A Stochastic Programming Model for Scheduling Call Centers with Global Service Level Agreements

Working Paper

Thomas R. Robbins • Terry P. Harrison
Department of Supply Chain and Information Systems , Smeal College of Business, Pennsylvania State University, University Park, PA

We consider the issue of call center scheduling in an environment where arrivals rates are highly variable, aggregate volumes are uncertain, and the call center is subject to a global service level constraint. This paper is motivated by work with a provider of outsourced technical support service where call volumes exhibit significant day of week and time of day variability, and are also subject to random shocks; external events that create a significant spike in call volume. The outsourcing contract specifies a service level agreement that must be satisfied over an extended period of a week or month. We formulate the problem as a mixed integer stochastic program. Our model has two distinctive features. Firstly, we combine the server sizing and staff scheduling into a single optimization program. Secondly, we explicitly recognize the uncertainty in period-by-period arrival rates. We show that the stochastic formulation in general calculates a higher cost optimal schedule than a model which ignores variability, but that the expected cost of this schedule is lower. We conduct extensive experimentation to compare the solution of the stochastic program with the deterministic program based on mean valued arrivals. We find that in general the stochastic model provides a significant reduction in the expected cost of operation. The stochastic model also allows the manager to make informed risk management decisions by evaluating the probability that the Service Level Agreement will be achieved.

Working Paper

1 Introduction

A call center is a facility designed to support the delivery of some interactive service via telephone communications; typically an office space with multiple workstations manned by agents who place and receive calls (Gans, Koole *et al.* 2003). Call centers are a large and growing component of the U.S. and world economy; by 2008 the United States will employ an estimated 2.1 million call center agents (Aksin, Armony *et al.* 2007). Large scale call centers are technically and managerially sophisticated operations and have been the subject of substantial academic research. Call center applications include telemarketing, customer service, help desk support, and emergency dispatch.

Staffing is a critical issue in call center management as direct labor costs often account for 60-80% of the total operating budget of a call center (Aksin, Armony *et al.* 2007). This paper addresses the scheduling problem in a call center with highly variable and uncertain arrival rates. The work is directly related to a research project with a provider of outsourced technical support delivered via globally distributed call centers. The company provides both tier 1 (help desk) and tier 2 (desk-side) support. The bulk of their business, and the focus of this research, is on the inbound call center operation. This operation involves providing help desk support to large corporate and government entities. While the scope of services varies from account to account, many accounts are 24 x 7 support and virtually all accounts are subject to some form of Service Level Agreement (SLA). There are multiple types of SLA, but the most common specifies a minimum level of the Telephone Service Factor (TSF). A TSF SLA specifies the proportion of calls that must be answered within a specified time. For example, an 80/120 SLA specifies that 80% of calls must be answered within 120 seconds. A very important point is that the service level applies to an extended period, typically a week or month. The SLA does not define requirements for a day or an hour. Therefore, the desk is often staffed so that at some times the service level is underachieved, sometimes overachieved, and is on target for the entire month.

The key challenge involved with staffing this call center is meeting a fixed SLA with a variable and uncertain arrival rate pattern. The number of calls presented in any ½ hour period is highly variable with multiple sources of uncertainty. A seasonal pattern exists with volumes declining from Monday to Friday. In addition to day of week seasonality these call centers also experience very significant time of day seasonality. Volume tends to dip down around the lunch break, but a second peak occurs in the

Working Paper

afternoon; though the afternoon peak is typically lower volume then the morning peak. While this basic arrival pattern exists on most business days, there is significant stochastic variability in the call pattern from day to day. The following graph shows call volume over an eight week period for a particular project. The inner region represents the minimum volume presented in each period, while the overall envelope is the maximum volume presented in each period. The outer region then represents the variability over this eight week period.

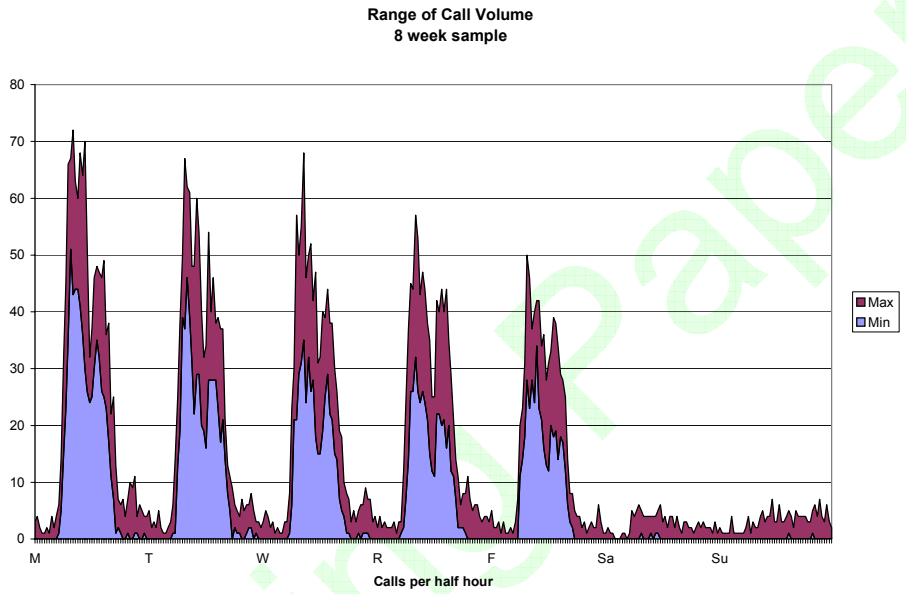

Figure 1-1 Range of Call Volume

This particular desk operates 24x7 and we see that the volume during the overnight hours is quite low. Volume ramps up sharply in the morning with a major surge of calls between 7 and 11 AM. Volume tends to dip down around the lunch break, but a second peak occurs in the afternoon,; though the afternoon peak is typically lower volume than the morning peak. The staffing challenge in this call center is to find a minimal cost staffing plan that achieves a global service level target with a high probability. The schedule must obviously be locked in before arrival rate uncertainty is revealed.

Throughout this analysis we will evaluate the models using three test problems based on specific outsourcing projects. Project J is a corporate help desk for a large industrial company averaging about 750 calls a day where the volatility of call volume is relatively low. Project S is help desk that provides support to workers in a large national retail chain. Call volume on this desk is about 2,000 calls a day. Because this desk supports users in retail stores, as opposed to a corporate office, the daily seasonality of

Working Paper

call volume is quite different. This company is making major changes in its IT infrastructure and as such call volume is very volatile and difficult to forecast. Project O is a help desk that provides support to corporate and retail site users of another retail chain. This is a smaller desk with about 500 calls a day, where call volume is fairly volatile and shocks are relatively common. We also examine various scheduling options. At one extreme we only allow workers to be assigned to five eight hour shifts per week. At the opposite extreme we allow a wide range of part time schedules. We allow for a total of five different flexibility options (A-E) which are summarized in the Appendix in Tables 9-1 and 9-2.

2 Literature

There is a large body of literature addressing manpower scheduling problems. An early work by Dantzig (Dantzig 1954) addressed scheduling toll booth operators. Dantzig formulated his model as a weighted set covering problem with known staffing requirements; the objective being to find the minimal cost covering from a set of available schedules. In the weighted set covering approach the staffing levels in each time period are calculated exogenously and are defined as hard constraints that must be satisfied in any feasible schedule. A similar model was applied to call center scheduling in (Henderson and Berry 1976). Due to the large number of potential schedules, especially when breaks are explicitly scheduled, much of the early research focused on solution algorithms. (Segal 1974) showed that without considering breaks the problems could be solved as a network flow problem in polynomial time. However when breaks are scheduled explicitly the problem becomes NP Hard (Garey and Johnson 1979).

Many early papers focused on heuristic algorithms. The Henderson and Berry model (Henderson and Berry 1976) applies two types of heuristics. The first heuristic reduces the number of shift types, scheduling against only a reduced set of schedules referred to as the *working subset*. The second approximation is the scheduling algorithm; the authors use three different scheduling heuristics. Another stream of research attacks the problem using an implicit scheduling approach. Implicit scheduling models generally use a two-phased approach, generating an overall schedule in the first phase, and then placing breaks in the second phase. Implicit scheduling approaches are addressed in (Bechtold and Jacobs 1990), (Thompson 1995) and (Aykin 1996). (Thompson 1995) includes a summary of related papers and then develops a Doubly-Implicit Shift Scheduling Model (DISSM). Several other papers addresses related problems (Brusco and Johns 1996; Brusco and Jacobs 1998; Brusco and Jacobs 2000). A succinct overview of a two-stage approach to scheduling in a call center environment is provided in section 12.7 of

Working Paper

(Pinedo 2005). This model is motivated by a call center application where resource requirements are defined exogenously and breaks must be scheduled.

(Milner and Olsen 2005) examine contract structures in call centers with service level agreements. In particular they examine call centers where agents can serve multiple customers – some with service level contracts and some without. In a related paper (Baron and Milner 2006) examine optimal staffing under various service level agreements. These papers classify serviced level agreements as individual based (IB), Period Based (PB) or Horizon Based (HB). Individually based SLAs assess a financial penalty for every customer not served within the specified service level. The PB-SLA specifies penalties for each time period in which the service level target is not achieved. Periods are defined as intervals over which the arrival rate can be considered constant – typically 15 or 30 minute intervals. The HB-SLA specifies penalties for service level shortcomings over an extended period such as a week or month. In this paper we examine scenarios where a HB-SLA has been specified with the horizon specified as one week.

(Ren and Zhou 2006) optimal staffing levels under various contracting models for outsourced call centers. This paper extends the notion of supply chain coordination to call centers and defines a call center contract as coordinating if it yields the maximum profit and allows an arbitrary split of profit between the user and call center. The profit is the difference between "revenue" earned by satisfactorily resolving the call and the cost of service delivery, consisting of staffing costs and imputed costs for abandonment, waiting and loss of goodwill. The paper examines both the level of staffing and "effort; e.g. training, selective hiring under four contract types. They demonstrate that when effort is observable and contractible a pay-per-call-resolved plus cost sharing contract can coordinate the supply chain. When effort is not observable a partnership contract can coordinate the supply chain.

Most call center scheduling models, including the implicit break scheduling models, implement a hard constraint for service level on a period by period basis – implicitly enforcing a period based SLA. Scheduling for a Period Based SLA is straightforward using the Stationary Independent Period by Period (SIPP) approach. The SIPP approach is described in detail in (Green, Kolesar *et al.* 2001) but essentially the day is divided into short periods, typically 15 or 30 minutes. In each period the arrival rate is assumed to be constant and performance is assumed to be independent of the performance in other periods. In each period a queuing model, often the Erlang C model, is used to calculate the staffing level required to

Working Paper

achieve the service level requirement. Given the vector of staffing level requirements a set covering integer program is used to schedule shifts. This two phased approach splits the task into a server sizing task, based on queuing models, and a staff scheduling task, based on discrete optimization.

A few models are formulated to solve a global service level requirement, i.e. an HB-SLA. It is our experience that outsourcing contracts typically specify global performance requirements. (Koole and van der Sluis 2003) attempt to develop a staffing model that optimizes a global objective. Their model uses a local search algorithm and to ensure convergence to a global optimum they require agent schedules with no breaks, and assume no abandonment. Their model also assumes a time varying, but known arrival rate. (Cezik and L'Ecuyer 2007) solve a global service level problem using simulation and integer programming. They use simulation to estimate service level attainment and integer programming to generate the schedule. The IP model generates cuts via sub-gradient estimation calculated via simulation. The model solves the sample average problem and therefore ignores arrival rate uncertainty, but it does allow for multiple skills. This model is a an extension of the model presented in (Atlason, Epelman *et al.* 2004). In a related paper (Avramidis, Chan *et al.* 2007) use a local search algorithm to solve the same problem. A related model is presented in (Avramidis, Gendreau *et al.* 2007). (Fukunaga, Hamilton *et al.* 2002) describe a commercial scheduling application widely used for call center scheduling. Global service level targets are modeled as soft constraints while certain staffing restrictions are modeled as hard constraints. The algorithm uses an artificial intelligence based search heuristic.

(Atlason, Epelman *et al.* 2007) develop an algorithm that combines server sizing and staff scheduling into a single optimization problem. This model focuses on the impact that staffing in one time period can affect performance in the subsequent period, a fact ignored in SIPP models. The algorithm utilizes discrete event simulation to calculate service levels under candidate staffing models, and a discrete cutting plane algorithm to search for improving solutions. This approach has the advantage of making accurate service level calculations for call centers with nonstationary arrival patterns, but this come at a high computational cost. The analysis is limited to single day scheduling with no breaks and therefore has only 13 feasible tours. This algorithm also enforces a Period Based SLA.

Each of these models either assume that the arrival pattern is known, or they schedule against a mean value arrival pattern. (Bassamboo, Harrison *et al.* 2005) develop a model that attempts to minimize the cost of staffing plus an imputed cost for customer abandonment for a call center with multiple

Working Paper

customer and server types when arrival rates are variable and uncertain. They solve the staffing and routing problems using an LP based method that is *asymptotically optimal*. (Harrison and Zeevi 2005) solve the staffing problem for call centers with multiple call types, multiple agent types, and uncertain arrivals using a fluid approximation. Their model seeks to minimize a deterministic staffing cost function along with a penalty cost associated with abandonment. Their approach models the staffing problem as a multidimensional newsvendor model and solves it through a combination of linear programming and simulation. (Whitt 2006) allows for arrival rate uncertainty as well as uncertain staffing, i.e. absenteeism when calculating staffing requirements. (Robbins, Medeiros *et al.* 2006) show that uncertainty can cause significant deviation from targeted performance metrics in call center applications. Each of these models incorporate arrival rate uncertainty into the server sizing step, but do not explicitly address the staff scheduling step. No model we know of explicitly accounts for arrival rate uncertainty when generating the call center schedule.

The model presented in our paper seeks to allow for arrival rate uncertainty while simultaneously integrating the server sizing and staff scheduling steps. We do this through a model formulated a stochastic integer program that includes a piecewise linear approximation of the service level curve.

3 Problem Formulation and Solution Approach

In this model we attempt to find a minimal cost staffing plan that satisfies a global service level requirement. Our model estimates the number of calls that meet the service level requirement in each period by making a piecewise linear approximation to the TSF curve; the curve that relates the number of agents to a given service level for a given arrival rate. We generate the linear approximation of the TSF curve using an Erlang A model which estimates abandonment rates based on an exponential patience distribution. Since the model allows for abandonment, it remains valid if the arrival rate exceeds the service rate whereas the Erlang C model becomes undefined in this condition and the queue size become infinite. Because of the high level of variability in the support desk environment, arrival rates will often exceed service rates, at least temporarily. This happens if the call center experiences unplanned spikes in arrivals. It may also happen by design for short periods of (known) high demand. In a corporate support desk a large surge of demand typically occurs in the morning. It is common practice to allow queues to build during this period. In some of our empirical analysis we found that over 80% of the abandonment in a 24 hour day occurred in a two hour window.

Working Paper

We formulate the model as a two stage mixed integer stochastic program. In the first stage staffing decisions are made and in the second stage call volume is realized and we calculate SLA attainment.

We formulate a model with the following definitions:

Sets

I: time periods
J: possible schedules
K: scenarios
H: points in a linear approximation

Deterministic Parameters

c_j: cost of schedule j
a_{ij}: indicates if schedule j is staffed in time i
g: global SLA goal
m_{ikh}: slope of piecewise TSF approximation h in period i of scenario k
b_{ikh}: intercept of piecewise TSF approximation h in period i of scenario k
p_k: probability of scenario k
μ_i: minimum number of agents in period i
r: per point penalty cost of TSF shortfall

Decision Variables

x_j: number of resources assigned to schedule j

State Variables

y_{ik}: number of calls in period i of scenario k answered within service level
S_k: TSF shortfall in scenario k

Stochastic Parameters

n_{ik}: number of calls in period i of scenario k

The model can then be expressed as

$$\min \sum_{j \in J} c_j x_j + \sum_{k \in K} p_k r S_k \tag{3.1}$$

subject to

$$y_{ik} \le m_{ikh} \sum_{j \in J} a_{ij} x_j + b_{ikh} \qquad \forall i \in I, k \in K, h \in H \tag{3.2}$$

$$\sum_{i \in I} n_{ik} S_k \ge \sum_{i \in I} (g n_{ik} - y_{ik}) \qquad \forall k \in K \tag{3.3}$$

$$y_{ik} \le n_{ik} \qquad \forall i \in I, k \in K \tag{3.4}$$

$$\sum_{j \in J} a_{ij} x_j \ge \mu_i \qquad \forall i \in I \tag{3.5}$$

$$x_j \le m_j \qquad \forall j \in J \tag{3.6}$$

$$x_j \in \mathbb{Z}^+, y_{ik} \in \mathbb{R}^+, S_k \in \mathbb{R}^+ \qquad \forall i \in I, k \in K, j \in J \tag{3.7}$$

Working Paper

The objective of this model (3.1) is to minimize the total cost of staffing plus the expected penalty cost associated with failure to achieve the desired service level. The optimization occurs over a set K of sample realizations of call arrivals. Constraint (3.2) defines the variable y_{ik} as the number of calls answered within the SLA target in period i of scenario k based on a convex linear approximation of the TSF curve. Constraint (3.3) calculates the TSF proportional shortfall; the maximum of the percentage point difference between the goal TSF and achieved TSF and zero. Constraint (3.4) limits the calls answered within the SLA target to the total calls received in the period. Constraint (3.5) defines the minimum number of agents in any period. Constraint (3.6) sets an upper limit on the number of agents assigned to each schedule. Constraint (3.7) defines the non-negativity and integer conditions for program variables.

The size of the model, and therefore the computation effort required to solve it, is driven in large part by two factors; the number of potential schedules (J) and the number of scenarios (K). The number of integer variables is equal to the number of schedules, while the number of continuous variables is equal to the product of the number of scenarios and the number of time periods, plus the number of scenarios.

In this analysis we are creating schedules for a week (with explicit breaks between shifts, but not within shifts.) In simple cases where we allow only 5 day a week, 8 hour shifts the number of possible schedules is 576. In more complex cases where we have a wider range of full and part time schedule options we have 3,696 schedules. We will investigate the number of scenarios required in the next section, but 50 scenarios is not unreasonable. This implies the requirement to solve models with 3,696 integer variables and over 16,000 continuous variables.

This program (3.1)-(3.7) is solved over some set of sample outcomes from the statistical model of call arrival patterns. Multiple approaches are available for generating simulated arrival patterns. A thorough analysis is provided in (Avramidis, Deslauriers *et al.* 2004). For our test problems we use a straight forward two stage algorithm. Details of the algorithm are presented in the appendix in Figure 9-1.

Working Paper

3.1 Solution Algorithm

This model is formulated as a MIP and as such can be solved by an implicit enumeration (branch and bound) algorithm. Branch and bound works well for smaller problems, but tends to become less effective as problem size increases. To facilitate the solution of large scale problems we implemented a version of the L-Shaped decomposition algorithm (Birge and Louveaux 1997). Our decomposition method is a direct implementation of this method, adapted for a discrete first stage. We decompose the problem into a master problem where the staffing decision is made, and a series of sub-problems where the TSF shortfall is calculated for each scenario.

Let v denote the major iterations of the algorithm. Also let E_{ik}^{v} and e_{ik}^{v} denote the coefficient of the cut generated in iteration k. The master problem is then defined as

$$\min \sum_{j \in J} c_j x_j + \theta^v \tag{3.8}$$

subject to

$$\theta^v \geq \sum_{k \in K} p_k E_{ik}^{v} \sum_{j \in J} a_{ij} x_j + e_{ik}^{v} \qquad \forall i \in I, v \tag{3.9}$$

$$\sum_{j \in J} a_{ij} x_j \geq \mu_i \qquad \forall i \in I \tag{3.10}$$

$$x_j \leq m_j \qquad \forall j \in J \tag{3.11}$$

$$x_j \in \mathbb{Z}^+, \theta^v \in \mathbb{R}^+ \qquad \forall j \in J \tag{3.12}$$

In this problem θ^v represents an estimate of the TSF shortfall penalty term. Let (x^v, θ^v) be an optimal solution. For each realization of the random vector $k = 1, \ldots, K$ we then solve the following subproblem

$$\min rS_k \tag{3.13}$$

subject to

$$y_{ik} \leq m_{ikh} \sum_{j \in J} a_{ij} x_j^v + b_{ikh} \qquad \forall i \in I, k \in K, h \in H \tag{3.14}$$

$$\sum_{i \in I} n_{ik} S_k \geq \sum_{i \in I} (g n_{ik} - y_{ik}) \qquad k \in K \tag{3.15}$$

$$y_{ik} \leq n_{ik} \qquad \forall i \in I, k \in K \tag{3.16}$$

Working Paper

$$x_j^v \in \mathbb{Z}^+, y_{ij} \in \mathbb{R}^+, S_k \in \mathbb{R}^+ \qquad \forall i \in I, j \in J, k \in K \tag{3.17}$$

We use the dual variables from the solution of the set of subproblems to improve the approximation of the penalty term. Let $\pi 1_{ikh}^v$ be the dual variables associated with (3.14), $\pi 2_k^v$ the dual variables associated with (3.15), and $\pi 3_{ik}^v$ the dual variables associated with (3.16). We then calculate the following parameters used for cut generation:

$$E_{ik}^{v+1} = \sum_{i \in I} \sum_{h \in H} \pi 1_{ikh}^v m_{ikh} \sum_{j \in J} a_{ij} x_j^v$$

$$e_k^{v+1} = \sum_{i \in I} \left[\pi 3_{ik}^v n_{ik} + \sum_{h \in H} \pi 1_{ikh}^v b_{ikh} n_{ik} \right] - \pi 2_k^v g \sum_{i \in I} n_{ik}$$

We use these values to generate a constraint of the form (3.9). Set $v = v + 1$, add the constraint to the master program and iterate. The algorithm solves the master program then solves each sub-program for the fixed staffing level defined in the master solution. Based on the solution of the sub-problems, each iteration adds a single cut to the master problem. These cuts create an outer linearization of the penalty function (Geoffrion 1970).

The solution of the master problem provides a lower bound on the optimal solution, while the average of the subproblem solutions provides an upper bound. In our implementation we solve the LP relaxation of the master until an initial tolerance level on the optimality gap is achieved and we then reapply the integrality constraints. We continue to iterate between the master MIP and the subprogram LPs until a final tolerance gap is achieved.

Whereas the branch and bound approach solves a single large MIP, the decomposition solves a large number of relatively small LPs and a small number of moderately sized MIPs. A representative instance with 100 scenarios required 30 major iterations, thereby requiring the solution of the master problem 30 times, and the subproblem 3,000 times. The master was solved as an LP relaxation 26 times and as a MIP 4 times.

The advantage of the decomposition approach is that solution time will tend to increase as an approximately linear function of the number of scenarios, while the branch and bound algorithm will increase as a nonlinear function of the number of scenarios.

Working Paper

The following graph illustrates the convergence of the L-Shaped decomposition algorithm for a particular instance with 384 schedules and 100 scenarios.

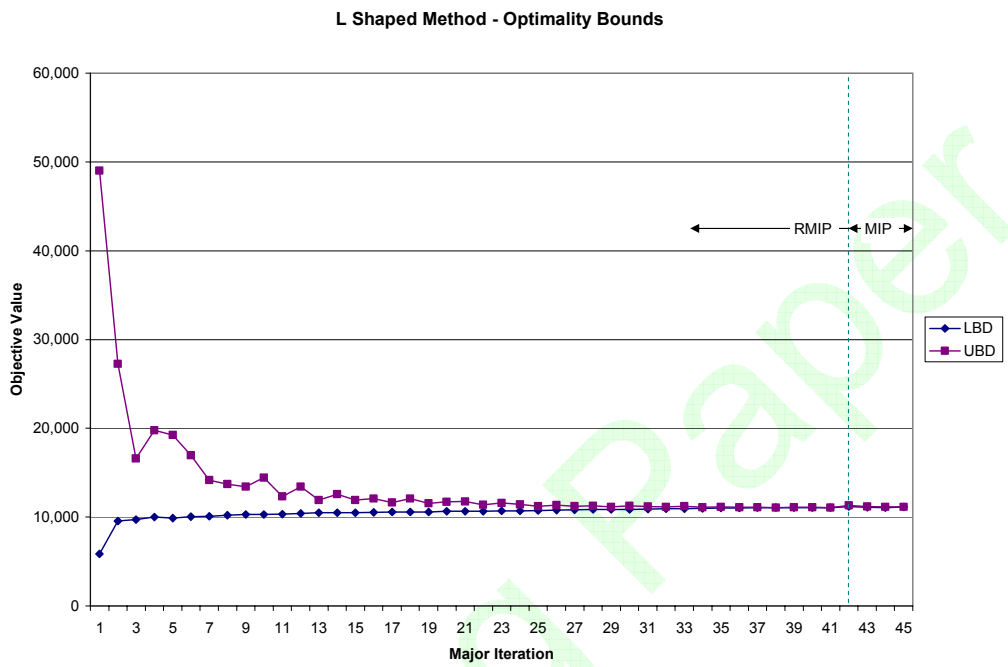

Fig 3-3 Convergence of the L-Shaped Algorithm

As is the case with a branch and bound algorithm relatively good bounds are found in the first few iterations. Convergence then slows as each successive iteration cuts a smaller area from the feasible region of (3.8) - (3.12). In this particular case the relaxation was solved 41 times and the MIP was solved four times. A slight shift in the bounds occurs when the integrality constraints are reapplied in iteration 42.

3.2 Post Optimization Analysis

The solution of (3.1) - (3.7) is the optimal solution of the sample path problem. We denote this solution as z_n^*, where n is the number of scenarios used to calculate the solution. This is a biased estimate of the solution to the true problem; that is the problem evaluated against the continuous distribution of arrival rates. We denote the true solution as z^*. (Mak, Morton *et al.* 1999) show that the expected bias in the solution is decreasing in sample size

Working Paper

$$E[z_n^*] \le E[z_{n+1}^*] \le z^*$$

From a practical perspective a key decision is determining the number of scenarios to use in our optimization. As we increase the number of scenarios the solution becomes a better approximation of the true solution, but the computational cost of finding that solution increases.

To aid in this process we perform a post optimization evaluation of the candidate solution using a Monte Carlo bounding process described in (Mak, Morton *et al.* 1999). Denote the solution to the sample problem as $\hat{x}$. We then solve the subprogram (3.13) to (3.17) using $\hat{x}$ as the candidate solution, to obtain the expected cost of implementing this solution. In this analysis we solve the subprogram with n_u equal 500 scenarios generated independently from the scenarios used in the optimization. The solution to the subprogram gives us an upper bound on the true solution, while the solution to the original problem z_n^* is a lower bound.

To obtain better bounds on the true optimal solution we may choose to solve the original problem multiple times, each with independently generated scenarios. Denote the number of batches (sets of scenarios) used to solve the original problem as n_ℓ and the sample variance of the objective as $s_\ell(n_\ell)$. Similarly we calculated the sample variance of the expected outcome of the candidate solution against the n_u evaluation scenarios. We can then define the following standard errors

$$\tilde{\varepsilon}_u = \frac{t_{n_u-1,\alpha} s_u(n_u)}{\sqrt{n_u}}$$

$$\tilde{\varepsilon}_\ell = \frac{t_{n_\ell-1,\alpha} s_\ell(n_\ell)}{\sqrt{n_\ell}}$$

Where $t_{n_u-1,\alpha}$ is a standard *t*-statistic, i.e. $P\left\{T_n \le t_{n_u-1,\alpha}\right\} = 1-\alpha$. We can now define an approximate $(1-2\alpha)$ confidence interval on the optimality gap as

$$\left[0, \left[\bar{U}(n_u) - \bar{L}(n_\ell)\right]^+ + \tilde{\varepsilon}_u + \tilde{\varepsilon}_\ell\right]$$

Note that we take the positive portion of the difference between the upper and lower bounds because it is possible, due to sampling error, that this difference is negative. This procedure allows us to generate a statistical bound on the quality of our solution.

Working Paper

4 TSF Approximation and SIPP

4.1 Overview

This model attempts to generate a schedule that meets a Service Level Agreement (SLA) at a minimal cost. For the sake of this analysis, we assume that the SLA is defined based solely on the TSF. In order to do so effectively the optimization program must estimate the service level that will be achieved for any staffing plan for each realization of calls. In this section we outline the approach used to estimate the TSF and document the assumptions used in developing this estimate. We then attempt to validate the estimate using a discrete event simulation model.

4.2 Basic TSF Calculations

The basic model used to estimate the service level in this analysis is the Erlang A model. The Erlang A model is a widely accepted model for call center systems with non-negligible abandonment rate. Erlang A assumes calls arrive via a Poisson process with rate λ and are served by a set of homogeneous agents with an exponentially distributed service time with mean $1/\mu$. If no agent is available when the call arrives it is placed in an infinite capacity queue where it waits for the next available agent. Each caller has a patience level which are iid draws from an exponential distribution with mean $1/\theta$. If a caller is not served by the time her patience expires she hangs up. The call center is also assumed to have infinite capacity so no calls are blocked.

In steady state, the staffing decision then involves forecasting the arrival rate λ_i and setting the staff level based on the Erlang A approximation. The result is a nonlinear S-shaped curve that for a fixed arrival rate, relates the achieved service level to the number of agents staffed.

4.3 Piecewise Linear Approximation

The TSF curve is neither convex nor concave over the full range of staffing. For very low staffing levels, where performance is very poor, the curve is convex and we experience increasing efficiency from incremental staffing. For higher staffing levels the curve becomes concave and the impact of incremental staffing becomes decreasing. Note that the area of convexity corresponds to very poor system

Working Paper

performance, an area where we do not plan to operate. In addition, embedding this function in our optimization model would create a non-convex optimization problem. To address this problem we create a piecewise linear approximation to the TSF curve as shown Figure 4-1.

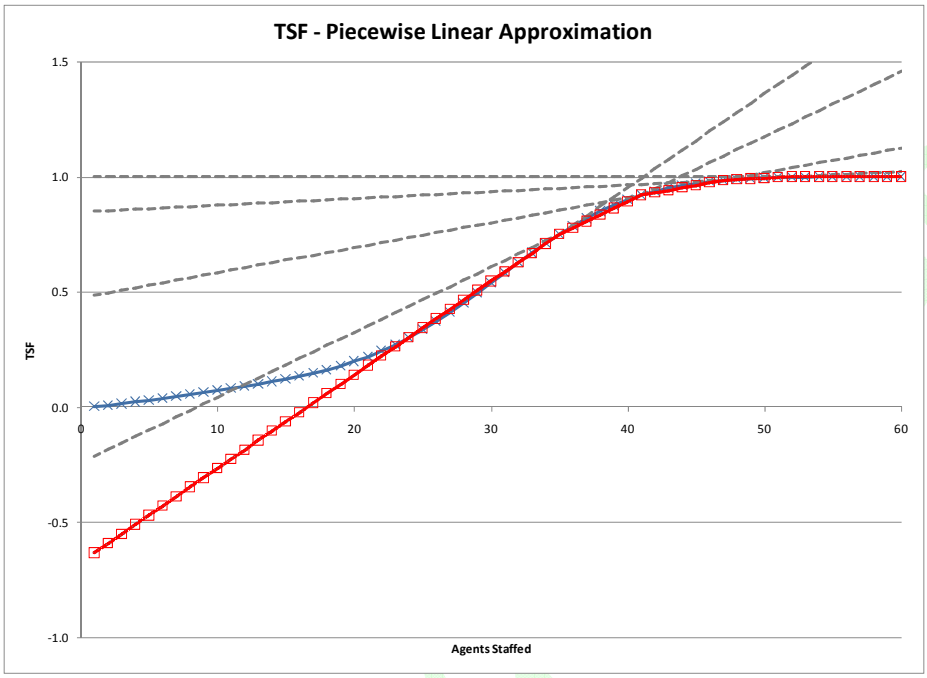

Figure 4-1 Convex Approximation of TSF

In this graph the straight lines represent the individual constraints, and the piecewise linear function is our approximation of the nonlinear curve. This graph has five linear segments, including a horizontal segment at a service level of 100%. The optimization model requires that the TSF is less than each line segment. The optimization process will force these constraints to be binding. The piecewise linear approximation and the true TSF curve are very close for staffing levels above 20. For very low staffing levels the linear approximation will overly penalize performance, potentially calculating a negative TSF level. Our assumption is that we are almost always operating in the higher performance region; we constrain the problem so that expected performance in any period is over some minimal threshold level of say 50%. Only in the case of very large shocks will we ever be driven into the poor performance region.

4.4 Non Stationary Arrivals

We estimate the TSF in each period using a standard representation of the Erlang A model. However, these equations are based on limiting behavior in steady state. For the most part, our analysis is concerned with nonstationary transient behavior. In our analysis we use a Stationary Independent Period

by Period (SIPP) approximation. The SIPP approach is described in more detail in (Green, Kolesar *et al.* 2001). Essentially in this approach we divide each day into 48 periods of 30 minutes each. We then estimate the average number of calls received in that period, set the arrival rate appropriately, and assume steady state behavior is quickly achieved in that period. We therefore assume that the Erlang A model can be used to estimate system performance in each 30 minute period using average arrivals. In applying the SIPP approach the arrival rate is assumed to change discontinuously at the start of each 30 minute period.

Clearly these assumptions have the potential to introduce significant error and the literature suggests several modifications of this approach, namely the SIPP Max and SIPP Mix approaches, both of which attempt to adjust for performance bias in the standard SIPP approach. We tested these alternatives by simulating the call center and found that the standard SIPP method gave the best approximation over the range of problems studied (Robbins 2007). A summary of the approximation for each of the three test projects is provide in Figure 4-2.

	Project		
Optimization	J	O	S
Scheduled Hours	1,160	1,080	2,760
Expected TSF	83.2%	81.5%	78.2%
Std. Dev of TSF	2.6%	2.9%	9.7%
Simulation			
Expected TSF	81.50%	80.99%	79.33%
Std. Dev of TSF	2.70%	3.22%	4.74%
Error (Opt-DES)	-1.72%	-0.54%	1.15%
Error in Std Dev of Sim TSF	-0.64	-0.17	0.24

Figure 4-2 TSF Approximation Results

4.5 Scenario Based TSF Approximation

The TSF calculations defined above are based on the call volume in each 30 minute period , n_{ik} , which is a random variable. The TSF calculations are therefore dependent on the sample path and must be included in the scenario generation algorithm. A comprehensive algorithm for scenario generation will then generate a simulated call volume and the associated set of linear constraints depicted in Figure 4-1. The details of this algorithm are provided in the appendix, in Figure 9-2. In addition to the individual scenario information, parameters for the minimum agent level constraint (3.5) must be generated. This is a straightforward procedure detailed in Figure 9-3. The scenario generation algorithm described above is

Working Paper

written in VB.Net. A 100 scenario file is generated in a few seconds on a desktop computer. Overall, the scenario generation time is negligible as compared to solution time for the stochastic program.

5 Cost and Service Level Tradeoffs

In a deterministic optimization approach to call center scheduling we set a performance target for some metric and then find the minimal cost schedule that satisfies that constraint; i.e. we implement the service level requirement as a *hard* constraint. In a stochastic setting the solution criteria are more complex. Given that the call volume, and therefore the service level, is random, the performance target can only be expressed in probabilistic terms. The resulting schedule will achieve the stated performance target with some probability. Given the nature of arrival variability it is not practical, or desirable, to generate a schedule that will always achieve the service level target as this schedule would be prohibitively expensive; i.e. we wish to implement the service level requirement as a *soft* constraint.

In our formulation we express the degree of certainty indirectly by assigning a financial penalty to a service level shortfall. By adjusting the performance penalty factor r, we adjust the degree of certainty associated with meeting the target. We now analyze the relationship between the penalty rate, the cost of service delivery, and the confidence associated with the performance target. Our model applies two performance constraints. Constraint (3.5) defines a minimum staff level in each period, which in our test cases we set to the minimum of a global minimum staffing level and the staffing level required to achieve some minimal performance level at expected volumes. (In our test problems we require that at least 2 agents are scheduled at all times. We also require that at expected volumes we achieve a minimum 50% TSF in each period.) If the penalty rate is set to zero the penalty term drops out of the objective function and constraint (3.5) becomes binding. As we increase the penalty rate the scheduled staff levels will increase to balance the cost of staffing and the expected penalty cost associated with TSF shortfalls.

In the following tables we show the result of an experiment to evaluate the impact of various penalty rates. For each project we test five design points (DPs) each with a different penalty rate. In each case we solve the stochastic problem five times, each with an independent batch of 50 scenarios. We then evaluate that solution against an independently generated set of 500 scenarios to determine the expected

Working Paper

outcome of implementing the candidate solution. The model is solved with the constraint that all schedules are full time (40 hours) - here we use schedule B defined in table 9-2.

		Average				Standard Deviation			
DP	Penalty Rate	Labor Cost	Expected Outcome	Average TSF	Confidence	Labor Cost	Expected Outcome	Average TSF	Confidence
1	0	8,800	8,800	60.5%	0.0%	0	0	0.00%	0.00%
2	25,000	10,800	11,008	80.6%	61.6%	0	18	0.16%	2.73%
3	50,000	10,880	11,249	81.0%	65.7%	179	40	1.16%	12.71%
4	75,000	11,120	11,332	82.6%	82.9%	179	28	1.11%	11.35%
5	100,000	11,120	11,419	82.7%	83.1%	179	127	1.11%	11.74%
6	150,000	11,200	11,458	83.1%	87.9%	0	36	0.30%	2.74%
7	200,000	11,200	11,504	83.1%	88.8%	0	56	0.23%	2.36%
8	250,000	11,200	11,597	83.1%	89.0%	0	72	0.31%	2.30%

Table 5-1 Cost and Service Level Tradeoffs – Project J

		Average				Standard Deviation			
DP	Penalty Rate	Labor Cost	Expected Outcome	Average TSF	Confidence	Labor Cost	Expected Outcome	Average TSF	Confidence
1	0	20,880	20,880	52.5%	0.0%	179	179	0.82%	0.00%
2	25,000	22,880	26,869	64.1%	1.9%	179	23	0.71%	1.00%
3	50,000	26,160	29,280	75.2%	41.1%	358	31	1.07%	7.26%
4	75,000	26,800	30,677	77.0%	53.2%	283	59	0.71%	4.76%
5	100,000	27,920	31,801	79.5%	67.3%	769	118	1.42%	6.45%
6	150,000	29,040	33,554	81.5%	76.1%	1,152	89	1.72%	5.03%
7	200,000	30,480	34,801	83.7%	80.9%	1,481	343	2.20%	6.47%
8	250,000	31,920	35,662	85.7%	84.4%	1,559	392	2.26%	4.23%

Table 5-2 Cost and Service Level Tradeoffs – Project S

		Average				Standard Deviation			
DP	Penalty Rate	Labor Cost	Expected Outcome	Average TSF	Confidence	Labor Cost	Expected Outcome	Average TSF	Confidence
1	0	8,240	8,240	54.2%	0.0%	219	219	1.49%	0.00%
2	25,000	10,800	11,705	76.8%	27.2%	0	37	0.17%	1.52%
3	50,000	11,360	12,294	79.9%	62.0%	219	37	0.97%	11.80%
4	75,000	11,600	12,736	80.6%	71.6%	0	58	0.33%	3.72%
5	100,000	11,600	13,022	80.9%	74.2%	0	46	0.21%	1.89%
6	150,000	12,000	13,595	82.5%	86.2%	0	21	0.17%	2.49%
7	200,000	12,000	14,127	82.4%	86.0%	0	112	0.36%	3.40%
8	250,000	12,320	14,591	83.1%	89.3%	179	72	0.71%	2.30%

Table 5-3 Cost and Service Level Tradeoffs – Project O

Working Paper

The following figures show the same data graphically. In the first set of graphs we show how confidence and average service level vary with the penalty rate

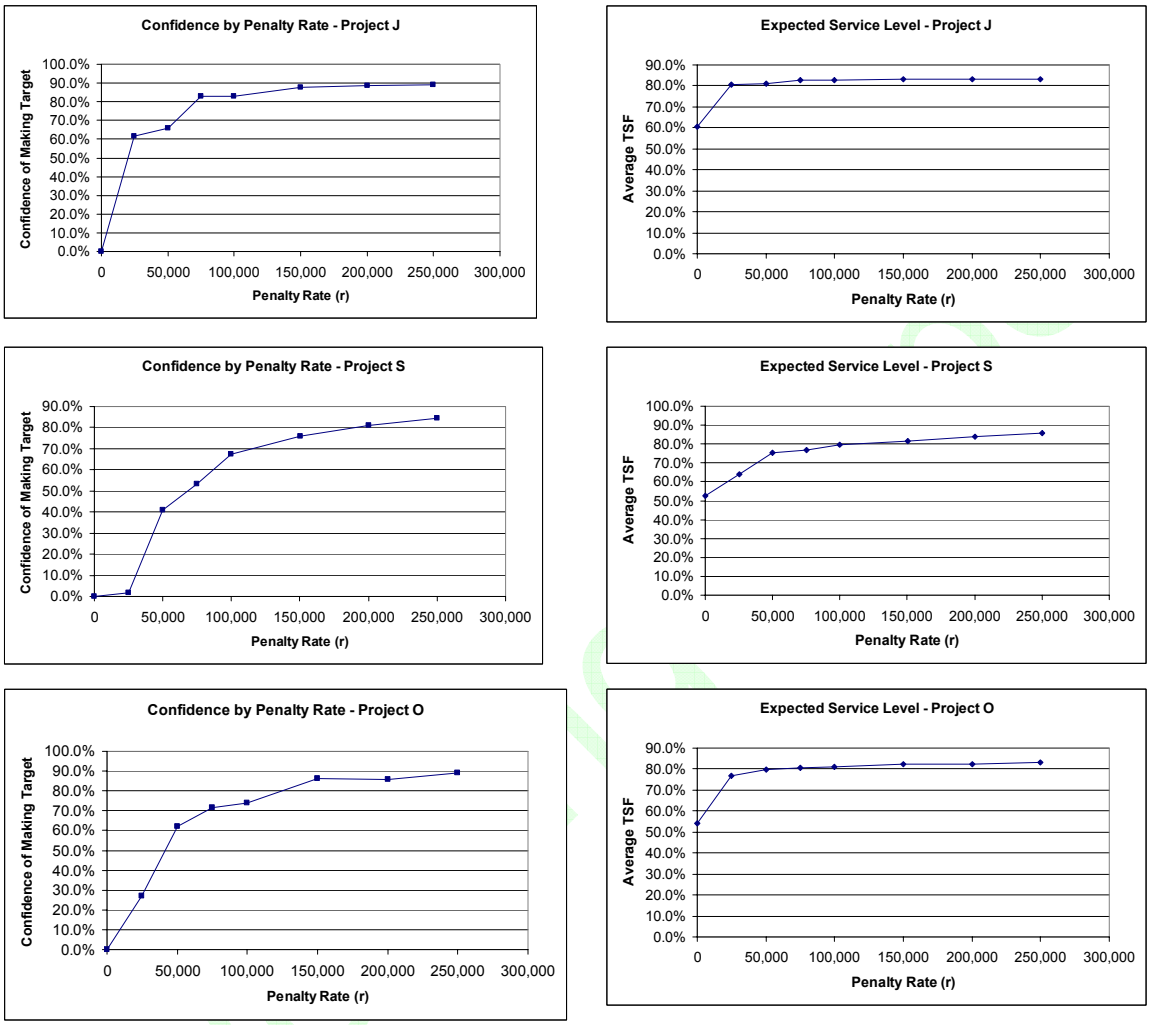

Figure 5-1 Confidence and Expected Service Level as a function of Penalty Rate

For each project, the panel on the left shows the confidence level of the resulting solution, i.e. the proportion of the evaluation scenarios in which the performance target was achieved. The panel on the right shows the corresponding expected service level associated with the candidate solution.

In all cases low penalties result in a zero confidence and an expected TSF near 60%[1]. As the penalty rate increases the expected TSF begins to increase as additional staffing is added to offset shortfall penalties. Both factors increase rapidly and then level off as it becomes increasingly expensive to meet the service levels in the tail of the arrival rate distribution. It is interesting to note that each project requires a different penalty rate to achieve a desired confidence level. Project S which has the largest staff levels and a high degree of variability, requires penalty rates in the range of 200,000 (2,000 per percentage point shortfall) to schedule with 80% plus confidence. Project O, a smaller project with moderate variability, plateaus with penalty rates around 100,000. Project J is a stable project stabilizes with penalty rates above 75,000.

The call center manager seeks to minimize the cost of staffing, while maximizing the probability of achieving the target service level. These two goals are clearly in conflict and the manager must decide how to balance cost and risk; *a decision obscured in a deterministic optimization approach*.

In the following graphs we recast the data from Figure 4-10 to illustrate this tradeoff. On the left side we see the confidence level of achieving the performance target as a function of staffing cost, and on the right we see the expected service level as a function of staffing cost.

Working Paper

[1] This model requires that the service level is at least 50% in every period based on expected volumes. In order to achieve that level in the busiest period staffing is set such that the service level is above 50% in subsequent periods. This is due to the constraint of scheduling agents to full time shifts.

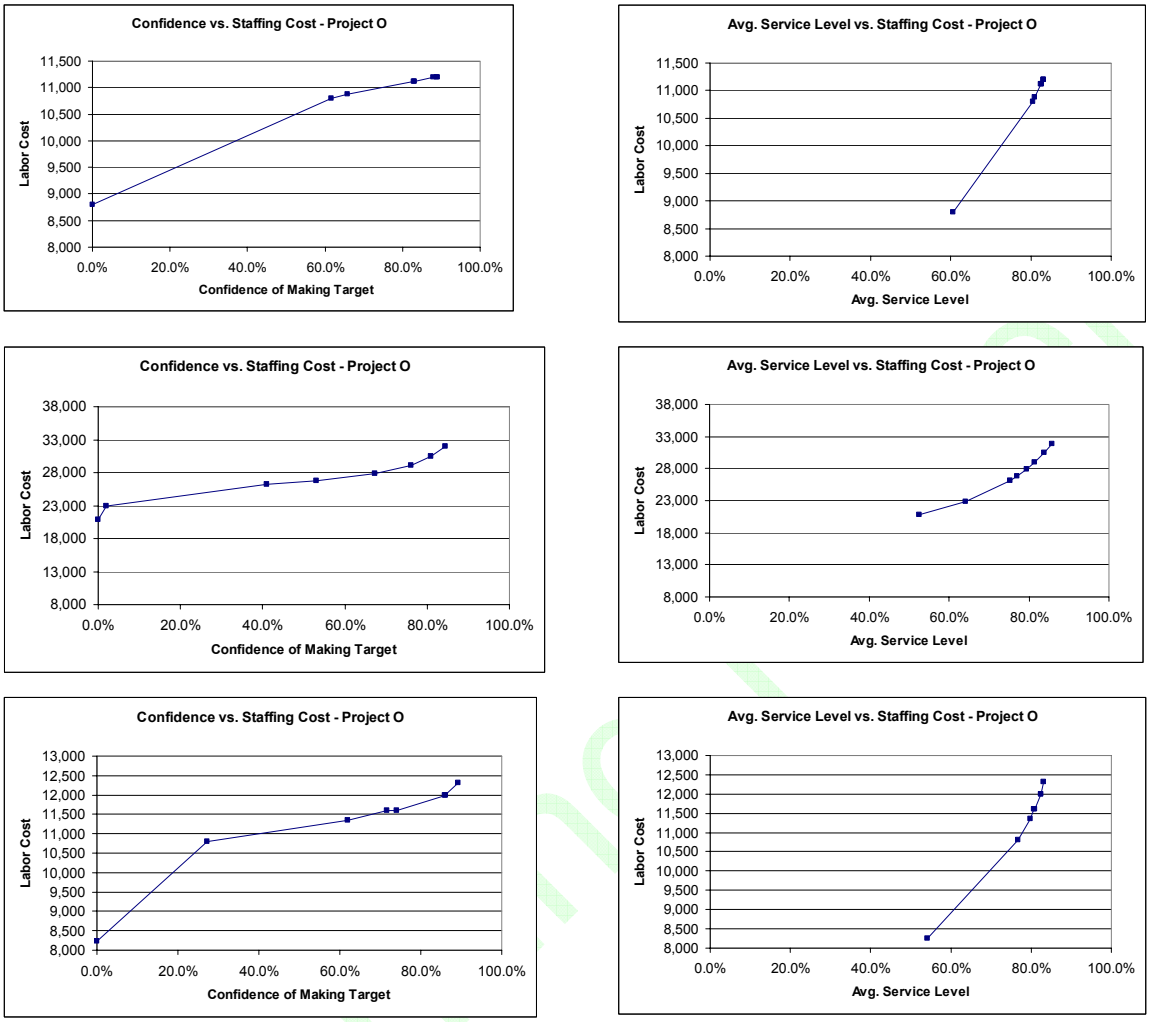

Figure 5-2 Confidence and Expected Service Level

The managerial implications here are important. When making day to day staffing decisions managers must make decisions about how much risk of missing the service level target they are willing to tolerate. Conversely, they decide how much insurance to buy in the form of excess capacity. In most situations managers must make these decision based on intuition. Our model operationalizes this decision by assigning a financial penalty to the possibility of failing to meet the service level target.

6 The Impact of Variability and VSS

6.1 Overview

The solution of the mean value program generates a biased estimate of the true cost of implementing the proposed solution. Solving a stochastic program reduces that bias, and the bias declines with the number of scenarios, going to zero as the number of scenarios goes to infinity (Mak, Morton *et al.* 1999). The expected cost of implementing the stochastic solution is lower than the cost of implementing the mean value solution, or stated differently we can lower the expected cost of operating the system by explicitly considering variability in our optimization problem. This reduction in cost is known as the Value of the Stochastic Solution (VSS). It is easily shown that VSS is a nonnegative quantity, (Birge 1982; Birge and Louveaux 1997)

The following figure depicts the relationship of the various costs.

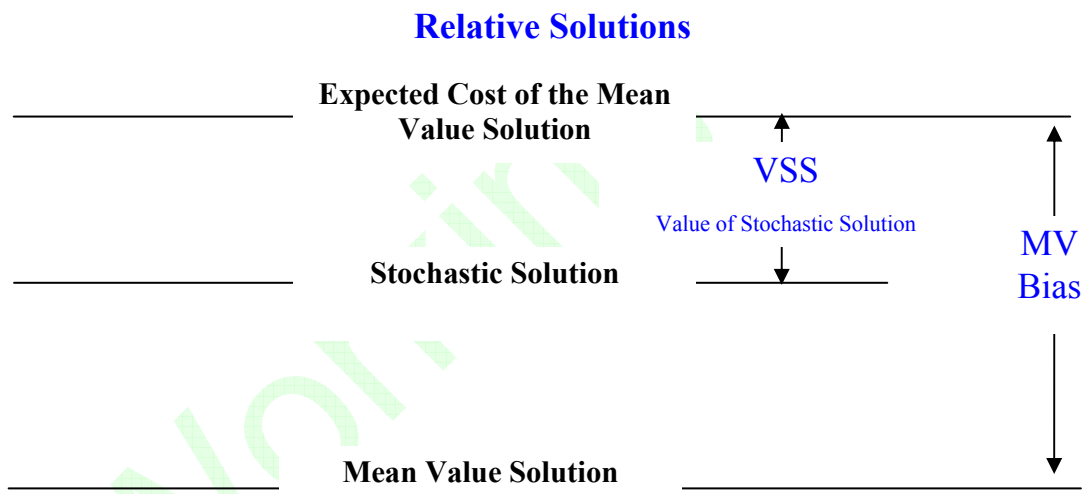

Figure 6-1 Relative Cost of Optimal Solutions

6.2 VSS and Solution Convergence

In this section we estimate the bias and VSS for three test projects for various scenario levels. At each scenario level we generate five independent batches and solve the program once for each batch. The

expected outcome is found by evaluating that solution against 500 evaluation scenarios. The following table summarizes the results.

Project	Scenarios	Direct Cost	Calculated Optimum	Expected Outcome	Solution Bias	VSS	VSS %	Confidence Level
Project J	MV	10,020	10,081	12,838	2,758			1.6%
	10	10,824	10,959	11,253	295	1,585	12.3%	63.5%
	25	10,848	11,044	11,146	121	1,693	13.2%	70.6%
	50	10,868	11,044	11,108	64	1,730	13.5%	74.4%
	100	10,884	11,075	11,092	36	1,747	13.6%	76.8%
Project S	MV	23,200	23,240	34,860	11,620			14.0%
	10	25,400	25,710	28,663	2,953	6,197	17.8%	56.2%
	25	26,720	27,376	27,540	193	7,320	21.0%	84.6%
	50	26,440	27,280	27,496	303	7,364	21.1%	81.2%
	100	26,260	27,069	27,337	304	7,523	21.6%	81.5%
Project O	MV	8,820	8,820	13,855	5,035			69.9%
	10	10,488	10,717	11,079	361	2,776	20.0%	80.2%
	25	10,500	10,844	11,009	199	2,846	20.5%	80.5%
	50	10,388	10,872	10,993	125	2,862	20.7%	80.1%
	100	10,520	10,879	10,956	77	2,899	20.9%	80.8%

Table 6-1 Solution Bias and VSS

In each case we find substantial bias in the Mean Value Solution and find substantial value from implementing the stochastic solution. On the moderately variable project J the stochastic program reduces expected cost by 13%. On the more variable projects S and O, the stochastic solution reduces cost by over 20%. Also note that the stochastic solution provides a higher confidence that the performance target will be achieved.

6.2.1 Sampling Bounds

In Section 3, we showed that the average solution to the stochastic program provides a point estimate on the lower bound on the true optimal solution, while the average expected outcome of the candidate solution forms a point estimate of the upper bound of the true optimal. In Figure 6-2 we plot the point estimate of the upper and lower solution bounds for project J at multiple scenario levels, estimated using five batches at each scenario level.

Working Paper

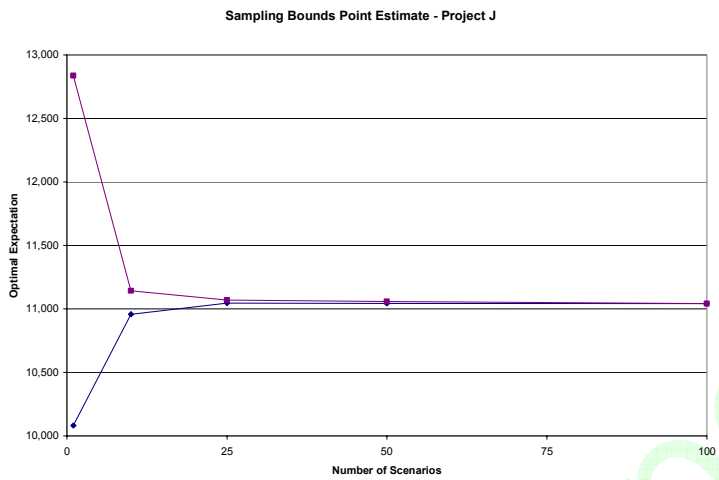

Figure 6-2 Point Estimate of Bounds

In Figure 6-3 we plot the 90% confidence interval on the magnitude of the optimality gap

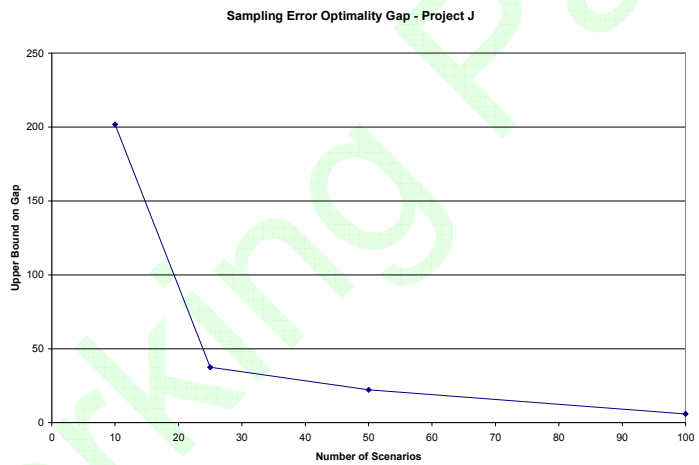

Figure 6-3 Optimality Gap

These graphs show that the mean value problem exhibits significant bias, but that even with a moderate number of scenarios, and a few batches, we are able to generate fairly tight bounds on the true optimal value. The data suggests that solving the problem with as few as 25 scenarios provides reasonably good results, while a 50 or 100 scenario model gives us a tighter bound that may be useful when trying to make detailed comparisons between alternatives.

For each project listed in Table 6-1 the stochastic program lowers overall expected cost by increasing direct labor. It is somewhat paradoxical that stochastic programs provide better results by calculating worse objective functions. The intuition is however straightforward; *deterministic optimization programs*

Working Paper

assume away uncertainty and therefore do not adequately hedge for variability; incremental staffing is added in periods with relatively high volumes and high variability.

7 Comparison with the Common Practice

7.1 Introduction

Throughout this paper we have analyzed a model that includes abandonment and arrival rate uncertainty. Neither of these conditions are included in industry standard models; "*common practice uses the M/M/N (Erlang C) queuing model to estimate the stationary system performance of short – half hour or hour – interval.*" (Gans, Koole *et al.* 2003) p.92. (Fukunaga, Hamilton *et al.* 2002) describe a commercial system deployed at over 800 call centers in which "*agent requirements are computed by applying the well-known Erlang-C formula*". Furthermore, standard industry practice is to make staffing decisions based on a period by period (local) service level requirement; "*each half hour interval's forecasted* λ_i *and* μ_i *give rise to a target staffing level for the period. ... determination of optimal set of schedules can then be described as the solution to an integer program*" (Gans, Koole *et al.* 2003) p.93.

In section 6.2 we showed that ignoring arrival rate uncertainty leads to verifiably more expensive solutions, on an expected cost basis, than models which account for variability. In this section we compare the stochastic Erlang A model to the commonly applied mean value arrival rate Erlang C model.

7.2 Weighted Set Covering Model

The standard approach described above generates a set of fixed staffing requirements in each period, and then attempts to find the lowest cost schedule to satisfy these requirements. The resulting integer program is a standard weighted set covering problem which can be expressed as

$$\min \sum_{j \in J} c_j x_j$$

Working Paper

subject to

$$\sum_{j \in J} a_{ij} x_j \geq b_i, \quad \forall i \in I$$

$$x_{ij} \in \mathbb{Z}^+$$

Where c_j is the cost of the schedule j, x_j is the number of resources assigned to the jth schedule, and a_{ij} is the mapping of schedules to time periods.

7.3 Locally Constrained Erlang C Model

We refer to the standard approach described in (Gans, Koole *et al.* 2003) as the locally constrained Erlang C model because it uses Erlang C to generate a hard local constraint in each period. We use a straightforward approach to generate the local constraints detailed in Figure 9-4.

The general problem with this approach is the constraint created by the per period service level requirement, coupled with the requirement to schedule resources in shifts. The peak staffing level is set by the peak arrival period, and depending on the length of the arrival peak, and the length of the flexibility of the staffing model, a substantial amount of excess capacity may be created in other periods. We refer to the extra capacity created in other periods as the *deadweight loss*; the loss in schedule efficient (man-hours scheduled) due to shift constraints. The magnitude of the deadweight loss will be a factor of the flexibility of the available set of schedules. With more flexible staffing options, the weighted set covering algorithm can match the requirement more closely.

Consider the examples shown in the following two graphs. In each graph the inner region defines the requirements generated for the set covering problem. The envelope of the graph represents the total staffing assigned by solving the set covering problem. The outer region therefore represents the excess capacity assigned above and beyond what was specified.

Working Paper

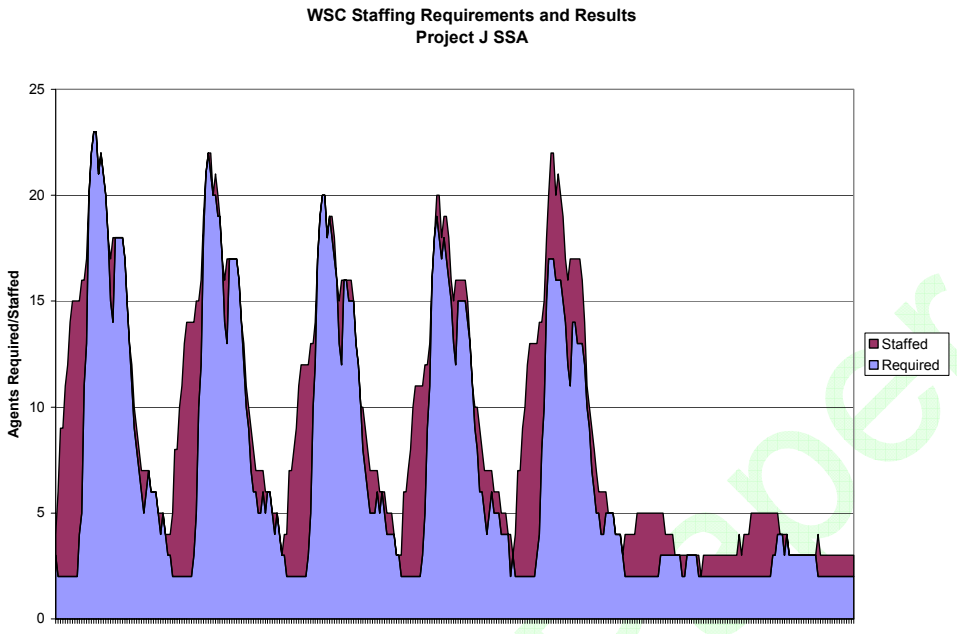

Figure 7-1 WSC Excess Staffing – Project J Schedule Set A

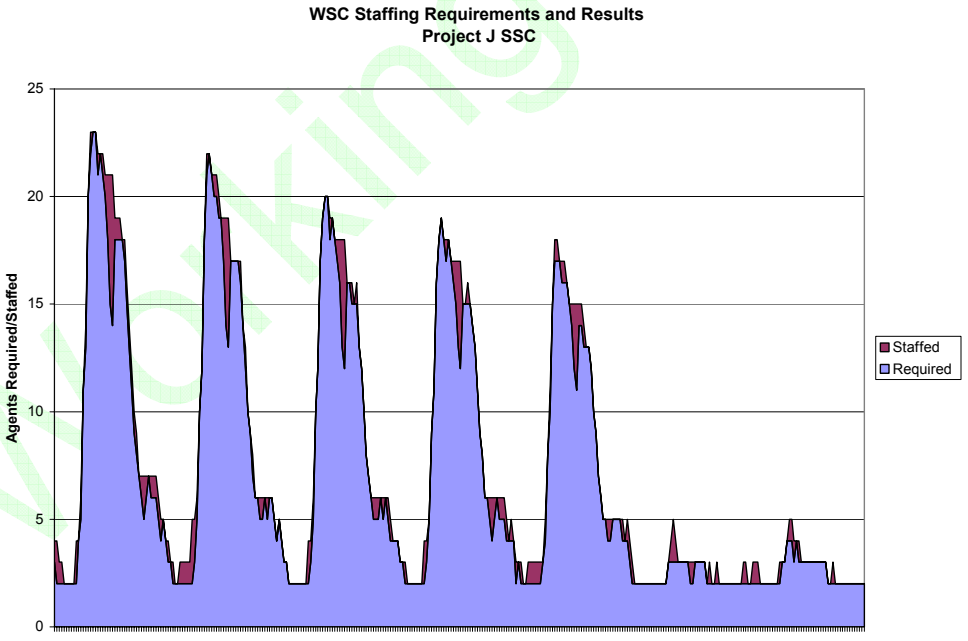

Figure 7-2 WSC Excess Staffing – Project J Schedule Set C

Working Paper

In Figure 4-24 we can only assign resources to full time 4x8 schedules and so the set covering is poor. The graph shows a significant amount of overstaffing throughout the course of the week. In Figure 4-25 we have the option of 4x10 and 4x8 shifts so we can match the required demand much more closely.

To quantify the impact we run a locally constrained Erlang C model for each of the three test projects for each of the five schedule sets. The per-period constraints are set so that the service level with expected volumes is at least 80% in every thirty minute period. In the following table we compare the results of this analysis with the results of the stochastic schedules generated in Section 4.6.

	Locally Constrained Erlang C						SCCS - Erlang A							
	Direct Labor	Expected Penalty	Expected Outcome	Average TSF	DWL	DWL %	Direct Labor	Expected Penalty	Expected Outcome	Average TSF	Direct Labor Savings		Expected Savings	
Project J														
Sched A	16,000	0	16,000	91.8%	4,055	34%	11,280	380	11,660	81.1%	4,720	29.5%	4,340	27.1%
Sched B	13,200	0	13,200	91.0%	1,255	11%	10,800	439	11,239	80.4%	2,400	18.2%	1,961	14.9%
Sched C	12,880	0	12,880	90.4%	935	8%	10,944	291	11,235	81.3%	1,936	15.0%	1,645	12.8%
Sched D	12,500	0	12,500	89.5%	555	5%	10,844	259	11,103	81.5%	1,656	13.2%	1,397	11.2%
Sched E	12,300	0	12,300	89.2%	355	3%	10,720	299	11,019	81.3%	1,580	12.8%	1,281	10.4%
Project S														
Sched A	38,000	1,565	39,565	91.6%	8,340	28%	30,960	4,345	35,305	83.2%	7,040	18.5%	4,260	10.8%
Sched B	32,800	3,847	36,647	88.0%	3,140	11%	30,320	4,408	34,728	83.7%	2,480	7.6%	1,919	5.2%
Sched C	32,320	4,184	36,504	87.4%	2,660	9%	30,384	4,349	34,733	83.6%	1,936	6.0%	1,772	4.9%
Sched D	30,900	4,820	35,720	86.1%	1,240	4%	30,092	4,493	34,585	83.5%	808	2.6%	1,135	3.2%
Sched E	30,980	4,796	35,776	86.2%	1,320	4%	30,096	4,499	34,595	83.5%	884	2.9%	1,181	3.3%
Project O														
Sched A	13,600	384	13,984	85.7%	2,180	19%	11,600	843	12,443	80.2%	2,000	14.7%	1,542	11.0%
Sched B	12,400	514	12,914	83.4%	980	9%	11,360	897	12,257	80.1%	1,040	8.4%	656	5.1%
Sched C	12,160	544	12,704	83.0%	740	6%	11,296	982	12,278	79.5%	864	7.1%	426	3.4%
Sched D	11,980	592	12,572	82.4%	560	5%	11,352	858	12,210	80.2%	628	5.2%	362	2.9%
Sched E	11,880	624	12,504	82.1%	460	4%	11,316	910	12,226	79.9%	564	4.7%	278	2.2%

Table 7-1 Comparing the Stochastic and Local Erlang C Schedules

The data confirms that the excess staffing is high for 4x8 staffing but decreases quickly with more flexible scheduling options. It also shows that this is a more significant problem for project J, which has a strong seasonality pattern, that for either Project S or O. The set covering approach tends to overstaff the project and achieves expected service levels higher than those achieved in the stochastic model. However, because the set covering model considers only the expected value and not the variance of arrivals, it is less effective at hedging than the stochastic model. Consider the case of schedule D for project S. The deterministic model has an expected service level of 86.2%, versus the goal of 80%, but still an expected penalty cost of $4,700. The stochastic model on the other hand has an expected service level of 82.9%, 3.3% lower, but an expected penalty only slightly higher at 5,080.

In all cases the stochastic model yields a lower direct labor cost and a lower expected cost of operation. The benefit of using the stochastic model is most significant when arrivals have a strong

Working Paper

seasonal pattern, as in Project J, or when workforce flexibility is low. With 4x8 only staffing the stochastic model provides at least 10.8% reduction in operating costs.

7.4 Globally Constrained Erlang C Model

In the previous section we showed that the stochastic model based on the Erlang A model provides lower cost solutions than the locally constrained Erlang C model discussed in the literature. An alternative approach is to use a deterministic Erlang C model, ignoring abandonment and uncertainty as in the previous model, but optimizing to global vs. local constraints. While this approach is not presented in the literature as far as we know, it is a natural simplification of the stochastic model we have analyzed so far. Because the model is deterministic, it assumes arrival rates are known, it will in general be easier to solve then the stochastic model. Ignoring abandonment will tend to increase recommended staffing, but ignoring uncertainty will tend to decrease staffing. It may be the case that under some circumstances these errors will cancel each other out and we can achieve good solutions at a lower computational cost.

The method for formulating and solving these problems is a straightforward implementation of the model (3.1) - (3.7). We solve a mean value version of the problem. The major change is that the coefficients for constraints (3.3) and (3.5) are calculated based on the Erlang C model. We still require a minimum of two agents staffed at all times, and a minimum service level at expected volume in every period of at least 50%.

We solve a version of this problem for each of the three projects for each scheduling option. Since the model is deterministic there is no need to solve multiple batches. To evaluate the expected cost of implementing the solution we continue to evaluate the resulting schedule against the stochastic Erlang A model. We assume that the Erlang A model with uncertain arrivals is the correct model and the objective of this analysis is to determine the error introduced by using a Globally Constrained Erlang C model.

Working Paper

The results of this analysis are shown in the following table:

	Globally Constrained Erlang C				SCCS - Erlang A							
	Direct Labor	Expected Penalty	Expected Outcome	Average TSF	Direct Labor	Expected Penalty	Expected Outcome	Average TSF	Direct Labor Savings		Expected Savings	
Project J												
Sched A	14,000	20	14,020	88.6%	11,280	380	11,660	81.1%	2,720	19.4%	2,360	16.8%
Sched B	12,000	2	12,002	87.1%	10,800	439	11,239	80.4%	1,200	10.0%	763	6.4%
Sched C	11,760	5	11,765	86.3%	10,944	291	11,235	81.3%	816	6.9%	530	4.5%
Sched D	11,600	7	11,607	86.3%	10,844	259	11,103	81.5%	756	6.5%	504	4.3%
Sched E	11,580	26	11,606	85.8%	10,720	299	11,019	81.3%	860	7.4%	587	5.1%
Project S												
Sched A	35,200	953	36,153	87.3%	30,960	4,345	35,305	83.2%	4,240	12.0%	848	2.3%
Sched B	30,400	5,412	35,812	84.8%	30,320	4,408	34,728	83.7%	80	0.3%	1,084	3.0%
Sched C	30,160	5,426	35,586	84.7%	30,384	4,349	34,733	83.6%	-224	-0.7%	854	2.4%
Sched D	29,340	6,080	35,420	83.6%	30,092	4,493	34,585	83.5%	-752	-2.6%	835	2.4%
Sched E	29,320	6,050	35,370	83.7%	30,096	4,499	34,595	83.5%	-776	-2.6%	775	2.2%
Project O												
Sched A	11,600	976	12,576	79.9%	11,600	843	12,443	80.2%	0	0.0%	133	1.1%
Sched B	11,200	1,305	12,505	78.5%	11,360	897	12,257	80.1%	-160	-1.4%	247	2.0%
Sched C	11,120	1,394	12,514	78.3%	11,296	982	12,278	79.5%	-176	-1.6%	236	1.9%
Sched D	10,960	1,442	12,402	78.0%	11,352	858	12,210	80.2%	-392	-3.6%	192	1.5%
Sched E	11,080	1,421	12,501	78.1%	11,316	910	12,226	79.9%	-236	-2.1%	276	2.2%

Table 7-2 Comparing the Stochastic and Global Erlang C Schedules

8 Conclusions and Future Research

In this paper we examined the issue of short term shift scheduling for call centers for which it is important to meet a service level commitment over an extended horizon. While the analysis focused exclusively on a TSF based SLA, the model could easily be adapted to support other forms of an SLA; such as abandonment rate or average speed to answer. The model was designed to recognize the uncertainty in arrival rates and was formulated as a mixed integer two stage stochastic program. Although difficult to solve, we showed the model is tractable and can be solved in a reasonable amount of time. We showed that uncertainty is of real concern in call centers, and that it has a real impact on scheduling decisions.

In Section 6-2 we showed the Value of the Stochastic Solution for this model is substantial; ranging from 12.3% to over 21%. The clear implication is that for this model formulation ignoring variability is a costly decision; however most models in practice ignore both uncertainty and abandonment. The implication is that one should not introduce abandonment into the model without also considering uncertainty. In Section 7 we compared this model with the common practice of scheduling to a local Erlang C constraint; that is scheduling based on a model that ignores abandonment and uncertainty but requires the service level target is achieved in every period. Comparing our model to this common practice we again found our model achieves lower cost results; ranging from 2.4% to 27%. The basic implication here is that the Erlang C model sometime achieves good results, since the abandonment

Working Paper

and uncertainty assumptions create counter balancing errors. However the stochastic model always achieves a better solution and in many practical cases a substantially better result. This is particularly true when the flexibility of the workforce is limited to full or near full time shifts and the set covering approach introduces considerable slack in the schedule.

Finally we compared this model to a Globally Constrained Erlang C model. It's rather obvious that one should expect a better result from a global constraint. This model gives superior results as compared to the local constrained Erlang C, but again our stochastic model outperforms this model in every case, by as little as 1% but by as much as 16%. The overall conclusion is that compared to the alternative methods analyzed here, the stochastic model will always give a lower cost of operation schedule, and sometime this difference can be substantial. This is a basic property of stochastic programming in general, but in this analysis we have shown that the difference is significant in real world cases.

In addition to providing a lower cost solution, the model presented in this paper addresses the scheduling problem from a fundamentally different perspective. In the standard set covering approach the service level constraint is a hard constraint, it must be satisfied and any candidate schedule either achieve the service level requirement or does not. But in reality the service level is a random variable and we will achieve the SLA target with some probability. The analysis in Section 5 examines this explicitly and addresses the trade-offs that managers must make in terms of cost and the confidence of achieving the service level. Our analysis shows how the cost of operation increases nonlinearly with the desired confidence level. This trade-off is obscured in the deterministic setting.

In future research this model can be easily extended to use different queuing assumptions, for example model that relax the requirement for exponential talk times. The trade off of solution precision and computational effort is also an area for future research, examining the impact of changing the convergence parameters discussed in Section 3.1. We will also investigate the implicit scheduling of breaks.

Working Paper

9 Appendix - Algorithms

1. Generate a call volume for each day of the week using the mean and standard deviation specified for the day.
2. For each time period in each day generate a random proportion of call volume based on the specified mean and standard deviation for the time period.
3. Normalize the time period proportions so that they add to 1 for each day.
4. Calculate the per period call volume by multiplying the daily total by the time period proportion.

Figure 9-1 Simulated Call Generation Algorithm

5. Generate a week of call volume using the algorithm shown in Figure 9-1 and calculate the associated per period arrival rate.
1. For a given call volume, select $h+1$ probability levels for estimating points on the TSF curve. (In practice we use values of .3, .72, .9, .98, and .995 for all periods with call volumes of at least 5. Different values are used for lower call volumes to maintain a concave approximation.)
2. Calculate the staff level required to achieve the target probabilities defined in Step 3.
3. Recalculate the TSF for the integral staffing level calculated in Step 4. We now have $h+1$ staff level probability pairs on the TSF curve.
4. Calculate the slope (m_{ikh}) and intercept (b_{ikh}) for each pair of adjacent points found in Step 5.
5. Generate a scenario that includes the per period call volumes (n_{ik}) and h pairs of slope and intercept parameters for each period in the planning horizon.

Figure 9-2 Scenario Based TSF Approximation Approach

1. Define w, the worst case acceptable expected service level, and n_{min} the overall minimum number of agents to be staffed at any time[2].
2. Repeat Steps 2 – 6 for each period i
3. Determine the expected call arrival rate
4. Calculate the staff level n_w required to achieve the worst case expected service level defined in step 1.
5. Calculate $\mu_i = \lceil \min(n_w, n_{\min}) \rceil$, the minimum agents to staff in period i.
6. Write out μ_i in a GAMS compatible format.

Figure 9-3 Minimum Staff Level Constraint Generation

1. Calculate the average volume in each 30 minute period of the week.
2. Using the volumes calculated in Step 1, determine the number of agents required to achieve the target service level in each 30 minute period by performing a search.
3. Set the period staffing requirement to the maximum of the number calculated in Step 2 and the global minimal staffing requirement.
4. Using the resulting vector of staffing requirements as the requirement parameter b_i in the IP **Error! Reference source not found. - Error! Reference source not found.**.

[2] Throughout this analysis we use a worst case TSF of 50% and a minimum staffing level of 2.

Working Paper

Figure 9-4 Local Constraint Generation

Pattern	Description
5 x 8	5 days a week, 8 hours a day (40 hr week)
4 x 10	4 days a week, 10 hours a day (40 hr week)
4 x 8	4 days a week, 8 hours a day (32 hr week)
5 x 6	5 days a week, 6 hours a day (30 hr week)
5 x 4	5 days a week, 4 hours a day (20 hr week)

Table 9-1 Shift Patterns

Pattern	Schedule Types Included	Feasible Schedules
A	5x8 only	336
B	5x8, 4x10	1,680
C	5x8, 4x10, 4x8	3,024
D	5x8, 4x10, 4x8, 5x6	3,360
E	5x8, 4x10, 4x8, 5x6, 5x4	3,696

Table 9-2 Scheduling Patterns

Working Paper

10 References

Aksin, Z., M. Armony and V. Mehrotra 2007. The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research. Working Paper 61p.

Atlason, J., M. A. Epelman and S. G. Henderson 2004. Call center staffing with simulation and cutting plane methods. *Annals of Operations Research* p. 333-358.

Atlason, J., M. A. Epelman and S. G. Henderson 2007. Optimizing Call Center Staffing Using Simulation and Analytic Center Cutting-Plane Methods. *Management Science* **Forthcoming** p. 15.

Avramidis, A. N., W. Chan and P. L'Ecuyer 2007. Staffing multi-skill call centers via search methods and a performance approximation. Working Paper p.

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

Avramidis, A. N., M. Gendreau, P. L'Ecuyer and O. Pisacane 2007. Simulation-Based Optimization of Agent Scheduling in Multiskill Call Centers. *2007 Industrial Simulation Conference*.

Aykin, T. 1996. Optimal Shift Scheduling with Multiple Break Windows. *Management Science* **42**(4) p. 591-602.

Baron, O. and J. M. Milner 2006. Staffing to Maximize Profit for Call Centers with Alternate Service Level Agreements. Working Paper 33p.

Bassamboo, A., J. M. Harrison and A. Zeevi 2005. Design and Control of a Large Call Center: Asymptotic Analysis of an LP-based Method. Working Paper 51p.

Bechtold, S. E. and L. W. Jacobs 1990. Implicit Modeling of Flexible Break Assignments in Optimal Shift Scheduling. *Management Science* **36**(11) p. 1339-1351.

Birge, J. R. 1982. The Value of the Stochastic Solution in Stochastic Linear Programs, with Fixed Recourse. *Mathematical Programming* **24** p. 314-325.

Birge, J. R. and F. Louveaux 1997. *Introduction to stochastic programming*, Springer. New York.

Brusco, M. J. and L. W. Jacobs 1998. Personnel Tour Scheduling When Starting-Time Restrictions Are Present. *Management Science* **44**(4) p. 534-547.

Brusco, M. J. and L. W. Jacobs 2000. Optimal Models for Meal-Break and Start-Time Flexibility in Continuous Tour Scheduling. *Management Science* **46**(12) p. 1630-1641.

Brusco, M. J. and T. R. Johns 1996. A sequential integer programming method for discontinuous labor tour scheduling. *European Journal of Operational Research* **95**(3) p. 537-548.

Cezik, M. and P. L'Ecuyer 2007. Staffing Multiskill Call Centers via Linear Programming and Simulation. Working Paper 34p.

Dantzig, G. B. 1954. A Comment on Edie's "Traffic Delays at Toll Booths". *Journal of the Operations Research Society of America* **2**(3) p. 339-341.

Fukunaga, A., E. Hamilton, J. Fama, D. Andre, O. Matan and I. Nourbakhsh 2002. Staff Scheduling for Inbound Call Centers and Customer Contact Centers. *Eighteenth National Conference on Artificial Intelligence*, Edmonton, Alberta, Canada.

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

Garey, M. R. and D. S. Johnson 1979. *Computers and intractability: a guide to the theory of NP-completeness*, W. H. Freeman. San Francisco.

Geoffrion, A. M. 1970. Elements of Large-Scale Mathematical Programming: Part I: Concepts. *Management Science* **16**(11, Theory Series) p. 652-675.

Green, L. V., P. J. Kolesar and J. Soares 2001. Improving the SIPP Approach for Staffing Service Systems That Have Cyclic Demands. *Operations Research* **49**(4) p. 549-564.

Harrison, J. M. and A. Zeevi 2005. A Method for Staffing Large Call Centers Based on Stochastic Fluid Models. *Manufacturing & Service Operations Management* **7**(1) p. 20-36.

Henderson, W. B. and W. L. Berry 1976. Heuristic Methods for Telephone Operator Shift Scheduling: An Experimental Analysis. *Management Science* **22**(12) p. 1372-1380.

Koole, G. and E. van der Sluis 2003. Optimal shift scheduling with a global service level constraint. *IIE Transactions* **35** p. 1049-1055.

Mak, W.-K., D. P. Morton and R. K. Wood 1999. Monte Carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters* **24**(1-2) p. 47-56.

Milner, J. M. and T. L. Olsen 2005. Service Level Agreements in Call Centers: Perils and Prescriptions. Working Paper 31p.

Pinedo, M. 2005. *Planning and scheduling in manufacturing and services*, Springer. New York, NY.

Ren, Z. J. and Y.-P. Zhou 2006. Call Center Outsourcing: Coordinating Staffing Level and Service Quality. *Management Science (forthcoming)*.

Robbins, T. R. 2007. Managing Service Capacity Under Uncertainty - Unpublished PhD Dissertation (http://www.personal.psu.edu/faculty/t/r/trr147). Working Paper 241p.

Robbins, T. R., D. J. Medeiros and P. Dum 2006. Evaluating Arrival Rate Uncertainty in Call Centers. *Proceedings of the 2006 Winter Simulation Conference*, Monterey, CA.

Segal, M. 1974. The Operator-Scheduling Problem: A Network-Flow Approach. *Operations Research* **22**(4) p. 808-823.

Thompson, G. M. 1995. Improved Implicit Optimal Modeling of the Labor Shift Scheduling Problem. *Management Science* **41**(4) p. 595-607.

Whitt, W. 2006. Staffing a Call Center with Uncertain Arrival Rate and Absenteeism. *Production and Operations Management* **15**(1) p. 88-102.

Working Paper