \end{bmatrix}$.
- $A_{35} = -\mathbf{I}_{\mathcal{K}}$.
- $A_{45} = [p_1, p_2, \ldots, p_{\mathcal{K}}]$.
- $A_{51} = -\mathbf{I}_{\mathcal{J}} \otimes \mathbf{1}_{\mathcal{K}}$.
- $A_{52} = \begin{bmatrix} \mathbf{I}_{\mathcal{K}} \otimes \mathbf{1}_{\mathcal{H}_1}^T & & & \\ & \mathbf{I}_{\mathcal{K}} \otimes \mathbf{1}_{\mathcal{H}_2}^T & & \\ & & \ddots & \\ & & & \mathbf{I}_{\mathcal{K}} \otimes \mathbf{1}_{\mathcal{H}_{\mathcal{J}}}^T \end{bmatrix}$.
- The other A_{ij}'s are all zero matrices.

For condition R1), let $\mathbf{y} = \begin{bmatrix} \mathbf{x} \\ \mathbf{z} \\ \boldsymbol{\alpha}^{(1)} \\ \boldsymbol{\alpha}^{(2)} \\ \boldsymbol{\alpha}^{(3)} \end{bmatrix}$, $\mathbf{y} \geq 0$, $\mathbf{y} \neq 0$.

$$[A_{41}, A_{42}, A_{43}, A_{44}, A_{45}]\mathbf{y} \leq 0 \Rightarrow \boldsymbol{\alpha}^{(3)} = 0.$$

$$[A_{31}, A_{32}, A_{33}, A_{34}, A_{35}]\mathbf{y} \leq 0 \text{ and } \boldsymbol{\alpha}^{(3)} = 0 \Rightarrow \boldsymbol{\alpha}^{(1)} = 0, \boldsymbol{\alpha}^{(2)} = 0.$$

If $\mathbf{z} = 0$, then there must be $\mathbf{x} \neq 0$ and $\mathbf{cy} = -\sum_j c_j x_j < 0$, and R1 holds.

If $\mathbf{z} \neq 0$,

$$[A_{51}, A_{52}, A_{53}, A_{54}, A_{55}]\mathbf{y} \leq 0 \Rightarrow \sum_{h \in \mathcal{H}_j} z_{jkh} \leq x_j, j \in \mathcal{J}, k \in \mathcal{K}.$$

Then,

$$\begin{aligned}
\mathbf{cy} &= -\sum_j c_j x_j - \sum_k p_k \sum_{j,h} d_{jh} z_{jkh} \\
&= -\sum_j c_j \sum_k p_k x_j - \sum_k p_k \sum_{j,h} d_{jh} z_{jkh} \\
&\leq -\sum_j c_j \sum_k p_k \sum_h z_{jkh} - \sum_k p_k \sum_{j,h} d_{jh} z_{jkh} \\
&= -\sum_{j,k,h} p_k z_{jkh} [c_j + d_{jh}] \\
&< 0,
\end{aligned}$$

and R1 holds.

For condition R2), let $\boldsymbol{\rho} = (\boldsymbol{\rho}^{(1)}, \boldsymbol{\rho}^{(2)}, \boldsymbol{\rho}^{(3)}, \rho^{(4)}, \boldsymbol{\rho}^{(5)})$, $\boldsymbol{\rho} \geq 0$.

$$\boldsymbol{\rho}[A_{11}^T, A_{21}^T, A_{31}^T, A_{41}^T, A_{51}^T]^T \geq 0 \Rightarrow \boldsymbol{\rho}^{(1)} = 0, \boldsymbol{\rho}^{(2)} = 0, \boldsymbol{\rho}^{(5)} = 0.$$

If $\rho^{(4)} = 0$,

$$\boldsymbol{\rho}[A_{15}^T, A_{25}^T, A_{35}^T, A_{45}^T, A_{55}^T]^T \geq 0 \Rightarrow \boldsymbol{\rho}^{(3)} = 0,$$

then $\boldsymbol{\rho} = 0$ and R2 automatically holds.

If $\rho^{(4)} \neq 0$, then $\boldsymbol{\rho}\mathbf{b} = \rho^{(4)}\alpha^*\bar{\lambda} > 0$, and R2 holds.

□

2.2.4 A More Generalized Model for the Simple Stochastic Program

We now consider a more generalized format of the simple stochastic program and derive the stability results.

Let $P(\lambda)$ be a probability distribution of $\lambda = (\lambda_1, \ldots, \lambda_I) \in \Lambda := \mathbb{R}_+^I$. $f(n, \lambda)$ is the abandonment proportion function under staffing level n and arrival rate λ, where $n \in \mathbb{Z}_+$ and $\lambda \in \mathbb{R}_+$. We extend the definition of $f(., \lambda)$ onto $\mathbb{R}_+$ for any λ by linearly interpolating adjacent points in $n \in \mathbb{Z}_+$. And we refer to $f(n, \lambda)$ as the interpolated functions on $\mathbf{R}_+ \times \mathbf{R}_+$ from now on.

Our stochastic program is

$$\begin{array}{lll} \min & \sum_j c_j x_j & \\ \text{subject to} & & \\ & \sum_i \int \lambda_i f(n_i(x), \lambda_i) \mathrm{d}P(\lambda_i) & \leq \alpha^* \sum_i \int \lambda_i \mathrm{d}P(\lambda_i), \\ & x & \in \mathbb{R}_+^J. \end{array}$$

where $n_i(x) = \sum_j a_{ij} x_j$ is the staffing level on the interval i under the schedule vector $x = (x_1, \ldots, x_J)$, $a_{ij} = 1$ if schedule j has an agent working on the interval i and $a_{ij} = 0$ otherwise.

We rewrite the above formulation in a more general form as follows:

$$\min \left\{ F_0(x) : x \in X, \int_\Lambda F_1(x, \lambda) \mathrm{d}P(\lambda) \leq 0 \right\}, \tag{2.26}$$

where

$$X = \mathbb{R}_+^J,$$

$$F_0(x) = F_0(x, \lambda) = \sum_j c_j x_j,$$

$$F_1(x, \lambda) = \sum_i \lambda_i [f(n_i(x), \lambda_i) - \alpha^*].$$

Denote the set of all Borel probability measures on Λ by $\mathcal{P}(\Lambda)$, the feasible set of (2.26) by $\mathcal{X}(P)$, the optimal value by $\vartheta(P)$ and the solution set of (2.26) by $X^*(P)$, i.e.,

$$\begin{aligned} \mathcal{X}(P) &:= \left\{ x \in X : \int F_1(x, \lambda) \mathrm{d}P(\lambda) \leq 0 \right\}, \\ \vartheta(P) &:= \inf \left\{ F_0(x) : x \in \mathcal{X}(P) \right\}, \\ X^*(P) &:= \left\{ x \in \mathcal{X}(P) : F_0(x) = \vartheta(P) \right\}. \end{aligned}$$

For any nonempty and open subset $U \subset \mathbb{R}^J$, consider the following sets,

$$\begin{array}{rcl} \mathcal{F}_U & := & \{F_j(x,.) : x \in X \cap \mathbf{cl}U, j = 0,1\}, \\ \mathcal{P}_{\mathcal{F}_U}(\Lambda) := \mathcal{P}_{\mathcal{F}_U} & := & \{Q \in \mathcal{P}(\Lambda) : -\infty < \int\limits_{\Lambda} \inf\limits_{x \in X \cap r\mathbb{B}} F_j(x,\lambda)\mathrm{d}Q(\lambda) \text{ for each } r > 0 \\ & & \quad \text{and} \sup\limits_{x \in X \cap \mathbf{cl}U} \int\limits_{\Lambda} F_j(x,\lambda)\mathrm{d}Q(\lambda) < \infty \text{ for } j = 0,1\}, \end{array}$$

For any $Q \in \mathcal{P}_{\mathcal{F}_U}(\Lambda)$, denote

$$\begin{array}{rcl} \mathcal{X}_U(Q) & := & \left\{ x \in X \cap \mathbf{cl}U : \int\limits_{\Lambda} F_1(x,\lambda)\mathrm{d}Q(\lambda) \le 0 \right\}, \\ \vartheta_U(Q) & := & \inf\{F_0(x) : x \in \mathcal{X}_U(Q)\}, \\ X_U^*(Q) & := & \{x \in \mathcal{X}_U(Q) : F_0(x) = \vartheta_U(Q)\}. \end{array}$$

Then we have the following theorem.

Theorem 2 *Let $P(\lambda)$ be a probability distribution of λ such that $\mathbf{E}_P(\lambda) < \infty$. Then $X^*(P)$ is non-empty and bounded. Let U be an open bounded neighborhood of $X^*(P)$.*

Furthermore, assume that the sequence of probability distributions $P_n(\lambda)$ satisfies the following conditions:

1. *P_n is weakly convergent to P.*
2. *$\sup\limits_n \left\{ \int\limits_{\lambda} (\sum\limits_i \lambda_i)^{1+\epsilon}\mathrm{d}P_n(\lambda) \right\}$ is bounded for some $\epsilon > 0$.*

Then the sequence $(\vartheta_U(P_n))$ converges to $\vartheta(P)$, and

$$\lim_{n\to\infty} \sup_{x \in X_U^*(P_n)} \mathrm{d}(x, X^*(P)) = 0.$$

Proof

We first show that $f(.,.)$ has the following properties:

- $f(n,\lambda)$ is continuous and strictly decreasing w.r.t. n for any $\lambda > 0$. $f(n,\lambda) \to 0$, as $n \to \infty$, $f(0,\lambda) = 1$ for any $\lambda > 0$.
- $f(n,\lambda)$ is continuous w.r.t. λ for all $n \in \mathbb{Z}_+$ and $\lambda > 0$. Then $f(n,\lambda) := (\lceil n \rceil - n) f(\lfloor n \rfloor, \lambda) + (n - \lfloor n \rfloor) f(\lceil n \rceil, \lambda)$ is continuous w.r.t. λ for all $n \geq 0$ and $\lambda > 0$.
- $\dfrac{\partial f}{\partial_+ n}(n,\lambda) := f(\lfloor n{+}1 \rfloor, \lambda) - f(\lfloor n \rfloor, \lambda) < 0$ is continuous w.r.t. λ, and $\dfrac{\partial f}{\partial_- n}(n,\lambda) := f(\lceil n \rceil, \lambda) - f(\lceil n-1 \rceil, \lambda) < 0$ is continuous w.r.t. λ, for all $\lambda > 0$, $n \geq 0$.

We use Theorem 5 and Theorem 6 in Romisch (2003) to prove our theorem. In the following, we verify that the conditions of Theorem 5 and Theorem 6 in Romisch (2003) hold.

Our formulation (2.26) is of the same form as (1.1) in Romisch (2003), where X is closed, Λ is a closed subset of $\mathbb{R}^I$. Next we show that the functions F_j are random lower semicontinuous functions for $j = 0, 1$. Consider the epigraphical mapping $\lambda \mapsto \mathbf{epi}F_j(.,\lambda) := \{(x,r) : F_j(x,\lambda) \leq r\}$. When $j = 0$, it is obvious that this epigraphical mapping is closed-valued and measurable. When $j = 1$, $F_1(x,\lambda)$ is continuous w.r.t. λ, so $F_1(x,.)$ is measurable for any fixed x. For any limit point $(\hat{x},\hat{r})$ of $\mathbf{epi}F_1(.,\lambda)$, there exists a sequence $(x_n, r_n) \in \mathbf{epi}F_1(.,\lambda)$ such that $(x_n, r_n) \to (\hat{x},\hat{r})$ as $n \to \infty$. Notice that $F_1(x_n,\lambda) \leq r_n$, and $F_1(.,\lambda)$ is continuous w.r.t. x, then $F_1(\hat{x},\lambda) \leq \hat{r}$. So the limit point $(\hat{x},\hat{r}) \in \mathbf{epi}F_1(.,\lambda)$, and $\mathbf{epi}F_1(.,\lambda)$ is closed. The σ-filed on $\mathbb{R}_+^J \times \mathbb{R}$ can be generated by the sets of the following form

$$[0, x_1'] \times \ldots \times [0, x_J'] \times (-\infty, r'].$$

Since $F_1(x,\lambda)$ is continuous and monotone w.r.t. $x_j, j = 1, \ldots, J$ and λ,

$$([0,x_1'] \times \cdots \times [0,x_J'] \times (-\infty, r'])^{-1} = ([0,x_1'] \times \cdots \times [0,x_J'] \times \{r'\})^{-1} = (\{x_1'\} \times \cdots \times \{x_J'\} \times \{r'\})^{-1}.$$

$(\{x_1'\} \times \cdots \times \{x_J'\} \times \{r'\})^{-1}$ is measurable because $F_1((x_1', \ldots, x_J'), .)$ is measurable.

Then the epigraphical mapping is measurable when $j = 1$. Hence by definition, the functions F_j are random lower semicontinuous functions for $j = 0, 1$.

We now verify the conditions of Theorem 5 in Romisch (2003).

For any open bounded set U, we have the followings

$$\inf_{x \in X \cap r\mathbb{B}} F_0(x) > -\infty, \quad \forall r > 0.$$

$$\int_\Lambda \inf_{x \in X \cap r\mathbb{B}} F_1(x, \lambda) \mathrm{d}P(\lambda) \geq \int_\Lambda \sum_i \lambda_i(-\alpha^*) \mathrm{d}P(\lambda) > -\infty.$$

$$\sup_{x \in X \cap \mathbf{cl}U} F_0(x) = \sup_{x \in X \cap \mathbf{cl}U} c \cdot x < \infty.$$

$$\sup_{x \in X \cap \mathbf{cl}U} \int_\Lambda F_1(x, \lambda) \mathrm{d}P(\lambda) \leq \int_\Lambda \sum_i \lambda_i [1 - \alpha^*] \mathrm{d}P(\lambda),$$

then $P \in \mathcal{P}_{\mathcal{F}_U}$.

Next we show that $X^*(P)$ is non-empty and bounded. Denote

$$g(x) := \int \sum_i \lambda_i [f(n_i(x), \lambda_i) - \alpha^*] \mathrm{d}P(\lambda),$$

and $g(x)$ is continuous w.r.t. x since $f(, .\lambda)$ is continuous w.r.t. x and by dominated convergence theorem. Furthermore,

$$\lim_{d \to \infty} g(d\overrightarrow{1}) = -\alpha^* \int \sum \lambda_i \mathrm{d}P(\lambda) < 0,$$

by dominated convergence theorem. Then there exists some d' such that $g(d'\overrightarrow{1}) < 0$ and $X(P)$ is non-empty. In addition, $X(P) = \{x : g(x) \leq 0\}$ is a closed set because $g(x)$ is continuous. If $X(P)$ has only one element, then $X^*(P) = X(P) \neq \emptyset$. If $X(P)$ has two or more elements, then let $x_1, x_2 \in X(P)$, such that $cx_1 \leq cx_2$.

$$\begin{aligned} X(P) &= [X(P) \cap \{x : cx \leq cx_2\}] \cup [X(P) \cap \{x : cx > cx_2\}] \\ &:= X' \cup X'' \end{aligned}$$

where X', X'' are disjoint, and X' is non-empty, closed, and bounded.

$$\mathcal{V}(P) := \inf\{cx : x \in X(P)\} = \inf\{cx : x \in X'\}.$$

There exists some $x \in X'$ such that $cx = \inf\{cx : x \in X'\}$. Then $cx = \mathcal{V}(P)$ and $X^*(P)$ is non-empty and bounded.

$F_0(x) = cx$ is a linear function about x, so it is Lipschitz continuous. So the second condition of Theorem 5 in Romnisch (2003) is satisfied.

For the third condition in Theorem 5 in Romnisch (2003). Notice that $g(x)$ is continuous and non-increasing w.r.t. $x_j, j = 1, \ldots, J$.

$$g(0) = \int \sum_i \lambda_i (1 - \alpha^*) \mathrm{d}P(\lambda) = (1 - \alpha^*) \sum_i \mathbf{E}(\lambda_i) > 0.$$

$$\lim_{x_j \to \infty, i \in \mathcal{J}} g(x) = \int \sum \lambda_i (0 - \alpha^*) \mathrm{d}P(\lambda) = -\alpha^* \sum \mathbf{E}(\lambda_i) < 0,$$

by dominated convergence theorem (because $f(n, \lambda) \to 0$ as $n \to \infty$). Denote $c_0 = g(0) > 0$, and $c_\infty = \lim_{x_j \to \infty, i \in \mathcal{J}} g(x) < 0$. For any $x_0 \in X = \mathbb{R}_+^J$, define

$$h_{x_0}(d) \quad := \quad \begin{cases} g(d \cdot x_0), \quad 0 \leq d \leq 1 \\ g(x_0 + (d-1) \cdot \overrightarrow{1}), \quad d \geq 1, \end{cases}$$

where $d \in [0, \infty)$. Then $h_{x_0}(d)$ is continuous, non-increasing w.r.t. d, and $h_{x_0}(0) = c_0$ and $h_{x_0}(\infty) = c_\infty$, for all $x_0 \in X$. Let $\epsilon > 0$, such that $\epsilon < \min\{|c_0|, |c_\infty|\}$. There exists $d_\epsilon > 1$, such that

$$h_0(d_\epsilon) = -\epsilon.$$

Notice that $n_i(x_0 + (d_\epsilon - 1)\overrightarrow{1}) \geq n_i(\overrightarrow{0} + (d_\epsilon - 1)\overrightarrow{1})$ for all $x_0 \in X$, then we have $h_{x_0}(d_\epsilon) \leq h_0(d_\epsilon) = -\epsilon$, for all $x_0 \in X$. Then

$$h_{x_0}([0, d_\epsilon]) \supseteq h_0([0, d_\epsilon]) \supseteq [-\epsilon, \epsilon], \quad \forall x_0 \in X.$$

For any $\bar{x} \in X^*(P)$, consider a set $A := \{x' \in X : x' = x + d\overrightarrow{1}, \text{where } x \in X, ||x - \bar{x}|| \leq \epsilon, d \in [0, d_\epsilon]\}$. The set A is bounded, then there exists $n' \in \mathbb{Z}_+$ such that $n_i(A) \subset [0, n'], i \in \mathcal{I}$.

Let λ', λ'' be such that $0 < \lambda' < \lambda'' < \infty$. Then $\dfrac{\partial f}{\partial_+ n}(n, \lambda) < 0$, and $\dfrac{\partial f}{\partial_- n}(n, \lambda) < 0$, for all λ with $\lambda' \leq \lambda \leq \lambda''$, and for all $n \in [0, n']$. Since $\dfrac{\partial f}{\partial_+ n}$ and $\dfrac{\partial f}{\partial_- n}$ are continuous w.r.t. λ for any fixed n, then

$$a_n := \sup_{\lambda \in [\lambda', \lambda'']} \left\{ \frac{\partial f}{\partial_+ n}(n, \lambda), \frac{\partial f}{\partial_- n}(n, \lambda) \right\} < 0, \quad \forall n \in [0, n'].$$

Notice that $f(n, \lambda)$ is the linear interpolated function w.r.t. n for any fixed λ, then

$$\begin{aligned} a &:= \sup_{n \in [0,n'], \lambda \in [\lambda', \lambda'']} \left\{ \frac{\partial f}{\partial_+ n}(n, \lambda), \frac{\partial f}{\partial_- n}(n, \lambda) \right\} \\ &= \sup_{n \in [0,n'] \cap \mathbb{Z}, \lambda \in [\lambda', \lambda'']} \left\{ \frac{\partial f}{\partial_+ n}(n, \lambda), \frac{\partial f}{\partial_- n}(n, \lambda) \right\} \\ &= \max_{n \in [0,n'] \cap \mathbb{Z}} \{a_n\} \\ &< 0 \end{aligned}$$

Thus we have

$$f(n_1, \lambda) - f(n_2, \lambda) \geq |a|(n_2 - n_1)$$

for all $n_1 \in [0, n']$ and $n_2 \in [0, n']$ such that $n_1 \leq n_2$, and for all $\lambda \in [\lambda', \lambda'']$.

For any $x \in X$ and y with $||x - \bar{x}|| < \epsilon$ and $|y| < \epsilon$, If $g(x) \leq y$, then $x \in \mathcal{X}_y(P)$ and $\mathrm{d}(x, \mathcal{X}_y(P)) = 0$. If $g(x) > y$, then there exists some d with $1 < d \leq d_\epsilon$ such that

$y = h_x(d) = g(x + (d-1)\overrightarrow{1})$. Then

$$\begin{aligned}
& \max\{0, \int F_1(x,\lambda)\mathrm{d}P(\lambda) - y\} \\
= \;& g(x) - y \\
= \;& g(x) - g(x + (d-1)\overrightarrow{1}) \\
= \;& \int \sum_i \lambda_i \left[f(n_i(x), \lambda_i) - f(n_i(x + (d-1)\overrightarrow{1}), \lambda_i) \right] \mathrm{d}P(\lambda) \\
\geq \;& \int\limits_{\{\lambda:\lambda_i \in [\lambda', \lambda''], i \in I\}} \sum_i \lambda_i \left[f(n_i(x), \lambda_i) - f(n_i(x + (d-1)\overrightarrow{1}), \lambda_i) \right] \mathrm{d}P(\lambda) \\
\geq \;& \int\limits_{\{\lambda:\lambda_i \in [\lambda', \lambda''], i \in I\}} \sum_i \lambda_i |a| [n_i(x + (d-1)\overrightarrow{1}) - n_i(x)] \\
= \;& \int\limits_{\{\lambda:\lambda_i \in [\lambda', \lambda''], i \in I\}} \sum_i \lambda_i |a| (d-1) n_i(\overrightarrow{1}) \mathrm{d}P(\lambda) \\
:= \;& a^*(d-1) \\
= \;& a^*(d-1)\frac{||\overrightarrow{1}||}{||\overrightarrow{1}||} \\
= \;& \frac{a^*}{||\overrightarrow{1}||}\mathrm{d}(x, x + (d-1)\overrightarrow{1}),
\end{aligned}$$

where

$$a^* = |a| \cdot \int\limits_{\{\lambda:\lambda_i \in [\lambda', \lambda''], i \in I\}} \sum_i \lambda_i n_i(\overrightarrow{1}) \mathrm{d}P(\lambda) > 0.$$

Thus we have

$$\mathrm{d}(x, x + (d-1)\overrightarrow{1}) \leq \tilde{a} \cdot \max\left\{0, \int F_1(x,\lambda)\mathrm{d}P(\lambda) - y\right\},$$

where $\tilde{a} = \frac{||\overrightarrow{1}||}{a^*}$ only depends on ϵ. Also notice that

$$x + (d-1)\overrightarrow{1} \in \mathcal{X}_y(P),$$

then we have

$$\mathrm{d}(x, \mathcal{X}_y(P)) \leq \tilde{a} \cdot \max\left\{0, \int F_1(x,\lambda)\mathrm{d}P(\lambda) - y\right\}.$$

And the third condition of Theorem 5 is satisfied.

Next we show the condition in Theorem 6 that $\mathcal{F}_U^R$ is a P-uniformity class for large $R > 0$,is valid. Sufficient conditions for $\mathcal{F}$ being a P-uniformity class is that: $\mathcal{F}$ is uniformly bounded and it holds that $\mathrm{P}(\{\lambda : \mathcal{F} \text{ is not equicontinuous at } \lambda\}) = 0$.

U is an open bounded neighborhood of $X^*(P)$. Then $F_0(x)|_{X\cap\mathbf{cl}U} = \sum\limits_j c_j x_j|_{X\cap\mathbf{cl}U}$ is bounded. In addition,

$$\sum_i \lambda_i(-\alpha^*) \le F_1(x,\lambda) \le \sum_i \lambda_i(1-\alpha^*), \quad \forall x \in X.$$

Then the class of truncated functions $\mathcal{F}_U^R = \mathcal{F}_U|_{|\lambda|\le R}$ is uniformly bounded.

U is an open bounded set, then there exists $\bar{n} \in \mathbb{Z}$ such that $n_i(U) \subseteq [0,\bar{n}], \forall i$. Let $\bar{N} = [0,\bar{n}] \cap \mathbb{Z}$. Since $f(n,.)$ is continuous with respect to λ, then for any $\lambda > 0$, $\epsilon > 0$, and $n \in \bar{N}$, there exists $\delta_n > 0$, such that $|f(n,\lambda') - f(n,\lambda)| < \epsilon$ holds for any λ' such that $|\lambda' - \lambda| < \delta_n$. Then $|f(n,\lambda') - f(n,\lambda)| < \epsilon$ holds for any λ' such that $|\lambda' - \lambda| < \delta := \min\limits_{n\in\bar{N}}\{\delta_n\}$ and for all $n \in \bar{N}$. Since $f(n,\lambda)$ is the linear interpolation of $f(n,\lambda)|_{\{n\in\mathbb{Z}_+\}}$ for any fixed λ, then $|f(n,\lambda') - f(n,\lambda)| \le \max\{|f(\lfloor n\rfloor,\lambda') - f(\lfloor n\rfloor,\lambda)|, |f(\lceil n\rceil,\lambda') - f(\lceil n\rceil,\lambda)|\} < \epsilon$ holds for all λ' such that $|\lambda' - \lambda| < \delta$ and for all $n \in [0,\bar{n}]$. Hence we have showed that

$$\{f(n,.) : n \in [0,\bar{n}]\}$$

is equicontinuous at all $\lambda > 0$. The identity function $I(\lambda) = \lambda$ is continuous, so we have

$$\{I(.)f(n,.) : n \in [0,\bar{n}]\}$$

is equicontinuous at all $\lambda > 0$. Since $n_i(x)$ is a linear function about x and $n_i(U) \in [0,\bar{n}]$, then

$$\{I(.)f(n_i(x),.) : x \in X \cap \mathbf{cl}U, i \in \mathcal{I}\}$$

is equicontinuous at all $\lambda > 0$. Since $F_1(x, \lambda) = \sum_i I(\lambda_i)[f(n(x), \lambda_i) - \alpha^*]$, then

$$\{F_1(x, .) : x \in X \cap \mathbf{cl}U\}$$

is equicontinuous at all $\lambda > 0$. Then $\mathcal{F}_U^R$ is equicontinuous at all $\lambda > 0$, and we have

$$\mathrm{P}\{\mathcal{F}_U^R \text{ is not equicontinuous at } \lambda\} = 0.$$

We now verify the condition in Theorem 6 that $\mathcal{F}_U$ is a uniformly integrable with respect to $\{P_n : n \in \mathbb{N}\}$. $\mathcal{F}_U$ is uniformly integrable if the moment condition

$$\sup_{n \in \mathbb{N}} \sup_{F \in \mathcal{F}_U} \int |F(\lambda)|^{1+\epsilon} \mathrm{d}P_n(\lambda) < \infty$$

holds for some $\epsilon > 0$.

$$\mathcal{F}_U = \{F_0(x) : x \in X \cap \mathbf{cl}U\} \cup \{F_1(x, .) : x \in X \cap \mathbf{cl}U\}$$

where $\{F_0(x) : x \in X \cap \mathbf{cl}U\} \subseteq \mathcal{F}_U$ is uniformly bounded by some R_0. With the assumption of our theorem,

$$\sup_n \int (\sum \lambda_i)^{1+\epsilon} \mathrm{d}P_n(\lambda)$$

is bounded by some R_1 for some $\epsilon > 0$ and notice that,

$$|F_1(x, \lambda)|^{1+\epsilon} \leq \max\{\alpha^*, 1 - \alpha^*\}^{1+\epsilon} (\sum_i \lambda_i)^{1+\epsilon},$$

we have

$$\begin{aligned}
& \sup_n \sup_{F \in \mathcal{F}_U} \int |F(\lambda)|^{1+\epsilon} \mathrm{d}P_n(\lambda) \\
\leq \;& \sup_n \max \left\{ R_0, \max\{\alpha^*, 1 - \alpha^*\}^{1+\epsilon} \int (\sum \lambda_i)^{1+\epsilon} \mathrm{d}P_n(\lambda) \right\} \\
\leq \;& R_0 \vee (\max\{\alpha^*, 1 - \alpha^*\}^{1+\epsilon} R_1) \\
< \;& \infty
\end{aligned}$$

Hence all the conditions in Theorem 5 and Theorem 6 hold, and we have proved our theorem.

□

2.3 Discussion and Future Work

The current definition of $\tilde{\alpha}$ in Section 2.2.2 is natural and straightforward, which however does not ensure stability of the optimal solutions under arbitrary perturbations of model parameters. The LP (2.5) is similar to the example of Robinson (1977) (Page 443) in that the primal constraints are not regular although the dual constraints are. It's demonstrated in Robinson (1977) that the LP with such constraints can behave very badly indeed, even due to rounding parameters. Nevertheless, perturbations of the parametric forecasts actually are not "arbitrary" perturbations, which at least guarantees the existence of optimal solutions. Thus the framework we applied from Robinson (1977) can be too "general" in our concerned context.

Future work includes further exploring the stability features of LP (2.5), which requires extra knowledge in math programming theories. And more generally, one may also consider using other definitions of $\tilde{\alpha}$.

Statistical features of the parametric forecasts are to be examined (Problem P1). In particular, we are interested in proving consistency property of the parametric forecasts with and without early stage updates. Meanwhile, the stability of IP's with respect to LP relaxations are to be analyzed (Problem P2.1).

3 Forecasting and Staffing Call Centers with Multiple Arrival Streams

In this chapter we are concerned with forecasting and staffing call centers with multiple uncertain arrival customer streams. In the statistical forecasting stage, Ibrahim and L'Ecuyer (2012) built linear mixed models to jointly forecast the arrival counts of two arrival streams. They performed numerical comparison between their bi-stream models and single-stream models on real call center data, but failed to explain the fact that their bi-stream models had no solid improvement over the single-stream models, although there is significant dependence between the two streams. In the staffing/scheduling stage, many operations research/management papers use skill based routing to deal with multiple uncertain arrival stream problem, among which Gurvich et al. (2010) proposed a chance-constraint approach.

In the following sections, we combine the statistical forecasting stage and the operational staffing stage together to form a complete solution to the staffing problem with multiple uncertain arrival streams. Our work is unique because:

- We are the first, to the best of our knowledge, to fill the gap between the forecasting stage and the staffing stage by completely solving the multiple arrival stream staffing problem. In particular, the benefits of combining the two stages include: we gain operational assessment of the statistical forecasting models (traditional forecasting papers only provide statistical evaluations for the point forecast accuracy), which is more informative in practice; we gain realistic assessment of the staffing policy, which helps making the staffing policy suitable for real life situations.

- We provide theoretical and numerical analysis showing how the benefit of incorporating inter-stream dependence varies by the type and strength of the inter-stream dependence, in both forecasting and staffing stages. The way we consider the inter-stream dependence makes more general sense compared with Ibrahim and L'Ecuyer (2012), and their numerical results are natural outcomes of our findings.

3.1 Literature Review

We now review the relevant literature on forecasting and staffing call centers. Most important are good call center review articles including Gans et al. (2003) and Aksin et al. (2007).

There are many papers related to forecasting call arrivals and arrival rate uncertainty. Recent work includes Avramidis et al. (2004), Brown et al. (2005), Weinberg et al. (2007) and Shen and Huang (2008). More relevantly, Aldor-Noiman et al. (2010) proposed an additive Gaussian linear mixed effect model for a single arrival stream. Ibrahim and L'Ecuyer (2012) consider a similar additive mixed effect model and extend it to incorporate two arrival streams. In their models, the transformed count is decomposed into the day-of-week effect, within-day time interval effect, the interaction term, a random daily effect and a Gaussian error as follows:

$$X^i_{d,t} = \alpha^i_{w_d} + \beta^i_t + \tau^i_{w_d,t} + \gamma^i_d + e^i_{d,t}. \tag{3.1}$$

They achieve the forecasts by modeling the random effect γ_d through an AR(1) time series structure. In their additive model, all the time intervals in a forecast day have the same random effect. The major difference between our multiplicative model and their additive model is that in the multiplicative model, a time interval has a random effect which is proportional to its arrival rate magnitude. In this way, the time intervals with fewer average arrivals would have a less random effect and vice versa.

Ibrahim and L'Ecuyer (2012) consider two ways of modeling the dependence between two arrival streams. In their first model, they assume dependence on the random daily effect. In particular, the random effect γ_d^i in model 3.1 only depends on its own first order lag term γ_{d-1}^i, and the white noise of the time series γ_d^1 and γ_d^2 are correlated. In their second model the dependence is modeled by the correlation between the within day error term $e_{d,t}^1$ and $e_{d,t}^2$. In our formulation, we model the dependence in a more general form, which covers but not limited to, both the above types of dependence. We also theoretically demonstrate that their first bivariate model provides no benefits in point forecasts, which coheres their numerical results.

Recent papers in operations management 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. (2006b,a), 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.

3.2 Statistical Methodology

In this section, we develop statistical models to forecast multiple-stream arrival volumes. We consider a multiplicative format for the intra-day arrival volume profile. Particularly, we use regression techniques to decompose the arrival volume profile of each stream on a certain day into the product of daily total arrival rate and proportion profile of the corresponding day-of-week. We then apply vector autoregressive time series model to forecast the vector daily total arrival rates. Distributional forecasts of both arrival rate and count is obtained. We discuss our estimation and forecasting procedure. We also discuss and compare alternative models.

3.2.1 Forecasting Model

Denote the number of customer types (or arrival streams) as I. For each arrival stream, say i, we observe the number of calls during time period t on day d, for $t = 1, \ldots, T$ and $d = 1, \ldots, D$. For example, the time period can be every quarter hour or half hour during the business day. Denote the number of arrivals as $N_{d,t}^{(i)}$, for

$i = 1, \dots, I$.

We model $N_{d,t}^{(i)}$ as Poisson($\lambda_{d,t}^{(i)}$) where the *random* arrival rate $\lambda_{d,t}^{(i)}$ depends on customer type i, time period of the day t and the day d (most likely through day of the week, say w_d).

To begin with, we apply the squareroot transformation to normalize the arrival counts. By now, this transformation has become common in the call center forecasting literature. Denote

$$X_{d,t}^{(i)} = \sqrt{N_{d,t}^{(i)} + \frac{1}{4}} \sim \mathrm{N}\left(\sqrt{\lambda_{d,t}^{(i)}}, \sigma_{(i)}^2\right).$$

We then consider the following forecasting model for the square-root-transformed counts $X_{d,t}^{(i)}$:

$$\begin{cases} X_{d,t}^{(i)} = \sqrt{\lambda_{d,t}^{(i)}} + \epsilon_{d,t}^{(i)}, \quad \epsilon_{d,t} = (\epsilon_{d,t}^{(1)}, \epsilon_{d,t}^{(2)}, \dots, \epsilon_{d,t}^{(I)})^T \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma), \\ \theta_{d,t}^{(i)} \equiv \sqrt{\lambda_{d,t}^{(i)}} = u_d^{(i)} \ f_{w_d,t}^{(i)}, \\ \mathbf{u}_d - \alpha_{w_d} = \mathbf{A}(\mathbf{u}_{d-1} - \alpha_{w_{d-1}}) + \mathbf{z}_d, \quad \mathbf{z}_d = (z_d^{(1)}, z_d^{(2)}, \dots, z_d^{(I)})^T \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Omega), \\ f_{w_d,t}^{(i)} \geq 0, \quad \sum_{t=1}^{T} f_{w_d,t}^{(i)} = 1, \end{cases} \tag{3.2}$$

where w_d is day-of-week of day d, $\mathbf{u}_d = (u_d^{(1)}, u_d^{(2)}, \dots, u_d^{(I)})^T$ is the vector daily total arrival rate of all customer streams (on the square-root scale), $\alpha_{w_d} = (\alpha_{w_d}^{(1)}, \alpha_{w_d}^{(2)}, \dots, \alpha_{w_d}^{(I)})^T$ is the adjustment of daily total arrival rate (on the square-root scale) for the day of week, $\mathbf{A} = (a_{i'j'})_{I\times I}$ is the auto-regressive coefficient matrix, $f_{w_d,t}^{(i)}$ is the intraday rate proportion for the t^{th} time interval for customer type i that also depends on the corresponding day of week, $\Omega = (\Omega_{rl})_{I\times I}$ and $\Sigma = (\Sigma_{rl})_{I\times I}$ are the covariance matrices.

Our model is the multivariate extension of the forecasting model in Noah et al.. It can be understood in the following way. The square-root transformed data approximately follows multivariate Gaussian distribution (Brown et al. 2005). On the square-root transformed scale, the arrival rate profile for any customer type i on day d ($\theta_d^{(i)} \equiv (\theta_{d,1}^{(i)}, \theta_{d,2}^{(i)}, \ldots, \theta_{d,T}^{(i)}) \equiv (\sqrt{\lambda_{d,1}^{(i)}}, \sqrt{\lambda_{d,2}^{(i)}}, \ldots, \sqrt{\lambda_{d,T}^{(i)}})$) is assumed to have a multiplicative format, which is the product of the daily total rate $u_d^{(i)}$ and the intraday proportion profile of the corresponding day-of-week $(f_{w_d,1}^{(i)}, f_{w_d,2}^{(i)}, \ldots, f_{w_d,T}^{(i)})$. The vector daily total rate of all customer types $\mathbf{u}_d$ follows a first-order vector autoregressive time series model, after removal of the day-of-week effect α_{w_d}.

We model the dependence among arrival streams via Σ, $\mathbf{A}$ and Ω in our formulation. Next we specify a particular 2-dimension case for Model 3.2, that is when $I = 2$, to explain how the dependence is modeled. The 2-d detailed equation is as follows:

$$\begin{cases} X_{d,t}^{(i)} = \sqrt{\lambda_{d,t}^{(i)}} + \epsilon_{d,t}^{(i)}, \quad i = 1, 2, \\ \begin{pmatrix} \epsilon_{d,t}^{(1)} \\ \epsilon_{d,t}^{(2)} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right), \\ \theta_{d,t}^{(i)} \equiv \sqrt{\lambda_{d,t}^{(i)}} = u_d^{(i)} \, f_{w_d}^{(i)}, \quad i = 1, 2, \\ f_{w_d}^{(i)} \geq 0, \quad \sum_{t=1}^{T} f_{w_d}^{(i)} = 1, \\ \begin{pmatrix} u_d^{(1)} - \alpha_{w_d}^{(1)} \\ u_d^{(2)} - \alpha_{w_d}^{(2)} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} u_{d-1}^{(1)} - \alpha_{w_{d-1}}^{(1)} \\ u_{d-1}^{(2)} - \alpha_{w_{d-1}}^{(2)} \end{pmatrix} + \begin{pmatrix} z_d^{(1)} \\ z_d^{(2)} \end{pmatrix}, \\ \begin{pmatrix} z_d^{(1)} \\ z_d^{(2)} \end{pmatrix} \overset{\text{i.i.d.}}{\sim} \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix} \right). \end{cases} \tag{3.3}$$

In equation 3.3, there are three types of dependence between the two streams. For example, the arrival stream of customer type 1 depends on the arrival stream of customer type 2 in the following three ways:

- Type (a) dependence: the arrival count of customer type 1: $X_{d,t}^{(1)}$ depends on the arrival count of customer type 2: $X_{d,t}^{(2)}$ of the same day and the same time interval. Particularly for our formulation, $(X_{d,t}^{(1)}, X_{d,t}^{(2)})^T$ follows a bivariate normal distribution with a correlation $r \equiv \Sigma_{12}/\sqrt{\Sigma_{11}\Sigma_{22}}$, given the arrival rate $(\theta_{d,t}^{(1)}, \theta_{d,t}^{(2)})^T$.

- Type (b) dependence: the daily total arrival rate of customer type 1: $u_d^{(1)}$ depends on the daily total arrival rate of customer type 2 of the previous day: $u_{d-1}^{(2)}$. The direction and strength of the dependence is carried by a_{12}.
- Type (c) dependence: the daily total arrival rate of customer type 1: $u_d^{(1)}$ depends on the daily total arrival rate of customer type 2 of the same day: $u_d^{(2)}$. Particularly for our formulation, $(u_d^{(1)}, u_d^{(2)})^T$ follows a bivariate Gaussian distribution with a correlation $\rho \equiv \Omega_{12}/\sqrt{\Omega_{11}\Omega_{22}}$, given the daily total rate of the previous day: $(u_{d-1}^{(1)}, u_{d-1}^{(2)})^T$.

Ibrahim and L'Ecuyer (2012) used a different formulation (additive instead of multiplicative), but we could still compare our model with theirs on how dependence between the two streams is modeled. In their paper, they consider two types of inter-stream dependence: correlation of the daily rate between two streams of the same day, which coincides our Type (c) dependence; correlation of the count between two streams of the same day and same time interval, which coincides our Type (a) dependence. However, their paper did not consider Type (b) dependence, which is crucial in reducing forecasting error as we'll discuss about later.

3.2.2 Forecasting Error

Let y denote a random variable and let $\xi_n, n = 1, 2, \ldots,$ denote a series of random variables. Let $\Gamma_y = \mathrm{Var}(y)$, $\Gamma_n = \mathrm{Cov}(y, \xi_n)$, $\mathbf{\Gamma}_{(n)} = (\Gamma_1, \Gamma_2, \ldots, \Gamma_n)^T$, $\boldsymbol{\xi}_{(n)} = (\xi_1, \xi_2, \ldots, \xi_n)^T$, $\boldsymbol{\mu}_{(n)} = \mathrm{E}(\boldsymbol{\xi}_{(n)})$, $m_{s,l} = \mathrm{Cov}(\xi_s, \xi_l)$, $s, l = 1, 2, \ldots,$ $\mathbf{M}_{(n)} =$

$\mathrm{Cov}(\boldsymbol{\xi}_{(n)}) = (m_{s,l})_{n\times n}$, $\mathbf{m}_{(n+1)} = (m_{1,n+1}, m_{2,n+1}, \ldots, m_{n,n+1})^T$. We assume $\mathbf{M}_{(n)}$ is non-singular, $n = 1, 2, \ldots$.

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

$$\begin{pmatrix} y \\ \boldsymbol{\xi}_{(n)} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_y \\ \boldsymbol{\mu}_{(n)} \end{pmatrix}, \begin{pmatrix} \Gamma_y & \boldsymbol{\Gamma}_{(n)}^T \\ \boldsymbol{\Gamma}_{(n)} & \mathbf{M}_{(n)} \end{pmatrix}\right).$$

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

$$y|\boldsymbol{\xi}_{(n)} \sim \mathrm{N}\left(\tilde{\mu}_n, \tilde{\Gamma}_n\right), \quad n = 1, 2, \ldots,$$

where

$$\tilde{\mu}_n = \mu_y + \boldsymbol{\Gamma}_{(n)}^T \mathbf{M}_{(n)}^{-1} (\boldsymbol{\xi}_{(n)} - \boldsymbol{\mu}_{(n)}),$$

$$\tilde{\Gamma}_n = \Gamma_y - \boldsymbol{\Gamma}_{(n)}^T \mathbf{M}_{(n)}^{-1} \boldsymbol{\Gamma}_{(n)}.$$

Notice that

$$\mathrm{Var}(\xi_{n+1}|\boldsymbol{\xi}_{(n)}) = m_{n+1,n+1} - \mathbf{m}_{(n+1)}^T \mathbf{M}_{(n)}^{-1} \mathbf{m}_{(n+1)} \geq 0.$$

And assume ξ_{n+1} is non-redundant with $\boldsymbol{\xi}_{(n)}$, then

$$\mathrm{Var}(\xi_{n+1}|\boldsymbol{\xi}_{(n)}) = m_{n+1,n+1} - \mathbf{m}_{(n+1)}^T \mathbf{M}_{(n)}^{-1} \mathbf{m}_{(n+1)} > 0.$$

Thus the forecasting variance reduced by introducing one more variable ξ_{n+1} is given by

$$\begin{aligned}
\Delta_{n+1} &:= \tilde{\Gamma}_n - \tilde{\Gamma}_{n+1} \\
&= -\boldsymbol{\Gamma}_{(n)}^T \mathbf{M}_{(n)}^{-1} \boldsymbol{\Gamma}_{(n)} + \boldsymbol{\Gamma}_{(n+1)}^T \mathbf{M}_{(n+1)}^{-1} \boldsymbol{\Gamma}_{(n+1)}^T \\
&= -\boldsymbol{\Gamma}_{(n)}^T \mathbf{M}_{(n)}^{-1} \boldsymbol{\Gamma}_{(n)} + (\boldsymbol{\Gamma}_{(n)}^T, \Gamma_{n+1}) \begin{pmatrix} \mathbf{M}_{(n)} & \mathbf{m}_{n+1} \\ \mathbf{m}_{n+1}^T & m_{n+1,n+1} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\Gamma}_{(n)} \\ \Gamma_{n+1} \end{pmatrix} \\
&= \frac{(\Gamma_{n+1} - \boldsymbol{\Gamma}_{(n)}^T \mathbf{M}_{(n)}^{-1} \mathbf{m}_{(n+1)})^2}{m_{n+1,n+1} - \mathbf{m}_{(n+1)}^T \mathbf{M}_{(n)}^{-1} \mathbf{m}_{(n+1)}} \qquad (3.4) \\
&\geq 0.
\end{aligned}$$

Then we discuss the forecasting variance reduced by introducing more arrival streams in our forecasting model. In Model 3.2, we consider to forecast $u_d^{(1)}$ based on $u_{d-1}^{(1)}, u_{d-1}^{(2)}, \dots, u_{d-1}^{(I)}$. Equation 3.4 shows that including more streams in forecast process will sometimes reduce the forecasting error. The benefits depends on many factors. Under the autoregressive structure in Model 3.2, Δ_n is a function of $\mathbf{A}$ and Ω, $n = 2, 3, \dots, I$.

For example, to compare the forecasting variance between bivariate method (when $I = 2$) with univariate method (when $I = 1$), denote the joint normal distribution of $(u_d^{(1)}, u_{d-1}^{(1)}, u_{d-1}^{(2)})^T$ as:

$$\begin{pmatrix} u_d^{(1)} \\ u_{d-1}^{(1)} \\ u_{d-1}^{(2)} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \begin{pmatrix} \Gamma_{11} & \Gamma_{12} & \Gamma_{13} \\ \Gamma_{21} & \Gamma_{22} & \Gamma_{23} \\ \Gamma_{31} & \Gamma_{32} & \Gamma_{33} \end{pmatrix} \right)$$

where $\Gamma_{sl} = \Gamma_{ls},\ s, l = 1, 2, 3,\ \Gamma_{11} = \Gamma_{22},\ \mu_1 = \mu_2$. Let $\rho_{sl} := \Gamma_{sl}/\sqrt{\Gamma_{ss}\Gamma_{ll}}$. By equation 3.4 the variance reduction of bivariate method from univariate method is

$$\begin{aligned} \Delta_2 &= \mathrm{Var}(u_d^{(1)} | u_{d-1}^{(1)}) - \mathrm{Var}(u_d^{(1)} | u_{d-1}^{(1)}, u_{d-1}^{(2)}) \\ &= \Gamma_{11} \frac{(\rho_{13} - \rho_{12}\rho_{23})^2}{1 - \rho_{23}^2} \geq 0. \end{aligned} \tag{3.5}$$

A simplest situation is when $a_{12} = a_{21} = 0$ (that is, when there is no Type (b) dependence), then $\Delta_2 = 0$, which means there is no need of considering bivariate forecasting method to improve point forecast in such a case. When $a_{21} \neq 0$ and $a_{12} \neq 0$, the expression of 3.5 is non-zero but has a very complicated form. We'll later on discuss this issue in a simulation study. Hence we see that considering Type (b) dependence is essential in reducing point forecast error. The first bivariate model in Ibrahim and L'Ecuyer (2012) did not provide substantial improvement in point forecasting, and the reason might be that their model didn't consider the Type(b) dependence between two streams.

3.2.3 Forecasting Procedure

The forecasting procedure is performed in two stages. First, we obtain the estimates of $u_d^{(i)}$, $f_{w_d,t}^{(i)}$ and Σ for $i = 1, \ldots, I, t = 1, \ldots, T, d = 1, \ldots, D$, through iterations of General Least Square (GLS) regression with the following steps.

- Denote the estimates of $u_d^{(i)}$, $f_{w_d,t}^{(i)}$ and Σ in the m^{th} iteration by $\hat{u}_d^{(i),(m)}$, $\hat{f}_{w_d,t}^{(i),(m)}$ and $\hat{\Sigma}^{(m)}$, respectively.
- Before the iteration starts, initialize the estimates $\hat{f}_{w_d,t}^{(i),(0)}$, $\hat{u}_d^{(i),(0)}$ and $\hat{\Sigma}^{(0)}$ through simple calculations and Ordinary Least Squares (OLS) regression. $\hat{f}_{w_d,t}^{(i),(0)}$ is estimated by taking the proportion of the transformed counts of time interval t out of all transformed counts in the same day-of-week:

  $$\hat{f}_{w_d,t}^{(i),(0)} = \frac{\sum\limits_{d':w_{d'}=w_d} X_{d',t}^{(i)}}{\sum\limits_{d':w_{d'}=w_d} \sum\limits_{t'} X_{d',t'}^{(i)}}. \tag{3.6}$$

  Then fit OLS to get $\hat{u}_d^{(i),(0)}$:

  $$X_{d,t}^{(i)} = u_d^{(i)}\ \hat{f}_{w_d,t}^{(i),(0)} + \epsilon_{d,t}^{(i)}, \quad \epsilon_{d,t}^{(i)} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2), \qquad \mathrm{i} = 1, 2, \ldots, \mathrm{I},\ \mathrm{d} = 1, 2, \ldots, \mathrm{D},\ \mathrm{t} = 1, 2, \ldots, \mathrm{T}.$$

  As is assumed, $\epsilon_{d,t} = (\epsilon_{d,t}^{(1)}, \epsilon_{d,t}^{(2)}, \ldots, \epsilon_{d,t}^{(I)}) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma)$, we calculate the model residual to get $\hat{\Sigma}^{(0)}$.
- In the m^{th} iteration: fit the following GLS regression model with linear constraints to get the updating estimate $\hat{f}_{w_d,t}^{(i),(m)}$:

  $$\begin{cases} X_{d,t}^{(i)} = \hat{u}_d^{(i),(m-1)}\ f_{w_d,t}^{(i)} + \epsilon_{d,t}^{(i)}, \quad \epsilon_{d,t} = (\epsilon_{d,t}^{(1)}, \epsilon_{d,t}^{(2)}, \ldots, \epsilon_{d,t}^{(I)}) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \hat{\Sigma}^{(\mathrm{m}-1)}), \\ \sum\limits_{t} f_{w_d,t}^{(I)} = 1. \end{cases}$$

  Get the update of $\hat{\Sigma}^{(m)}$ using the model residual covariance matrix. Then fit the following GLS regression model to get the update $\hat{u}_d^{(i),(m)}$:

  $$X_{d,t}^{i} = u_d^{i}\ \hat{f}_{w_d,t}^{i,(m)} + \epsilon_{d,t}^{i}, \quad \epsilon_{d,t} = (\epsilon_{d,t}^{1}, \epsilon_{d,t}^{2}, \ldots, \epsilon_{d,t}^{I}) \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \hat{\Sigma}^{(\mathrm{m})})$$

Again, update $\hat{\Sigma}^{(m)}$ by the model residual covariance matrix.

- The iteration process stops when $\hat{f}^{(i),(m)}_{w_d,t}$ converges in terms of m. Say, stop at the M^{th} iteration such that

$$\sqrt{\frac{\sum\limits_{t=1}^{T}(\hat{f}^{(i),(M-1)}_{w_d,t} - \hat{f}^{(i),(M)}_{w_d,t})^2}{T}} < 10^{-8}, \qquad \forall w_d, i.$$

$\hat{u}^{(i),(M)}_d$ and $\hat{f}^{(i),(M)}_{w_d,t}$ are the final estimates for $u^{(i)}_d$ and $f^{(i)}_{w_d,t}$, respectively. Then we use the residual covariance matrix to estimate $\hat{\Sigma}$ width d.f. $= DT - D - w^* \cdot (T-1)$, where w^* is the number of working days in one week.

We refer to the above forecasting procedure as "GLS" method for later discussion.

Remarks: Fitting GLS regression model is very time-consuming. Alternatively we consider fitting Ordinary Least Square (OLS) regression instead of GLS in the iterations. The alternative process provides estimates that are quite close to that of GLS method. On our call center data set, the relative difference on $\sqrt{\frac{\sum\limits_t f^{(i)}_{w_d,t}}{T}}$ is around 0.02% between the two estimation methods. And using OLS is faster and easier to program. The estimation procedure of OLS method follows these steps:

- Denote the estimates of $u^{(i)}_d$ and $f^{(i)}_{w_d,t}$ in the m^{th} iteration by $\hat{u}^{(i),(m)}_d$ and $f^{(i),(m)}_{w_d,t}$, respectively.
- Before the iteration starts, initialize $\hat{f}^{(i),(0)}_{w_d,t}$ by Equation 3.6
- In the m^{th} iteration, fit the following OLS model to get the update $\hat{u}^{(i),(m)}_d$:

$$X^{(i)}_{d,t} = u^{(i)}_d \ \hat{f}^{(i),(m-1)}_{w_d,t} + \epsilon^{(i)}_{d,t}, \qquad \epsilon^{(i)}_{d,t} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2).$$

Fit OLS model to get the update $\hat{f}^{(i),(m)}_{w_d,t}$:

$$X^{(i)}_{d,t} = \hat{u}^{(i),(m)}_d \ f^{(i)}_{w_d,t} + \epsilon^{(i)}_{d,t}, \qquad \epsilon^{(i)}_{d,t} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \tilde{\sigma}^2).$$

Normalize $\hat{f}^{(i),(m)}_{w_d,t}$ by a multiplier such that $\sum\limits_t \hat{f}^{(i),(m)}_{w_d,t} = 1$.

- The iteration process stops at the M^{th} iteration such that

$$\sqrt{\frac{\sum_{t=1}^{T}(\hat{f}_{w_d,t}^{(i),(M-1)} - \hat{f}_{w_d,t}^{(i),(M)})^2}{T}} < 10^{-8}, \qquad \forall w_d, i.$$

  $\hat{u}_d^{(i),(M)}$ and $\hat{f}_{w_d,t}^{(i),(M)}$ are the final estimates for u_d^i and $f_{w_d,t}^i$. And we use the residual covariance matrix in the last iteration to estimate Σ with d.f. $= DT - D - w^*(T-1)$.

We refer the above estimation method as "OLS" method for later discussion.

Once we get the estimates of $\hat{\mathbf{u}}_d$, $\hat{f}_{w_d,t}^{(i)}$ and $\hat{\Sigma}$, $d = 1, 2, \ldots, D, t = 1, \ldots, T, i = 1, \ldots, I$, from the first stage multiplicative estimation procedure, we use the vector time series model to obtain a distributional forecast for the daily total rates $\mathbf{u}_{D+h}$ in day h in the future , which is Gaussian. In particular, we first estimate the day-of-week effect for the daily total rates:

$$\hat{\alpha}_{w_d}^{(i)} = \frac{\sum\limits_{d':w_{d'}=w_d} u_{d'}^{(i)}}{\sum\limits_{d':w_{d'}=w_d} 1}.$$

Let $\hat{\alpha}_{w_d} = (\hat{\alpha}_{w_d}^{(1)}, \ldots, \hat{\alpha}_{w_d}^{(I)})^T$. We then apply the vector time series model as follows:

$$\hat{\mathbf{u}}_d - \hat{\alpha}_{w_d} = \mathbf{A} \cdot (\hat{\mathbf{u}}_{d-1} - \hat{\alpha}_{w_d}) + \mathbf{z}_d, \quad \mathbf{z}_d \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Omega), \quad d = 2, \ldots, D. \tag{3.7}$$

We fit model 3.7 in R using the function "ar" and denote the estimated coefficient matrix and covariance matrix by $\hat{\mathbf{A}}$ and $\hat{\Omega}$, respectively. Then the point forecast for the daily total rate on day $D + h$ is given by

$$\hat{\mathbf{u}}_{D+h} = \hat{\alpha}_{w_{D+h}} + \hat{\mathbf{A}}^h \cdot (\hat{\mathbf{u}}_D - \hat{\alpha}_{w_D}), \tag{3.8}$$

with the forecast error

$$\sum_{h'=1}^{h} \hat{\mathbf{A}}^{h'-1} \mathbf{z}_{D+h'},$$

where $\mathbf{z}_{D+h'} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \hat{\Omega})$. In particular, the covariance matrix of the forecast error of $\hat{\mathbf{u}}_{D+h}$ is

$$\hat{\Omega}^{(D+h)} = \sum_{h'=0}^{h-1} \hat{\mathbf{A}}^{h'} \hat{\Omega} \hat{\mathbf{A}}^{h'}. \tag{3.9}$$

3.2.4 Distributional Forecast

Given the mean in equation 3.8 and variance in equation 3.9, and under the Gaussian assumption, the distributional forecast for daily arrival rate $\mathbf{u}_{D+h}$ is

$$\mathbf{u}_{D+h} \sim \mathrm{N}\left(\hat{\mathbf{u}}_{D+h}, \hat{\Omega}^{(D+h)}\right).$$

Let $\hat{\Omega}^{(D+h)} = (\hat{\Omega}_{rl}^{(D+h)})_{I \times I}$, $\hat{\mathbf{F}}_{D+h,t} := \mathrm{diag}\{\hat{f}_{w_{D+h},t}^{(1)}, \ldots, \hat{f}_{w_{D+h},t}^{(I)}\}$, then the distributional forecast for the arrival rate vector $\theta_{D+h,t} := (\theta_{D+h,t}^{(1)}, \ldots, \theta_{D+h,t}^{(I)})^T = \hat{\mathbf{F}}_{D+h,t}\mathbf{u}_{D+h}$ is as follows:

$$\theta_{D+h,t} \sim \mathrm{N}\left(\hat{\mathbf{F}}_{D+h,t}\hat{\mathbf{u}}_{D+h},\ \hat{\mathbf{F}}_{D+h,t}\hat{\Omega}^{(D+h)}\hat{\mathbf{F}}_{D+h,t}\right).$$

Particularly, the forecast mean for $\theta_{D+h,t}^{(i)}$ is $\hat{f}_{w_{D+h},t}^{(i)}\hat{u}_{D+h}$, $i = 1, \ldots, I$, and the forecast covariance between $\theta_{D+h,t}^{(i)}$ and $\theta_{D+h,t}^{(i')}$ is $\hat{f}_{w_{D+h},t}^{(i)}\hat{f}_{w_{D+h},t}^{(i')}\hat{\Omega}_{ii'}^{(D+h)}$, $i, i' = 1, 2, \ldots, I$.

Let $\mathbf{X}_{D+h,t} := (X_{D+h,t}^{(1)}, \ldots, X_{D+h,t}^{(I)})^T$ and $\epsilon_{D+h,t} := (\epsilon_{D+h,t}^{(1)}, \ldots, \epsilon_{D+h,t}^{(I)})$. Notice that

$$\mathbf{X}_{D+h,t} = \theta_{D+h,t} + \epsilon_{D+h,t}.$$

And also notice that the estimated distribution for $\epsilon_{D+h,t}$ is

$$\epsilon_{D+h,t} \sim \mathrm{N}(0, \hat{\Sigma}).$$

Then the distributional forecast for the arrival count vector $\mathbf{X}_{D+h,t}$ is

$$\mathbf{X}_{D+h,t} \sim \mathrm{N}(\hat{\mathbf{F}}_{D+h,t}\hat{\mathbf{u}}_{D+h},\ \hat{\mathbf{F}}_{D+h,t}\hat{\Omega}^{(D+h)}\hat{\mathbf{F}}_{D+h,t} + \hat{\Sigma}).$$

In particular, the forecast mean of $X^{(i)}_{D+h,t}$ is $\hat{f}^{(i)}_{w_{D+h,t}}\hat{u}_{D+h}$, $i = 1, \ldots, I$, and the forecast covariance between $X^{(i)}_{D+h,t}$ and $X^{(i')}_{D+h,t}$ is $\hat{f}^{(i)}_{w_{D+h,t}}\hat{f}^{(i')}_{w_{D+h,t}}\hat{\Omega}^{(D+h)}_{ii'}+\hat{\Sigma}_{ii'}$, $i, i' = 1, 2, \ldots, I$.

3.2.5 Performance Measures

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

$$\text{RMSE}_d = \sqrt{\frac{1}{T}\sum_t (\hat{N}_{d,t} - N_{d,t})^2},$$

$$\text{MRE}_d = \frac{100}{T}\sum_t \frac{|\hat{N}_{d,t} - N_{d,t}|}{N_{d,t}}.$$

To assess the distributional forecast, we define the coverage probability and the width of the 95% confidence interval of day d as follows:

$$\text{COVER}_d = \frac{1}{T}\sum_t \text{I}(\hat{N}^{(2.5)}_{d,t} \leq N_{d,t} \leq \hat{N}^{(97.5)}_{d,t})$$

$$\text{WIDTH}_d = \frac{1}{T}\sum_t (\hat{N}^{(97.5)}_{d,t} - \hat{N}^{(2.5)}_{d,t}),$$

where $\hat{N}^{(q)}_{d,t}$ is the q^{th} percentile of the distributional forecast for $N_{d,t}$, I(.) is the indicator function. Good forecasting model is supposed to have the coverage probability close to the nominal value (95%) and narrow confidence interval.

3.3 Staffing Algorithm

Multiple stream staffing problem could be dealt with skill-based routing strategies, which assign the “most suitable” agent to an incoming call instead of simply choosing the next available agent. Recently, skill-based routing papers attempt to account for the uncertain arrival rate while making staffing and scheduling policies. Among those, Gurvich et al. (2010) proposed a multiple stream chance-constraint optimization approach providing staffing levels that meets the uncertain demand in a pre-chosen probability level. In their approach, they assume the existence of some forecasting distributions for the arrival rates before the staffing planning process can be taken. In this paper we adopt their method to explore the operational effect of simultaneously modeling multiple arrival streams instead of independently modeling each of them.

3.3.1 The Chance Constraint Formulation

Consider the call center with I customer classes have J server pools. Set $\mathscr{I} = \{1, \ldots, I\}$ and $\mathscr{J} = \{1, \ldots, J\}$. Servers in the same pool have the same skills in terms of the set of customer classes they are capable of serving. 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 $N = (N_1, N_2, \ldots, N_J)^T$ where N_j denotes the number of agents on schedule from server pool j. 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. During the time interval, class-i customers arrive according to a stationary Poisson process with rate Λ_i, where $\Lambda = (\Lambda_1, \ldots, \Lambda_I)$ is a multivariate random variable following a certain distribution.

If there is no available agents upon the arrival of a customer, he is queued. Customers are served in a First Come First Serve manner and we allow the customers to abandon the queue with an exponential patience rate ϕ_i for customer class i. Let $a_i(\lambda, N, \pi)$ denote the long run fraction of class-i customers that abandon before being serviced when the arrival rate is λ, the staffing vector is N and the routing rule is π.

Quality-of-service constraints: Given a risk level $\delta > 0$, pre-specified threshold proportion of abandonments ψ_i for customer type i, and a random arrival rate $\Lambda = (\Lambda_1, \Lambda_2, \ldots, \Lambda_I)$, the QoS constraint is given by

$$\mathbf{P}\left(\Lambda : a_i(\Lambda_i, N, \pi) \leq \psi_i,\ i \in \mathscr{I}\right) \geq 1 - \delta.$$

The staffing problem: Assume that agents of pool j incur a cost c_j, $c = (c_1, \ldots, c_J)$. Our objective is find the staffing vector that minimizes the staffing cost subject to the QoS constraint. The optimization problem is given by:

$$\begin{aligned} \min\quad & c \cdot N \\ \text{s.t.}\quad & \mathbf{P}\left(\Lambda : a_i(\Lambda_i, N, \pi) \leq \psi_i,\ i \in \mathscr{I}\right) \geq 1 - \delta. \\ & N \in \mathbb{Z}_+^J,\ \pi \in \Pi. \end{aligned} \tag{3.10}$$

Analytical solution of the above optimization problem 3.10 might not be approachable. When the arrival rate is perfectly known, a static-planning problem (SPP) is often used to provide first-order approximations for the optimization problem 3.10. Particularly, given the arrival rate vector $\Lambda = (\Lambda_1, \Lambda_2, \ldots, \Lambda_I)$, the SPP is given by:

$$\begin{aligned} \min\quad & c \cdot N \\ \text{s.t.}\quad & \sum_{j \in J(i)} \mu_{ij} v_{ij} \geq \Lambda_i (1 - \psi_i),\ i \in \mathscr{I}, \\ & \sum_{i \in I(j)} v_{ij} \leq N_j,\ j \in \mathscr{J} \\ & N \in \mathbb{R}_+^J,\ v \in \mathbb{R}_+^{I \times J}. \end{aligned}$$

When the arrival rate is a random variable with a certain distribution, Gurvich et al. use a random static-planning problem (RSPP) whose optimal solution provides a lower bound of the optimal values of 3.10. The RSPP is given as follows:

$$\begin{aligned} \min \quad & c \cdot N \\ \text{s.t.} \quad & \mathbf{P}(\Lambda \in \mathcal{B}(N)) \geq 1 - \delta, \\ & N \in \mathbb{R}_+^J, \end{aligned} \tag{3.11}$$

where

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

And they also proved any feasible staffing vector N for 3.11 is necessarily feasible for 3.10.

We should notice that the optimal solution of the RSPP 3.11 might not be feasible for the original formulation 3.10. The output of the RSPP also provides a set of Λ which has a probability measure of at least $1 - \delta$, and on which the QoS constraint is met.

To solve the RSPP, they used a discrete approximation of the random arrival rate Λ 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 our paper. In particular, independent samples of size K are generated from the distribution of Λ and each sample point is assigned the same probability $1/K$. Denote the k^{th} sample by $\Lambda(k) = (\Lambda_1(k), \dots, \Lambda_I(k))^T$. Hence the sample based RSPP is given by:

$$
\begin{aligned}
\min \quad & c \cdot N \\
\text{s.t.} \quad & \sum_{j \in J(i)} \mu_{ij} v_{ij}^k \geq y_k \Lambda_i(k)(1 - \psi_i), \quad i \in \mathcal{I}, k = 1, \dots, K, \\
& \sum_{i \in I(j)} v_{ij}^k \leq N_j, \quad j \in \mathcal{J}, k = 1, \dots, K, \\
& \sum_k y_k \geq K(1 - \delta), \\
& N \in \mathbb{R}_+^J, \quad y_k \in \{0, 1\}, v^k \in \mathbb{R}_+^{I \times J}, k = 1, \dots, K.
\end{aligned}
\tag{3.12}
$$

The optimal solution of 3.12 includes a staffing vector $\hat{N}$ as well as a set of $\Lambda_i(k)$'s, using which they generated a set of staffing frontier $\mathcal{F}$. The support area of $\mathcal{F}$ defined by $\mathcal{M}(\mathcal{F}) \equiv \bigcup_{\lambda' \in \mathcal{F}} \{\lambda : \lambda \leq \lambda'\}$, has a probability measure of at least $1 - \delta$.

Then starting from the vector $\hat{N}$, they used a simulation based approach to search feasible solutions for 3.10 on the staffing frontier $\mathcal{F}$, which will give the final optimal staffing vector N^*.

3.3.2 Sampling Process

Notice that we only observe the realized counts instead of the true arrival rate. And also notice the fact that the first order approximation of the chance-constraint is in essence that the number of customers to be served is less than the number of customers that the system is capable of serving with a pre-specified confidence probability. Thus we make use of the distributional forecast of the counts instead of the arrival rates.

3.3.3 Performance Measures

Given a distributional forecast of $\Lambda = (\Lambda_1, \dots, \Lambda_I)^T$, a staffing vector N can be obtained with the above algorithm. Given the realization of Λ, we are able to evaluate the performance of the staffing vector and further the distributional forecast. Let $C = (C_1, C_2, \dots, C_I)^T$ denote the realized Poisson count, and it's assumed that $C_i \sim \text{Poisson}(\Lambda_i)$, and Λ has a certain distributional forecast.

We consider the following performance measures. Denote $v(C, N)$ as the violation indicator function for the QoS constraint. $v(C, N) = 1$ indicates the QoS constraint is violated when the realized count is C under the staffing vector N, and accordingly $v(C, N) = 0$ indicates the QoS constraint is satisfied. Denote $s(C, N)$ as the magnitude of the violation if there is one, under realization C and staffing vector N. Notice that $s(C, N) = 0$ when $v(C, N) = 0$ and $s(C, N) > 0$ when $v(C, N) = 1$. 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 T' intervals. Let $C^{(t)}$ and $N^{(t)}$ denote the observed counts and staffing vector for the t^{th}, respectively. Then the violation probability is defined as

$$\text{v.prob} = \frac{1}{T'} \sum_{t=1}^{T} v(C^{(t)}, N^{(t)}).$$

3.3.4 Operational Staffing Algorithm Setup

We first consider one operational set-up (M-design) for dealing with two arrival streams: two pools of dedicated servers that only handle one customer type and one pool of flexible servers that handles both types of customers.

3.4 Simulation Study

In this section, we consider different scenarios of dependence among streams. And we evaluate the benefits of using multiple-stream method under those different scenarios. We focus on two streams, that is, $I = 2$.

3.4.1 Simulation Set-up

In our multiple stream formulation, the dependence between the two streams is modeled through Σ, Ω and $\mathbf{A}$. In particular, $r \equiv \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}}$ and $\rho \equiv \frac{\Omega_{12}}{\sqrt{\Omega_{11}\Omega_{22}}}$ depicts the daily-total correlation and time-interval correlation between the two types, respectively. a_{12} and a_{21} sculpture the daily-total dependence on the other queue's past day information. We're wondering how multiple-stream method performs when the type of dependence changes, that is, under different values of r, ρ, a_{12} and a_{21}.

Moreover, the effect of ρ on each time interval is very weak according to real data estimates, because the variation of daily-totals Ω after distributed to each time interval is considerably small, compared with Σ. So we deliberately omit the dependence in Ω in our simulation.

With the above facts, we consider the following set-ups for generating scenarios:

- Use the real data estimates as $f^{(i)}_{w_d,t}$ and α_{w_d}.
- Set $\Omega = ((300, 0)^T, (0, 120)^T)^T$, $\Sigma_{11} = 0.8$, $\Sigma_{22} = 0.6$, $a_{11} = 0.6$, $a_{22} = 0.4$. These numbers are chosen according to real data estimates.
- Vary the strength and direction of interval dependence r (that is, Type (a) dependence described in Section 3.2.1). In particular, let $r = -0.8, -0.4, 0, 0.4, 0.8$.

- Vary the strength and direction of past day dependence on the other queue a_{12} and a_{21} (that is, Type (b) dependence described in Section 3.2.1). In particular, let $a_{12} = -0.5, -0.25, 0, 0.25, 0.5$ and $a_{21} = -0.3, -0.15, 0, 0.15, 0.3$.

Hence we have $5 \times 5 \times 5 = 125$ different scenarios for the triple (r, a_{12}, a_{21}). In each scenario, we simulate two-stream call arrivals according to model 3.2 for 300 days. We mainly want to see whether or how accounting for dependence between two streams help improve forecasting and staffing performance. Hence we consider the following two forecasting methods:

- MU1: fit Model 3.2 separately on each stream. That is, omit the dependence between the two customer arrival streams and consider each queue as independent. Then use the OLS estimation method.
- MU2: fit Model 3.2 on the two streams simultaneously and use the OLS estimation method.

Notice that we need the estimation mechanism to be computationally efficient since we have to perform the estimation process many times in the simulation study. Thus we use the OLS estimation method instead of the GLS estimation method.

3.4.2 Forecasting Comparison

In this section, we compare the forecasting performance between MU2 and MU1. In particular, we examine to what degree the multiple-stream method MU2 outperforms the single-stream method MU1 at each scenario.

With each method we perform the rolling forecast 200 times, in each rolling step using the past 100 days information to fit the model and forecast the count profile.

In each rolling step we record the performance measures such as RMSE and MRE for the count. Since we also know true arrival rate, we could also calculate and record the RMSE for the arrival rate at each rolling step. To compare the point forecast in each scenario, we denote the mean count-RMSE of the 200 rolling forecasts for MU2 as $\overline{\text{RMSE}}_{MU2}$ and similarly denote the mean count-RMSE of the 200 rolling steps for MU1 as $\overline{\ \text{RMSE}}_{MU1}$. We then calculate the relative mean count-RMSE reduction by MU2 from MU1, as

$$\Delta_{\overline{RMSE}} = \frac{\overline{\text{RMSE}}_{MU1} - \overline{\text{RMSE}}_{MU2}}{\overline{\text{RMSE}}_{MU1}}.$$

Similarly, we are also able to calculate relative mean rate-RMSE reduction by MU2 from MU1, which we denote as $\tilde{\Delta}_{\overline{RMSE}}$.

Figure 3.1 displays the count-RMSE reduction $\Delta_{\overline{RMSE}}$ for all the simulation scenarios. There are $5 \times 2 = 10$ plots in the figure. Each column corresponds to a customer type, which we refer to as Type A and Type B in the figure. Each row corresponds to a different value of r. In each plot, there are 25 lattices referring to different pairs of (a_{12}, a_{21}), where a_{12} varies in horizontal direction and a_{21} varies in vertical direction. Colors in lattices show which method performs better, where magenta indicates MU2 is better and cyan indicates MU1 is better. The color intensity indicates the magnitude of improvement on the other method. Figure 3.2 displays the rate-RMSE reduction $\tilde{\Delta}_{\overline{RMSE}}$ for all the simulation scenarios. The results are generalized in the following

- Larger value of a_{12} leads to larger improvement of multiple-stream method in forecasting Type A. Or, stronger dependence on Type B's past information leads to better point forecasts for Type A in multiple-stream method.
- Larger value of a_{21} leads to larger improvement of multiple-stream method in forecasting Type B. Or, stronger dependence on Type A's past information

leads to better point forecasts for Type B in multiple-stream method.

- There is no clear relationship between the magnitude/sign of r and the improvement of multiple-stream method in terms of the accuracy of point forecast(RMSE).
- The magnitude of improvement on arrival rate is noticeably larger than that on arrival counts.

Similar figures can be generated to display the relative improvement on MRE. They exhibit the same patterns as that for RMSE and we omit to show them in the paper.

3.4.3 Staffing Comparison

In this section, we compare the staffing effect between the multiple-stream method and single-stream method under different scenarios. To assess the staffing effect of any forecast distribution, we first input the forecast distribution to the staffing algorithm and generate a staffing vector. We then use the true counts to evaluate the staffing vector, and record the performance measure $v(.,.)$, $s(.,.)$ and $c(.)$ as described in section 3.3.3. Since it takes a while for the staffing program to generate the staffing vectors for one day and to evaluate them, we pick only 10 from the 125 scenarios and use the first 100 rolling forecasts in each scenario for the staffing test. In particular, we choose scenarios in the upper right corner and in the lower left corner for each r in figure 3.1.

To compare the staffing performance between MU2 and MU1, we calculate the daily staffing cost and daily shortage for each rolling experiment and we perform paired t test on those two measures. We also calculate the average violation probability for each method and the p-value for testing the proportion difference.

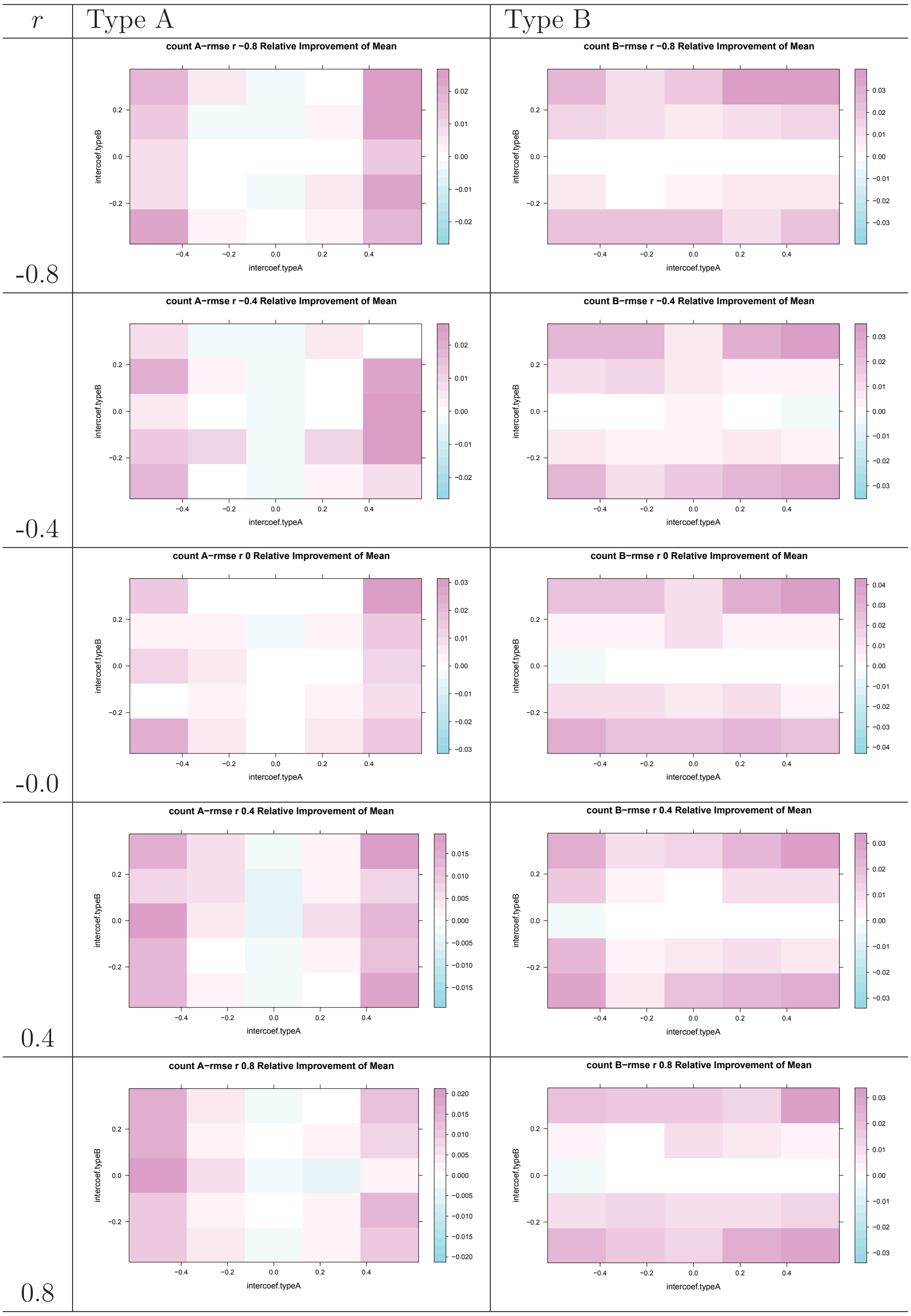
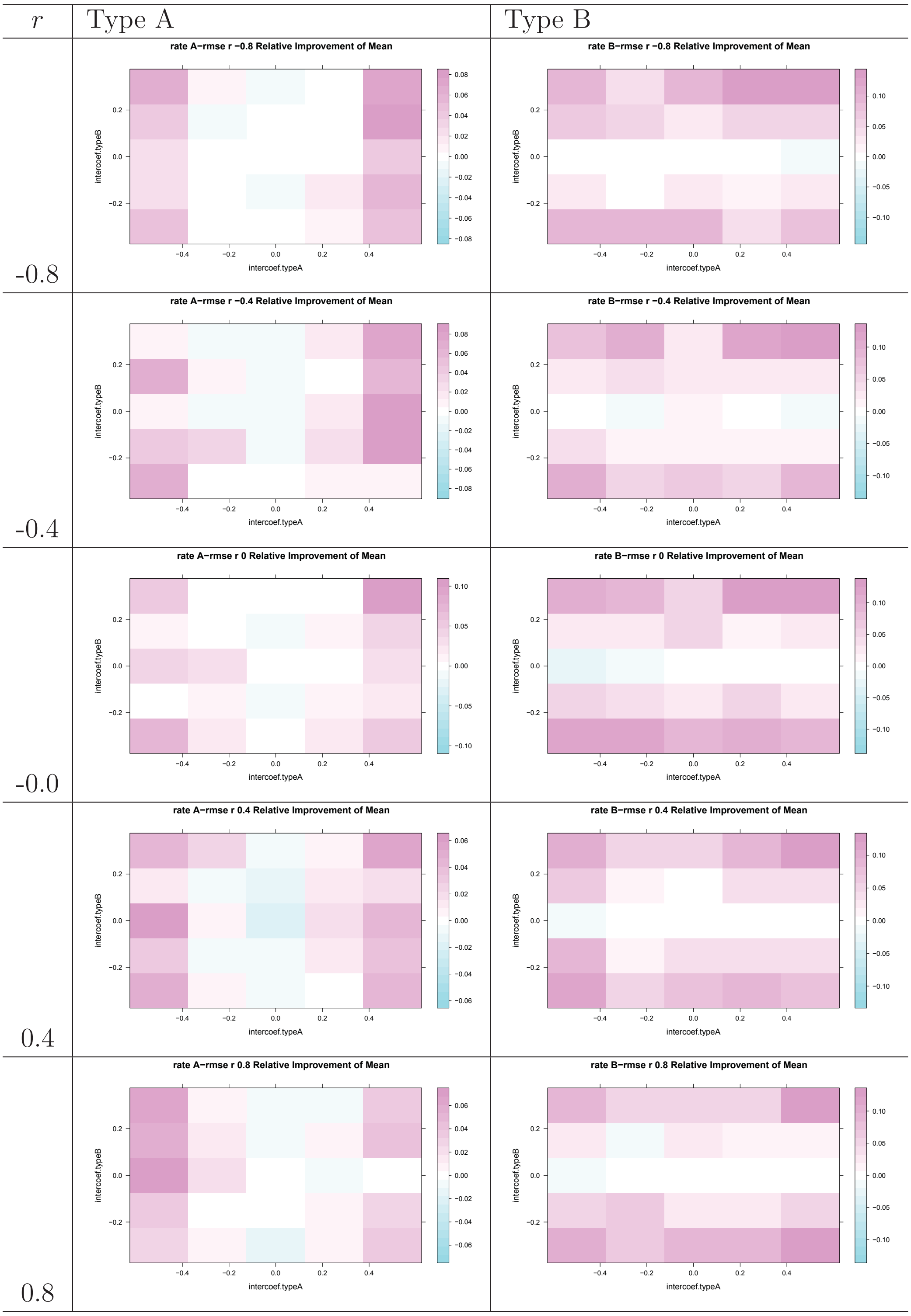


Figure 3.1: *Simulation comparison on the count RMSE between MU2 and MU1. Warm color indicates the superior of MU2 in forecasting accuracy of the counts, and cold color indicates the inferior of MU2 in forecasting accuracy of the counts.*



rate A−rmse r −0.8 Relative Improvement of Mean

rate B−rmse r −0.8 Relative Improvement of Mean

rate A−rmse r −0.4 Relative Improvement of Mean

rate B−rmse r −0.4 Relative Improvement of Mean

rate A−rmse r 0 Relative Improvement of Mean

rate B−rmse r 0 Relative Improvement of Mean

rate A−rmse r 0.4 Relative Improvement of Mean

rate B−rmse r 0.4 Relative Improvement of Mean

rate A−rmse r 0.8 Relative Improvement of Mean

rate B−rmse r 0.8 Relative Improvement of Mean

Figure 3.2: *Simulation comparison on the rate RMSE between MU2 and MU1. Warm color indicates the superior of MU2 in forecasting accuracy of the rates, and cold color indicates the inferior of MU2 in forecasting accuracy of the rates.*

Table 3.1 displays the daily staffing comparison between the two methods under each scenario. Our results confirm the conflict between lower staffing cost and better service quality, as the method with lower violation probability or shortages cost more and vice versa. When the two streams are positively correlated, that is, $r > 0$, the multiple-stream method MU2 is closer to meet the chance-constraint while MU1 is under-staffing. When the two streams are negatively correlated, the results are more complicated. When $r = -0.8$, MU1 is over-staffing and MU2 is under-staffing. When $r = -0.4$, MU1 is closer to the target violation probability $\delta = 0.05$. And when $r = 0$, the performance of MU2 and MU1 are statistically same and both of them are under-staffing.

r	a_{12}	a_{21}	Mean Daily Cost			Mean Daily Shortages			Vio. Prob.		
			MU2	MU1	p	MU2	MU1	p	MU2	MU1	p
(-0.8,	-0.5,	-0.3)	20463	21236	0	9.283	5.710	0.0005	0.0774	0.0353	0
(-0.8,	0.5,	0.3)	20761	21495	0	8.707	6.566	0.0017	0.0641	0.0338	0
(-0.4,	-0.5,	-0.3)	20690	21101	0	10.22	7.717	0.0584	0.0691	0.0506	0.0006
(-0.4,	0.5,	0.3)	20618	20906	0	11.64	10.43	0.2197	0.0788	0.0650	0.0137
(0,	-0.5,	-0.3)	21253	21278	0.5004	9.478	10.82	0.0752	0.0612	0.0641	0.3084
(0,	0.5,	0.3)	21656	21697	0.1356	10.37	10.83	0.4314	0.0621	0.0603	0.3807
(0.4,	-0.5,	-0.3)	20947	20722	0	11.72	15.56	0	0.0644	0.0876	0
(0.4,	0.5,	0.3)	21025	20832	0	9.106	10.69	0.0012	0.0559	0.0632	0.1001
(0.8,	-0.5,	-0.3)	21699	21250	0	11.03	18.63	0	0.0635	0.0994	0
(0.8,	0.5,	0.3)	21856	21358	0	13.48	22.05	0	0.0697	0.1076	0

Table 3.1: *Comparison of the staffing performance between MU1 and MU2.*

Figure 3.3 and 3.4 compares the interval results of the staffing experiment between the two methods. Each panel is a scatter plot of mean interval values of the two methods. The first column plots the mean cost, the second column plots the mean shortages and the third column plots the violation probability. Each row corresponds to one of the 10 scenarios we pick. Similar messages go with Table 3.1.

In the above experiments, both the violation probability and the daily cost are different between the two methods. Next we let the two methods have the same

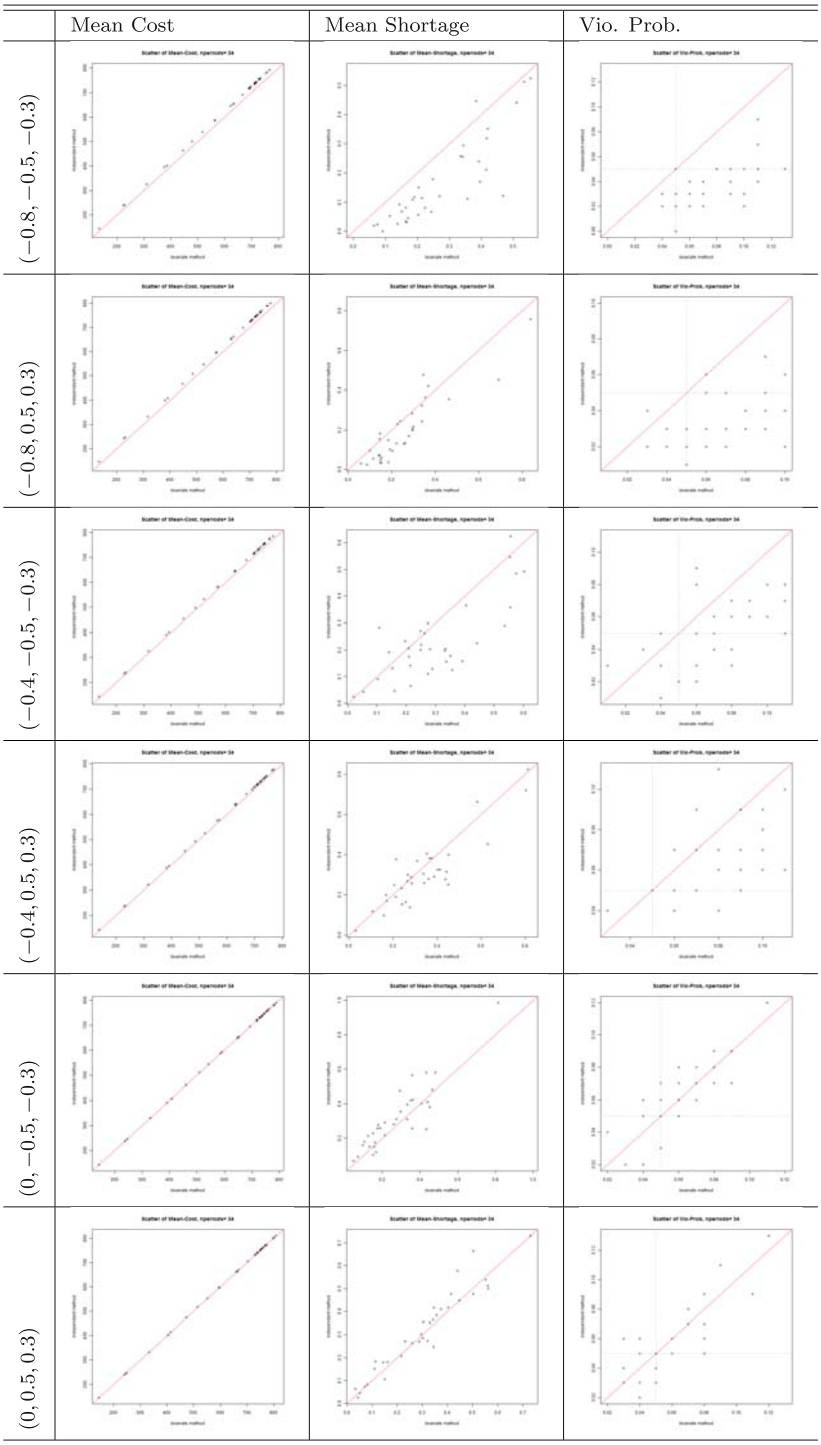


Figure 3.3: *Staffing performance comparison. Each panel is the scatter plot of mean interval values between MU1(vertical) and MU2(horizontal). Each row corresponds to a specific scenario, determined by* (r, a_{12}, a_{21}). *Each column corresponds to a performance measure (see column titles).*

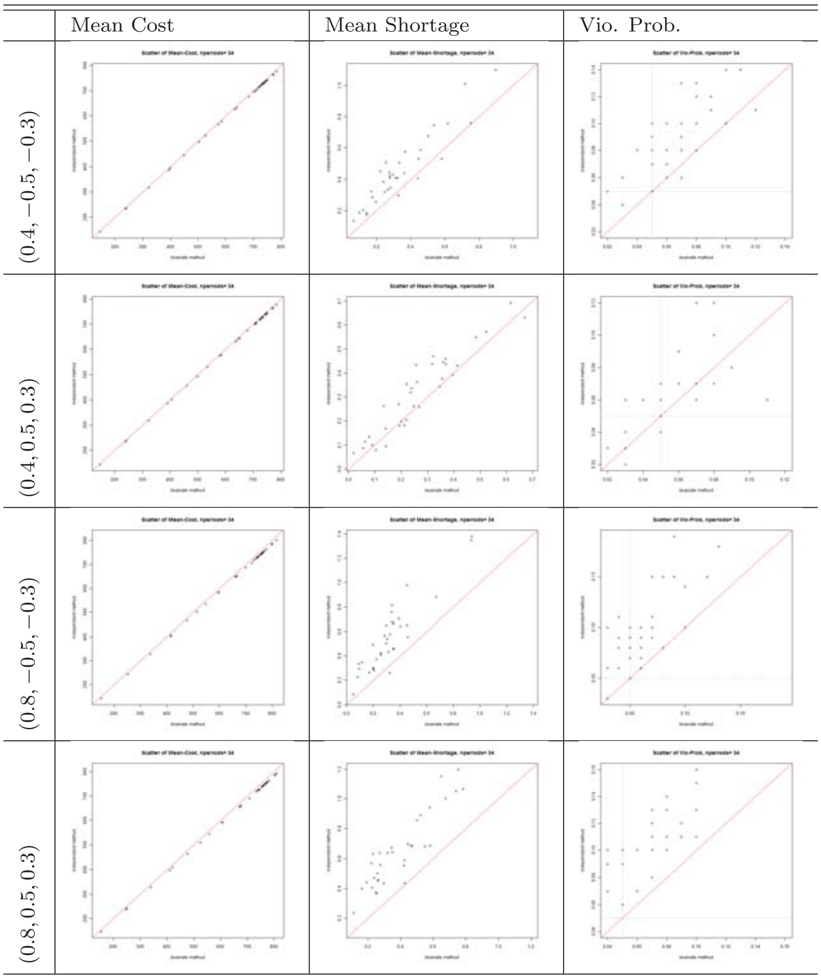


Figure 3.4: *Staffing performance comparison. Each panel is the scatter plot of mean interval values between MU1(vertical) and MU2(horizontal). Each row corresponds to a specific scenario, determined by (r, a_{12}, a_{21}). Each column corresponds to a performance measure (see column titles).*

violation probability and see whether multiple-stream method saves money in terms of staffing cost. We choose the scenario with strong negative correlation and past day dependence: $(r = -8, a_{12} = -0.5, a_{21} = -0.3)$. By giving proper parameters to the program, we get the staffing results from MU2 and MU1, whose violation probability is 0.0485 and 0.0488 respectively. To test the proportion difference we perform z-test. The p-value is 0.955 when testing the proportion difference between the two methods. The p-value is 0.694 and 0.753 for MU2 and MU1 respectively, when testing the proportion difference with the target value 0.05. We see that MU2 and MU1 statistically have the same violation probability.

We conduct paired t-test for the daily staffing cost and daily staffing shortages. Table 3.2 shows the t-tests results. We see that MU2 incurs less staffing cost and also less staffing shortages.

Figure 3.5 shows the paired t-test results of staffing cost between MU2 and MU1 for each time interval. We see clearly that for each time interval, MU2 is saving money.

Violation Probability				
MU2	MU1	z-test p-value		
$p_{(2)}$	$p_{(1)}$	$p_{(2)} \neq 0.05$	$p_{(1)} \neq 0.05$	$p_{(2)} \neq p_{(1)}$
0.0485	0.0488	0.694	0.753	0.955
Daily Staffing Cost				
Mean		Paired t-test		
MU2	MU1	p-value	lower	upper
20737	20964	0	-321.0	-132.1
Daily Service Quality Shortage				
Mean		Paired t-test		
MU2	MU1	p-value	lower	upper
5.607	7.720	0.007	-3.648	-0.578

Table 3.2: *Comparison between MU2 and MU1.*

3.5 Real Call Center Data

3.5.1 Background of the Data

Our data were collected at an Israel telecom call center. There are several service queues: Private customers, Business customers, Technical Support customers and some other minor queues. Among those Private customers and Business customers are the two main streams that take up 30% and 18% of the overall incoming calls,

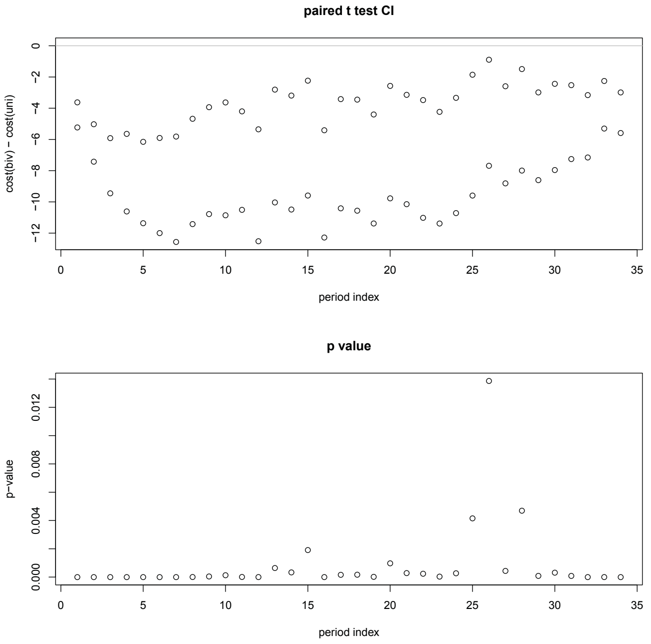


Figure 3.5: *Upper: paired t-test C.I. between MU2 and MU1 for each interval. Lower: paired t-test p-values.*

respectively.

In this section, we apply our multiple-stream forecasting and staffing method on this real call center data. And we focus our analysis on the two main streams Private and Business, and keep in mind that our method can be applied to more streams.

Our data range from 06/19/2004 to 04/14/2005, which contains 300 days. For both types of queues, the call center is open everyday and mainly operates from 7:00 am to midnight. Fridays and Saturdays have very low volume compared with the other weekdays, so we focus on the weekdays from Monday to Thursday, which includes 215 days. For each day we divide the 17 working hours into 34 half-hour time intervals, and record the count of the arriving calls during each time interval.

Figure 3.6 plots the call center data for Private customers on the transformed scale. The left panel displays the call volume profiles for each day. The middle shows average arrival volume for each day-of-week. And the right plots the daily-total arrival volumes along the days. Similar plots for the Business customers are displayed in Figure 3.7. The two arrival streams exhibit similar patterns in both the within day profiles and daily total volumes. Both of the two streams have two peaks in the with-day profile around 13:00 and 18:00 and Sunday has the highest volume compared with other day-of-week's. For daily totals, both of the streams have an increasing trend in the first 80 days and then go down till around 150 days and increase again. Hence we expect the two streams are dependent of each other.

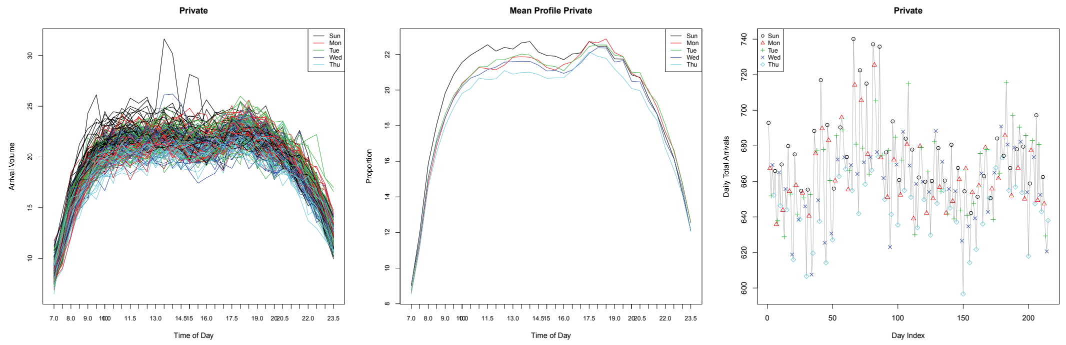


Figure 3.6: *Plot for Private customers. Left: arrival volume profiles on transformed scale. Middle: the mean arrival volumes for each day-of-week. Right: Daily total arrivals on transformed scale.*

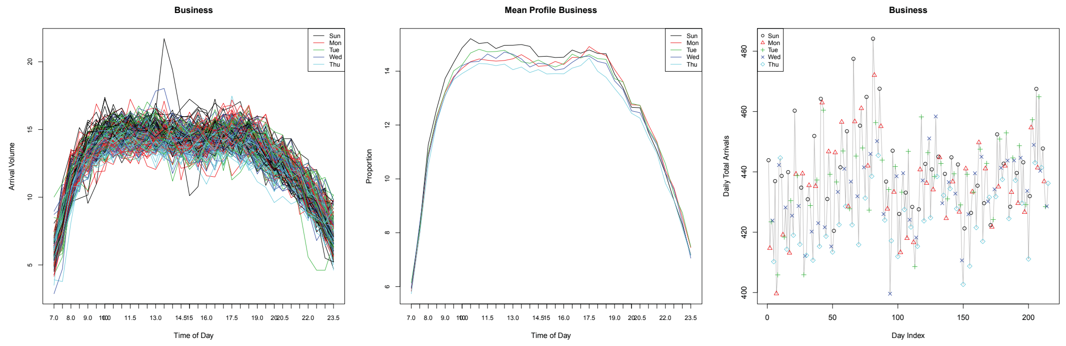


Figure 3.7: *Plot for Business customers. Left: arrival volume profiles on transformed scale. Middle: the mean arrival volumes for each day-of-week. Right: Daily total arrivals on transformed scale.*

Figure 3.8 displays the dependency between the two arrival streams. The left panel is the scatter plot of the daily total arrival volumes on the transformed scale with day-of-week effect removed. The correlation between the two time series is around 0.72 which is fairly strong. The right panel of Figure 3.8 is the scatter plot for the interval call volumes between the two customer types after we remove both the day-of-week effect and the interval effect. We also exclude 14 outliers to make the plot, and observe a moderately strong correlation of 0.38.

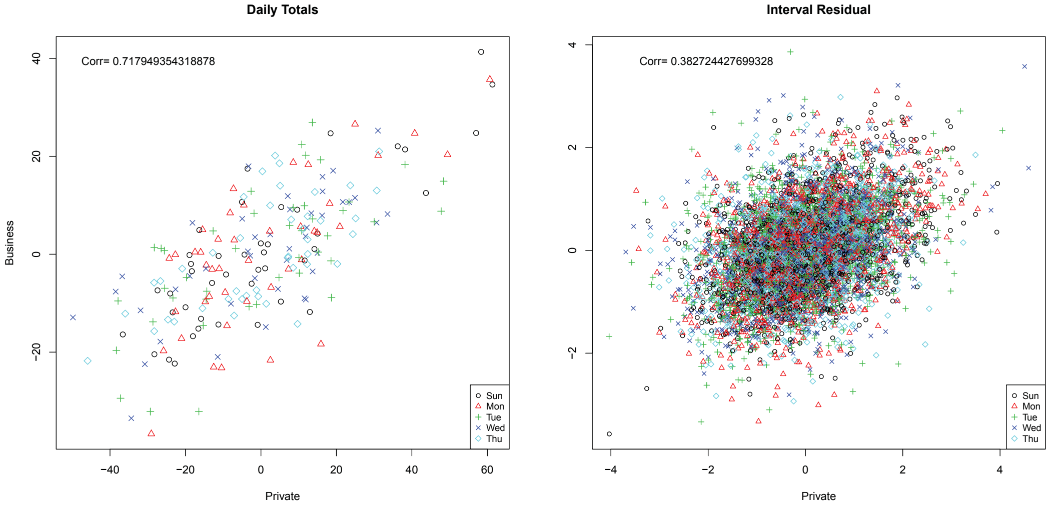


Figure 3.8: *Dependence between Private and Business customers. Left: scatter plot for daily totals on the transformed scale with day-of-week effect removed. Right: scatter plot of interval residual volumes on the transformed scale with day-of-week and interval effect removed.*

We refer to the Private queue as customer type 1 and the business queue as customer type 2, then fit Model 3.2 on our data where $I = 2$. Some estimates are given in Table 3.3. We see that both the daily-total and interval dependence between the two customer types is strong with the correlation around 0.67. The corresponding closest scenario in the simulation study is $(r = 0.8, a_{12} = 0, a_{21} = -0.15)$.

$\Sigma = (\Sigma_{sl})_{2\times 2}$			
Σ_{11}	Σ_{22}	Σ_{12}	correlation
0.8114	0.6273	0.4759	0.6671
$\Omega = (\Omega_{sl})_{2\times 2}$			
Ω_{11}	Ω_{22}	Ω_{12}	correlation
310.3	122.2	133.9	0.6875
$\mathbf{A} = (a_{sl})_{2\times 2}$			
a_{11}	a_{12}	a_{21}	a_{22}
0.6037	0.0922	-0.1018	0.4149

Table 3.3: *Some estimates of Model 3.2 on real data.*

3.5.2 Numerical Comparison

In this section we perform rolling forecast and staffing to compare the performance between different forecasting methods. Besides MU1 and MU2 as stated in section 3.4.1, we also consider the following two forecasting methods:

- HA: use the historical average of the same day-of-week as the forecast. Details are as follows.

$$X_{dt}^{(i)} = \bar{X}_{w_d t}^{(i)} + \epsilon_{dt}^{(i)}, \quad \epsilon_{dt}^{(i)} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2),$$

  where

$$\bar{X}_{w_d t}^{(i)} = \frac{1}{|\{d' : w_{d'} = w_d\}|} \sum_{d' : w_{d'} = w_d} X_{d't}^{(i)}.$$

- MU2G: fit Model 3.2 on the two streams simultaneously and use the GLS estimation method.

With each method, we use the first 100 continuous days to generate distributional forecast of the arrival profiles for day 101 and record the RMSE, MRE, COVER and WIDTH as described in section 3.2.5. Then we move our data window one day ahead, and repeat the forecasting process. The rolling experiment is performed 115 times, thus we have for each method 115 records of performance measures including RMSE, MRE, COVER and WIDTH.

Table 3.4 gives a summary of the four measures on point forecast from 115 rolling experiment for each method. We see that all the other forecasting methods beat the historical method (HA) in RMSE, MRE and WIDTH. And there is no strong evidence for us to select a forecasting method which gives the "best" point forecast since their performance varies on different measures and queues.

RMSE												
	Private Queue						Business Queue					
Method	Min.	Q1	Median	Mean	Q3	Max.	Min.	Q1	Median	Mean	Q3	Max.
HA	22.03	32.81	37.81	40.41	45.67	84.38	12.99	18.64	20.94	21.53	23.39	36.64
MU1	20.43	30.99	**36.80**	38.53	43.93	80.24	12.81	18.26	**20.07**	20.98	**22.14**	36.84
MU2	**20.35**	**30.43**	37.10	38.50	43.90	**79.99**	**12.76**	18.24	20.25	21.06	22.49	**36.37**
MU2G	20.33	30.45	37.14	**38.47**	**43.87**	80.05	**12.76**	**18.23**	20.11	**21.04**	22.48	36.42
MRE												
	Private Queue						Business Queue					
Method	Min.	Q1	Median	Mean	Q3	Max.	Min.	Q1	Median	Mean	Q3	Max.
HA	4.678	7.749	9.060	9.506	11.03	20.37	6.824	9.828	11.73	11.96	13.69	21.82
MU1	**4.424**	**7.313**	8.930	8.887	**9.987**	**13.96**	**6.378**	**9.602**	**11.52**	11.68	13.32	17.41
MU2	4.429	7.382	8.774	8.858	10.210	**13.96**	6.454	9.623	11.56	11.64	13.29	17.45
MU2G	4.425	7.365	**8.747**	**8.849**	10.170	13.97	6.453	9.624	11.53	**11.63**	**13.26**	**17.36**
Coverage Probability												
	Private Queue						Business Queue					
Method	Min.	Q1	Median	Mean	Q3	Max.	Min.	Q1	Median	Mean	Q3	Max.
HA	0.6765	0.9265	0.9706	0.9437	1	1	0.7353	0.9118	0.9706	0.9404	0.9706	1
MU1	0.7059	0.9265	0.9706	0.9435	1	1	0.7647	0.9118	0.9706	0.9425	0.9706	1
MU2	0.7059	0.9118	0.9706	0.9422	1	1	0.7647	0.9118	0.9706	0.9409	0.9706	1
MU2G	0.7059	0.9118	0.9706	0.9430	1	1	0.7647	0.9118	0.9706	0.9412	0.9706	1
95% Confidence Width												
	Private Queue						Business Queue					
Method	Min.	Q1	Median	Mean	Q3	Max.	Min.	Q1	Median	Mean	Q3	Max.
HA	135.8	143.4	162.3	160.7	174.9	188.8	77.05	81.80	85.27	84.79	87.62	92.97
MU1	**130.2**	**139.3**	150.3	152.0	165.1	178.2	76.68	**80.61**	83.09	83.02	85.27	91.24
MU2	**130.2**	**139.3**	**150.1**	151.7	**164.5**	178.1	76.59	80.62	**82.64**	**82.64**	84.43	90.81
MU2G	**130.2**	139.4	**150.1**	**151.6**	**164.5**	**177.9**	**76.54**	**80.64**	82.64	82.65	**84.41**	**90.80**

Table 3.4: *Comparison of 115 rolling forecasts of RMSE, MRE, Coverage Probability and Confidence Width.*

Next we conduct paired t-tests to compare the RMSE, MRE and WIDTH between any two of the methods. Table 3.5 shows the p-values of the paired t-tests. The entry in row r and column l is the p-value for paired t-test between the method of row r and the method of column l, with the alternative hypothesis being "the value of method in row r is less than the value of method in column l". We see that the other three methods are always better than HA. For most of the time, MU2G is better than MU2. The confidence width of bivariate methods MU2 and MU2G are shorter than that of univariate method MU1. There is no solid evidence to conclude that bivariate method is more accurate in providing point forecast than univariate method.

RMSE								
	Private Queue				Business Queue			
	HA	MU1	MU2	MU2G	HA	MU1	MU2	MU2G
HA		0.9978	0.9974	0.9976		0.9982	0.9906	0.9929
MU1	0.0022		0.5899	0.7173	0.0018		0.1534	0.2066
MU2	0.0026	0.4101		0.9893	0.0094	0.8466		0.9855
MU2G	0.0024	0.2827	0.0107		0.0071	0.7934	0.0145	

MRE								
	Private Queue				Business Queue			
	HA	MU1	MU2	MU2G	HA	MU1	MU2	MU2G
HA		0.9999	1.0000	1.0000		0.9952	0.9974	0.9980
MU1	0.0001		0.9142	0.9602	0.0048		0.8898	0.9391
MU2	0.0000	0.0858		0.9903	0.0026	0.1102		0.9640
MU2G	0.0000	0.0398	0.0097		0.0020	0.0609	0.0360	

WIDTH								
	Private Queue				Business Queue			
	HA	MU1	MU2	MU2G	HA	MU1	MU2	MU2G
HA		1.0000	1.0000	1.0000		1.0000	1.0000	1.0000
MU1	0.0000		1.0000	1.0000	0.0000		1.0000	1.0000
MU2	0.0000	0.0000		0.9999	0.0000	0.0000		0.0686
MU2G	0.0000	0.0000	0.0001		0.0000	0.0000	0.9314	

Table 3.5: *Paired t-test p-values.*

Then we consider the distributional forecast among the above methods. HA, MU1 and AD consider all queues as independent of each other and they provide independent distributional forecast. MU2 and MU2G account for the dependence between queues and provide a multivariate normal distribution as the forecast distribution for $X_{D+h,t}$.

Figure 3.9 displays the density of the forecasted correlation between $X^{(1)}_{D+h,t}$ and $X^{(2)}_{D+h,t}$ in the 115 rolling experiment, using method MU2. We see that the two arrival streams are strongly correlated so multiple-stream method provides a more accurate distributional forecast. The method MU2G provides similar results since its estimates are very close to those of MU2.

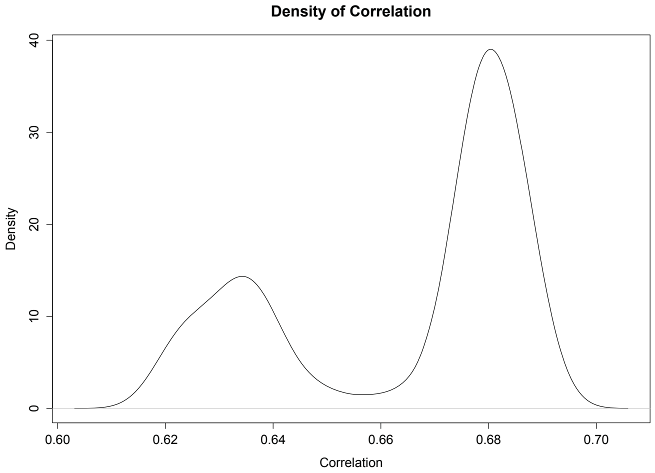


Figure 3.9: *Density plot of the forecasting count correlation between the two customer types.*

In the forecasting stage, the performance of multiple-stream methods and single stream methods are close in providing point forecasts, while the multiple-stream methods generate more accurate distributions. And as a result, the staffing policies differ with different input distributions.

Next we compare the consequential operational effects between single-stream method (MU1) and multiple-stream method (MU2) by implementing the chance-constraint staffing algorithm. We set the violation probability δ to be 0.05 in the

QoS constraint and use the set-ups described in section 3.3.4. At each rolling experiment, we give the staffing algorithm the forecast distribution of the counts. Then we produce a staffing vector and test how it works by recording the measures described in section 3.3.3 for each time interval. For each time interval, we have 115 records of $v(.,.)$, $s(.,.)$ and $c(.)$. Then we are able to compare the mean of cost $c(.)$, the mean of shortage $s(.,.)$ and the violation probability between multiple-stream method and single stream method for each time interval.

Figure 3.12 compares the staffing performance between single stream method (MU1) and multiple stream method (MU2). In each plot, a point corresponds to a time interval and the y-coordinate is the mean of the values of 100 rolling experiments. The left plot compares the mean staffing cost, where we see that MU2 results in higher cost. The middle plot compares the mean shortage, where we observe that the shortage of single stream method is always larger than that of multiple-stream method except for 3 intervals. And in the right panel we see that the violation probability of single-stream method is always larger than that of multiple-stream method. On average, the violation probability of MU1 is 0.0862 and the violation probability of MU2 is 0.0698 which is closer to the target value 0.05. The p-value of the difference between the two violation probabilities is 0.0069. Our results suggest that with multiple-stream forecasts we are more likely to meet the service quality constraint in staffing.

We also compare the daily staffing cost and daily shortages between MU1 and MU2 via paired t-test. Table 3.6 lists the mean values of daily staffing cost and mean values daily shortages, as well as p-values of the two-sided paired t-tests. We see that MU2 leads to less quality shortages but costs more.

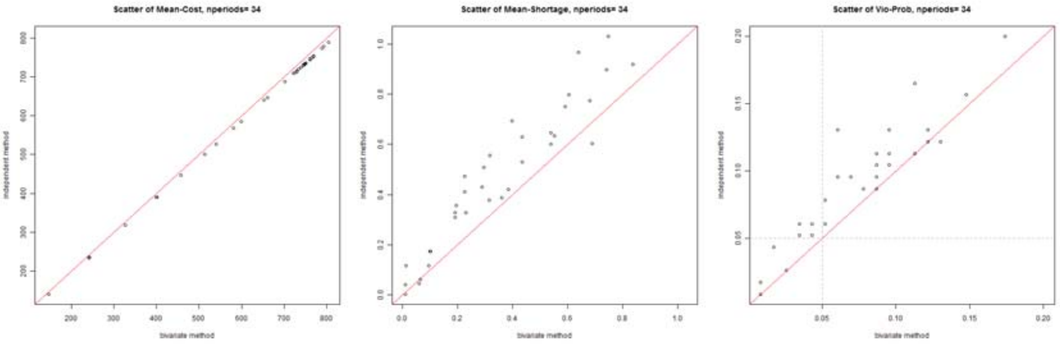


Figure 3.10: *Staffing performance comparison in rolling experiments. Left: Mean of cost for each time interval. Middle: mean of shortage for each time interval. Right: violation probability for each tme interval. Horizontal axis: multiple stream method. Vertical axis: single-stream method.*

Violation Probability				
MU2	MU1	z-test p-value		
$p_{(2)}$	$p_{(1)}$	$p_{(2)} \neq 0.05$	$p_{(1)} \neq 0.05$	$p_{(2)} \neq p_{(1)}$
0.0698	0.0862	0	0	0.0069
Daily Cost				
Mean		paired t-test		
MU2	MU1	p-value	lower	upper
21453	**21000**	0	427.6	478.4
Daily Shortage				
Mean		paired t-test		
MU2	MU1	p-value	lower	upper
12.10	16.08	0	-5.787	-2.173

Table 3.6: *Results of two-sided paired t-test on daily statistics between method MU2 and MU1.*

3.5.3 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 3.11, 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).

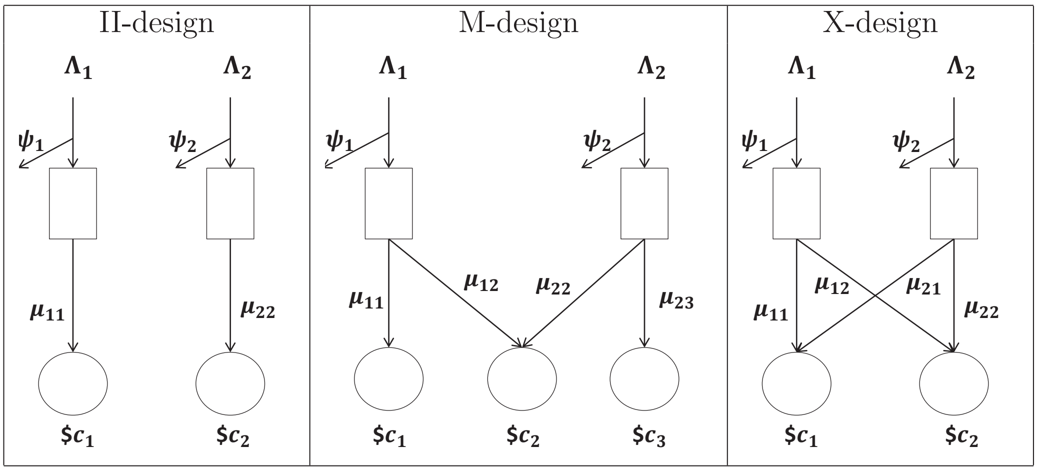


Figure 3.11: *Staffing designs.*

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 3.11, 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 3.11;
- service rate of the cross-trained servers when they handle their non-primary customer class, i.e. μ_{12} and μ_{21} of the X-design in Figure 3.11.

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 3.7, where the M-designs are arranged in the order of decreasing system flexibility. The violation probability target δ is 0.05, the allowed abandonment proportion ψ_i is 0.04 for both customer classes, and the service rate μ_{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.

Design	δ	ψ_i	μ_{ij}	c_j	Setting Index
M	0.05	$\psi_1 = \psi_2 = 0.04$	$\mu_{11} = \mu_{12} = \mu_{22} = \mu_{23} = 1$	$c_1 = c_3 = 1, c_2 = 1.1$	M1.1
				$c_1 = c_3 = 1, c_2 = 1.3$	M1.3
				$c_1 = c_3 = 1, c_2 = 1.5$	M1.5
				$c_1 = c_3 = 1, c_2 = 1.7$	M1.7
				$c_1 = c_3 = 1, c_2 = 2.0$	M2.0
X	0.05	$\psi_1 = \psi_2 = 0.04$	$\mu_{11} = \mu_{22} = 1, \mu_{12} = \mu_{21} = 0.8$	$c_1 = c_2 = 1$	X0.8
			$\mu_{11} = \mu_{22} = 1, \mu_{12} = \mu_{21} = 0.6$		X0.6
			$\mu_{11} = \mu_{22} = 1, \mu_{12} = \mu_{21} = 0.4$		X0.4
II	0.05	$\psi_1 = \psi_2 = 0.04$	$\mu_{11} = \mu_{22} = 1$	$c_1 = c_2 = 1$	II

Table 3.7: *System parameter settings for the II-design, M-design and X-design.*

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 3.12 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. MU1 and MU2 perform similarly regarding point forecast accuracy as shown in Table 3.4. 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 3.7, in the order of decreasing system flexibility. The values of δ, ψ_i, μ_{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).

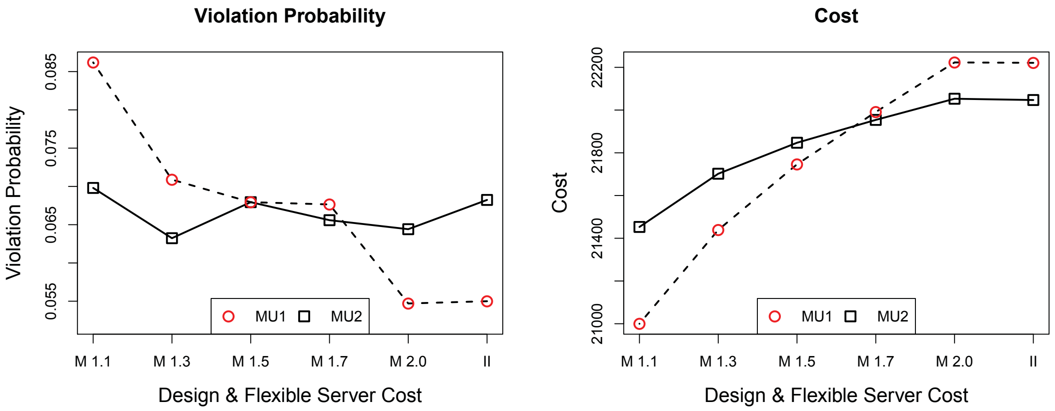


Figure 3.12: *Real data: daily staffing comparison among the II-design and various M-designs with different flexible server costs.*

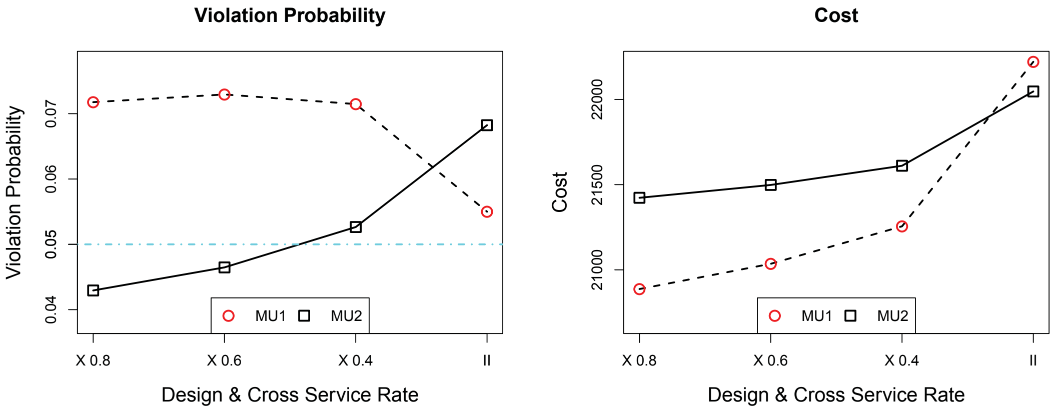


Figure 3.13: *Real data: daily staffing comparison among the II-design and various X-designs with different cross service rates.*

Figure 3.13 presents the daily staffing comparison between MU2 and MU1 under the X-design with varying cross-service rates. We can understand Figure 3.13 similarly as Figure 3.12: 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.

3.5.4 Comparison with Existing Bivariate Forecasting Models

To our best knowledge, Ibrahim and L'Ecuyer (2012) 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 3.6 and 3.7, we think a multiplicative structure is more realistic. A similar preference has been stated in Shen and Huang (2008) based on their numerical studies. Moreover, Ibrahim and L'Ecuyer (2012) 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 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 3.5.3, 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 3.8 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.

	M1.1 design	X0.8 design
MU2	0.086	**0.065**
BME2	0.083	0.084

Table 3.8: *Staffing comparison in violation probability between MU2 and BME2.*

Method	Min.	Q1	Median	Mean	Q3	Max.
MU2	**0.030**	**0.040**	**0.050**	**0.056**	**0.070**	**0.150**
BME1	205.6	590.0	725.6	805.5	933.1	2250
BME2	333.1	521.4	605.4	648.6	713.7	2306

Table 3.9: *Computation time comparison in seconds.*

Finally, we compare the computation time of the three methods. Table 3.9 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.

3.6 Discussion

3.6.1 Testing the Staffing Algorithm

In this section we test the staffing algorithm. We consider $I = 2$ and input true distribution to the algorithm and see how it performs under different strengths of inter-stream dependence. In particular, we set a bivariate Gaussian distribution for each time interval of one day as the truth and simulate with-in day count profiles for 100 times from that particular distribution. Hence we have 100 days of simulated within-day profile data as our realization. For each simulated day, we give the staffing algorithm the true bivariate distribution along with the simulated within-day profile as the realized counts to evaluate the staffing vector. In the meanwhile, we also input only the marginal distribution to examine how the algorithm performs when we deliberately omit the inter-stream dependence. Particularly, we use the following set up to determine the true distribution:

- We only consider one day, thus d is fixed here. The daily total rate $(u_d^{(1)}, u_d^{(2)})$, within-day proportion profile $f_{w_d,t}^{(1)}, f_{w_d,t}^{(2)}$, $t = 1, 2, \ldots, 34$ and the marginal variance $\Sigma_{11}, \Sigma_{22}, \Omega_{11}, \Omega_{22}$ are chosen based on real data estimates.
- Set $\delta = 0.05$, $\rho = 0$.
- Vary the inter-stream correlation r from -0.9 to 0.9 with a resolution 0.225

Table 3.10 shows the violation probabilities and corresponding p-values of the multi-stream input and independent input. We see how the violation probability diverge from the target value 0.05 if we omit the dependence among queues. We also see that the staffing algorithm performs more consistent when the correlation is larger. Figure 3.14 consents the above point, which plots the violation probabilities against the correlation for the true distributions. The correlation between the violation probability and bivariate correlation is -0.705 with p-value = 0.034.

r	v.prob		p-value		
	Multi-Stream	Single-Stream	z-test p-value		
	$p_{(2)}$	$p_{(1)}$	$p_{(2)} \neq 0.05$	$p_{(1)} \neq 0.05$	$p_{(2)} \neq p_{(1)}$
-0.900	0.0585	0.0256	0.0224	0	0
-0.675	0.0571	0.0294	0.0295	0	0
-0.450	0.0579	0.0424	0.0336	0.0408	0.0032
-0.225	0.0585	0.0494	0.0224	0.8749	0.0961
0	0.0576	0.0571	0.0407	0.0590	0.9169
0.225	0.0550	0.0612	0.0905	0.0014	0.1381
0.450	0.0582	0.0774	0.0275	0	0.0009
0.675	0.0556	0.0832	0.0674	0	0
0.900	0.0535	0.0929	0.3450	0	0

Table 3.10: *Violation probability and corresponding p-values for true multi-stream distribution input and independent single-stream input.*

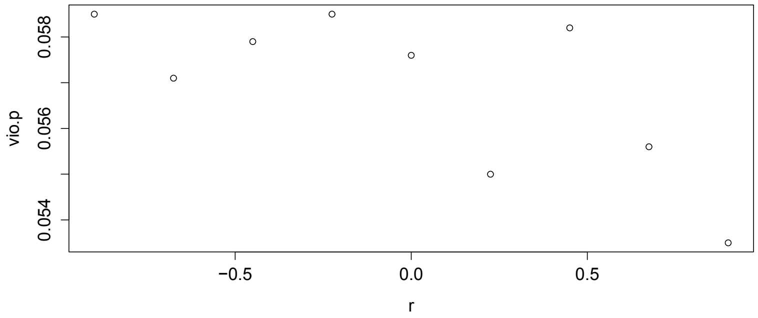


Figure 3.14: *Plot of violation probabilities against correlation for true distributions.*

3.6.2 Alternative Estimation Method

Instead of splitting the forecasting process into two stages, we also considered fitting a Gaussian mixed effect model to obtain the forecasts through an integrated procedure. Due to the multiplicative format, iterations are also required to achieve the final forecasts. Specifically, the steps below can be followed:

- Initiate $\hat{f}_{d,t}^{(i),(0)}$ by Equation 3.6.
- In the m^{th} iteration, fit linear mixed effect model

  $$\begin{cases} X_{d,t}^i = \gamma_d^i \, \hat{f}_{w_d,t}^{i,(m-1)} + \alpha_{w_d,t}^i \, \hat{f}_{w_d,t}^{i,(m-1)} + \epsilon_{d,t}^i \\ \epsilon_{dt} \equiv (\epsilon_{dt}^1, \epsilon_{dt}^2, \ldots, \epsilon_{dt}^I)^T \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Sigma) \\ \gamma_d = \mathbf{A}\gamma_{d-1} + \mathbf{z}_d, \quad \gamma_d \equiv (\gamma_d^1, \ldots, \gamma_d^I)^T, \quad \mathbf{z}_d \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \Omega). \end{cases}$$

  This model provides the estimates $\hat{\gamma}_d^{i,(m)}$, $\hat{\alpha}_{w_d,t}^{i,(m)}$, $\hat{\mathbf{A}}^{(m)}$, $\hat{\Sigma}^{(m)}$ and $\hat{\Omega}^{(m)}$ and thus the distributional forecast for $X_{D+h,t}^i$. Then update and normalize $\hat{f}_{w_d,t}^{i,(m)}$.
- Terminate the iteration until $\hat{f}_{w_d,t}^{i,(m)}$ converges in terms of m.

We implement both the additive and multiplicative mixed effect models in SAS, using the mixed procedure. However, solving the above linear mixed effect models is very computationally intense. We perform 100 times rolling test of our model with OLS estimation method and the mixed effect models on a real call center data set, using historical data of previous 50 days in each rolling step. We then record the RMSE, MRE, COVER and WIDTH for each rolling step. It takes minutes for the additive linear mixed effect model to perform one rolling test while our model with OLS estimation method requires less than a second. The multiplicative mixed effect model does not converge. We then perform paired t-test on the performance measures with the alternative hypothesis: the linear mixed effect model is performing better. Table 3.11 lists the test p-values. And we see that our model with OLS estimation method is no inferior to the linear mixed effect model.

We also tried the two-stage alternative model in Aldor-noiman et al. to save time but convergence problem arises. Thus we didn't adopt the mixed effect estimation

RMSE	MRE	WIDTH
0.9491	0.3207	0.1357

Table 3.11: *p-values of the paired t-tests. The alternative: additive linear mixed model is better.*

methods in our analysis due to its computational intensity.

3.7 Conclusion and Future Work

Our paper provided some insights on forecasting and staffing call centers with multiple uncertain arrival streams. We developed statistical models to forecast multiple stream arrivals, which is reliable in forecasting accuracy and computationally fast. We theoretically discussed the benefits of considering the dependence among multiple streams.

We implemented and tested the chance-constraint optimization staffing approach with sample based approximation by Gurvich et al. (2010). We showed there always was deviation between achieved service quality and objective service quality using this approach, and the deviation decreased as the correlation among different streams increased.

We combined our forecasting method and the chance-constraint staffing approach and formed an entire solution to forecast and staff call centers with multiple uncertain arrival streams, in the presence of dependence among streams. We compared our multiple-stream solution with an alternative single stream solution which ignored the inter-stream dependence. We tested both solutions on a real call center data set and showed accounting for dependence among streams provided more accurate forecast and the following staffing vector better met the quality of service target. Simulation experiments showed how benefits of the multiple stream solution varied by types and strength of dependence among streams. In particular, when two streams are positively correlated, both multiple stream and single stream solutions are under-staffing while multiple stream approach is closer to the QoS constraint. When two streams are negatively correlated, multiple stream solution saves money while providing the same service level.

A further research step is to develop more efficient and accurate estimation method in the forecasting process. Notice that the GLS estimation method is providing more accurate forecasts than the OLS estimation method in a very small degree, while costing much more time. Also notice that our forecasting procedure includes two stages. An undivided estimation process which produces estimates along with the forecasts at the same time, and computationally practicable is desired.

Besides exploring the benefits of incorporating dependence among multiple arriving streams, it might also be interesting to study the benefits of cross train agents in the multiple stream staffing context. We only tested the effects of considering dependence in the arrival process, under one staffing set-up. Further research includes exploiting the operational effects of cross training agents.

There are two other factors worth considering. One is to develop more scientific approximation method to discretize arrival distribution. We showed that sample-based approximation method in the staffing algorithm is not steady, especially when the value of correlation is small. Possible explanations could lie in the randomness of the discrete sample of the staffing algorithm, the chance fluctuation in the simulated realizations and the shape of the distribution. The Gaussian quadrature method whose samples match the original random variable for the first $2K-1$ moments might be useful for ruling out randomness. The other factor is to consider other staffing algorithms rather than the chance constraint approach to test the operational effects of incorporating inter-stream dependence.

4 Agent Heterogeneity

In this chapter we consider agent heterogeneity in terms of service efficiency and service quality. Traditional operations management papers often omit one or more of the following facts about agent heterogeneity:

- service efficiency and service quality vary by different agent,
- service efficiency and service quality vary by time for the same agent,
- service attributes (such as service time) might have different configurations.

Failing to consider agent heterogeneity might lead to improper anticipation of system performance, and further result in inefficient operational policies.

Some settings in recent operational programs can be suitable for taking into account agent heterogeneity. For example, the consideration of multiple agent pools allows agents to have multiple service efficiency/quality levels, and short-run staffing/scheduling algorithms allow agents' service attributes to change over time. However, these programs have no adequate input to be applied in practice, because there are rare papers providing methods for estimating or forecasting agent performance in the presence of agent heterogeneity. In this chapter, we are aiming to provide some methodologies to evaluate and forecast agent performance, and to address factors affecting agent performance.

The proxies considered for agent performance are agent service time (corresponding to service efficiency) and issue resolution probability (corresponding to service quality), respectively. In the first section we will conduct detailed learning curve analysis on agent service times and in the second section we will develop an issue resolution estimator and discuss some factors affecting issue resolution probability.

4.1 Heterogeneity in Service Times: Agent Learning Curve Modeling

Learning effects have been studied extensively in the past. See Yelle (1979) for a literature review on this topic. Recent developments include Bailey (1989), Nembhard and Uzumeri (2000), Shafer et al. (2001) and Nembhard and Osothsilp (2002), etc. However, this literature focuses on worker learning in the manufacturing/production context, while scarce literature has investigated agent learning in call centers. Our work supplements their work by providing a comprehensive study on the learning curves of a large group of call center agents. To the best of our knowledge, this work is the first to do so.

We view agents as accumulating experience on a day-to-day basis. In this section, we consider different learning-curve models and compare the performance of their in-sample estimates, as well as the accuracy of their out-of-sample predictions.

4.1.1 Four Learning-Curve Models

We assume that service times of an agent follow lognormal distributions. For an arbitrary agent, let y_{jk} denote the service time of the kth call during the jth day over this agent's tenure, and n_j be the total number of calls served by this agent during the jth day. Define $z_{jk} = \log(y_{jk})$.

We consider the following three parametric models and one nonparametric model to capture the agent learning effect.

Model 1

$$z_{jk} = a + b\log(j) + \epsilon_{jk}, \quad \epsilon_{jk} \sim N(0, \sigma_j^2);$$

Model 2

$$z_{jk} = a + b\log(\frac{j}{j+\gamma}) + \epsilon_{jk}, \quad \epsilon_{jk} \sim N(0, \sigma_j^2);$$

Model 3

$$z_{jk} = a + b\frac{j}{j+\gamma} + \epsilon_{jk}, \quad \epsilon_{jk} \sim N(0, \sigma_j^2);$$

Model 4

$$z_{jk} = f(j) + \epsilon_{jk}, \quad f(\cdot) \text{ is a smooth function, } \epsilon_{jk} \sim N(0, \sigma_j^2).$$

The first model adapts the learning curve equation in Yelle (1979). Model 2 and Model 3 have a common interesting feature: the mean log service time approaches some limit as the training period approaches infinity. This feature conforms to the conjecture that the service rate eventually stabilizes and will not improve further after the agent has taken calls for a sufficiently long period of time. Unlike the first three models, Model 4 is nonparametric and assumes the least amount of structure on the underlying learning curve.

Below, we briefly discuss how we estimate the parameters for each model, using our data. To keep the models parsimonious, we assume homoscedasticity: $\sigma_j^2 = \sigma^2$. For the nonparametric Model 4, we estimate the smooth function $f(\cdot)$ using the smoothing spline technique Green and Silverman (1994), implemented via the function *smooth.spline* in the R package. For the three parametric models, we estimate the model parameters using maximum likelihood.

In particular, we illustrate the estimation procedure using Model 3, as it involves some constraints. Let $z = \{z_{jk}\}$ and $\theta = (a, b, \gamma, \sigma)$. Then the likelihood function

$L(\theta|z)$ can be expressed as

$$L(\theta|z) = \prod_{j,k} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z_{jk}-a-bj/(j+\gamma))^2}{2\sigma^2}}.$$

Thus, the log likelihood is

$$\log(L(\theta|z)) = \sum_{j,k} \left\{ -\log(\sqrt{2\pi}) - \log(\sigma) - \frac{(z_{jk} - a - b\frac{j}{j+\gamma})^2}{2\sigma^2} \right\}, \tag{4.1}$$

which is to be maximized with respect to (a, b, γ, σ). Note that one must have $\gamma > -1$, since $j \geq 1$. Equivalently, we solve the following optimization problem:

$$\min \sum_{j,k} \left\{ \log(\sigma) + \frac{(z_{jk} - a - b\frac{j}{j+\gamma})^2}{2\sigma^2} \right\} + (-\gamma - \delta)^+ M. \tag{4.2}$$

In the last term of (4.2), the constant δ is a number smaller than but very close to 1, say 0.999, and M is a very large number, say 10^8. This term applies a very large penalty to the optimization objective if $\gamma \leq -1$ and does not affect the objective otherwise, so it ensures that $\gamma > -1$. The criterion is then optimized numerically. Model 2 can be fitted similarly. The last term in (4.2) is not needed for Model 1.

4.1.2 Learning Patterns of Agents

We fit the four learning models to a group of 129 agents, whose records suggest that they are common agents with no previous work experience in the call center. Remarkably, we find that there exists a variety of learning patterns among these agents. They exhibit mainly the following three patterns: (1) always learn, (2) never learn, and (3) learning and forgetting interwoven throughout the whole tenure. The majority of the agents possess the third pattern. For ease of presentation, we name these three learning patterns: the **optimistic** case, the **pessimistic** case and the **common** case, respectively. To illustrate these behaviors, we select two agents from each case and display their estimated learning curves below.

The Optimistic Case: agents always learn

Figure 4.1 plots the learning curves for two agents: Agent 33156 and Agent 33235, respectively. On both panels of the figure, the x-axis is the duration of the agent's working experience (in days), and the y-axis displays the mean log service time for that day. The dots are the average log service times calculated from the data. More precisely, for day j on the x-axis, the value of the corresponding dot on the y-axis is calculated as

$$\bar{z}_{j\cdot} = \frac{1}{n_j} \sum_{k=1}^{n_j} z_{jk}.$$

The four curves show the estimates for the mean log service time, given by the four models that we considered. See the legend within each panel for a detailed description.

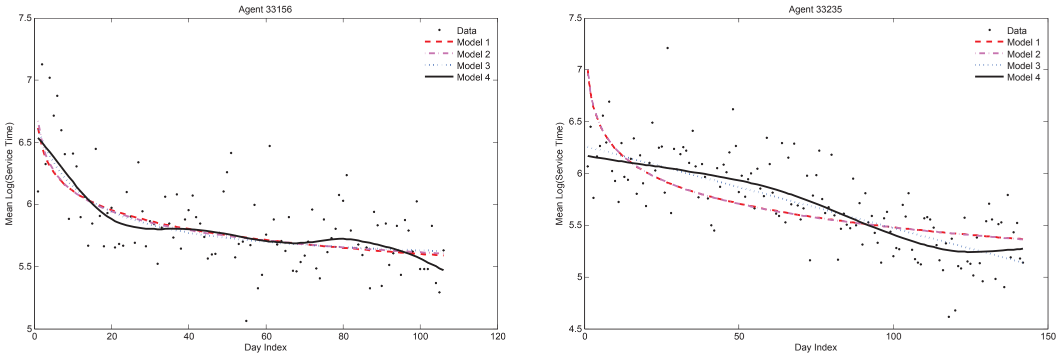


Figure 4.1: *Learning curves for the optimistic case*

From Figure 4.1, we deduce that, throughout these two agents' tenures, their mean log service times are decreasing. This implies that they are always learning and are getting faster on their jobs as they work longer. However, their learning rates seem to be decreasing, and the learning curve becomes flatter, which suggests that the purely log-log linear learning curve (Model 1) is too simple to capture the underlying behavior.

The Pessimistic Case: agents never learn

Not all agents learn during their working period, and in Figure 4.2, we show two agents who never learn. As we see, Agent 74527 is getting slower as he works longer, and Agent 76859 seems to maintain a stable service rate throughout her tenure.

From the plot of Agent 74527, one also observes that the nonparametric spline model is more sensitive to the short-term trend of the service rate. In particular, this agent's mean service time has a significant leap around day 130, which is captured nicely by the nonparametric model. The three parametric models are too rigid to fit such a dramatic change.

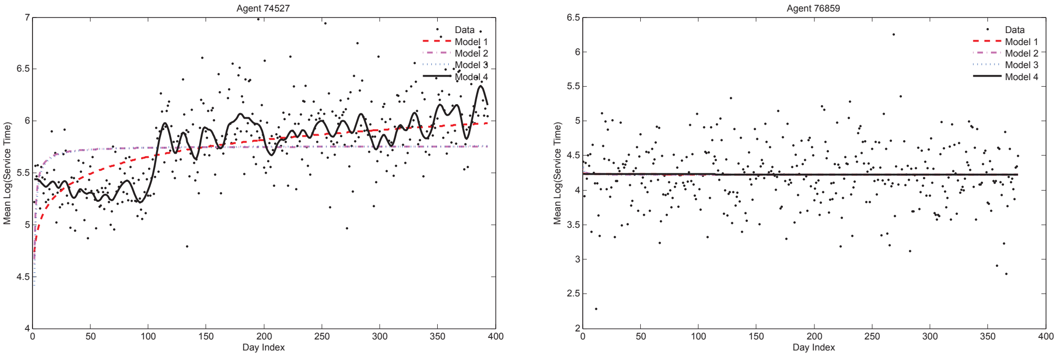


Figure 4.2: *Learning curves for the pessimistic case*

The Common Case: agents may learn as well as forget

We observe that most agents do not have a monotone learning curve: their log mean service times are "zigzagging" throughout their working period; for such a behavior, the nonparametric model captures much better the trend of the mean log service time. Figure 4.3 depicts the learning curves of two such agents.

Agent 33146's mean log service time is decreasing during his first 100 working days and afterwards has two significant leaps. The first leap starts at around day 110 and reaches a peak at around day 150. After that, the log service time starts to decrease. The second jump begins at about day 220 and arrives at the apex at about

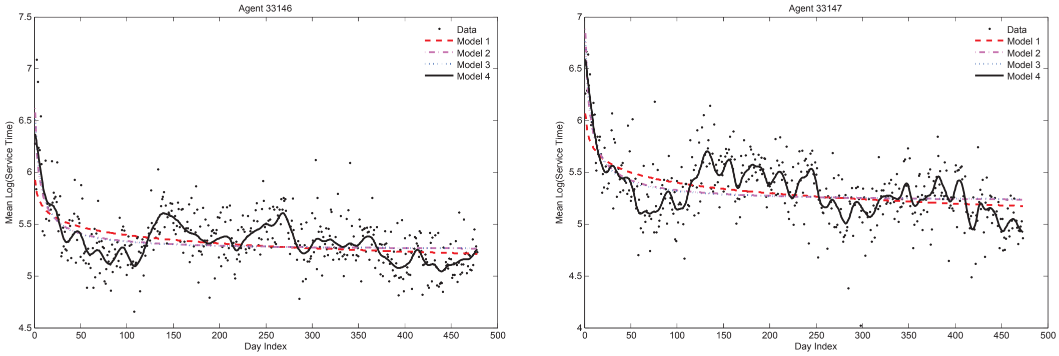


Figure 4.3: *Learning curves for the common case*

day 270, after which the log service time keeps decreasing until the end.

Agent 33147's learning curve is similar. As we see in the right panel of Figure 4.3, her mean log service time first drops, then starts to jump at around day 100, reaches its peak around day 130, then begins to decrease slowly until day 240, makes a sharp drop between day 240 and day 280, and finally seems to stabilize, from day 280 onwards.

Based on the above results, we conclude that agents' learning curves can differ significantly. However, we note that the above analysis uses only the service times of the calls; other factors may explain the leaps and bumps observed in the learning curves; however, we do not have access to them.

4.1.3 Out-of-sample Prediction of Service Rate

As existing simulation results suggest, managers need statistical models that can sensitively monitor the service rates of individual agents; otherwise, the call center may end up being overstaffed or understaffed. The analysis in Section 4.1.2 compares the in-sample performance of four learning models. In addition, we calculate below the out-of-sample prediction errors of the service rate, using the four models.

We incorporate a rolling-window prediction procedure. For a given agent, the schematic algorithm of the out-of-sample prediction exercise is as follows:

Algorithm for Computing Prediction Errors

For $j = 6$ to n, with n being the length of the agent's tenure (in days):

- Fit Model i using data from the 1st day to the $(j-1)$th day, for $i = 1, 2, 3, 4$;
- Predict the mean log service time of the agent on the jth day, denoted as $\hat{z}_{ij}$, using the fitted learning model;
- Predict the mean service rate of the agent on the jth day as $\hat{\mu}_{ij} = e^{-(\hat{z}_{ij}+\hat{\sigma}_i^2/2)}$ where $\hat{\sigma}_i$ is the estimated standard deviation for the measurement error in Model i;
- Estimate μ_j, the "true" mean service rate of the agent on the jth day, calculating it as the reciprocal of the mean service time of the calls answered on that day;
- Calculate the prediction errors (PE) and relative prediction errors (RPE) of the service rates on the jth day as

$$PE_{ij} = \hat{\mu}_{ij} - \mu_j, \tag{4.3}$$

and

$$RPE_{ij} = \frac{\hat{\mu}_{ij} - \mu_j}{\mu_j} \times 100\%. \tag{4.4}$$

End For

After reviewing the out-of-sample prediction performance of the models, we conclude that the nonparametric learning model is more sensitive and effective in monitoring the changes in agent service rate, no matter what the agent's learning pattern. Hence, among the four approaches tested, the nonparametric model is the most robust. To illustrate this observation, below we plot in Figures 4.4 and 4.5 the prediction errors for Agent 33146 and 33147. As shown in Section 4.1.2, these two agents have the most intricate learning patterns among the six agents plotted there.

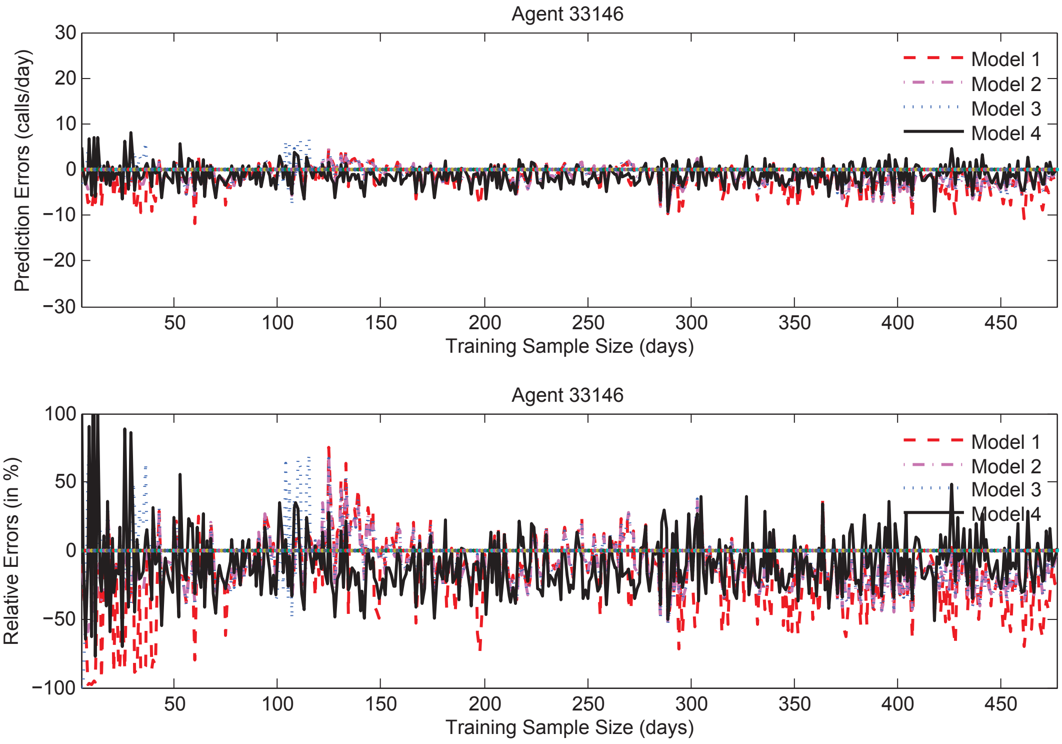


Figure 4.4: *Prediction errors for Agent 33146*

In each panel, the x-axis is the number of historical working days used to fit the learning-curve models for the agent; the y-axis shows either the prediction error or the relative prediction error. The curves for the four models are plotted using different colors, as indicated in the panel legend. From these plots, we observe that, over the full tenure of both agents, the performance of the nonparametric spline model is the best and the most stable. In particular, from Figure 4.3, we observe

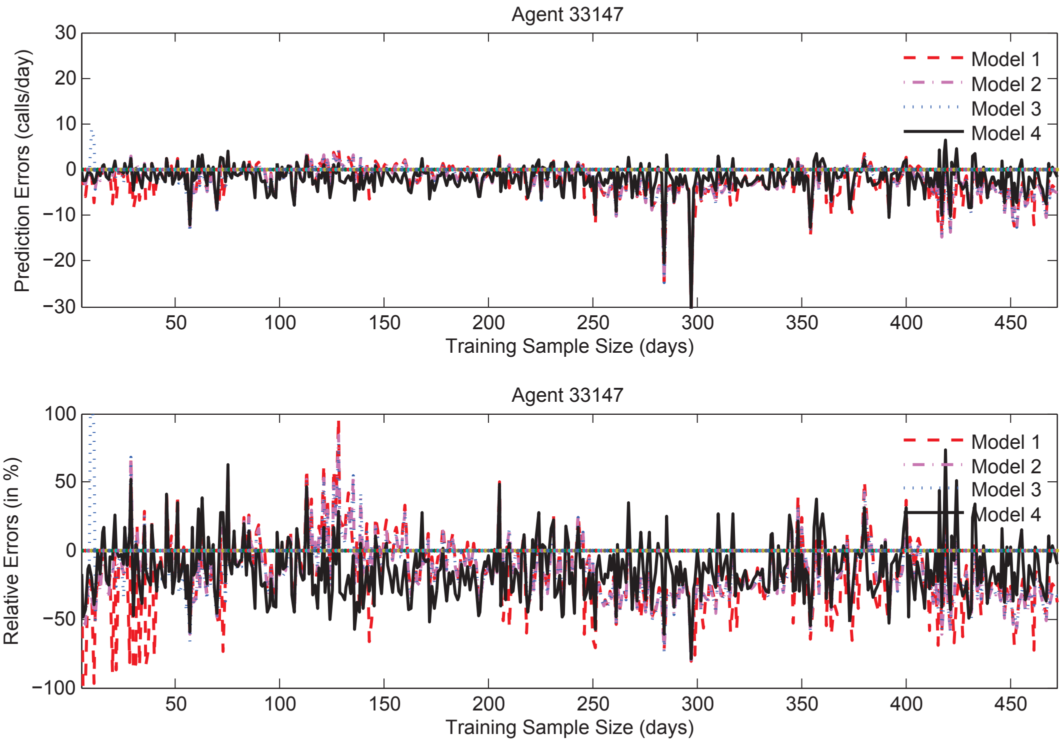


Figure 4.5: *Prediction errors for Agent 33147*

that the mean log service time of Agent 33146 has a clear jump from day 110 to day 150. Correspondingly, in Figure 4.4, we see that the prediction errors of using the spline model are much closer to zero during this time period. Similar observations can be made for the periods between day 220 and day 270 for Agent 33146, and between day 100 to day 130 for Agent 33147. In addition, the spline model is also more sensitive to drops in mean log service time (i.e., service rate jumps), during the period between day 240 to day 280 for Agent 33147. These observations imply that the nonparametric model is the most sensitive one in capturing agent service-rate changes.

4.2 Heterogeneity in Service Quality: Issue Resolution Rate Analysis

This section is based on empirical analysis of nine million call by call data at a U.S. telephone service call center. Selected results are shown in the following section.

4.2.1 Exploratory Data Analysis

Exploratory analysis of the nine million call-by-call data reveals the following unusual phenomenons. Before going to the results, we define a term:

Definition 1 *Agent-release Behavior: The interaction between the customer and the call center is terminated by agent hanging up the call.*

Many empirical studies suggest talk-time distribution follows log-normal distribution (Brown et al. (2005)). However, the talk-time distribution of agent-released calls exhibits bi-modal pattern (Figure 4.6), indicating many calls are released by agents at around 30 seconds, which is unusual.

Another unusual finding comes from the within-day agent-released call profile (Figure 4.7). Commonly the count profile of the same day-of-the-week follows the same pattern, as shown in the left plot in Figure 4.7. However, there are several spikes in the count profile of agent-released calls (middle plot in Figure 4.7), indicating agents hang up more calls than usual. The right plot in Figure 4.7 is the proportion profile of agent-released calls. We also see several curves stands out with unusual high agent release rate. We are interested in identifying those spikes(abnormal agent behavior) and finding the factors triggering them.

The above results show unusual agent release behavior in terms of talk-time distribution and spikes. In next section we will demonstrate the impact of agent release

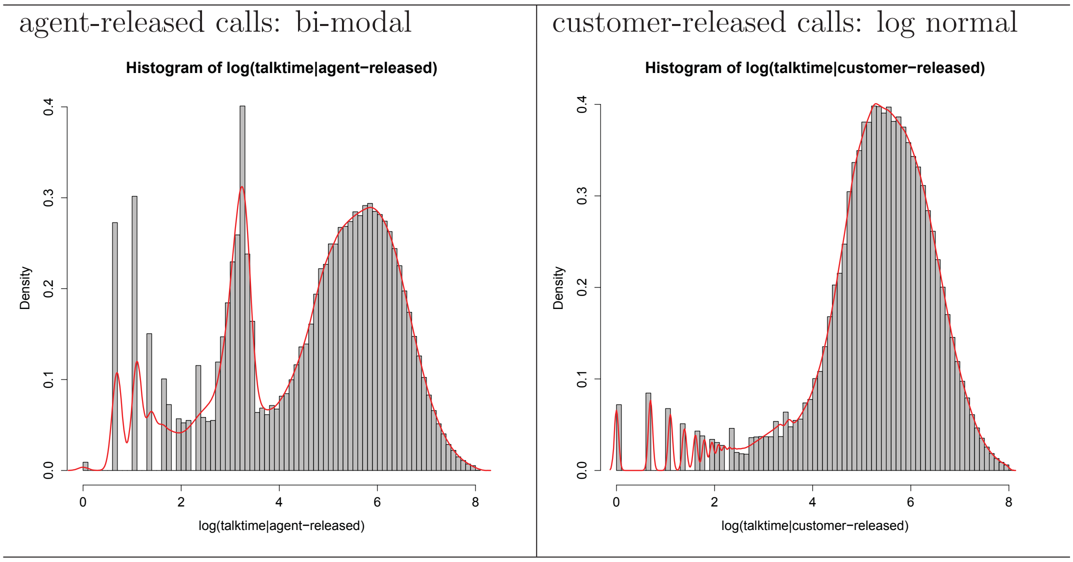


Figure 4.6: *Talk-time distribution. Left: agent-released calls (bimodal pattern is unusual). Right: customer-released calls (expected pattern).*

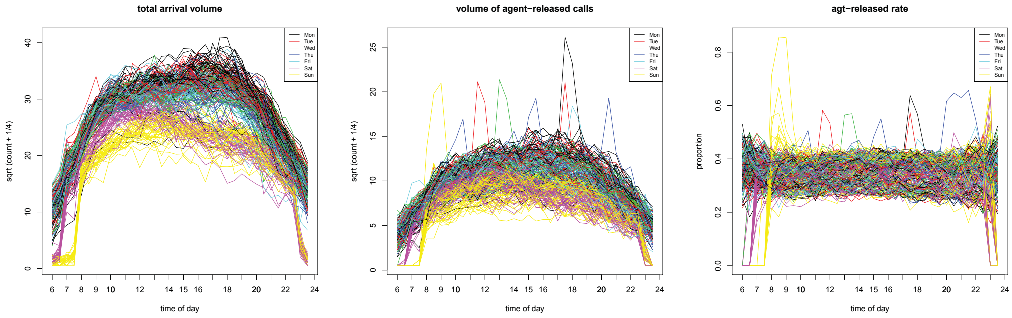


Figure 4.7: *Left: within-day profile of total arrivals. Middle: within-day profile of agent-released calls. Right: within-day profile of agent-released rate.*

behavior: lower issue resolution rate.

4.2.2 Issue Resolution Estimation

After noticing the abnormal agent release behavior, one may ask: are the problems really solved when agents hang up calls? To answer this question, we look into Issue Resolution Rate(IR).

Definition 2 *Issue Resolution Rate (IR): the probability that the customer's problem is solved by the end of the call.*

IR is an important call center performance measure. Traditionally it's obtained by conducting customer survey which is expensive (additional agent effort) and not reliable(non-response bias). In this section we derive an IR estimator from operational data which is free and reliable and we compare the IR between agent-released calls and customer released calls. We find the IR of agent-released calls is at least 10.72 percent point less than that of customer-released calls, which informs agent-release behavior will negatively affect service quality.

Intuitively, the one-time-service customers would never call in again after their problems are solved. So the IR should be correlated with the one-time call proportion, which can be extracted from the operational data. We construct a graphic model which reflects the whole service around a call to build up the dependency (Figure 4.8). Based on our model, the one-time call proportion is decomposed in the form

$$\theta = q \cdot p_1 \cdot p_2 + (a-1) \cdot q \cdot p_1^2 + p_1 - p_1 \cdot p_2.$$

Then the IR is

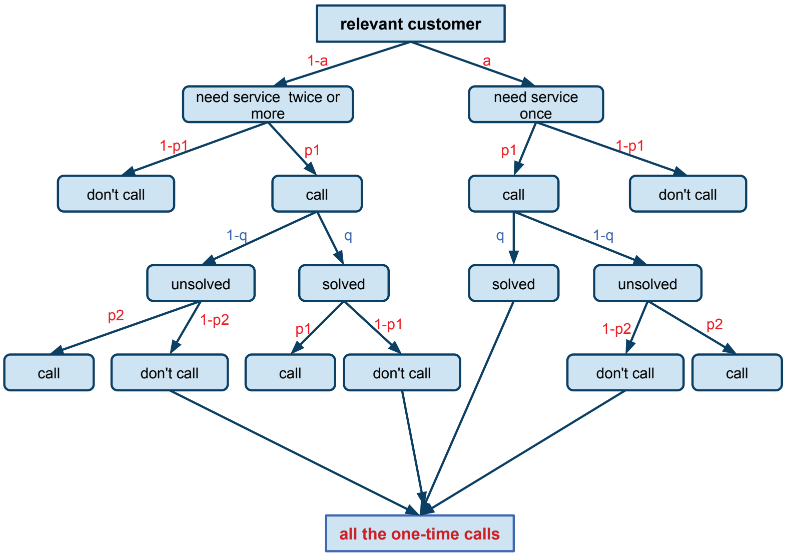
$$q = \frac{\theta - p_1 + p_1 p_2}{p_1 p_2 - (1-a) p_1^2}.$$

Next we compare IR between agent-released calls and customer released calls. We assume $p_1 = p_2 = p$, meaning customers consistently care about their problem, then

$$q = \frac{\theta - p + p^2}{a p^2}.$$

Noticing that a and p are independent of agent, we have

$$q|_{c-released} - q|_{a-released} = \frac{\theta|_{c-released} - \theta|_{a-released}}{a \cdot p^2}.$$



a probability(a relevant customer needs one-time service). Independent of agent.

p_1 probability(a customer will call in given he needs to solve a new problem). Independent of agent.

p_2 probability(a customer will call in given he needs to solve an existing&unfinished problem). Independent of agent.

q issue resolution rate, i.e. probability(the problem will be solved by the end of the call). Depends on whether the call is released by agent.

θ one-time call rate, i.e. proportion of the calls made by one-time callers.

Figure 4.8: *Graphic network model for estimating issue resolution rate.*

Rewrite the above as:

$$\delta_q = \frac{\delta_\theta}{a \cdot p^2}.$$

We obtain following estimates from the data:

$$\hat{\delta}_\theta = 3.96\%, \quad S.D.(\hat{\delta}_\theta) = 0.0577\%$$

$$\hat{a} = 36.93\%, \quad S.D.(\hat{a}) = 0.0219\%$$

Then

$$\hat{\delta}_q = \frac{\hat{\delta}_\theta}{\hat{a} \cdot p^2} = \frac{10.72\%}{p^2}.$$

Applying Delta method, we get

$$S.D.(\hat{\delta}_q) \approx \frac{\sqrt{\frac{\hat{\sigma}_1^2}{\hat{a}^2} + \frac{\hat{\delta}_\theta^2 \hat{\sigma}_2^2}{\hat{a}^4} - \frac{2\hat{\delta}_\theta \hat{\sigma}_1 \hat{\sigma}_2}{\hat{a}^3} \cdot \rho}}{p^2} = \frac{\sqrt{2.445 \cdot 10^{-6} - 1.987 \cdot 10^{-7} \rho}}{p^2} \leqslant \frac{0.1626\%}{p^2}$$

where $\rho \doteq \text{Correlation}(\hat{\delta}_\theta, \hat{a}) \in [-1, 1]$.

Thus we come to the conclusion that the issue resolution rate of agent-released calls is at least 10.72 percent points less than that of customer-released calls and the estimate is statistically accurate.

We then look at agent heterogeneity in issue resolution. We use one-time call proportion as an auxiliary estimator to rank agents' issue resolution probabilities and compare our estimates with survey-based estimates.

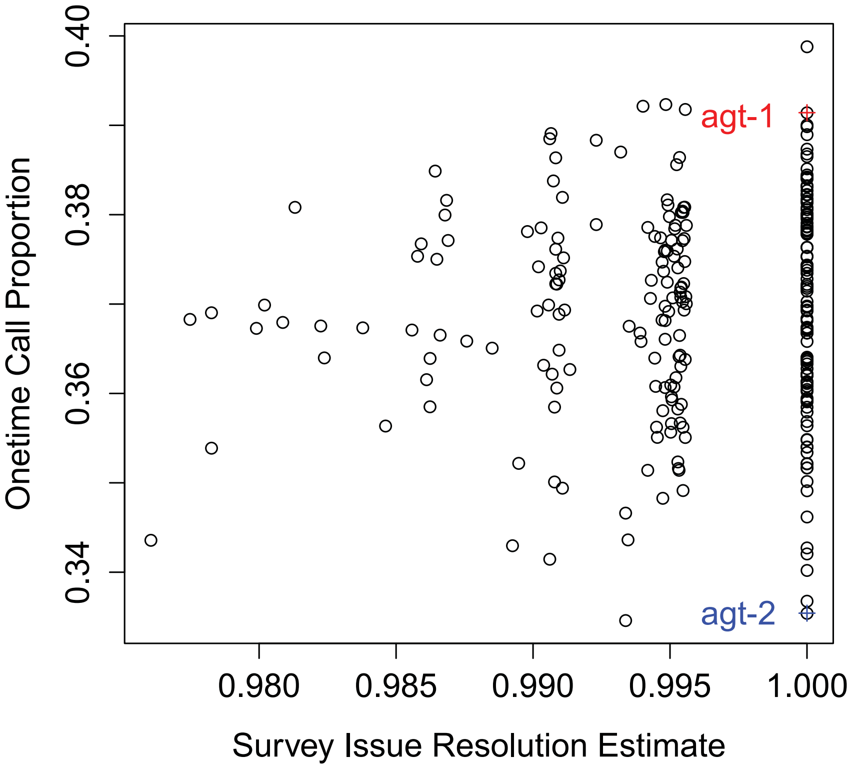


Figure 4.9: *Comparison between data-driven and survey-driven estimates.*

Figure 4.9 plots our data-driven estimates against the survey driven estimates. As some agents have only handled a small number of calls, we only keep those agents who have handled more than 5000 calls to insure that our estimates are accurate. Totally there are 262 agents in the figure. Also keep in mind that our data-driven estimates are relative IRP's to compare between different agents. From the right graph we see that our data-driven method provides different IRP estimates even if the survey-estimates are all the same.

Take two agents for example: Agent-1 and Agent-2 that both have survey-IRP 1. Based on our data-driven estimates, Agent-1 has a higher IRP compared to Agent-2 while the customer survey indicates that both of them have solved all their problems. We carefully examine all the calls they have handled and find that Agent-2 answered 7507 calls from Split 700 totally, among which 5444 calls are lastly handled by him/her(not transferred to other agents). Agent-1 answered 5684 calls from Split 700, among which 3983 are lastly handled by hime/her(not transferred to other agents). For those calls lastly handled by them, 24.50% of them call back within an hour for Agent-2, and 18.56% call back within an hour for Agent-1, which results are consistent with our data-driven IRP estimates. As it's very unlikely that all the within-1hour callbacks are for a different problem to be solved and Agent-1 and Agent-2 have different 1hour callback rate, the survey estimated IRP 1 for both of them can be misleading.

4.2.3 Abnormality Detection for Agent Release Burst

As we have found out that agent release behavior leads to less issue resolution, unusually frequent agent release behavior would bring down the system performance in both service quality (customers' problems are not well solved) and operational efficiency (customers with unsolved problems will call in again requesting additional resource). In this section we are to identify those unusual high agent-release rate (spikes) in Figure 4.7 and explore factors triggering those spikes.

Hubert et al. (2005) introduce a Robust Principle Component Analysis (R-PCA) method which is able to deal with outlying observations and they also provide methods to classify the outliers. As the spike curve is one type of outliers, we borrow strength from R-PCA to develop our spike detection method.

Our method follows four steps (Figure 4.10):

- Step 1: use the first robust principal component to de-trend the data.
- Step 2: calculate the robust standard deviation of de-trended data.

$$\sigma_j = \frac{\text{Median}(\{|x_{i,j}|\}_{i=1}^{D})}{0.6745} \tag{4.5}$$

- Step 3: smooth the standard deviation $\{\sigma_j\}_{j=1}^{J}$ using local polynomial regression

$$\hat{\sigma}_j = f(j) \tag{4.6}$$

- Step 4: use the threshold value to mark the spikes

$$\text{Thr} = 4\hat{\sigma}_j$$

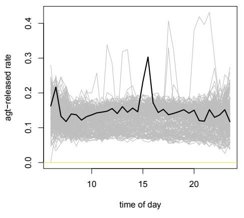
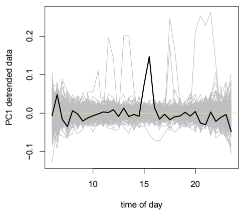
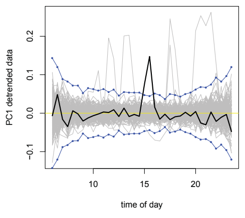
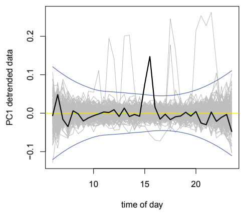
raw data	de-trended data	threshold by Equation 4.5	threshold by Equation 4.6
			

Figure 4.10: *Illustration of our spike detection method.*

We compare our method to other methods (not necessarily designed for spike detection since to the best of our knowledge there is no existing spike detection method that handles functional data with background trend). As the spike is one type of outlier, we may use the outlier identifying method in Hubert et al. (2005) as the comparable methods to identify the spike. We consider the following methods:

- meth1: our method described above from Step 1 to Step 4
- meth2: same as meth1 instead of skipping Step 3 and using the threshold in Equation 4.5
- rpca1: the outlier identifying method provided by Hubert et al. (2005) using one principal component.
- rpca2: the outlier identifying method provided by Hubert et al. (2005) using two principal components.

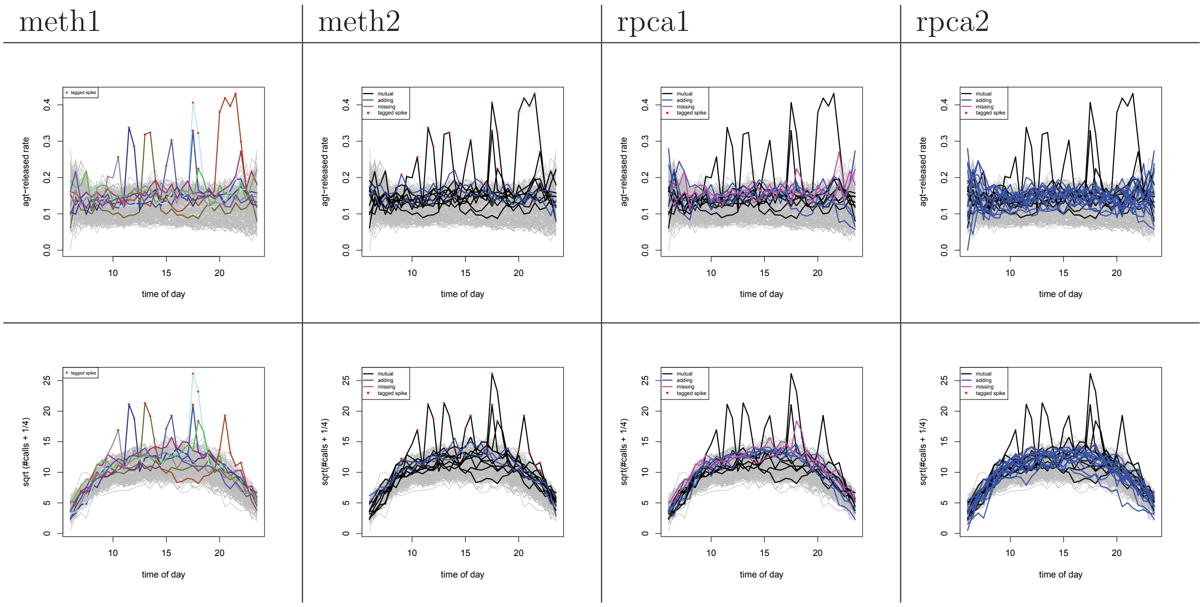


Figure 4.11: *Methods comparison.*

We apply the above four methods on the call center agent-released rate data. Figure 4.11 demonstrates the comparison results. The highlighted curves are the identified outliers by the corresponding method. Red points are the detected spikes.

We use METH1 as benchmark to compare with the other 3 methods. The left 2 plots in Figure 4.11 show the spikes detected by METH1. The next 3 column of plots are the comparison between METH1 and the other 3 methods.

The black curves are the mutual findings from METH1 and the other method. The pink curves are the outliers detected by METH1 but not detected by the other method. The blue curves are the outliers detected by the other method but not detected by METH1. We see that RPCA1 detects several non-spike curves and fails to detect two spike curves compared with METH1 which indicates it does not work well for the functional data. And for RPCA method, the more RPC's used, the more non-spike curves detected. Based on the comparison plots, we see that METH1 is the most robust and accurate to identify the spikes.

With the spike curves identified we are able to study the factors triggering the spikes. We dichotomize the spike curves identified by METH1 and use it as response variable. Then we consider the exploratory variables listed in Table 4.1 and fit logistic regression model. After variable selection by likelihood ratio test, the remaining variables are listed in Table 4.2.

variable Name	variable Description
RAV	relative arrival volume/workload (the centered data by removing the median arrival profile volume)
DIFF	the difference of RAV between the current time interval and its previous time interval
DIFF.PREV	the difference of RAV between previous time interval and one more time interval ahead
SRAV	the sign of RAV (whether the workload is above average)
SRAV.PREV	the sign of the RAV of previous time interval
INT	the index of the time interval
DIFF*SRAV.PREV	the interaction term of DIFF and SRAV.PREV

Table 4.1: *Factors considered for the agent-release rate spikes*

The chi=square test $1 - pchisq(103.99, 300) \approx 1$ indicates that our model is

Coefficients	Estimate	Std.Error	P-value
Intercept	-3.6663	0.3911	<2e-16
RAV	0.3311	0.1303	0.01103
DIFF	0.4978	0.1722	0.00384
DIFF.PREV	0.4277	0.1543	0.00558
DIFF:I(SRAV.PREV==1)	-0.5925	0.2538	0.01958
Null deviance: 136.92 on 305 degrees of freedom			
Residual deviance: 103.99 on 301 degrees of freedom			

Table 4.2: *Significant variables after variable selection.*

statistically plausible and $1 - pchisq(136.92 - 103.99, 305 - 301) \approx 0$ indicates that our model is significantly better than the NULL model. Figure 4.12 displays the ROC plot of the final logistic model. The area under the curve is 0.812, which demonstrates a good fit.

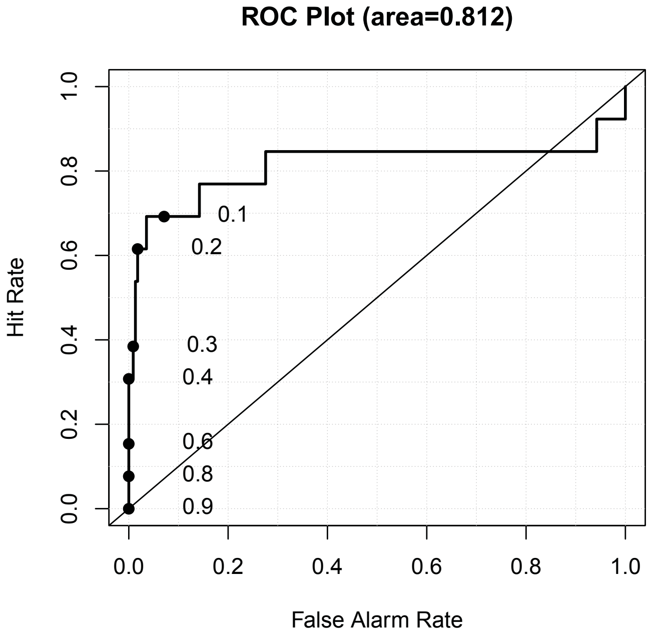


Figure 4.12: *ROC plot for the logistic regression model.*

The regression model implies the following factors correlate with the occurrences of a spike:

- the relative workload of current time interval
- the increment of relative workload of the current time interval.
- the increment of relative workload of previous time interval.
- the increment of relative workload of the current time interval given that the workload of previous time interval is below average.

In a word, our findings reveal that the agents are sensitive to the increasing workload.

BIBLIOGRAPHY

Aksin, 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. D. Feigin, A. Mandelbaum. 2010. Workload forcasting for a call center: Methodology and a case study. *Annals of Applied Statistics* .

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

Bailey, C.D. 1989. Forgetting and the learning curve: a laboratory study. *Management Science* **35**(3) 340–352.

Bassamboo, A., J.M. Harrison, A. Zeevi. 2006a. Design and control of a large call center: Asymptotic analysis of an lp-based method. *Operations Research* **54**(3) 419–435.

Bassamboo, A., J.M. Harrison, A. Zeevi. 2006b. Dynamic routing and admission control in high volume service systems: Asymptotic analysis via multi-scale fluid limits. *Queueing Systems* **52**(3-4) 249–285.

Bassamboo, A., A. Zeevi. 2009. On a data-driven method for staffing large call centers. *Operations Research* **57**(3) 714–726.

Bertsimas, D., X.V. Doan. 2010. Robust and data-driven approaches to call centers. *European Journal of Operational Research* **207**(2) 1072–1085.

Birge, J.R., F. Louveaux. 1997. *Introduction to stochastic programming*. Springer Verlag.

Brown, L. D., 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.

de Vericourt, Francis, Yong-Pin Zhou. 2005. Managine response time in a call-routing problem with service failure.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management* **5**(2) 79–141.

Gans, Noah, Haipeng Shen, Yong-Pin Zhou. 2012. Parametric stochastic programming models for call-center workforce scheduling .

Garnett, O., A. Mandelbaum, M. Reiman. 2002. Designing a call center with impatient customers. *Manufacturing & Service Operations Management* **4**(3) 208–227.

Green, P.J., BW Silverman. 1994. *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman & Hall/CRC.

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 & Service Operations Management* **7**(1) 20–36.

Hubert, M., P.J. Rousseeuw, K.V. Branden. 2005. Robpca: A new approach to robust principle component analysis. *Technometrics* **47**(1).

Ibrahim, Rouba, Pierre L'Ecuyer. 2012. Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models .

Liao, S., G. Koole, C. Van Delft, O. Jouini. 2012. Staffing a call center with uncertain non-stationary arrival rate and flexibility. *OR spectrum* 1–31.

Mehrotra, V., O. Ozlük, R. Saltzman. 2010. Intelligent procedures for intra-day updating of call center agent schedules. *Production and Operations Management* **19**(3) 353–367.

Nembhard, D.A., N. Osothsilp. 2002. Task complexity effects on between-individual learning/forgetting variability. *International Journal of Industrial Ergonomics* **29**(5) 297–306.

Nembhard, D.A., M.V. Uzumeri. 2000. Experiential learning and forgetting for manual and cognitive tasks. *International journal of industrial ergonomics* **25**(4) 315–326.

Perry, O., W. Whitt. 2009. Responding to unexpected overloads in large-scale service systems. *Management Science* **55**(8) 1353–1367.

Robbins, T. R., D. J. Medeiros, T. J. Harrison. 2010. Cross training in call centers with uncertain arrivals and global service level agreements. *International Journal of Operations and Quantitative Management* **16** 307–329.

Robbins, T.R., T.P. Harrison. 2010. A stochastic programming model for scheduling call centers with global service level agreements. *European Journal of Operational Research* **207**(3) 1608–1619.

Robinson, S.M. 1977. A characterization of stability of linear programming.

Romisch, W. 2003. Stability of stochastic programming problems. *Handbooks in OR and MS* **10**.

Shafer, S.M., D.A. Nembhard, M.V. Uzumeri. 2001. The effects of worker learning, forgetting, and heterogeneity on assembly line productivity. *Management Science* **47**(12) 1639–1653.

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

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 Statis-*

tical Association forthcoming.

Williams, A. C. 1963. Marginal values in linear proramming.

Yelle, L.E. 1979. The learning curve: Historical review and comprehensive survey. *Decision Sciences* **10**(2) 302–328.