Scheduling A Call Center Utilizing A Stochastic Model

Ditila Ekmekçiu
Department of Finance
Teleperformance Albania, AMS shpk
Tirana, Albania
e-mail: ditila.ekmekciu@gmail.com

***Abstract*—We examine the issue of call center scheduling in an environment where arrival rates are very variable, cumulative volumes are uncertain, and the call center is conditioned by a global service level constraint. This article is inspired by the work with an Albanian call center where call volumes express significant variability and uncertainty. The contracts with all the clients specify a Service Level Agreement that should be achieved over a specific period of a week or month. We define the problem as a mixed-integer stochastic program. Our model has two unique characteristics. First of all, we integrate the server sizing and staff scheduling steps into a single optimization program. Secondly, we clearly identify the uncertainty in time-by-time arrival rates. We demonstrate that the stochastic formulation, generally, calculates a higher cost optimal schedule instead of a model that ignores variability, but that the estimated cost of this schedule is lower. We operate expanded experimentation to compare the solutions of the stochastic program with the deterministic programs, considering mean valued arrivals. We discover that, generally, the stochastic model gives an important reduction in the expected cost of operation. The stochastic model also permits the operations manager to make informed risk management decisions by taking into account the probability that the Service Level Agreement will be achieved.**

Keywords—call center; stochastic model programming; cost optimization; scheduling; uncertainty

I. Introduction

A call center is composed by a set of resources (generally staff, computers and telecommunication supplies) that guaranties the delivery of services through the phone. The working environment of a typical large call center could be imagined as very big room with numerous open-space workstations, in which people with earphones sit in front of computer terminals, providing services/products to “unseen” customers (Ekmekçiu, 2015). During the last years the growth of the call center’s industry has been very high. In particular in Albania, the number of outsourcing call centers is more than 300 with 20.000 employees (Mapo online, Call-center: Biznesi që vijon lulëzimin). Large range call centers are technologically and organizationally sophisticated operations and have been the object of important academic research. Call centers can be categorized into many different dimensions: functionality (help desk, emergency, telemarketing, teleselling, technical support, market research, information providers, etc.), size (from a few to several thousands of agent workstations), geography (single- or several locations, that can be domestic, nearshore or offshore), agents qualifications (low-skilled or highly-trained, single or multiskilled etc.), industry (telecommunication, finance, TV and internet providers, travel, marketing, sport etc.), line of business (inbound, outbound, backoffice etc.), and so on (Ekmekçiu et al 2015). Staffing is a crucial issue in call center management due to the fact that direct labor costs usually account for 70–80% of the total operating budget of a call center (Aksin et al., 2007). This work addresses the scheduling problem in a call center with very variable and uncertain arrival rates. The paper is precisely related to a research project with an outsourced Albanian call center. This operation includes providing help desk support to big corporate entities. While the range of services differs from account to account, many accounts demand 24/7 support and virtually all accounts depend on some shape of Service Level Agreement. There are different types of SLAs, but the most typical determines a minimum level of the Telephone Service Factor. A TSF SLA determines the fraction of calls that should be answered within a given time. For example, a 90/110 SLA specifies that 90% of calls must be answered within 110 seconds. A very significant point is that the service level is applied to an extended period, generally a week or month. Consequently, the workstation is generally staffed so that at some times the service level is under obtained, sometimes over obtained, and reaches the target for the entire extended period (week or month). The main challenge implicated with staffing this call center is meeting a certain, fixed SLA with a variable and uncertain arrival rate pattern. During this analysis, we will evaluate the models using three test problems applied on specific outsourcing projects. Project A is a corporate help desk for a large company handling in average about 750 calls a day, where the unpredictability of call volume is relatively low. Project B is a help desk that provides technical support to workers in a retail chain. The volume of the calls of this project is about 2000 calls a day. Because of the fact that it supports users in retail stores, as the opposite of a corporate office, the daily seasonality of call volume is really different.

This company is making important changes in its IT infrastructure and because of that call volume is really volatile and difficult to forecast. Project C is a help desk that provides support to corporate and retail site users of another retail chain. This is a smaller project with about 500 calls a day, where call volume is moderately volatile and shocks are almost ordinary. We analyze different scheduling options, too. At one extreme, we only permit agents to be assigned to five 8-hour shifts per week. At the opposite extreme, we permit a wide range of part time schedules. We permit for a total of five distinctive flexibility options (A-E) that are summarized in the Appendix in Tables A.1 and A.2.

II. LITERATURE REVIEW

There is a big number of literature addressing call center issues. Gans et al. (2003) gives a detailed and complete review of the literature. Another review of the call center literature is provided by Aksin et al. (2007). Call center works cover a large range of topics and contain a number of OR methodologies that include queuing theory, optimization, and simulation.

The first problem we address in this work is shift scheduling. The main approach to this issue was first defined in a paper by Dantzig (1954), which addressed scheduling toll booth agents. Dantzig developed his model like a weighted set covering issue with noted staffing requirements; the goal being to find the slightest cost involving from a group of available schedules. In this approach, the staffing levels in each time period are calculated externally and are determined as hard restrictions that should be satisfied in each feasible schedule. Segal (1974) demonstrated that without considering breaks the problem could be resolved as a network flow problem in polynomial time. Nevertheless, when breaks are scheduled in an explicit way the problem becomes NP Hard (Garey and Johnson, 1979). Because of the large number of possible schedules, particularly when breaks are explicitly scheduled, many of the early research focused on solution algorithms.

A lot of early papers concentrated on heuristic algorithms. Henderson and Berry (1976) apply two kinds of heuristics. The first one decreases the number of shift types, scheduling against only a decreased set of schedules indicated as the working subset. The second approximation is the scheduling algorithm, in which the authors use three distinctive scheduling heuristics. Another flow of research attacks the problem utilizing an absolute scheduling approach. Absolute scheduling models use two sets of decision variables; one to appoint break less shifts, the other to fit breaks. Implicit scheduling approaches are forwarded in Bechtold and Jacobs (1990), Thompson (1995) and Aykin (1996). Many other articles address similar problems (Brusco and Jacobs, 1998, 2000). A brief examination of the approach in two stages to scheduling in a call center environment is given in Section 12.7 of Pinedo (2005).

Customer service is a substantial consideration in call centers, and a lot of centers are subject to SLAs. Milner and Olsen (2008) analyze contract structures in call centers with SLAs. Baron and Milner (2006) analyze optimal staffing under various SLAs. These works classify SLAs as Individual Based, Period Based, or Horizon Based. IB-SLAs determine a financial penalty for every customer not served inside the specific service level. The PB-SLA indicates penalties for each time period during which the service level objective is not achieved. Periods are determined as intervals during which the arrival rate may be considered constant – in general 15 or 30 minute intervals. The HB-SLA indicates penalties for service level deficiency over an extended period such as a week or month. In this paper we analyze scenarios where a HB-SLA has been determined with the horizon specified as a week.

Most call center scheduling models in the literature implement a hard restriction for service level on a time-by-time basis. Scheduling for a PB-SLA is unequivocal using the Stationary Independent Period by Period (SIPP) approach. This approach is explained in detail in Green et al. (2001), but basically the day is divided into short periods, generally 15 or 30 minutes. In every period, the arrival rate is pretended to be constant and performance is pretended to be independent of the performance in other periods. In every period, a queuing model, usually the Erlang C model, is utilized to calculate the staffing level necessary to achieve the service level necessity. Then, a set covering integer program is utilized to schedule shifts. This two phased approach divides the job into a server sizing issue, depending on queuing models, and a staff scheduling task, based on discrete optimization.

Some models are formulated to resolve a global service level requirement, like an HB-SLA. Based on our experience, outsourcing contracts usually specify an HB-SLA, and all of the projects we analyzed were subject to this kind of SLA. Koole and van der Sluis (2003) pursue to develop a staffing model that optimizes a global target based on an HB-SLA. Their model uses a local search algorithm, and to guarantee union to a global optimum they necessitate agent schedules with no breaks, and consider no abandonment. Their model also considers a time varying, but also known, arrival rate. Cezik and L'Ecuyer (2007) resolve a global service level problem utilizing simulation and integer programming. They apply simulation to estimate service level achievement and integer programming to produce the schedule. The IP model produces cuts via sub gradient estimation calculated via simulation. The model resolves the sample average problem and as a result ignores arrival rate uncertainty, but it does permit for multiple skills. This model is a continuation of the model presented in Atlason et al. (2004). In an analogous paper Avramidis et al. (2007a) utilize a local search algorithm to resolve the same problem. A similar model is introduced in Avramidis et al. (2007b). Fukunaga et al. (2002) explain a commercial scheduling application extensively used for call center scheduling. Global service level objectives are modeled as soft constraints while fixed staffing restrictions are modeled as hard constraints. The algorithm utilizes an artificial intelligence depending search examining. Atlason et al. (2008) develop an algorithm that integrates server sizing and staff scheduling into a unique optimization problem. This model concentrates on the impact that staffing in one time period might have on performance in the subsequent period, a fact ignored in SIPP models. The algorithm uses discrete event simulation to find out the service levels under aspirant staffing models and a discrete cutting plane algorithm to search for bettering solutions. Any of these models either considers that the arrival rate per-period is identified or schedules versus the expected arrival rate. The problem of arrival rate uncertainty has been considered in several papers. Both big call center reviews (Gans et al., 2003; Aksin et al., 2007) have sections concerned to arrival rate uncertainty. Brown et al. (2005) complete a detailed empirical analysis of call center data. Even though they discover that a time-inhomogeneous Poisson process matches their data, they also discover that arrival rate is difficult to forecast and suggest that the arrival rate has to be modeled as a stochastic process. Different authors try to convince that call center arrivals pursue a doubly stochastic process, a Poisson process where the arrival rate is a random variable, too (Aksin et al., 2007; Whitt, 2006; Chen and Henderson, 2001). Arrival rate uncertainty might exist for many different reasons. Arrivals may show randomness bigger than that predicted by the Poisson process due to ignored variables; the weather might have an impact on emergency calls (Chen and Henderson, 2001), the condition of an organization's IT infrastructure might have an influence on support center calls (Robbins, 2007). Call volume is highly seasonal during a day, week, month and year (Robbins, 2007; Gans et al., 2003; Andrews and Cunningham, 1995). Call center managers try to take into account these factors when they develop forecasts, even though forecasts are subject to significant error. Robbins (2007) analyzed four months of weekday forecasts to real call volume for 11 call center projects. He discovered that the average forecast error surpasses 10% for 8 of 11 projects, and 25% for 4 of 11 projects. The standard deviation of the daily estimation to real ratio surpasses 10% for all 11 projects. Steckley et al. (2009) analyzed forecasted and actual volumes for 9 weeks of data taken from 4 call centers. They demonstrated that the forecasting errors are high and modeling arrivals as a Poisson process with the estimated call volume as the arrival rate may present significant error. Robbins et al. (2006) utilized simulation analysis to analyze the impact of forecast error on performance measures, showing the significant impact forecast error can have on system performance.

Some works direct staffing requirements when arrival rates are not certain. Bassamboo et al. (2005) evolved a model that tries to minimize the cost of staffing plus a supposed cost for customer abandonment for a call center with numerous customer and server kinds when arrival rates are uncertain and variable. They resolve the staffing and routing problems using a Linear Programming based method which is asymptotically optimal. Harrison and Zeevi (2005) utilize a fluid approximation to resolve the sizing problem for call centers with numerous call types, numerous agent types, and with uncertain arrivals. Their model looks for minimizing a deterministic staffing cost function along with a penalty cost related with abandonment. Their approach designs the staffing issue like a multidimensional newsvendor model and resolves it through a combination of linear programming and simulation. Whitt (2006) permits for arrival rate uncertainty and also uncertain staffing, like absenteeism when calculating staffing requirements. Steckley et al. (2004) analyze the kind of performance measures to utilize when staffing under arrival rate uncertainty. Any of these models combine arrival rate uncertainty into the server sizing step, but do not address the staff scheduling step, explicitly.

The model given in our paper pursues to allow for arrival rate uncertainty while at the same time merging the server sizing and staff scheduling phases. We make this through a model formulated as a stochastic integer program. The approach of stochastic programming is well described.

Birge and Louveaux (1997) is a text which analyzes the theory of stochastic programming and numerous solution algorithms, too.

A standard method of resolving stochastic programs is to resolve the sample path problem, solving the optimization problem versus a discrete set of samples that are referred as scenarios. Mak et al. (1999) discuss important statistical properties related with sample path optimization.

III. FORMULATION OF THE PROBLEM AND SOLUTION APPROACH

In this model, we try to find a staffing plan with minimal cost that satisfies a global service level requirement. Our model predicts the number of calls that reach the service level requirement in any period by doing a piecewise linear approximation to the TSF curve; the curve that correlates the number of agents to a specific service level for a specific arrival rate. In this section, we firstly present our formulation of the model, containing our approach for estimating service levels. Then we describe a process for solving large-scale integer program that emerges. At the end, we present a post-optimization approach to determine the quality of the arising solution, an important application in stochastic program.

A. Formulation

We formulate the model as a two phase, mixed-integer stochastic program. During the first phase, staffing decisions are made and during the second phase, call volume is accomplished and we calculate SLA achievement. We formulate a model with the definitions below:

Sets

J possible schedules

I time periods

K scenarios

H points in a linear approximation

Deterministic parameters

c_j cost of schedule j
a_{ij} displays if schedule j is staffed in time period i
g global SLA objective
m_{ikh} tilt 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
d_j maximum number of agents available for schedule j

Decision variables
x_j number of agents assigned to schedule j
State variables
y_{ik} number of calls in period i of scenario k answered within service level
S_k proportional TSF shortfall in scenario k

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

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

subject to $y_{ik} \leq n_{ik}\left(m_{ikh} \sum_{j\in J} a_{ij}\, x_j + b_{ikh}\right)$
□$i \in I,\ k \in K,\ h \in H,$ (3.2)

$\sum_{i\in I} n_{ik}\, S_k \geq \sum_{i\in I}(g n_{ik} - y_{ik})$
□$k \in K,$ (3.3)

$y_{ik} \leq n_{ik}$ □$i \in I,\ k \in K$ (3.4)

,
$\sum_{j\in J} a_{ij}\, x_j \geq \mu_i$ □$i \in I$ (3.5)

,
$x_j \leq d_j$ □$j \in J$ (3.6)

,
$x_j \in Z^+,\quad y_{ik} \in R^+,\quad S_k \in R^+$
□$i \in I,\ k \in K,\ j \in J.$ (3.7)

The goal of this model (3.1) is to minimize the sum of the total cost of staffing and the expected penalty cost related with failure to reach the wanted service level. The optimization happens over a set K of model achievements of call arrivals. Restriction (3.2) determines the variable yik as the number of calls answered within the SLA target in period i of scenario k depending on a convex linear approximation of the TSF curve shown in Fig. 3.1. Restriction (3.3) calculates the TSF proportional deficit, Sk: the maximum of the percentage point difference between the target TSF and obtained TSF or zero. Restriction (3.4) reduces the calls answered within the SLA target to the total calls received in the period. Restriction (3.5) determines the minimum number of agents in each period. The minimum agent level is fixed to the maximum of the global minimum number of agents needed by policy, generally two agents, and the staffing level needed to obtain a minimum service level at expected call volumes. In our test examples, the parameter dj is fixed to the maximum of two, and the number of agents that results in a service level of at least 50% at the average volume for the period. Restriction (3.6) defines an upper limit on the number of agents appointed to each schedule. The intention of this restriction is to limit the number of agents appointed to a schedule based on agent availability or readiness to work. Practically, this restriction also permits the call center manager to turn off certain schedules as he sees fit. Restriction (3.7) determines the non-negativity and integer conditions for program variables.

For a given planning horizon and scheduling interval, the size of the model, and as a result the computation struggle needed to solve it, is guided in large part by two factors; the number of possible schedules (J) and the number of scenarios (K). The number of integer variables is the same 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. A usual planning horizon is one week, and a usual interval is 30 minutes.

Expanding the planning horizon or diminishing scheduling interval will immediately grow the number of time periods (I) and indirectly increase the number of schedules (J) and as a result the resources needed to resolve the problem.

In this examination, we are developing schedules for a week (with specific breaks between shifts, but not within shifts.) In easy cases, where we permit only 5 day a week, 8-hour shifts, the number of probable schedules is 576. In cases where we have a vaster range of full and part time schedule options we get 3696 schedules. (The details are given in Table A.2.) We examine the number of scenarios required in the next section, but 50 scenarios are not irrational. This indicates the requirement to resolve models with 3696 integer variables and more than 16000 continuous variables

This program above (3.1)–(3.7) is resolved over some set of sample results from the statistical model of call arrival patterns. Multiple approaches are available for generating simulated arrival patterns. An accurate analysis is given in Avramidis et al. (2004). For our test problems, we utilize a simple two-stage algorithm comparable to the model in Weinberg et al. (2007). We use a multiphase, multiplicative model where the arrival rate is calculated as the product of the number of calls in a daily basis and the proportion of daily calls received in that time period, which

.

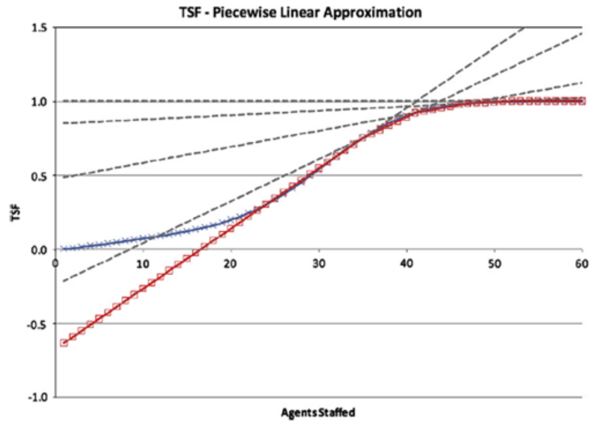


Figure. 3.1. Piecewise approximation of TSF

are random. In the Appendix in Fig. A.1, are given details of the algorithm, but it has to be noted that the scheduling algorithm is neither way dependent on the model of arrivals.

B. TSF approximation

The goal of our optimization model is finding the lowest cost staffing plan giving a service level restriction based on the TSF. In order to satisfy service level restrictions, the model has to estimate the service level that will be accomplished for a given staffing model and call pattern.

We complete this estimate by using a piecewise linear approximation of the TSF curve utilizing the Erlang A queuing model. This model is a vastly accepted model for call center systems with a non-insignificant abandonment rate. Erlang A takes for granted that calls arrive via a Poisson process with rate k and are handled by a group of homogeneous agents with an exponentially distributed service time with mean 1/μ. If no agent is available when the call comes, it is placed in an infinite capacity queue where it waits for the next available agent

Any customer that calls has a patience level, which are iid draws from an exponential distribution with mean 1/Ɵ. If a customer is not served by the time his patience expires, he hangs up. The call center is also pretended to have infinite capacity, as a result no calls are blocked. For each test project, we use forecasts derived from real call center data. In stable state, the staffing decision then includes forecasting the arrival rate ki and applying the staff level utilizing the Erlang A approximation. The result is a nonlinear S-shaped curve which, for a fixed arrival rate, correlates the obtained service level to the number of agents staffed.

The TSF curve is not convex and neither concave over the complete range of staffing. For very low staffing levels, where performance is very weak, the curve is convex, and in this case we experience increasing efficiency from incremental staffing. For higher staffing levels, the curve turns concave, and the impact of additional staffing becomes decreasing. Notice that the area of convexity coincides to very weak system performance, an area where we do not plan to operate. Additionally, embedding this function in our optimization model would create a nonconvex optimization problem. To direct this issue we develop a piecewise linear approximation to the TSF curve as demonstrated in Fig. 3.1.

In this chart, the straight lines represent the individual restrictions, and the piecewise linear function is our approximation of the nonlinear curve. This chart has five linear segments, including a horizontal segment at a service level of 100%. The optimization model demands that the TSF is less than any line segment. The piecewise linear approximation and the true TSF curve are very near for service levels higher than 25%. For very low staffing levels, the linear approximation will excessively penalize performance, potentially calculating a negative TSF level. The optimization process will force these restrictions to be binding and will force the TSF to be non-negative. Our assumption is that we are nearly every time operating in the greater performance region. In each of our test cases, we force the problem so that the performance expected in every period is higher than 50% via restriction (3.5).

C. Solution algorithm

Our model is defined with a limited number of call arrival patterns, and as a result can be expressed as a deterministic equivalent mixed integer program and being like that can be solved by an implicit enumeration (branch and bound) algorithm. Algorithms like branch and bound that ignore the special structure of a stochastic program, tend to become completely ineffective for large-range stochastic programs (Birge and Louveaux, 1997). A general approach for solving stochastic programs is to utilize the structure of the program across a decomposition algorithm (Birge and Louveaux, 1997). We realized a version of the L-Shaped decomposition algorithm adjusted for a discrete first stage. We break down the problem into a master problem where the staffing decision is done, and a series of subproblems where the TSF deficit is calculated for every scenario.

Let v designate the major iterations of the algorithm. Also let Eikv and eikv designate the coefficients of the cut produced in iteration k. The key problem is then determined as

min $\sum_{j\in J} c_j x_j + \Theta^v$ (3.8)

subject to $\Theta^v \geq \sum_{k\in K} p_k E_{ik}^v \sum_{j\in J} a_{ij} x_j + e_{ik}^v$

□i $\in$ I,v, (3.9)

$\sum_{j\in J} a_{ij} x_j \geq \mu_i$ □i $\in$ I, (3.10)

$x_j \leq d_j$ □j $\in$ J, (3.11)

$x_j \in Z^+$, $\Theta^v \in R^+$ □j $\in$ J. (3.12)

In this problem, Ɵv represents an estimate of the TSF deficit penalty term. Let (xv ,Ɵv) be an optimal solution. For each fulfillment of the random vector k = 1,...,K, we then solve the following subproblem

min rS_k (3.13)

subject to $y_{ik} \leq n_{ik}\left(m_{ikh} \sum_{j \in J} a_{ij} x_j^v + b_{ikh}\right)$

□i Є I, k Є K, h Є H, (3.14)

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

$y_{ik} \leq n_{ik}$ □i Є I, k Є K, (3.16)

$x_j \in Z^+$, $y_{ik} \in R^+$, $S_k \in R^+$ □i Є I, k Є K, j Є J, (3.17)

We utilize the dual variables from the solution of the set of sub-problems 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).

Then we calculate the following parameters utilized for cut generation:

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

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

We utilize these values to generate a restriction of the form (3.9). Set v = v + 1, add the restriction to the master program and iterate. The algorithm resolves the master program and then resolves every subprogram for the established staffing level determined in the master solution. Based on the solution of the subproblems, every iteration adds a single cut to the master problem. According to Geoffrion, 1970, these cuts generate an exterior linearization of the penalty function

The solution of the master problem gives a lower bound on the optimal solution, in the time that the average of the subproblem solutions gives an upper bound (Birge and Louveaux, 1997). In our application, we resolve the LP relaxation of the master till an initial tolerance level on the optimality gap is obtained, and after that we reapply the integrality restrictions. We go on with the iteration between the master MIP and the subprogram LPs up till an ultimate tolerance gap is obtained. While the branch and bound approach resolves a single large MIP, the decomposition resolves a large number of relatively small LPs and a small number of moderately sized MIPs and MIP relaxations. A typical instance with 100 scenarios needed 30 major iterations, as a result requiring the solution of the key problem 30 times and of the subproblem 3000 times. The master was resolved as an LP relaxation 26 times and as a MIP four times. (The Fig. 10-8 in the supplementary material demonstrates the convergence of the L-Shaped decomposition algorithm for a specific instance with 384 schedules and 100 scenarios.)

Like is the case with a branch and bound algorithm, nearly good bounds are found in the first few iterations. Convergence after that slows as any successive iteration cuts a smaller area from the feasible region of (3.8)–(3.12).

D. Post- optimization examination

The solution to the classic path formulation of a stochastic program is an approximation of the solution to the true optimization problem where parameters are random variables (Mak et al., 1999). A well-developed theory exists for determining the quality of simple path approximations based on Monte Carlo sampling techniques (Mak et al., 1999; Birge and Louveaux, 1997; Bayraksan and Morton, 2009). In this section, we describe a process by which we utilized this method to check the quality of the solution we achieve when evaluating against the sample of 25 arrival patterns.

The solution of (3.1)–(3.7) is the optimal solution of the sample path problem. We designate the goal value of this solution as zn* , where n is the number of scenarios utilized to calculate the solution. This is a partial estimate of the solution to the true problem; which is, the problem evaluated against the continuous distribution of arrival rates. We designate the goal of the true solution as z*. Mak et al. (1999) demonstrate that the expected bias in the solution is becoming lower in sample size

E$[z_n^*]$ ≤ E$[z_{n+1}^*]$ ≤ z^*

From a useful perspective, a key decision is establishing the number of scenarios to use in our optimization. As we expand the number of scenarios, the solution becomes a better approximation of the real solution, but the computational cost of finding that solution becomes higher.

To help in this process, we execute a post-optimization evaluation of the candidate solution utilizing a Monte Carlo bounding process explained in Mak et al. (1999). Designate the solution to the sample problem as $\hat{x}$. Then we resolve the subprogram (3.13)–(3.17) using $\hat{x}$ as the candidate solution, to achieve the expected cost of implementing this solution. In this analysis, we resolve the subprogram with n_u equal 500 scenarios produced independently from the scenarios utilized in the optimization. The solution to the subprogram gives us an upper bound on the true solution $\left(\bar{U}(n_u)\right)$, while the solution to the original problem, z_n^*, is a lower bound $\left(\bar{L}(n_l)\right)$.

To achieve better bounds on the true optimal solution, we can choose to resolve the original problem multiple times, each with separately generated scenarios. Designate the number of batches (sets of scenarios) utilized to resolve the original problem as n_l and the sample variance of the goal as $s_l(n_l)$. Likewise, we calculated the model variance of the expected result of the candidate solution versus the n_u evaluation scenarios. We may then determine the following standard errors

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

$$\tilde{\varepsilon}_l = \frac{t_{n_l-1,\alpha} s_l(n_l)}{\sqrt{n_l}},$$

where $t_{n_u-1,\alpha}$ is a standard t-statistic, like P{$T_n \le t_{n_u-1,\alpha}$} = 1-α. Now we can determine an approximate (1- 2α) confidence interval on the optimality gap as

$$[0, [\bar{U}(n_u) - \bar{L}(n_l)]^+ + \tilde{\varepsilon}_u + \tilde{\varepsilon}_l].$$

Notice that we take the positive portion of the discrepancy between the upper and lower restrictions because it is possible, because of sampling error, that this variation is negative. This procedure permits us to create a statistical bound on the quality of our solution. (You can find a graphical analysis of the optimality gap in Figs. 10-13 and 10-14 in the additional material.) In an optimization problem with 25 scenarios we obtained a gap of €50 on a schedule with a cost in exuberance of €11,000, a gap of less than 0.5%. According to this analysis, we concluded that resolving the stochastic program with 25 scenarios would give near optimal solutions. In our test cases, we utilized 50 scenarios except if noted contrarily

IV. COST AND SERVICE LEVEL COMPROMISE

In our model we control the certainty with which the target service level is obtained by assigning a financial penalty to a service level deficit. By adjusting the performance penalty factor, r, we fix the favorite degree of certainty related with meeting the target. While the penalty rate r can be set based on the contractual penalty for not achieving the service level, there is an additional implicit cost related with the perception of poor quality. Saying it differently, managers generally want to guarantee a higher probability of obtaining the service level than involved by the explicit penalty rate. Now we analyze the relationship between the cost of service delivery, the penalty rate and the confidence related with the performance target, like the probability that the service level target is obtained.

In a deterministic optimization approach to call center scheduling, we set a performance objective for some metric and after that find the minimal cost schedule that satisfies that restriction; like we implement the service level requirement as a hard restriction. In a stochastic setting, the call volume, and as a result the service level, is random, and the performance objective may only be expressed in probabilistic terms. Provided the nature of arrival variability, it is neither practical nor wanted to generate a schedule that will always obtain the service level objective as this schedule would be prohibitively expensive. As a result, we want to implement the service level requirement as a soft restriction.

In Tables 4.1–4.3, we illustrate the result of an experiment figuring out the impact of different penalty rates. For every project, we check eight design points,

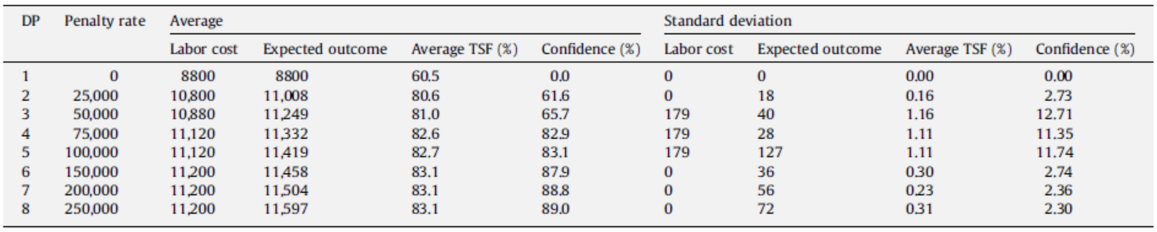
each one with a various penalty rate. (We can find the same data demonstrated graphically in Fig. 10-9 of the supplementary material.)

The scope of this experiment is to establish the penalty rate that should be utilized for each project to obtain a wanted confidence of achieving the service level objective. In any case, we resolve the stochastic problem five times, every time with an independent amount of 50 scenarios.

Then we evaluate each solution versus an independently generated set of 500 scenarios to forecast the expected outcome of implementing the candidate solution. The model is resolved with the restriction that all schedules are full-time (40 hours), utilizing schedule B determined in Table A.2.

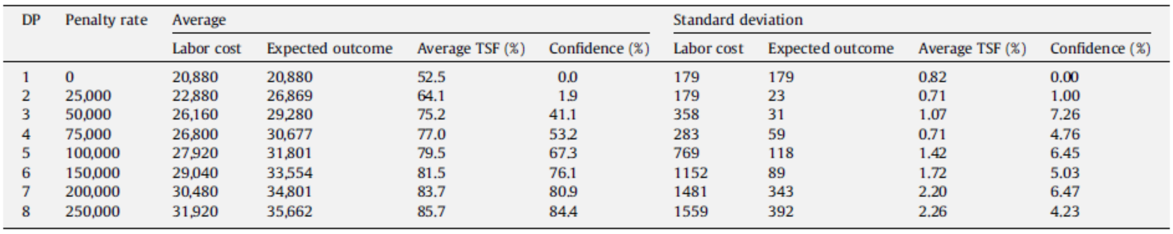
In each case, low penalties result in a zero confidence and an expected TSF near 60%. When the penalty rate gets higher, the expected TSF starts to increase as additional staffing is added to offset deficit penalties. Both factors increase quickly and then level off as it becomes increasingly expensive to reach the service levels in the tail of the arrival rate distribution. It is curious to note that each project needs a different penalty rate to obtain a desired confidence level. Project A that has the largest staff levels and a high degree of variability, needs penalty rates in the range of €200,000 (€2000 per percentage point deficit) to schedule with higher than 80% confidence. Project B, a smaller project with moderate variability, lands with penalty rates around 100,000. Project C, a stable project, balances with penalty rates at/or above 75,000. The call center manager searches to reduce in minimum the cost of staffing, meanwhile maximizing the probability of achieving the objective service level. These two targets are obviously in conflict and the manager should decide how to balance cost and risk: a decision that is hidden in a deterministic optimization approach

Table 4.1: Cost and service level compromises – Project A.



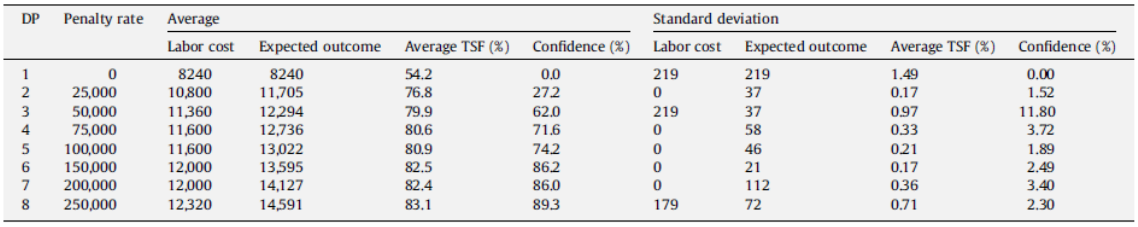
DP	Penalty rate	Average				Standard deviation			
		Labor cost	Expected outcome	Average TSF (%)	Confidence (%)	Labor cost	Expected outcome	Average TSF (%)	Confidence (%)
1	0	8800	8800	60.5	0.0	0	0	0.00	0.00
2	25,000	10,800	11,008	80.6	61.6	0	18	0.16	2.73
3	50,000	10,880	11,249	81.0	65.7	179	40	1.16	12.71
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 4.2: Cost and service level tradeoffs – Project B.



DP	Penalty rate	Average				Standard deviation			
		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	1152	89	1.72	5.03
7	200,000	30,480	34,801	83.7	80.9	1481	343	2.20	6.47
8	250,000	31,920	35,662	85.7	84.4	1559	392	2.26	4.23

Table 4.3: Cost and service level tradeoffs – Project C



DP	Penalty rate	Average				Standard deviation			
		Labor cost	Expected outcome	Average TSF (%)	Confidence (%)	Labor cost	Expected outcome	Average TSF (%)	Confidence (%)
1	0	8240	8240	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

.

The managerial implications here are substantial. When making day-to-day staffing decisions managers have to consider how much risk of missing the service level goal they are going to tolerate. Contrarily, they also decide how much insurance to buy in the form of exuberance capacity. In most of the situations, managers have to make these decision based on instinct. Our model operationalizes this kind of decision by assigning a financial penalty to the possibility of not meeting the service level objective

V. THE IMPACT OF VARIABILITY AND VALUE OF THE STOCHASTIC SOLUTION

A. *Overview*

The solution of the mean value program generates a partial estimate of the true cost of implementing the recommended solution. Resolving a stochastic program diminishes that partiality, and the partiality drops with the number of scenarios, tending to go to zero when the number of scenarios goes to infinity (Mak et al., 1999). The expected cost of implementing the stochastic solution is diminished versus the cost of implementing the mean value solution, or articulated in another way we may diminish the expected cost of operating the system by specifically considering variability in our optimization problem. This kind of reduction in cost is known as the Value of the Stochastic Solution (VSS). It is simply demonstrated that VSS is a zero or positive quantity (Birge and Louveaux 1997; Birge, 1982.). (The Fig. 10-11 in the supplementary material graphically describes the relationship of the distinct costs.) We calculate the VSS to define if there is benefit from resolving the stochastic version of the issue. During this section, we calculate the VSS to show that mean value solutions are optimistically biased, but not probable to obtain the desired service level, for any of our three projects.

B. *VSS and solution convergence*

In this section, we predict the bias and the VSS for the same three projects earlier examined for different scenario levels. At each scenario level, we produce five independent batches and resolve the program once for every batch. The expected outcome is identified by evaluating that solution against 500 evaluation scenarios. In Table 5.1 we find the summarized results.

In every case we identify considerable bias in the Mean Value Solution and find considerable value from implementing the stochastic solution. On the slightly variable project A, the stochastic program reduces the expected cost by 13%. On the more variable projects B and C, the stochastic solution reduces cost by over 20%. Notice that the stochastic solution gives a higher confidence that the performance objective will be obtained.

For every project filed in Table 5.1 the stochastic program reduces overall expected cost by increasing direct labor. It is somehow paradoxical that stochastic programs give better results by calculating worse target functions. However, the intuition is unequivocal; deterministic optimization programs take for granted away uncertainty and as a result do not sufficiently hedge for volatility; increasing staffing is added in periods with almost high volumes and high variability.

We demonstrated in Section 3 that the average solution to the stochastic program gives a point estimate on the lower bound of the real optimal solution, in the time that the average expected outcome of the candidate solution makes a point estimate of the upper bound of the real optimal. (The fig. 10-12 in the additional material plots the point forecast of the upper and lower solution restrictions. The fig. 10-13 plots the 90% certainty interval on the magnitude of the optimality gap.) These charts demonstrate that the mean value problem presents important bias, but that even with a temperate number of scenarios, and several batches, we are able to produce moderately compact bounds on the true optimal value.

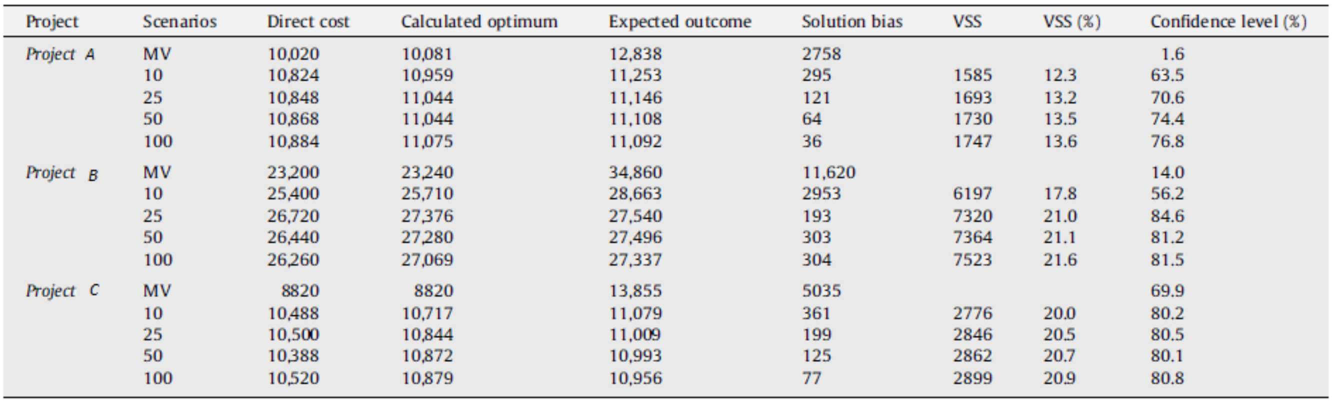
The data suggests that resolving the problem with as few as 25 scenarios gives reasonably good results, while a 50 or 100 scenario model provides us a tighter bound that can be useful when trying to make detailed comparisons between the choices.

VI. COMPERATIVE ANALYSIS

A. *Introduction*

Throughout this article, we have examined a model that includes abandonment and arrival rate uncertainty. None of these conditions is involved in a lot of industry standard models. As observed in Gans et al. (2003), ‘‘general practice uses the M/M/N (Erlang C) queuing model to predict the stationary system performance of short interval” p. 92. According to Fukunaga et al. (2002) is explained a commercial system used at over 800 call centers in which ‘‘agent requisites are calculated by practicing the well-known Erlang-C formula.” Additionally, common industry practice is making staffing decisions based on a time-by-time service level requirement, ‘‘each half hour interval’s estimated λi and μi give rise to a goal staffing level for the period.... determination of an optimal set of schedules may after that be explained as the resolution to an integer program” (Gans et al., 2003), p. 93. In Section 5.2, we demonstrated that ignoring arrival rate uncertainty drives to verifiable more costly solutions, on a wonted cost basis, than models that consider variability. During this section, we analyze the stochastic Erlang A model to the generally applied mean value arrival rate Erlang C model.

Table 5.1: Solution bias and VSS.



Project	Scenarios	Direct cost	Calculated optimum	Expected outcome	Solution bias	VSS	VSS (%)	Confidence level (%)
Project A	MV	10,020	10,081	12,838	2758			1.6
	10	10,824	10,959	11,253	295	1585	12.3	63.5
	25	10,848	11,044	11,146	121	1693	13.2	70.6
	50	10,868	11,044	11,108	64	1730	13.5	74.4
	100	10,884	11,075	11,092	36	1747	13.6	76.8
Project B	MV	23,200	23,240	34,860	11,620			14.0
	10	25,400	25,710	28,663	2953	6197	17.8	56.2
	25	26,720	27,376	27,540	193	7320	21.0	84.6
	50	26,440	27,280	27,496	303	7364	21.1	81.2
	100	26,260	27,069	27,337	304	7523	21.6	81.5
Project C	MV	8820	8820	13,855	5035			69.9
	10	10,488	10,717	11,079	361	2776	20.0	80.2
	25	10,500	10,844	11,009	199	2846	20.5	80.5
	50	10,388	10,872	10,993	125	2862	20.7	80.1
	100	10,520	10,879	10,956	77	2899	20.9	80.8

The common approach explained above creates a set of fixed staffing requirements in every period, and then pursues to find the lowest cost schedule to please these requirements. The integer program that results is a common weighted set covering problem that can be articulated as

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

subject to $\sum_{j\epsilon J} a_{ij} x_j \geq b_i$, □i Є *I*,

x_{ij} Є Z^+,

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

B. *Locally constrained Erlang C model*

We quote the standard approach as the locally constrained Erlang C model because it utilizes Erlang C to produce a hard restriction in every period as described in Gans et al. (2003). The general problem with this approach is the restriction generated by the per-period service level requirement, connected with the requirement to schedule resources in shifts. The maximum staffing level is set by the peak arrival period and, depending on the length of the arrival max and the length of the flexibility of the staffing model, a significant amount of exuberance capacity can be created in other periods because of shift restrictions. The magnitude of the exuberance capacity will be an element of the flexibility of the possible set of schedules. With more flexible staffing alternatives, the weighted set covering algorithm may match the requirement more approximately.

To quantify the impact, we run a locally constrained Erlang C model for each of the three projects for each of the five schedule sets.

The per-period restrictions are set so that the service level with the expected volumes is at least 80% in every 30 minute period, this way ensuring the global SLA of 80% is reached. In Table 6.1, we compare the results of this analysis with the results produced from solving the stochastic program. We resolve every project for each of the five levels of staffing flexibility determined in Table A.2.

The data confirms that the exuberance staffing is high for 5 x 8 staffing, but diminishes fast with more flexible scheduling alternatives. It also demonstrates that this is a more important problem for project A that has a strong seasonality pattern, than for either Project B or C. The set covering approach has the tendency to overstaff the project and obtains expected service levels higher than those achieved in the stochastic model.

Nevertheless, because the set covering model takes into consideration only the expected value and not the variance of arrivals, it is less efficient at hedging versus the stochastic model. Take into consideration the case of schedule D for project B. The deterministic model owns a wonted service level of 86.1%, against the target of 80%, but still has an expected penalty cost of €4820. On the other hand, the stochastic model has an expected service level of 83.5%, 2.6% lower, but an expected penalty only a little higher at €4493.

In each case, the stochastic model produces a lower direct labor cost and a lower expected cost of operation. The benefit of utilizing the stochastic model is most important when arrivals have a strong seasonal pattern, as in Project A, or when workforce flexibility is low. With 5 x 8 only staffing, the stochastic model provides at least 10.8% cut in operating costs.

C. *Globally constrained Erlang C model*

In the earlier section, we demonstrated that the stochastic model based on the Erlang A model gives lower cost solutions than the locally constrained Erlang C model discussed in the literature. Another approach is to utilize a deterministic Erlang C model, ignoring uncertainty and abandonment as in the earlier model, but optimizing to global against local restrictions. While this approach is not present in the literature for what we know, it is a natural simplification of the stochastic model we have examined until now. As the model is deterministic, it takes for granted arrival rates are known, and, it will, generally, be easier to resolve than the stochastic model. Ignoring abandonment will have the tendency to increase recommended staffing, but not considering uncertainty will tend to diminish staffing. It can be the case that under some circumstances these errors will erase each other out, and we can obtain good solutions at a lower computational cost.

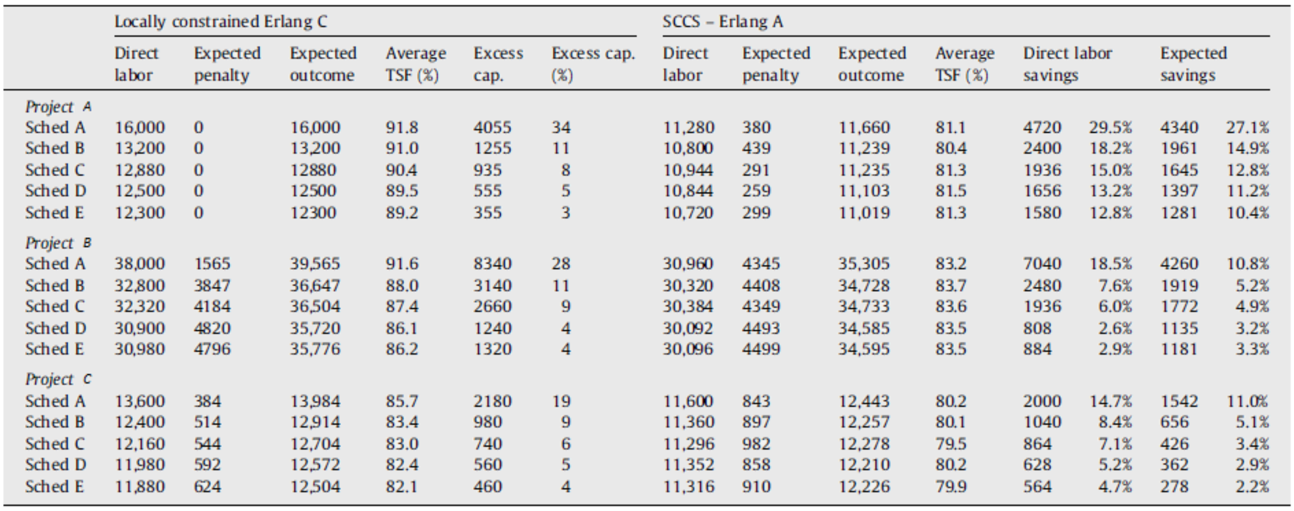
The method for formulating and resolving these problems is an unequivocal implementation of the model (3.1)–(3.7). We resolve a mean value version of the problem. The bigger change is that the coefficients for restrictions (3.3) and (3.5) are calculated based on the Erlang C model. We still need a minimum of two agents staffed at all times and a minimum service level at expected volume in every period of at least 50%.

We resolve this version of the problem for each of the three projects and for each scheduling option. Because the model is deterministic, we do not need to resolve multiple batches. To judge the expected cost of implementing the solution, we go on with the evaluation of the resulting schedule against the stochastic Erlang A model. We take for granted that the Erlang A model with uncertain arrivals is the correct model and the goal of this analysis is to establish the error presented by using a Globally Constrained Erlang C model. The results of this analysis are demonstrated in Table 6.2.

This analysis takes us to several interesting understandings. First of all, the stochastic model surpasses the global Erlang C model in all cases; in some cases this improvement is large and in others it is small. Knowing that the two models are scheduling to a global target, the difference is due to a better hedging strategy. In some cases the stochastic model schedules less hours, other times more.

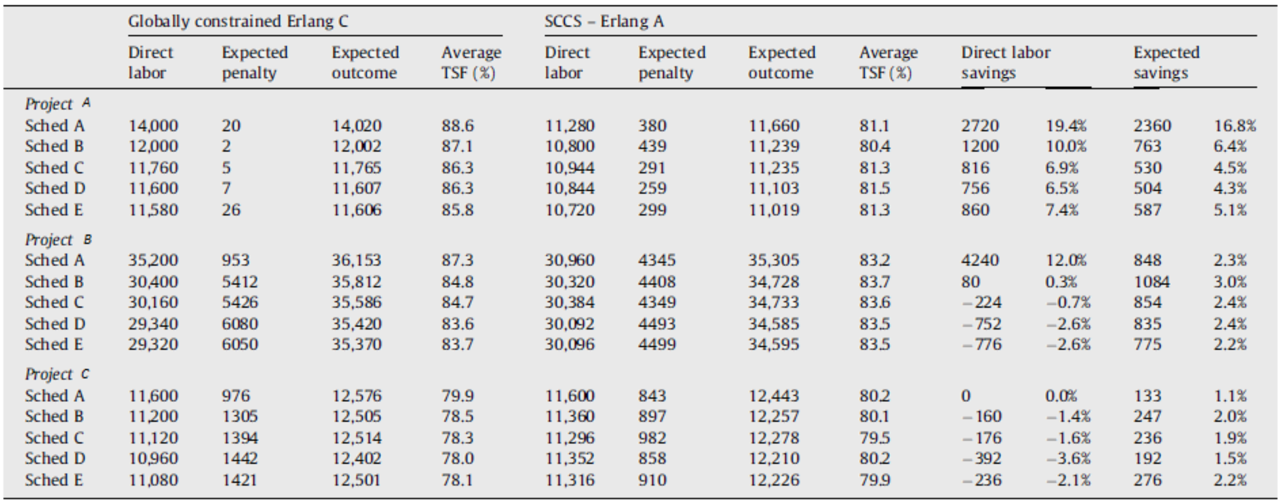
The second awareness is that the Mean Value Globally Constrained Erlang C model does much better than the Mean Value Globally Constrained Erlang A model, still under the hypothesis that the Erlang A model is correct. The GCEC model makes two simplifying

Table 6.1: Comparing the stochastic and local Erlang C schedules.



	Locally constrained Erlang C						SCCS – Erlang A							
	Direct labor	Expected penalty	Expected outcome	Average TSF (%)	Excess cap.	Excess cap. (%)	Direct labor	Expected penalty	Expected outcome	Average TSF (%)	Direct labor savings		Expected savings	
Project A														
Sched A	16,000	0	16,000	91.8	4055	34	11,280	380	11,660	81.1	4720	29.5%	4340	27.1%
Sched B	13,200	0	13,200	91.0	1255	11	10,800	439	11,239	80.4	2400	18.2%	1961	14.9%
Sched C	12,880	0	12880	90.4	935	8	10,944	291	11,235	81.3	1936	15.0%	1645	12.8%
Sched D	12,500	0	12500	89.5	555	5	10,844	259	11,103	81.5	1656	13.2%	1397	11.2%
Sched E	12,300	0	12300	89.2	355	3	10,720	299	11,019	81.3	1580	12.8%	1281	10.4%
Project B														
Sched A	38,000	1565	39,565	91.6	8340	28	30,960	4345	35,305	83.2	7040	18.5%	4260	10.8%
Sched B	32,800	3847	36,647	88.0	3140	11	30,320	4408	34,728	83.7	2480	7.6%	1919	5.2%
Sched C	32,320	4184	36,504	87.4	2660	9	30,384	4349	34,733	83.6	1936	6.0%	1772	4.9%
Sched D	30,900	4820	35,720	86.1	1240	4	30,092	4493	34,585	83.5	808	2.6%	1135	3.2%
Sched E	30,980	4796	35,776	86.2	1320	4	30,096	4499	34,595	83.5	884	2.9%	1181	3.3%
Project C														
Sched A	13,600	384	13,984	85.7	2180	19	11,600	843	12,443	80.2	2000	14.7%	1542	11.0%
Sched B	12,400	514	12,914	83.4	980	9	11,360	897	12,257	80.1	1040	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 6.2: Comparing the stochastic and global Erlang C schedules.



	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 A												
Sched A	14,000	20	14,020	88.6	11,280	380	11,660	81.1	2720	19.4%	2360	16.8%
Sched B	12,000	2	12,002	87.1	10,800	439	11,239	80.4	1200	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 B												
Sched A	35,200	953	36,153	87.3	30,960	4345	35,305	83.2	4240	12.0%	848	2.3%
Sched B	30,400	5412	35,812	84.8	30,320	4408	34,728	83.7	80	0.3%	1084	3.0%
Sched C	30,160	5426	35,586	84.7	30,384	4349	34,733	83.6	−224	−0.7%	854	2.4%
Sched D	29,340	6080	35,420	83.6	30,092	4493	34,585	83.5	−752	−2.6%	835	2.4%
Sched E	29,320	6050	35,370	83.7	30,096	4499	34,595	83.5	−776	−2.6%	775	2.2%
Project C												
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	1305	12,505	78.5	11,360	897	12,257	80.1	−160	−1.4%	247	2.0%
Sched C	11,120	1394	12,514	78.3	11,296	982	12,278	79.5	−176	−1.6%	236	1.9%
Sched D	10,960	1442	12,402	78.0	11,352	858	12,210	80.2	−392	−3.6%	192	1.5%
Sched E	11,080	1421	12,501	78.1	11,316	910	12,226	79.9	−236	−2.1%	276	2.2%

assumptions. First of all, it assumes away abandonment that causes the model to be overstaffed. The model also assumes away arrival rate uncertainty that takes us to understaffing. These two effects tend to compensate each other, indicating it cannot be reasonable to introduce abandonment unless arrival rate uncertainty is also taken into consideration.

VII. CONCLUSIONS AND FUTURE RESEARCH

In this paper, we examined the problem of short term shift scheduling for call centers for which it is important to reach a service level commitment over an extended horizon. While the analysis focused completely on a TSF based SLA, the model could without difficulty be adapted to hold up other forms of an SLA; like the abandonment rate or the average speed to answer. The model was designed to identify the uncertainty in arrival rates and was formulated as a mixed-integer two-stage stochastic program. Even though difficult to resolve, we demonstrated the model is manageable and may be resolved in a moderate amount of time. We also demonstrated that uncertainty is greatly appropriate and that it has an actual impact on scheduling decisions in call centers.

In Section 5.2, we demonstrated the VSS for this model is important; ranging from 12.3% to over 21%. The obvious association is that, for this model formulation, ignoring variability is a costly decision. Nevertheless, most models in practice ignore both uncertainty and abandonment. The implication is that one has not to introduce abandonment into the model without also considering uncertainty. In Section 6.2, we compared this model with the general practice of scheduling to a local Erlang C restriction; which is, scheduling based on a model that does not consider uncertainty and abandonment but requests the service level objective is obtained in each period. Comparing our model to this general practice, we again found our model obtains lower cost results, ranging from 2.4% to 27%. The basic implication in this case is that the Erlang C model sometimes obtains good results, probably because the abandonment and uncertainty assumptions create counter balancing errors. Nevertheless, the stochastic model always obtains a better solution, and in a lot of practical cases the results are considerably better. This is especially true when the flexibility of the workforce is limited to full- or near-full-time shifts and the set covering approach presents abundant loose in the schedule.

At the end, we compared this model to a Globally Constrained Erlang C model. This model provides better results as compared to the local constrained Erlang C, but still our stochastic model exceeds this model in each case, by as little as 1% but by as much as 16%. The general conclusion is that, compared to the alternative methods examined here, the stochastic model gives a lower cost of operation schedule, and sometimes this difference can be crucial. This is a basic property of stochastic programming generally, but in this analysis we have demonstrated that the difference is important in real world cases.

Besides providing a lower solution of the cost, the model described in this paper addresses the scheduling problem from an essentially different perspective. The service level restriction is a hard restriction, in the standard set covering approach, it has to be satisfied and each candidate schedule either obtains the service level requirement or it does not. But, in reality, the service level is a random variable and we will obtain the SLA goal with some probability. Our analysis checks this certainly and addresses the compromise that managers have to make in terms of cost and the confidence of obtaining the service level. Our analysis demonstrates that the cost of operation grows nonlinearly with the desired confidence level. This compromise is confusing in the deterministic setting.

In other future research, this model can be without difficulty expanded to utilize different queuing assumptions, like, that relax the requirement for exponential service times. The compromise of solution precision and computational achievement is an area for future research, too, analyzing the impact of modifying the convergence parameters examined in Section 3.1.

Appendix A. Algorithms and shift patterns

See Figs. A.1–A.4 and Tables A.1 and A.2.

1. Generate a call volumes 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 sum to 1 for each day.
4. Calculate the per-period call volume by multiplying the daily total by the time period proportion.

Fig. A.1. Simulated call generation algorithm.

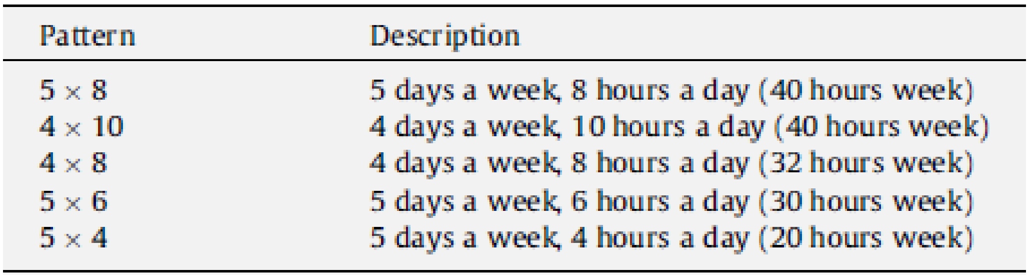
1. Generate a week of call volumes using the algorithm shown in Fig A.1 and calculate the associated per-period arrival rate.
2. For a given call volume select h+1 probability levels for estimating points on the TSF curve. (In practice we use values of .3, .72, .9, .98, and .995 for all periods with call volumes of ate least 5. Different values are used for lower call volumes to maintain a concave approximation.)
3. Calculate the staff level required to achieve the target probabilities defined in Step 2.
4. Recalculate the TSF for the integral staffing level calculated in Step 3. We now have h+1 staff level probability pairs on the TSF curve.
5. Calculate the slope (m_{ikh}) and intercept (b_{ikh}) for each pair of adjacent points found in Step 5.
6. 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.

Fig. A.2. Scenario based TSF approximation approach.

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. Use the resulting vector of staffing requirements as the requirement parameter b_i in the IP constraint (3.2).

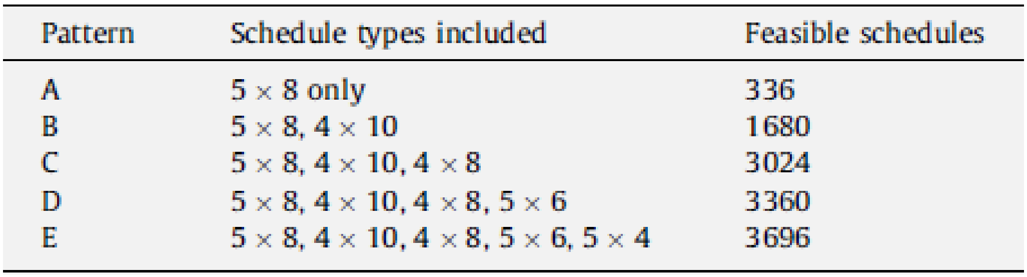
Fig. A.4. Local constraint generation

Table A.1 Shift patterns.



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

Table A.2 Scheduling patterns.



Pattern	Schedule types included	Feasible schedules
A	5 × 8 only	336
B	5 × 8, 4 × 10	1680
C	5 × 8, 4 × 10, 4 × 8	3024
D	5 × 8, 4 × 10, 4 × 8, 5 × 6	3360
E	5 × 8, 4 × 10, 4 × 8, 5 × 6, 5 × 4	3696

References

[1] Aksin, Z., Armony, M., Mehrotra, V., 2007. The modern call-center: A multi-disciplinary perspective on operations management research. Production and Operations Management 16 (6), 665–668.

[2] Andrews, B.H., Cunningham, S.M., 1995. L.L. Bean improves call-center forecasting. Interfaces 25 (6), 1–13.

[3] Atlason, J., Epelman, M.A., Henderson, S.G., 2004. Call center staffing with simulation and cutting plane methods. Annals of Operations Research 127, 333–358.

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

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

[6] Avramidis, A.N., Chan, W., L'Ecuyer, P., 2007a. Staffing Multi-skill Call Centers via Search Methods and a Performance Approximation. University of Montreal.

[7] Avramidis, A.N., Gendreau, M., L'Ecuyer, P., Pisacane, O., 2007b. Simulation-based optimization of agent scheduling in multiskill call centers. In: Fifth Annual International Industrial Simulation Conference (ISC-2007), Delft, The Netherlands.

[8] Aykin, T., 1996. Optimal shift scheduling with multiple break windows. Management Science 42 (4), 591–602.

[9] Baron, O., Milner, J.M., 2006. Staffing to maximize profit for call centers with alternate service level agreements. Operations Research 57 (3), 685–700.

[10] Bassamboo, A., Harrison, J.M., Zeevi, A., 2005. Design and control of a large call center: Asymptotic analysis of an LP-based method. Operations Research 54 (3), 419–435.

[11] Bayraksan, G., Morton, D.P., 2009. Assessing solution quality in stochastic programs via sampling. In: Informs 2009 Tutorials in Operations Research, pp. 102–122.

[12] Bechtold, S.E., Jacobs, L.W., 1990. Implicit modeling of flexible break assignments in optimal shift scheduling. Management Science 36 (11), 1339–1351.

[13] Birge, J.R., 1982. The value of the stochastic solution in stochastic linear programs, with fixed recourse. Mathematical Programming 24, 314–325.

[14] Birge, J.R., Louveaux, F., 1997. Introduction to Stochastic Programming. Springer, New York.

[15] Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Haipeng, S., Zeltyn, S., Zhao, L., 2005. Statistical analysis of a telephone call center: A queueing-science perspective. Journal of the American Statistical Association 100 (469), 36–50.

[16] Brusco, M.J., Jacobs, L.W., 1998. Personnel tour scheduling when starting-time restrictions are present. Management Science 44 (4), 534–547.

[17] Brusco, M.J., Jacobs, L.W., 2000. Optimal models for meal-break and start-time flexibility in continuous tour scheduling. Management Science 46 (12), 1630–1641.

[18] Brusco, M.J., Johns, T.R., 1996. A sequential integer programming method for discontinuous labor tour scheduling. European Journal of Operational Research 95 (3), 537–548.

[19] Cezik, M., L'Ecuyer, P., 2007. Staffing multiskill call centers via linear programming and simulation. Management Science 54 (2), 310–323.

[20] Chen, B.P.K., Henderson, S.G., 2001. Two issues in setting call centre staffing levels. Annals of Operations Research 108 (1), 175–192.

[21] Dantzig, G.B., 1954. A comment on Edie's "Traffic delays at toll booths". Journal of the Operations Research Society of America 2 (3), 339–341.

[22] Ekmekçiu, 2015. Optimizing a call center performance using queueing models – an Albanian Case. 5th International Conference - "Compliance of the Standards in South-Eastern European Countries with the Harmonized Standards of European Union", 15-16 June, 2015 Peja, Republic Of Kosovo.

[23] Fukunaga, A., Hamilton, E., Fama, J., Andre, D., Matan, O., Nourbakhsh, I., 2002. Staff scheduling for inbound call centers and customer contact centers. In: Eighteenth National Conference on Artificial Intelligence, Edmonton, Alberta, Canada.

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

[25] Garey, M.R., Johnson, D.S., 1979. Computers and Intractability: A Guide to the Theory of NP-completeness. W.H. Freeman, San Francisco.

[26] Geoffrion, A.M., 1970. Elements of large-scale mathematical programming: Part I: Concepts. Management Science 16 (11), 652–675 (Theory Series).

[27] Green, L.V., Kolesar, P.J., Soares, J., 2001. Improving the SIPP approach for staffing service systems that have cyclic demands. Operations Research 49 (4), 549–564.

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

[29] Henderson, W.B., Berry, W.L., 1976. Heuristic methods for telephone operator shift scheduling: An experimental analysis. Management Science 22 (12), 1372–1380.

[30] Koole, G., van der Sluis, E., 2003. Optimal shift scheduling with a global service level constraint. IIE Transactions 35, 1049–1055.

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

[32] Mapo online, Call-center: Biznesi që vijon lulëzimin

[33] Milner, J.M., Olsen, T.L., 2008. Service-level agreements in call centers: Perils and prescriptions. Management Science 54 (2), 238–252.

[34] Pinedo, M., 2005. Planning and Scheduling in Manufacturing and Services. Springer, New York, NY.

[35] Robbins, T.R., 2007. Managing Service Capacity Under Uncertainty. Unpublished PhD Dissertation, Pennsylvania State University, 240p. http://personal.ecu.edu/robbinst/ (accessed 01.04.10).

[36] Robbins, T.R., Medeiros, D.J., Dum, P., 2006. Evaluating arrival rate uncertainty in call centers. In: Proceedings of the 2006 Winter Simulation Conference, Monterey, CA.

[37] Segal, M., 1974. The operator-scheduling problem: A network-flow approach. Operations Research 22 (4), 808–823.

[38] Steckley, S.G., Henderson, W.B., Mehrotra, V., 2004. Service System Planning in the Presence of a Random Arrival Rate. Cornell University.

[39] Steckley, S.G., Henderson, S.G., Mehrotra, V., 2009. Forecast errors in service systems. Probability in the Engineering and Informational Sciences (23), 305–332.

[40] Thompson, G.M., 1995. Improved implicit optimal modeling of the labor shift scheduling problem. Management Science 41 (4), 595–607.

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

[42] Whitt, W., 2006. Staffing a call center with uncertain arrival rate and absenteeism. Production and Operations Management 15 (1), 88–102.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.