Staffing and shift-scheduling of call centers under call arrival rate uncertainty

Shuang Qing Liao

ÉCOLE CENTRALE DES ARTS ET MANUFACTURES
« ÉCOLE CENTRALE PARIS »

THÈSE

présentée par

LIAO Shuangqing

pour l'obtention du

GRADE DE DOCTEUR

Spécialité : Génie Industrielle

Laboratoire d'accueil : Génie Industrielle

SUJET : Dimensionnement des Centres d'Appels avec Incertitude sur les Paramètres d'Arrivées

soutenue le : 1 Juillet 2011

devant un jury composé de :

LISSER Abdel	**Président**
KARAESMEN AKSIN Zeynep	**Examinateurs**
VIAL Jean-Philippe	
DALLERY Yves	
KOOLE Ger	
VAN DELFT Christian	
JOUINI Oualid	

2011ECAP0027

Staffing and Shift-Scheduling of Call Centers under Call Arrival Rate Uncertainty

shuangqing LIAO

May 5, 2011

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Call centers have become more and more important for many large organizations. Brown et al. (2002) report that in 2002 more than 70% of all customer-business interactions were handled by call centers. They also report that call centers in the U.S. employ more than 3.5 million people, i.e., 2.6% of the workforce. Due to the importance of this industry, considerable literature has focused on the operations management of call centers, in particular on the following issues: demand forecasting, quality of service and call routing (often using queueing theory), and staffing and agents shift scheduling (using combinatorial optimization). We refer the reader to the comprehensive surveys of Gans et al. (2003) and Aksin et al. (2007). A central feature in call centers is the significant uncertainty in the number and length of calls or on the effective number of available agents. This randomness leads to performance measures which deviate from those predicted at the moment of planning (see Avramidis et al. (2004); Harrison and Zeevi (2005); Whitt (2006); Robbins (2007) and Green et al. (2007)).

Comparing to the traditional manufacturing operations, the service capacity of a call center mainly depends on the quantity and skills of human resources available, either through a direct employment, subcontracting, or through cooperation with other service firms. This heavy dependence on human resources implies that the Manpower Planning Problem is a key challenge for managers.

The Manpower Planing Problem for call centers contains three parts: the *resource acquisition*, the *resource deployment* and real-time updating and call routing (Aksin et al. (2007)). The resource acquisition determines the quantity and the time to hire the agents by a long-term view of demand for services. The resource deployment schedules the available agents based on a short-term forecasts of demand for services. After that the resource deployment decisions have been made, in real time, it is also possible to make shorter-term decisions like forecast updating, schedule updating, and real-time call routing.

Due to the complexity of the process of hiring and training agents which requires long lead times, resource acquisition decisions should be made several weeks and sometimes months ahead of time. Resource deployment decisions requires less lead time, and are usually made several weeks before the actual arrive of calls. The challenge of resource deployment plan is to minimize the cost, and at the same time closely match the supply with the uncertain demand of agent resources. Many studies of call forecasts show that both of the call arrival distributions and service time distributions vary over time (Aksin et al. (2007) and Gans et al. (2003)), consequently the demand for resources is highly variable. Both forecasting and queueing models are therefore important in modeling resource deployment decisions. A third activity which plays important role in resource deployment decisions is the scheduling planning, which determines the number of agents assigned to a range of shifts. The process of determining an optimal (or near-optimal) schedule is well known to have a significant combinatorial complexity. This is our major concern in this dissertation. Once the decisions of resource acquisition and resource deployment have been made, for a given day or week, some new information about forecasts and agent availability become available. One can use these new elements to update the call volume forecast and the agents schedules. This problem is also analyzed in this dissertation. Finally, at the time when calls arrive, queueing policies and real-time call routing are used to assign calls to the appropriate agents.

In this dissertation, we consider the staffing-scheduling problem. The staffing-scheduling planning problem aims to build an agent schedule that minimizes costs while achieving some predefined quality of service objectives. By dividing the scheduling horizon into several periods, for example periods with 30 minutes each, staffing decisions are usually made to determine the target staffing level for each period. These targets depend on both the quantity of work arriving (as estimated by the call volume forecasts), the duration (the forecasted mean service times) and the quickness the call center seeks to serves these customers (estimated by some function of the customer waiting time distribution). Once the forecasts and waiting time goals have been established, staffing formulations such as queueing performance evaluation models and simulation models are used to determine the targeted number of service resources, typically on a period-by-period basis. Taking the targeted number of service resources as inputs, giving the definition of shifts, the scheduling problem determines an optimal set of agents numbers for each shift. The traditional way to solve the scheduling problem is to formulate and solve a mathematical program to identify a minimum cost schedule while achieving the target staffing level or other labor requirements.

In what follows, we motivate our work and describe its contribution to practice and to the

literature of call centers.

1.1 Motivation

Call centers have emerged as the primary vehicle for firms to interact with consumers, transforming consumer service jobs once characterized by variety and personal relationships into routinized and high speed operations. Call centers are used to provide services in many areas and industries: banks, insurance companies, emergency centers, information centers, help-desks, tele-marketing and more. Technological development has allowed remote service delivery using various channels of telecommunication. The definition of a call center is continuously changing, but the core fundamentals of a customer making a call (via an inbound call or outbound call by phone, email, web site, fax or Interactive Voice Response) to a center (collection of resources) will remain constant. Call center, contact center or customer interaction center operate on identical principals of meeting customer needs in real-time or near real-time. Here, we consider a call center dealing with inbound calls and back office jobs such as emails.

Call centers can be broadly classified into two types: call centers with equally skilled agents and homogeneous calls and call centers with multiple queues and agent skills. Our concern in this thesis is single-skilled call centers.

A large quantity of literature, both in the context of call centers (see references in Gans et al. (2003)) and in more general contexts (see Ernst et al. (2004)), has extensively studies the scheduling problem. In the context of call centers, the arrival process of calls is usually assumed to be Poisson. In most of the existing scheduling problem models, the overall arrival rate is assumed to be known. However, Gans et al. (2003) point out that this is typically not the case. Rather, the arrival rate is predicted from historical data and the forecasts of arrival rate are not exact. Due to insufficient historical data upon which reliable estimates can be based, and unpredictable factors such as weather conditions, the arrival rate is not known with certainty, this is called *parameter uncertainty*. The authors in Gans et al. (2003) say “It can be risky to ignore arrival-rate uncertainty" and “Surprisingly, however, there is little work devoted to an exploration of how to accommodate uncertainty".

In Aksin et al. (2007), the authors also underline that reconsidering the scheduling problems under the more general assumption that arrival rates are random variables is very promising area that is just now beginning to receive attention from researchers.

In this dissertation, we consider the shift scheduling problems of a call center, in which we allow the mean arrival rate of calls to be uncertain. We model in this whole dissertation the arrival process of calls by a doubly non-stationary stochastic process, with random mean arrival

rates.

1.2 Description and Contribution

The goal of the present thesis is to contribute to the operations management research of call centers. We aim to enhance our understanding of such complex systems, so as we gain useful guidelines for the practitioners. We specifically address the analysis of four problems that take into account the important feature of uncertainty in the call arrival parameters. In what follows, we briefly provide the description of the models under consideration.

1. **Blending Single Shift Scheduling Problem**: In the first model, we consider a multi-period staffing problem in a single-shift call center. The call arrival process is assumed to follow a doubly non-stationary stochastic process with a random mean arrival rate. We consider a setting in which there exists some flexibility to modify in real-time (within the same day) the instantaneous capacity dealing with inbound calls. The alternative work for the employees is to handle the day's workload of back-office jobs. The flexibility arises from the fact that back-office jobs, which can be viewed as storable, can be answered at any time of the day, but they have to be treated within the same day, in overtime if necessary. The inbound calls in our model should be handled (almost) immediately, using a standard service level constraint (on average at least a given fraction of customers should wait less than a given threshold of time). This constraint has to be satisfied on a period-by-period basis. After closing the inbound calls channel, agents can recourse to work on overtime hours in order to handle eventual unfinished back-office jobs.

   The staffing problem is modeled as a cost optimization-based newsboy-type model. The cost criterion function includes the regular and overtime salary cost, a penalty cost for excessive waiting times for inbound calls. Our objective is to find the optimal staffing level which minimizes the total call center operating cost. We propose two solution methodologies: the stochastic and robust programming, to solve this problem, and give managerial insight on the trade-off between operation cost and service quality. Also, we analyze the impact of the flexibility offered by back-office workloads. We show that combining the two types of jobs offers flexibility, partially absorbing the undesirable effects of uncertainty in the arrival parameters.

2. **Multi Shifts Scheduling Problem with Recourse**: In the second model, we consider a multi-periodic multi-shift call center staffing problem, which decides an initial schedules before the beginning of the working day and allows real-time recourse actions to adjust

the initially scheduled staffing levels in reaction to realized deviations from arrival-rate forecasts.

In most of call center environment, scheduling decisions are typically taken one or two weeks ahead of time. However, many random elements such as the call arrival rate reveal only until that day has begun. It is then important to make real-schedule adjustment even though very little attention has been devoted to this issue, either for call centers or other types of service systems. We construct a two-stage model with the first-stage decision as initial schedules and the second-stage decision allowing the manager to make adjustment by increasing or decreasing staffing levels. The objective is to minimize the sum of the regular salary, the update adjustment cost and the penalty cost of agents shortfall (under-staffing).

We analyze a special case where all shifts are without break, thanks to which we find some very important and interesting property of the two-stage recourse problem. Two different solution approaches are considered: the classic two-stage stochastic program with recourse, and the modified robust optimization method with discrete recourse decisions. The efficiency and excellent performance of these two approaches are analyzed theatrically and illustrated through a numerical study based on real-life data. We also analyze the added advantage of using dynamic adjustment (update). We show that the update action reduces the operational cost and the under-staffing probability.

3. **Distributionally Robust Optimization Problem**: In the third model, besides the parameter uncertainty on arrival rate as presented above, we consider an additional type of uncertainty: the uncertainty on the probability distribution of a random parameter. The traditional way to take into account the parameter uncertainty is to assume a known probability distribution of this random parameter. However, in practice, the exact probability distributions are often unknown. In this model, we consider the case where the probability distribution of the random parameter is ambiguous and belongs to some probability distribution set.

   We consider a multi-periodic multi-shift static call center shift-scheduling problem. A random number of agents related with the ambiguous probability distribution is required to handle the inbound calls in each period. The assigned agents number is allowed to be less than that required, i.e., under-staffing, but the expected total under-staffing for the whole day should not exceed a certain limit, even for the worst case probability distribution belonging to the considered probability distribution set. The objective is to minimize the agents salary under condition of respecting the expected total under-staffing limit.

We propose an approach combining stochastic programming and distributionally robust optimization to optimize the operation cost, and show the necessity of taking into account the uncertainty on probability distribution of random parameters.

4. **Staffing-Scheduling Problem with Global Service Level and Update**: To the contrary of all the above models which consider period to period service levels, we consider in this last model a global service level objective for the whole day.

   For the above three models, we assumes that service level goals, or the targeted number of service resources, are *hard constraints* that must be met during each period. One could use the Erlang formula to determine the staffing level for each period, and then schedule for each shift a certain number of agents to obtain the staffing level for each period. However, the model with global service level combines both the staffing level determination and the shift scheduling steps. In general, comparing the two types of models, the one with hard constraints leads to more overcapacity in certain intervals than the model with global service level.

   We consider a multi-periodic multi-shift call center staffing problem with possibility to adjust the staffing level during the day. The call center's operational cost includes the initial staffing salary cost and the adjusted staffing cost. The achieved global service level is allowed to be less than the target global service level, but the expected service level shortfall should not exceed a certain limit. We construct two models to describe this problem and analyze the efficiency and performance of these two models. We then conduct a comparison study of different models: the above two models, the model with hard constraints and staffing update, and the static one with global service level constraint and without staffing update. The comparison shows the advantages of adding the update flexibility, and points out the impact of having a global service level constraint.

In Chapter 2 we provide the industrial background on call center workforce management and the basic mathematical tools dealing with uncertainty optimization. Also we review the relevant literature related to this thesis. In Chapters 3-6, we give the analysis of the four problems presented above, respectively. The manuscript ends with general conclusions and future research.

Chapter 2

Background and Literature Review

This chapter introduces and defines the different concepts used in this thesis, and reviews the related literature.

In Section 2.1, we give a brief introduction of the call center workforce management, such as the call center operation system, the staffing-scheduling problem and some possible ways to offer staff flexibility. We emphasize the uncertainty in arrival rate of inbound calls, and the definition of a special shift-setting which leads to total unimodular period-shift matrix.

In Section 2.2, we collect the main mathematical optimization facts and results on uncertainty optimization, upon which we build our results in the sequel. The most common way to treat uncertainty is stochastic optimization. A more recent approach called *Robust Optimization* takes a deterministic set-based view of uncertainty in optimization thus remains highly tractable. We present also the results in recourse in stochastic optimization and robust optimization.

Lastly, We review the literature related to the staffing-scheduling problem in call center management.

2.1 Call Center Workforce Management

2.1.1 Introduction to Call Center Operations

Many companies and organizations, in public and private sector, use call centers to communicate with their customer relationships. For some of the companies, such as banks and cellular operators, their call centers are the main channel for maintaining contact with their customers. In general, call centers are becoming a vital part of the service-driven society nowadays. As a result call centers have also become an object for academic research.

Besides the agents, call centers are also equipped with computers and telecommunication equipment which enable to the delivery of services via the telephones. Figure 2.1 provides a general architecture of the equipment. An inbound call connects from the public service telephone network (PSTN) to the call center's privately owned switch, the private automatic branch exchange (PABX or PBX), though a number of telephone lines often called trunk lines. At first, calls may be connected thought the PABX to an Interactive Voice Response unit (IVR) where the caller can use her keypad to select options and potentially provide data input to call center systems. If callers need to speak to an agent, the calls are handed from the IVR to an automatic call distributor (ACD). The ACD is a specialized switch to route calls from PABX to individual agent. Modern ACDs are highly sophisticated and they can monitor agent status, collect data, manage on hold queues and make complex routing decisions based on various criteria. Particulary in call centers employing *skills based routing*, the decision process to match callers and agents can turn out to be quite complex. In addition, Computer-telephone integration (CTI) can be used to improve the efficiency of routing process and agents work by using the callers' record information. In more sophisticated settings, CTI is used to integrate the customer relationship management (CRM) system, which track callers records and allow them to be used in operating decisions, such as suggesting cross-selling (Gans et al. (2003)).

Figure 2.2 depicts an operational scheme of a simple call center as a queueing system. The trunk lines connect calls to the center while a group of agents serve incoming calls. An arriving call that find all the trunk lines occupied receives a busy signal and is blocked from entering the system. Otherwise it is connected to the call center and occupies one of the free trunk lines. If some of the agents are available, the call is served immediately. Otherwise, it waits in queue for an agents to become available. Callers who become impatient hang up, or abandon, before getting into service. Some of the blocked and abandoned calls become retrials that attempt to reenter service. The remaining of them are lost. Finally, it is also possible that served caller may return to the system (Gans et al. (2003)).

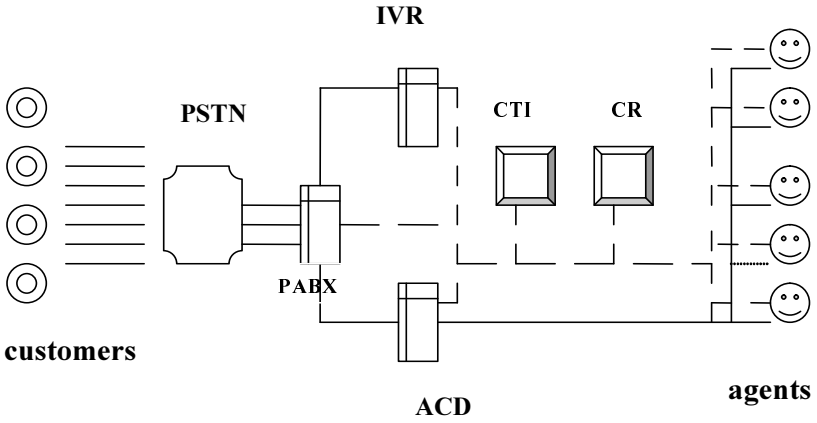

Figure 2.1: Schematic diagram of call-center technology

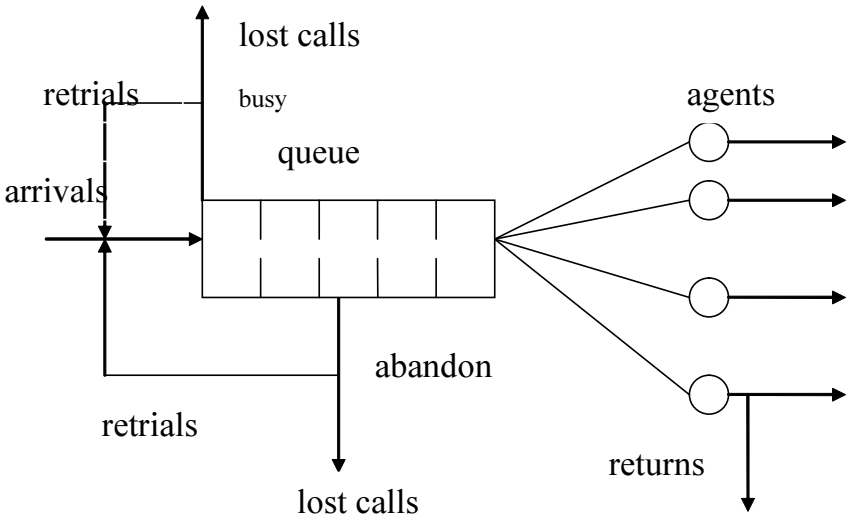

Figure 2.2: Operational scheme of a call center

Service quality is an important and complex issue related to evaluate the call center management. The notions of service quality most commonly tracked and managed by call centers could be the accessibility of agents (how long the callers have to wait to speak to an agent), the effectiveness of service encounters (whether need to rework), and the content of agents interaction with the callers (Gans et al. (2003)). A standard quality of service constraint is that ensures that $SL\%$ of customers wait less than AWT seconds, i.e., $\mathrm{P}\{\text{Wait} \leq AWT\} \geq SL$. For example, the 80/20 rule, where at least 80% of customers wait in queue less than 20 seconds. The service quality can be defined for a month, a day, an hour or a period even shorter.

The staffing cost is a major component in the operating costs of call centers. In Gans et al. (2003), the authors indicate that it represents 70% of the labor cost. An efficient staffing is thereafter crucial. Unfortunately, the uncertainty of arrivals make the staffing problem difficult. In addition to the usual uncertainty modelled by a stochastic process, there is indeed also, as mentioned in Chapter 1, uncertainty in the process parameters. Another source of uncertainty is the absenteeism of agents which highly affects the efficiency of the before-hand planned agents in order to meet the quality of service constraints. In this dissertation, we do not consider the latter type of uncertainty. We only consider the uncertainty in the inbound call arrival process parameter while allowing it to be non-stationary, i.e., varying in the day time. Most call center models in the literature assume a known, fixed arrival rate and ignore the issue of arrival rate uncertainty. In Gans et al. (2003), the authors say “Surprisingly, however, there is little work devoted to an exploration of how to accommodate uncertainty".

2.1.2 Uncertain and Non-Stationary Arrival Rate

Several characteristics of the arrival process of calls have been underlined in the recent call center literature. First, it has been observed that the total daily number of calls has an overdispersion relative to the classical Poisson distribution. Second, the mean arrival rate considerably varies with the time of day. Third, there is a strong positive correlation between arrival counts during the different periods of the same day. We refer the reader to Avramidis et al. (2004) and Brown et al. (2005) for more details.

In order to address uncertain and time-varying mean arrival rates coupled with significant correlations, we model the inbound call arrival process by a doubly stochastic Poisson process (see Avramidis et al. (2004); Harrison and Zeevi (2005), and Whitt (1999)) as follows. We assume that a given working day is divided into n distinct, equal periods of length T, so that the overall horizon is of length nT. The period length in practice is often 15 or 30 minutes. The mean arrival rate of calls during period i is denoted by Λ_i and is random. The stochastic process describing

the cumulative number of arrivals up to time t is defined by

$$A(t) = M(\sum_{i=0}^{\kappa} T\Lambda_i + (t - T\kappa)\Lambda_{\kappa+1} : \kappa = \lfloor t/T \rfloor), \tag{2.1}$$

where $M = (M(t) : 0 \leq t < \infty)$ is the unit rate Poisson process, and $\Lambda = (\Lambda_i : 0 \leq i \leq n)$ is the sequence of arrival rates, with $E[\sum_{i=1}^{n} \Lambda_i] < \infty$. By conditioning on an outcome of the average arrival rate in a given period, say λ, the process $A(\cdot)$ is therefore a rate-λ Poisson process during that period. Furthermore, using the modeling in Avramidis et al. (2004) and in Whitt (1999), we assume that the arrival rate Λ_i is of the form

$$\Lambda_i = \Theta f_i, \text{ for } i = 1, ..., n, \tag{2.2}$$

where Θ is a positive real-valued random variable. The random variable Θ can be interpreted as the unpredictable *busyness* of a day. A large (small) outcome of Θ corresponds to a busy (not busy) day. The constants f_i model the shape of the variation of the arrival rate intensity across the periods of the day. Formally, let us denote a sample value, for a given day, of the random variable Θ by the positive real value θ. Then, the corresponding replication of the arrival rate over period i for that day is $\lambda_i = \theta f_i$.

Using the data of a Dutch hospital, we determine the f_is of Monday by averaging on all Mondays of a year. In Figure 2.3, we plot in solid line f_i as a function of period i, and also plot in dashed line two examples of not busy and busy days. The example of the Dutch hospital we consider in this dissertation agrees with the observed experience in call centers. In most call centers we indeed have very significant time-of-day seasonality. The arrival rate at the beginning and at the end of the day is quite low. It ramps up sharply in the morning and tends to dip down around the lunch break, but a second lower peak occurs in the afternoon. Although there is a significant stochastic variability in the arrival rate from one day to another, there is a strong seasonal pattern across the periods of a given day.

2.1.3 The Call Center Staffing and Scheduling Problem

The call centers staffing and scheduling problem consists on building an agent schedule that minimizes costs while achieving some customer waiting time distribution objective. An efficient schedule balances firm and individual goals and constraints. As such, the key input of this staffing scheduling problem is the targeted staffing levels, which depend on the forecasted arrival call volume, the forecasted mean service time and the objective distribution of costumer waiting time. Simulation models and analytic queueing models(exact or approximation) are the two

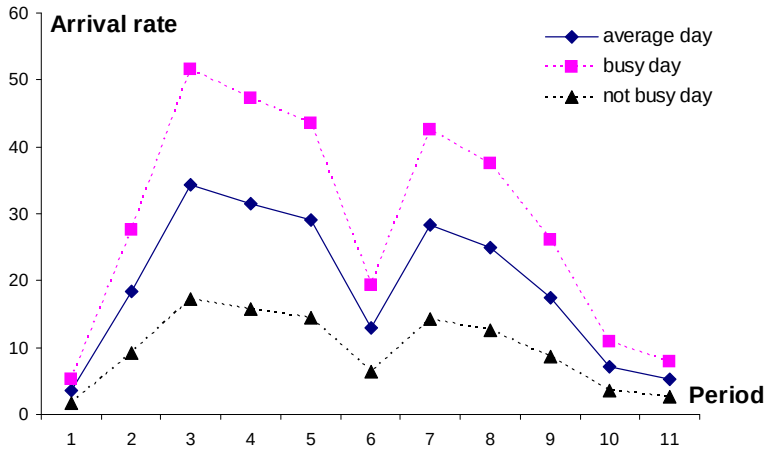

Figure 2.3: Arrival rate diagram

alternatives to evaluate performance. And they can be used to determine the target staffing levels once the forecasts on mean arrival rates and mean service time, the waiting time goals are established.

Staffing problem

Queueing analytical models are used to determine how many agents must be available to serve calls over a given period. The simplest M/M/N (Erlang C) queueing model is widely used to estimate stationary system performance of short-half-hour or hour-periods. A standard service level constraint is introduced for each time period, through which the waiting time is kept in convenient limits. For period i, let the random variable WT_i denote the waiting time of an arbitrary call. The probability distribution of the waiting time of calls is computed using the classical results of the Erlang C model. In doing so, the mean arrival rates and service rates are assumed to be constant in each period of the day. Also, it is reasonable to assume that the system achieves a steady state quickly within each period. It is known (see for example Gross and Harris (1998)) that for a given staffing level N which only handle inbound calls, one has for period i,

$$\begin{aligned} P\{WT_i \leq AWT \,|\, \theta\}(N) &= 1 - \left(\sum_{k=0}^{N-1} \frac{(\theta\, f_i/\mu)^k}{k!} + \frac{(\theta\, f_i/\mu)^N}{N!\,(1 - \frac{\theta\, f_i/\mu}{N})} \right)^{-1} \frac{(\theta\, f_i/\mu)^N}{N!\,(1 - \frac{(\theta\, f_i/\mu)}{N})}\, e^{-(N\mu - \theta\, f_i)\, AWT} \\ &= F_{\theta\, f_i}(N), \end{aligned} \tag{2.3}$$

where AWT represents the Acceptable Waiting Time (for example 20 seconds). For a given value of the objective service level in period i, say $SL_i\%$, and a given sample value of the arrival

rate, $\theta\, f_i$, this formula is used in the reciprocal way in order to compute the staffing level which guarantees the required service level,

$$N_i(\theta\, f_i) \quad = \quad F^{-1}_{\theta\, f_i}(SL_i). \tag{2.4}$$

Although the Erlang C formula is widely used and easily implemented, it does not give an intuitive insight on the size of agents number required and it can turn out to be highly inaccurate if underlying assumptions are violated (Gans et al. (2003)). A well known approximation of the Erlang C for heavy-traffic regimes, those in which agents utilization is high, is the square-root safety staffing approximation. Given the arrival rate λ_i for period i and service rate μ, this implies that the system's offered load in this period is given by $R_i = \frac{\lambda_i}{\mu}$. In their pioneering paper, Halfin and Whitt (1981) showed that when the offered load R_i is high, and an appropriate number of agents are employed, a system can achieve a high agent utilization and yet deliver a good service level by choosing the number of servers called square-root safety staffing $R_i + \beta\, \sqrt{R_i}$, where β is some fixed service grade related to a target delay probability $\mathrm{P}(\mathrm{WT}_i > 0)$ by the following expression:

$$P(WT_i > 0) \approx P(\beta) = [1 + \frac{\beta\, \Phi(\beta)}{\phi(\beta)}]^{-1}. \tag{2.5}$$

In the equation above, Φ and ϕ are the cumulative distribution and density functions of the standard normal distribution (mean=0, variance=1), respectively.

With this square-root safety staffing formula, the agents are highly utilized, answering calls almost 100% of the time. On the other hand, a large fraction of customers should receive no or just a small amount of waiting. This form of staffing gives rise to the so-called Quality and Efficiency Driven (QED) regime that has been extensively studied in the literature; see for example Borst et al. (2004) and the survey paper by Gans et al. (2003).

Erlang B is a loss model which assumes that the number of lines is equal to the number of agents (no queues) and incorporates blocking of the customers. Eralng B is often used to calculate the number of lines required in order to achieve a desired blocking probability. While Erlang C is a direct result of the assumption of zero abandonment, it is further developed to incorporate customer impatience in the Erlang A system. Detail reviews of queueing models of call centers are available in Koole and Mandelbaum (2002).

Shifts-scheduling Problem

Taking the results from the staffing problems as input, the shifts-scheduling problem deals with the assignment of a specific number of agents to detailed shifts. A shift denotes a set of periods during which an agent works over the course of the day.

The shifts-scheduling problem has been extensively analyzed in the literature, dating back to the set covering problem modeled by Dantzig (1954). Let J be the set of all the feasible work schedules, each of which dictates if an agent answers calls in period $i \in I$. For $i \in I$ and $j \in J$, we define the $|I| \times |J|$ matrix $\mathbf{A} = [a_{ij}]$, where

$$a_{ij} = \begin{cases} 1, & \text{if agents in schedule } j \text{ answer calls during period } i, \\ 0, & \text{otherwise.} \end{cases}$$

Each agent assigned to shift j gets a salary c_j for the day. Letting the decision variables x_j, $j \in J$ represent the numbers of agents assigned to the various shifts and let N_i, $i \in I$, denote the required number of agents of each period. The set covering problem can be expressed as

$$\begin{aligned} \text{Min} \quad & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} \quad & \mathbf{A}\,\mathbf{x} \geq \mathbf{N} \\ & \mathbf{x} \in \mathbb{Z}. \end{aligned} \tag{2.6}$$

The optimal solution of Equation (2.6) defines the number of agents assigned to each shift, and for each period the constraint on available agents is respected.

This shift-scheduling problem is known to be NP complete unless that each shifts is continuous with no breaks. It can be very difficult to solve for problems with many shifts and periods. In practice, these scheduling problems are not solved to optimality. The reasons are first the integer nature of the decision variables, and second the presence of breaks in the middle of the shifts.

A particular case is that each agent works over consecutive periods, without breaks. Then every column of matrix $\mathbf{A}$ has contiguous ones and this kind of matrix is totally unimodular. Totally unimodular matrices are of extreme importance in polyhedral combinatorics and combinatorial optimization. It is well known that if matrix $\mathbf{A}$ is totally unimodular and vector $\mathbf{N}$ is integral, every extreme point of the feasible region $\{\mathbf{x} \mid \mathbf{A}\mathbf{x} \geq \mathbf{N}\}$ is integral and thus the feasible region is an integral polyhedron. This implies that the linear program (LP) of Equaiton (2.6) is integral (has an integral optimum, when any optimum exists).

2.1.4 Flexibility in Staffing

Call centers may benefit from flexible staffing, i.e., the ability to adjust from one period to another staffing levels (and/or schedules) with observed traffic. Such flexibility may be attained using temporary operators, in addition to the permanent ones always available to provide service. The temporary operators may be either supervisors/managers or other operators who are on call. Another type of flexibility corresponds to the presence of different shifts for the operators. By combining such shifts, the operators capacity can be aligned with the time-period varying average workload. The third type of flexibility concerns an objective in terms of a global service level. It allows to achieve low service level during some periods of the day, and to achieve higher ones during the other periods. A last type of flexibility corresponds to the presence of blending of jobs, i.e., dealing with different types of jobs, with different admissible qualities of service. The key flexibility of problems with jobs blending comes from the fact that less urgent calls (as e-mails or calls with a possible callback) can be inventoried to some extent, to the contrary to other more urgent calls.

Blending different types of jobs

One of the strategies hedging against parameter uncertainty is to provide flexibility to change the staffing level upon short notice in response to unanticipated change in demand, as discussed in Whitt (1999). Flexibility staffing can be achieved by ensuring that the staff of the call center have alternative work. The forms of alternative works are various, such as training, after-call processing of previous calls and making outbound calls. But Whitt (1999) limits in obtaining reliable estimates of the mean and variance of the demand in the near future, in order to adapt the necessary staff.

Gans et al. (2003) present additionally the multimedia. Differences among media are deeper than differences among calls: One important difference is the natural time scales at which the various media must be responded to. Typically, telephone calls should be served immediately (within seconds or minutes), and should not be interrupted once started. E-mail and fax, on the other hand, can be delayed for hours or days. This time natural difference lead one to consider a blending center with inbound calls which requires high priority and emails or fax with lower priority. During the time intervals, agents who might be idle can become productive by handling low-priority work.

In Chapter 3 we construct a model with multimedia: the inbound call and the alterative work represented as emails which could be handled by idle agents. And we analyze the advantages of flexibility. In Robbins et al. (2008, 2007) the authors introduce an improvement on the overall

operating characteristics of the queueing system by adding a relatively small portion of cross trained workforce. This is an example of blending calls.

Real-time schedules adjustments

Another type of flexibility comes from real-time schedules adjustments which are made after agents have been hired and trained and agents schedules have been created. These adjustments are made on an intra-day basis to agents' schedules, once additional information about call volumes, absenteeism and all other activities such and training and meeting, have become available.

Mehrotra et al. (2010) point out four reasons why real-time schedules adjustment helps for successful call center management. We simply quote the following sentences:

"Most importantly, several researchers (Avramidis et al. (2004), Brown et al. (2005),Jongbloed and Koole (2001a),Shen and Huang (2008),Steckley et al. (2004) and Weinberg et al. (2007)) have recently identified significant correlation between arrivals in different time intervals within the same day, and have suggested methods for updating call forecasts on an intra-day basis; a primary purpose for such updated call forecasts is to provide support for real-time schedule adjustments. Secondly, given the lead time associated with schedule generation, many changes to employee availability can and do take place after the original schedules have been created. Thirdly, detecting how well the scheduled agent workforce actually matches the actual workload is often not possible for a given day until that day has begun, at which point responding to the incremental (positive or negative) demand may be crucial. Finally, managers regularly struggle with staffing tradeoffs, for while having too few agents on duty can lead to severe degradations in service quality, having too many agents results in low resource utilization and overspending of scarce financial resources."

In Chapter 4 we construct a model with information update and real-time schedules adjustment. Different to the two relevant work existing (Gans et al. (2009) and Mehrotra et al. (2010)), we employ totally unimodular property to obtain interesting result for the solving process. Moreover, we modify and apply the adjustable approach for this real-time schedules adjustment problem.

Global service level

Generally, papers treating agent scheduling problems consider service level constraints period by period, which is referred to as *hard constraint* by Koole and van der Sluis (2003). Using Erlang formula to determine the required staffing level for each time interval, and solving a setting cover integer problem, the solution leads in general to overcapacity in certain intervals. On average,

the service level is higher than required, and as a consequence the number of scheduled shifts is higher than necessary.

Koole and van der Sluis (2003) is the first to consider a global service level constraint, which is called *soft constraint*, where a larger number of employees in one time interval can compensate a shortage in another interval. The flexibility involves in allowing that intervals with a low service level to be compensated by intervals with high service levels: the objective is to reach on average at the correct service level.

In Chapter 6 we construct a model with a global service level constraint, in which the real-time schedules adjustments are allowed.

2.2 Uncertainty Optimization

Uncertainty can take two forms: (i) the parameters are constant but unknown. The randomness comes form the errors produced by our estimation of the parameter values, and (ii) the parameters themselves vary as a function of the states conditions and the decisions taken beforehand. In this thesis, we consider only the fist case, in which for a multi-stage problem, the random parameters are independent of the decisions of previous stages.

The filed of decision-making under uncertainty was pioneered by Dantzig (1955) and Charnes and Cooper (1959b) in the 1950's. They set foundation respectively for stochastic programming and optimization under probabilistic constraints. Both of these classes of problems share the same assumption that the probability distribution of the random variables are known exactly. Another stream of research call robust optimization which is pioneered by Soyster (1973), addresses the decision-making problem under uncertainty but without this probability assumption. In robust optimization, random variables are modelled as uncertain parameters belong to a convex uncertainty set and the decision-maker protect the system against the worst case in that set. In this section, we survey the primary research, on both stochastic programming and robust programming optimization.

Uncertainty can come in many different forms, and hence it is possible to model it in various ways. In a mathematical approach one formulates an objective function $f : \mathbb{R}^n \to \mathbb{R}$ which should be optimized (say minimized) subject to specified constraints. That is, one formulates a mathematical programming problem:

$$Min_{\mathbf{x}\in\mathbf{X}} f(\mathbf{x}). \tag{2.7}$$

The feasible set $\mathbf{X} \in \mathbb{R}^n$ is defined by a number of constraints $\{\mathbf{x} \in \mathbb{R}^n : g_i(\mathbf{x}), i \in I\}$. Inevitably

the objective and constraint functions depend on parameters, denoted by vector $\widetilde{\varepsilon} \in \mathbb{R}^d$. This indicates that $f(\mathbf{x}, \widetilde{\varepsilon})$ and $g_i(\mathbf{x}, \widetilde{\varepsilon}), i \in I$, can be viewed as functions of the decision vector $\mathbf{X} \in \mathbb{R}^n$ and the parameter vector $\widetilde{\varepsilon} \in \mathbb{R}^d$.

In the usual set-up, the parameters defining the optimization, are known deterministically. Even if we can find the optimal solution for this problem (for integer optimization, it is not guaranteed to get an optimal solution), the solution is not designed to be robust in perturbation in the feasible set. Ben-Tal and Nemirovski (2000) shows that even small perturbations of parameters make the usual optimal solution completely meaningless from a practical point of view. Stochastic programming and robust optimization, using different models and solution techniques, immune the problem against parameter uncertainty. In what follows, we discuss these two main streams of research.

2.2.1 Stochastic Optimization

Math programs which explicitly incorporate uncertainty in parameter values are known as stochastic programs. The notion of stochastic programming was first introduced in the 1950s, see Dantzig (1955).

Overview and Scenarios Construction

When the parameters are uncertain, but assumed to take values in some given set of possible values, one may seek to find solutions which are feasible for all choices of possible parameters and optimize a given objective function. Stochastic programming models assumes that the probability distributions governing the data are known or can be estimated. The objective is to find decisions which are feasible for all (or almost all) the possible parameter realizations and optimize the expectation of some function of the decisions and the random variables.

Coming back to Problem (2.7), suppose that $\widetilde{\varepsilon}$ takes values in set Ξ, with corresponding probability distribution P. We then formulate the following stochastic programming problem:

$$Min_{\mathbf{x}\in\mathbf{X}}\mathbb{E}[f(\mathbf{x})] = Min_{\mathbf{x}\in\mathbf{X}} \int_{\Xi} f(\mathbf{x}, \widetilde{\varepsilon}) dP(\widetilde{\varepsilon}). \tag{2.8}$$

A possible justification of this approach is as follows. If the process repeats itself, by the Law of Large Numbers, for a given solution $\mathbf{x}$, if we are supposed to solve the same problem under the same probability distribution many times, the average of the total cost will converge to the expectation $\mathbb{E}[f(\mathbf{x})]$. In that case Formulation (2.8) gives a best possible solution on average.

Next, we give a short review about the scenarios construction. In general, it is preferable that the number of constructed scenarios is relatively modest so that the obtained (linear) problem

can be solved within reasonable computational effort. A standard approach to generate scenarios is by discretization. That is, one discretes the continuous probability distributions into a finite number of points $\{\widetilde{\varepsilon}_k \in \Xi,\, k \in K\}$ with K the set of points. A positive weight p_k is assigned to each $\widetilde{\varepsilon}_k$, with $\sum_{k\in K} p_k = 1$. The discretized set $\{\widetilde{\varepsilon}_k, k \in K\}$ and the corresponding probabilities $\{p_k, k \in K\}$ can be viewed as a representation of the underlying probability distribution. With respect to this distribution, the integral Problem (2.8) is approximated by

$$Min_{\mathbf{x}\in\mathbf{X}}\mathbb{E}[f(\mathbf{x})] = Min_{\mathbf{x}\in\mathbf{X}} \sum_{k\in K} p_k\, f(\mathbf{x}, \widetilde{\varepsilon}_k). \tag{2.9}$$

Suppose that the components of the random vector $\widetilde{\varepsilon} \in \mathbb{R}^d$ are independent from each other. We construct scenarios by discretizing the probability distributions of each components into M possible values. Then the total number of scenarios is M^d. Such exponential growth of the number of scenarios makes the method of discretization very difficult, even for reasonable size (see Shapiro (2008)). The discretization approach may still work for two, may be three, or even four (parameter components) variables. It is not possible to use such a discretization approach with more than 10 variables, which would lead to a *curse of dimensionality.*

An alterative approach that reduces the scenarios to a manageable size is the Monte Carlo Simulation. Assume that the total number of scenarios is very large or even infinite, due to the exponential growth in the number of random parameters. Assume further that it is possible to generate a random sample $\{\widetilde{\varepsilon}^1, \widetilde{\varepsilon}^2, ..., \widetilde{\varepsilon}^N\}$ of N replications of the random vector $\widetilde{\varepsilon}$. That is, each sample $\widetilde{\varepsilon}^j, j = 1, .., N$, has the same probability distribution as $\widetilde{\varepsilon}$. Moreover, assuming that $\widetilde{\varepsilon}^j, j = 1, .., N$, are independent, the corresponding sample average function is

$$\hat{f}_N(\mathbf{x}) = N^{-1} \sum_{j=1}^{N} f(\mathbf{x}, \widetilde{\varepsilon}^j). \tag{2.10}$$

The function $\hat{f}_N(\mathbf{x})$ is random since it depends on the generated random sample. It is an approximation of the expected function $\mathbb{E}[f(\mathbf{x})]$ when $N \to \infty$ by the Law of Large Numbers. This motives to introduce the sample average approximation (SAA) problem:

$$Min_{\mathbf{x}\in\mathbf{X}}\{\hat{f}_N(\mathbf{x}) = N^{-1} \sum_{j=1}^{N} f(\mathbf{x}, \widetilde{\varepsilon}^j)\}. \tag{2.11}$$

Note that once the sample is generated, Problem (2.11) becomes a problem of the form of (2.9), with scenarios $\widetilde{\varepsilon}^j, j = 1, .., N$, and identical probabilities $p_j = \frac{1}{N}, j = 1, .., N$.

Two basic varieties of stochastic programs are chance constrained programs and recourse

programs. Chance constrained programs implement constraints with some confidence level, the higher the confident level is, the lower the probability of violation of constraint would be. Recourse programs on the other hand recognize two types of decisions, which are decisions that occur before uncertainty is revealed: (i)the (here and now) first-stage decisions, and (ii) decisions that occur after uncertainty has been revealed: the (wait and see) recourse decisions.

Recourse Programs

Recourse problems have been widely analyzed in the literature. A brief tutorial type introduction is provided in Higle (2005). Several excellent texts are also available that outline the structure and solution approaches for stochastic programming. Kall and Wallace (1994) is an excellent introduction that includes a survey of various solution techniques and algorithms. Birge and Louveaux (1997) is a thorough review of linear and non-linear stochastic programming, while Kall and Mayer (2005) focuses strictly on stochastic linear programs.

For a two stage stochastic recourse problem, the recourse decisions evolve the second time horizon. Adopting the notation from Birge and Louveaux (1997), the general stochastic linear programming problem can be expressed as

$$\begin{aligned} Min \quad & \mathbf{c}^T \mathbf{x} + E_\xi[\min \mathbf{q}(\omega)^T \mathbf{y}(\omega)] \\ \text{s.t} \quad & \mathbf{A}\,\mathbf{x} \geq \mathbf{b} \\ & \mathbf{T}(\omega)\,\mathbf{x} + \mathbf{W}\,\mathbf{y}(\omega) \geq \mathbf{h}(\omega) \\ & \mathbf{x} \geq 0, \mathbf{y}(\omega) \geq 0. \end{aligned} \tag{2.12}$$

The objective of the stochastic linear program (2.12) is to minimize the cost of the first-stage decision, plus the expected cost of the recourse decisions. The optimization is constrained by the first set of constraints that depend only on the deterministic first-stage variables, and the second set of constraints that depend on the recourse decisions ($\mathbf{y}(\omega)$) and may have random components. The stochastic program is typically solved relative to a finite set of scenarios, sample draws of the random vector ξ. If the number of sample outcomes is denoted by K, with probability p_k for each scenario, then we can write the stochastic program in extensive form as

$$\begin{aligned} Min \quad & \mathbf{c}^T \mathbf{x} + \sum_{k=1}^{K} p_k\, \mathbf{q}_k^T\, \mathbf{y}_k \\ \text{s.t} \quad & \mathbf{A}\,\mathbf{x} \geq \mathbf{b} \\ & \mathbf{T}_k\,\mathbf{x} + \mathbf{W}\,\mathbf{y}_k \geq h_k, k = 1, ..., K \\ & \mathbf{x} \geq 0, \mathbf{y}_k \geq 0, k = 1, ..., K. \end{aligned} \tag{2.13}$$

The extensive form program (2.13) is the deterministic equivalent of (2.12) with a finite set of outcomes, and as such can be written as a large linear program. The program can then be solved using the standard simplex algorithm for linear programs. However, as the number of realizations increases the size of the program can be quite large and difficult to solve. Some approach are proposed for solving large scale stochastic programs, such as the L-Sharped decomposition, Benders decomposition, stochastic decomposition, etc. We shall present in detail the Benders' decomposition later.

As for the multi-stage stochastic optimization, the recourse decisions evolve over some (usually finite) stages. This is even more complicate than the two stage recourse stochastic problem. Then multi-stage model can be expressed as:

$$\begin{aligned} Min \quad & \mathbf{c_1}^T \mathbf{x_1} + E_{\xi_2}[\min \mathbf{c_2}^T \mathbf{x_2}(\omega_2) + ... + E_{\xi_H}[\min \mathbf{c_H}^T \mathbf{x_H}(\omega_H)]] \\ \text{s.t} \quad & A\,\mathbf{x_1} \geq \mathbf{b_1} \\ & \mathbf{T}_1(\omega)\,\mathbf{x_1} + \mathbf{W}_2\,\mathbf{x_2}(\omega_2) \geq h_2(\omega) \\ & ... \\ & \mathbf{T}_{H-1}(\omega)\,\mathbf{x_{H-1}} + \mathbf{W}_H\,\mathbf{x_H}(\omega_H) \geq h_H(\omega) \\ & \mathbf{x_1} \geq 0, \mathbf{x_t}(\omega_t) \geq 0, t = 2, ..., H. \end{aligned} \tag{2.14}$$

In the above model (2.14), the decision $\mathbf{x_t}$ made in each stage t depend on the realized information vector (ω_t), which contains all the information observed in previous stages. A common way to represent these realizations is via a scenario tree. A scenario tree is a graph with a single root node at level 0, and branches to a series of nodes at level 1, with each node representing a possible realization of ω in this period. Again each node branches to a series of nodes (finite number) at the successor level. And each node has a single predecessor node. It is obvious that the size of the scenario tree grows very quickly. Suppose the stage number is T, with R_t realizations at each stage, the total number of scenarios of this scenario tree is

$$N = \Pi_{t=1}^{T} R_t \tag{2.15}$$

The size of scenarios increases non-linearly and grows very large either when the stage number or the scenario number of each stage are big. Figure 2.4 shows an example of a scenarios tree.

There are many studies on how to construct such scenarios trees in a reasonable and meaningful way. One possible approach is to use Monte Carlo technique to generate scenarios by

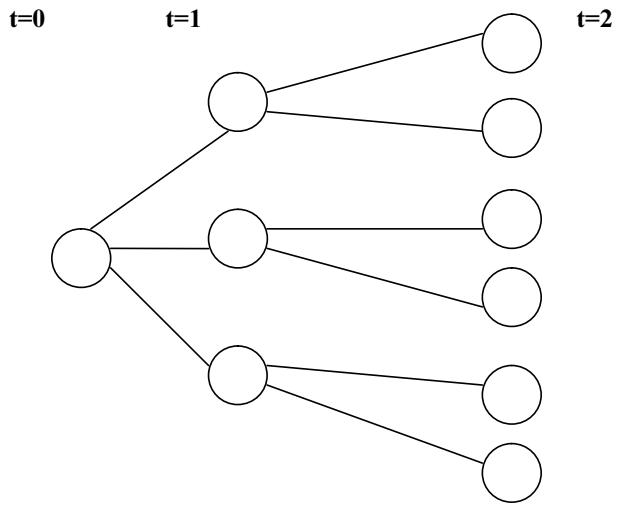

Figure 2.4: Scenario tree example

conditional sampling. This is an extension of SAA method to multi-stage setting. Noting that the total number of scenarios N is the product of constructed realization number R_t at each stage. In order to avoid the explosion of the number of scenarios, and reduce it to a manageable level, one may take $R_t = 1$ from a certain stage on. This means that we relax the assumptions that the parameters are random from this stage on. The reason for doing so, it is that for the multi-stage stochastic program, the most important is to compute the first-stage optimal decision. After that the first-stage decision has been applied, more information about the random parameter are available, the managers could re-optimize the decisions of the follow-up stages by using new information.

For a large-scaled multi-stage linear problem, several decomposing and partitioning techniques are proposed in the literature. In what follows, we give a summary of one of the well known techniques, namely, the Benders decomposition.

Benders' decomposition (Benders (1962)) is a well-known approach for solving two-stage linear models and also combinatorial optimization problems. There are many examples of successful applications of this methodology. For example, the large scale water resource management problem (Cai et al. (2001)) and the two-stage stochastic linear problem (Zhao (2001)).

The main idea of Benders' composition is to partition the model into two simpler problems: a master problem and a subproblem. The master problem is a relaxed version of the original problem, containing only a subset of the original variables and the associated constraints. The variables obtained in the master problem takes fixed values in the subproblem. They are used as linking variables.

Consider the following general formulation in order the illustrate the main idea of the Benders'

decomposition:

$$\begin{aligned} \text{Minimize} \quad & \mathbf{c^T x + d^T y} \\ \text{s.t.} \quad & \mathbf{A x + B y \geq b}, \\ & \mathbf{D y \geq e}, \\ & \mathbf{x, y \geq 0}. \end{aligned} \tag{2.16}$$

For the combinatorial optimization problem, Benders' decomposition also works for the case where $\mathbf{x}$ and $\mathbf{y}$ are continuous and integer decision vectors, respectively. Vectors $\mathbf{c}$ and $\mathbf{d}$ are associated with costs. Matrices $\mathbf{A}, \mathbf{B}, \mathbf{D}, \mathbf{b}$ and $\mathbf{e}$ are of appropriate dimensions. This problem can be written as:

$$\text{Min}_{\bar{\mathbf{y}} \in \mathbf{Y}} \ \{\mathbf{d^T y} + \text{Min}_{x \geq 0} \ \{\mathbf{c^T x} : \mathbf{A x \geq b - B \bar{y}}\}\}, \tag{2.17}$$

where $\mathbf{Y} = \{\mathbf{y} | \mathbf{D y \geq e}, \mathbf{y \geq 0}\}$. The Benders' decomposition subproblem is

$$\begin{aligned} \text{Min} \quad & \mathbf{c^T x} \\ & \mathbf{A x \geq b - B \bar{y}}, \\ & \mathbf{x} \geq 0. \end{aligned} \tag{2.18}$$

The dual version of this problem is

$$\begin{aligned} \text{Max} \quad & \mathbf{u^T (b - B \bar{y})} \\ & \mathbf{u A \leq c}, \\ & \mathbf{u} \geq 0. \end{aligned} \tag{2.19}$$

Let $\mathcal{F}$ the feasible set of the dual maximization problem (2.19). It should be noticed that this feasible set is independent of the values of $\mathbf{y}$. We assume that $\mathcal{F}$ is not empty for it would correspond to a primal problem either infeasible or unbounded. $\mathcal{F}$ is therefore composed of extreme points $\mathbf{u}^p$ for $(p = 1...P)$ and extreme rays r^q for $(q = 1...Q)$.

The solution of the dual problem (2.19) can be either bounded or unbounded. A feasible primal problem ends a bounded dual problem, in which case the solution is one of the extreme points $\mathbf{u}^p (p = 1...P)$. An unfeasible primal problem leads to the unbounded dual problem. In such a situation, there is a direction $\mathbf{r}^q$ for which $\mathbf{r}^q (\mathbf{b - B \bar{y}}) \geq 0$. This situation should be

avoided. A group of constraints

$$\mathbf{r}^q\,(\mathbf{b}-\mathbf{B}\,\bar{\mathbf{y}})\leq 0, q=1...Q, \tag{2.20}$$

restricts this unbounded situation for the dual maximum problem. With this restrictions, the maximum value of the dual problem falls on one of the extreme points of $\mathcal{F}$.

Since the primal and dual formulations can be interchanged according to duality theory, (2.17) can be rewritten as

$$\text{Min}_{\bar{\mathbf{y}}\in\mathbf{Y}}\ \{\mathbf{d^T}\,\mathbf{y}+\text{Max}_{u\geq 0}\ \{\mathbf{u}^T\,(\mathbf{b}-\mathbf{B}\,\bar{\mathbf{y}}):\mathbf{u}\,\mathbf{A}\leq\mathbf{c}\}\}, \tag{2.21}$$

which is equivalent to

$$\begin{aligned}\text{Min}_{\bar{\mathbf{y}}\in\mathbf{Y}}\quad & \{\mathbf{d^T}\,\mathbf{y}+\text{Max}\ \{\mathbf{u}^p\,(\mathbf{b}-\mathbf{B}\,\bar{\mathbf{y}}):p=1...P\}\},\\ \text{s.t.}\quad & \mathbf{r}^q\,(\mathbf{b}-\mathbf{B}\,\bar{\mathbf{y}})\leq 0, q=1...Q.\end{aligned} \tag{2.22}$$

Adding an auxiliary continuous variable z, the Benders' reformulation of (2.22) is:

$$\begin{aligned}\text{Min}\quad & \mathbf{d^T}\,\mathbf{y}+z\\ \text{s.t.}\quad & z\geq\mathbf{u}^p\,(\mathbf{b}-\mathbf{B}\,\bar{\mathbf{y}}),\ \ p=1...P,\\ & \mathbf{r}^q\,(\mathbf{b}-\mathbf{B}\,\bar{\mathbf{y}})\leq 0,\ \ q=1...Q,\\ & \mathbf{y}\in\mathbf{Y}, z\geq 0.\end{aligned} \tag{2.23}$$

The number of extreme points and extreme rays is usually extremely large. It is difficult to find out them all at once. Benders' proposes to generate the extreme points and extreme rays by iteration. Initially, only the following simplified master problem is solved:

$$\begin{aligned}\text{Min}\quad & \mathbf{d^T}\,\mathbf{y}+z\\ \text{s.t.}\quad & \mathbf{y}\in\mathbf{Y}, z\geq 0.\end{aligned} \tag{2.24}$$

This problem is a relaxed version of (2.23) and therefore the objective value is a lower bound to the original problem. The optimal value of $\bar{y}$ by solving this problem is used in the subproblem (2.18) or equivalently (2.19). This subproblem is solved and the results are either unbounded or bounded, consequently, an extreme ray $\mathbf{r}^q$ or an extreme point $\mathbf{u}^p$ is found. The sum of $\mathbf{d^T}\,\bar{\mathbf{y}}$ and the objective value of the subproblem (2.18) gives an upper bound of the original problem.

The master and the subproblem are solved iteratively, until the upper and lower bounds are

sufficiently close. The Benders' Decomposition algorithm can be stated as:

ALGORITHM OF BENDERS' DECOMPOSITION()

Initialization: ε (a small value), $\bar{y}$ (initial feasible integer solution), $LB(-\infty)$, $UB(\infty)$ $q = 0$, $p = 0$,,

If $UB - LB \geq \varepsilon$ Then

solve Problem (2.19)

If unbounded then

$q = q + 1$, get unbounded ray r^q,

add cut $\mathbf{r}^q(\mathbf{b} - \mathbf{B}\bar{\mathbf{y}}) \leq 0$ to master problem

Else

$p = p + 1$, get extreme point u^p,

add cut $z \geq \mathbf{u}^p(\mathbf{b} - \mathbf{B}\bar{\mathbf{y}})$ to master problem

$UB := \min\{UB, \mathbf{d^T}\bar{\mathbf{y}} + \mathbf{u}^p(\mathbf{b} - \mathbf{B}\bar{\mathbf{y}}\})$

end if

solve master problem

$LB := \mathbf{d^T}\mathbf{y} + z$.

end while

END ALGORITHM.

Chance Constraints Programs

An alternative aspect of stochastic programming is chance constraints programs introduced by Charnes and Cooper (1959a). To the contrary to the aspect of multi-stage models, chance constraints programs focus on constraints violation probability. A chance constrained model of a single stage problem is defined as follows.

$$\begin{aligned} \text{Min} \quad & \mathbf{c}^T\mathbf{x} \\ \text{s.t.} \quad & P(\mathbf{A}^T(\widetilde{\varepsilon})\,\mathbf{x} \geq \mathbf{b}(\widetilde{\varepsilon})) \geq 1 - \alpha \\ & \mathbf{x} \in \mathbf{X}, \end{aligned} \tag{2.25}$$

where $\mathbf{X}$ is a feasible set of x. In model(2.25), the probability that all the m linear constraints are jointly feasible is required to be more than $1 - \alpha$, with $\alpha \in [0, 1]$. Let m be the dimension of vector $\mathbf{b}(\widetilde{\varepsilon})$. If $m = 1$ then Problem (2.25) is called individual chance constrained problem; otherwise, it is called joint chance constrained problem. It is obvious to see that the latter is

more complicated than the former.

The intension of chance constraints is reasonable, in order to avoid the solutions to be too conservative and to make a trade off between the violation of constraints and the objective quality. Chance constrained problems are computationally intractable in general, see Nemirovski and Shapiro (2006). The only exception is by assuming that the random parameters following a multivariate normal distribution. Then the optimization model of an individual chance constrained problem becomes a second-order cone problem, which is tractable.

Consider the objective of model (2.25), let us put the objective into a chance constraint as follows,

$$s(\alpha) = \min\{s|\mathbf{Pr}\{\mathbf{c}^T\,\mathbf{x} \leq s\} \geq 1-\alpha)\}. \tag{2.26}$$

The critical threshold $s(\alpha)$ guarantees that the objective $\mathbf{c}^T\,\mathbf{x}$ is smaller than $s(\alpha)$ with a probability higher than $1-\alpha$. This is the definition of the popular risk measure in finance called Value at Risk (VaR), which is the maximum loss not exceeded with a given probability defined as the confidence level, over a given period of time. VaR is coherent only when it is based on the standard deviation of normal distributions. And it is difficult to optimize when calculated by generating scenarios, since VaR is non-convex, non-smooth and has multiple local extremes in this case.

An alternative and more attractive way to take into account the risk consists on bounding the conditional value at risk (CVaR), see Rockafellar and Uryasev (2002). CVaR is more consistent because of its sub-additivity and convexity. Ogryczak and Ruszczynski (2002) mention that the CVaR is a coherent risk measure which is computationally tractable in the framework of stochastic programming.

The CVaR is defined by

$$CVaR_\alpha(\mathbf{x}) = E(\mathbf{c}^T\,\mathbf{x}|\mathbf{c}^T\,\mathbf{x} \geq s(\alpha)). \tag{2.27}$$

Rockafellar and Uryasev (2002) proved that the minimization of $CVaR_\alpha(\mathbf{x})$ w.r.t. the decision vector $\mathbf{x}$ is the solution of a simple minimization problem given by

$$\rho_{1-\alpha}(\mathbf{c}^T\,\mathbf{x}) = \min_{\mathbf{x}} CVaR_\alpha(\mathbf{x}) = \min_{\mathbf{x},s}\{s + \alpha^{-1}E[\mathbf{c}^T\,\mathbf{x} - s)^+]\}. \tag{2.28}$$

Furthermore, the right-hand side of the optimization problem (2.28) is jointly convex in $(\mathbf{x}, s)$ if the cost function $\mathbf{c}^T\,\mathbf{x}$ is convex in $\mathbf{x}$.

From the definition, it is clear that VaR is always smaller or equal to CVaR, then minimizing CVaR also leads to near minimization of VaR. Similarly to CVaR minimization on objective

function, we can include CVaR in a constraint by adding the following constraint:

$$\rho_{1-\alpha}(\mathbf{b} - \mathbf{a}^T \mathbf{x}) \leq 0. \tag{2.29}$$

Nemirovski and Shapiro (2006) has established that the CVaR constraint (2.29) is the tightest convex approximation of the following individual chance constraint

$$P(\mathbf{b}(\widetilde{\varepsilon}) - \mathbf{a}^T(\widetilde{\varepsilon})\,\mathbf{x} \leq 0) \geq 1 - \alpha. \tag{2.30}$$

The CVaR constraints can be approximated by sampling average approximations, but its solutions may not be a safe approximation for the chance constraint. The same to the classical stochastic programming, sampling approximation requires full knowledge of the distribution of random parameters, which would not be always available. Chen et al. (2010) propose upper bounds for the CVaR constrains by using robust optimization on a varieties of uncertainty set. Thus the chance constrained problem is approximately solved.

2.2.2 Robust Optimization

Stochastic programming is widely used as a strong modelling tool when an accurate probability description of the random is available. However, the decision-maker does not necessarily have this perfect information in real-life application. A more recent approach to optimization under uncertainty, in which the uncertainty model is not stochastic, but rather deterministic and set based, is pioneered by Soyster (1973). In this work, the author proposes a linear optimization model to construct a solution which is feasible for all data belonging to a convex set. This model was widely deemed being too conservative in the sense that too much of optimality is lost in order to ensure robustness. In the late 1990s, a significant step forward for developing a theory for robust optimization was taken by research teams led by Ben-Tal and Nemirovski (1998, 1999, 2000) and Ghaoui and Lebret (1997); Ghaoui et al. (1998). These papers addressed the issue of overconservative by restricting the uncertain parameters to belong to ellipsoidal uncertainty set, which involves solving the robust counterpart of the nominal problem in the form of conic quadratic problem. A draw-back of the robust modelling framework with ellipsoidal uncertainty sets is that it increases the computational complexity of the problem. For example, the robust counterpart of a linear programming problem is non-linear, although it is a convex problem. More lately, Bertsimas and Sim (2003, 2004) and Bertsimas et al. (2004) propose a robust optimization approach based on polyhedral uncertainty sets. To the contrary to the approaches with ellipsoidal uncertainty sets, the robust counterpart they propose are linear optimization

problems. This approach is thus generalized to discrete optimization problems. The origin of robust optimization deals with static problems where all the decision variables are determined before any of the uncertainty parameters are realized. Ben-Tal et al. (2004) first extended the robust optimization to dynamic setting, where the decision-maker adjusts his strategy according to information revealed. They suggested an approximation to the linear adjustable robust counterpart called affinley adjustable robust counterpart. More recently, Bertsimas and Caramanis (2010) propose the finite adaptability which treats problems in which the second-stage decisions are discrete.

This section outline the main aspects of robust optimization. The first part focuses on the introduction of models and solving techniques concerning static robust optimization with polyhedral uncertainty sets. The second part describes the new results on the direction of dynamic robust optimization, incorporating the fact that information is revealed in stages.

Static Robust Optimization

Firstly, we present the robust optimization framework when all decisions should be decided before (or without) knowing the exact value taken by the uncertain parameters. We address later the case where the decision-maker can adjust his decisions according to the revealed information.

Given an objective $f_0(\mathbf{x})$ to optimize, subject to constraints $f_i(\mathbf{x}, \mathbf{u_i}) \geq 0$ with uncertain parameters, $\{\mathbf{u_i}\}$, the general optimization formulation is:

$$\begin{aligned} \min \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}, \mathbf{u_i}) \geq 0, \ \forall \mathbf{u_i} \in U_i, i = 1, ..., m, \end{aligned} \tag{2.31}$$

with $\mathbf{x}$ as vector of decision variables, $f_0, f_i : \mathbb{R}^n \to \mathbb{R}$ as functions, and uncertainty parameters $\mathbf{u}_i \in \mathbb{R}^k$ taking arbitrary values in the closed uncertainty sets $\mathbf{U}_i \in \mathbb{R}^k$.

The goal of (2.31) is to find the minimum cost solution $\mathbf{x}$ which is feasible for all realizations of the uncertain parameter $\mathbf{u}_i \in \mathbf{U}_i$. Computational tractability is an important issue of the robust optimization. In general, the robust version of a tractable optimization problem may not itself be tractable. The tractability of the robust counterpart of a problem depends on the structure of the nominal problem as well as the class of uncertainty set. It is well-known that some classes of optimization problems, including LP, QCQP, SOCP, SDP, and some discrete problems as well, have a robust counterpart formulation that is tractable. It is important to take care of the choice of the uncertainty set to ensure that tractability is preserved.

We now give a summary of the LP robust optimization with a polyhedron uncertainty set $\mathcal{A}$, as proposed by Bertsimas et al. (2004). Consider the following nominal linear optimization

problem:

$$\begin{aligned} \min \quad & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} \quad & \mathbf{A}\,\mathbf{x} \geq \mathbf{b}, \\ & \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{2.32}$$

Without loss of generality, we assume that the uncertainty affects only the matrix $\mathbf{A}$, and $\mathcal{X}$ is a polyhedron not subject to uncertainty. Noting that if the coefficient $\mathbf{c}$ in the objective is affected by uncertainty, we can simply add the constraint $Z - \mathbf{c}^T \mathbf{x} \geq 0$ and use the objective minimize Z. The reformulated problem then has the form of (2.32).

Feasibility of the solutions is a fundamental issue of this problem with random parameters. One feature of robust optimization is that it guarantees that every constraint is satisfied for any possible value of $\mathbf{A}$ in a given convex set $\mathcal{A}$. This leads to the following robust counterpart of Problem (2.32):

$$\begin{aligned} \min \quad & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} \quad & \mathbf{a}_i^T \mathbf{x} \geq \mathbf{b}_i, \ \forall\, i, \forall\, \mathbf{a}_i \in \mathcal{A}, \\ & \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{2.33}$$

where $\mathbf{a}_i$ is the ith vector of $\mathbf{A}$. It is more difficult to solve the robust problem (2.33) than the nominal one (2.32), since the former requires the minimum value of $\mathbf{a}_i^T \mathbf{x}$ with $\mathbf{a}_i \in \mathcal{A}$ be still equal or bigger than $\mathbf{b}_i$.

The uncertainty set $\mathcal{A}$ is defined as follows. In order to keep simplicity, it is assumed that each coefficient a_{ij} of the matrix $\mathbf{A}$ is affected by the uncertainty, and is modelled as a symmetric and bounded random variable $\tilde{a}_{ij}$ that takes values in $[\bar{a}_{ij} - \hat{a}_{ij}, \bar{a}_{ij} + \hat{a}_{ij}]$. It is also assumed that all the coefficients are independent from each other. The scaled deviation z_{ij} is defined as $z_{ij} = \frac{\tilde{a}_{ij} - \bar{a}_{ij}}{\hat{a}_{ij}}$, which obeys an unknown but symmetric distribution, and takes values in $[-1, 1]$. For each i, $\sum_{j=1}^{n} z_{ij}$ takes values in the interval $[0, n]$. As analyzed in Bertsimas and Thiele (2006), the true value of $\sum_{j=1}^{n} z_{ij}$ will take much less value than n since some parameters will exceed their point forecasts and others will fall below estimate, so z_{ij} tends to cancel each other out. This coincides with the fact that aggregate forecasts are more accurate than individual ones.

For each i, a parameter Γ_i is introduced as $\sum_{j=1}^{n} |z_{ij}|$, called *budget of uncertainty*. Then the

set $\mathcal{A}$ becomes:

$$\mathcal{A} = \{(\tilde{a}_{ij}) | \tilde{a}_{ij} = \bar{a}_{ij} + \hat{a}_{ij}\, z_{ij}, \forall i, j, \mathbf{z} \in U\}, \tag{2.34}$$

with uncertainty set U:

$$U = \{\mathbf{z} || z_{ij}| < 1, \forall i, j, \sum_{j=1}^{n} |z_{ij}| \leq \Gamma_i, \forall i\}. \tag{2.35}$$

This uncertainty set considers all parameters $\tilde{a}_{ij}$ such that belonging to the interval $[\bar{a}_{ij} - \hat{a}_{ij}, \bar{a}_{ij} + \hat{a}_{ij}]$, with the restriction that the total weight of deviation from $\bar{a}_{ij}$, summed across all realizations, may be no more than Γ_i. When $\Gamma_i = 0$, this set is the singleton $\bar{a}_{ij}$. At the other extreme, when $\Gamma_i = n$, it considers all uncertainty realizations in the range $[\bar{a}_{ij} - \hat{a}_{ij}, \bar{a}_{ij} + \hat{a}_{ij}]$.

Besides the tractability, the managerial insights support of choosing uncertainty set as (2.35) is as follows. From the point view of management, point forecasts are less meaningful than range forecasts, since the point forecasts are always wrong. And aggregate forecasts are more accurate than individual ones (see Bertsimas and Thiele (2006)). Thus the framework of robust optimization cooperate these managerial insights, the uncertainty parameters or variables are assumed to belong to an interval, and an additional constraint limits the maximum deviation of the aggregate forecasts from its nominal value.

The problem (2.33) can be then reformulated as:

$$\begin{aligned} \min \quad & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} \quad & \bar{\mathbf{a}}_i^T \mathbf{x} + \min_{\mathbf{z}_i \in \mathbf{U}_i} \sum_{j=1}^{n} \hat{a}_{ij}\, z_{ij}\, x_j \geq \mathbf{b}_i, \ \forall i, \\ & \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{2.36}$$

where $\mathbf{z}_i$ is the vector whose j-th element is z_{ij} and $\mathbf{U}_i$ is defined as

$$U_i = \{\mathbf{z_i} || z_{ij}| < 1, \forall j, \sum_{j=1}^{n} |z_{ij}| \leq \Gamma_i\}. \tag{2.37}$$

$\min_{\mathbf{z}_i \in \mathbf{U}_i} \sum_{j=1}^{n} \hat{a}_{ij}\, z_{ij}\, x_j$ for a given i is equivalent to

$$\begin{aligned} -\max \quad & \sum_{j=1}^{n} \hat{a}_{ij}\, z_{ij}\, |x_j| \\ \text{s.t.} \quad & |z_{ij}| < 1, \forall j, \\ & \sum_{j=1}^{n} |z_{ij}| \leq \Gamma_i. \end{aligned} \tag{2.38}$$

This equation is linear in the decision variables z_{ij}. The strong duality of problem (2.38) is as follows.

$$\begin{aligned} -\min \quad & (\sum_{j=1}^{n} q_{ij} + \Gamma_i\, p_i) \\ \text{s.t.} \quad & p_i + q_{ij} \geq \hat{a}_{ij}\, y_j, \ \ \forall\, j, \\ & -y_j \leq x_j \leq x_j, \forall\, j, \\ & p_i, q_{ij} \geq 0, \forall\, j. \end{aligned} \tag{2.39}$$

The robust problem is then reformulated as a linear programming problem:

$$\begin{aligned} \min \quad & \mathbf{c}^T\, \mathbf{x} \\ \text{s.t.} \quad & \bar{\mathbf{a}}_i^T\, \mathbf{x} - (\sum_{j=1}^{n} q_{ij} + \Gamma_i\, p_i) \geq \mathbf{b}_i, \ \ \forall\, i, \\ & p_i + q_{ij} \geq \hat{a}_{ij}\, y_j, \ \ \forall\, i, j, \\ & -y_j \leq x_j \leq x_j, \forall\, i, j, \\ & p_i, q_{ij} \geq 0, \forall\, i, j, \\ & \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{2.40}$$

Compared to Problem (2.32), which contains m constraints and n variables, Problem (2.40) has $n + m(n+1)$ variables and $n(m+2)$ constraints besides the nonnegativity ones, but keeps being LP.

The value assigned to budget of uncertainty Γ_i for each i reflects the decision-maker's attitude toward uncertainty. Despite the lack of information of the random matrix $\mathbf{A}$, we might ask for probabilistic guarantees for the robust solution as a function of the structure and size of the uncertainty set. Specifically, what is the probability of feasibility of the the robust solution in practice? This may become a guideline for the selection of the budget of uncertainty Γ_i.

For the constraints and uncertainty set defined above, Bertsimas et al. (2004) links the value of

the budget to the probability of constraints violation as follows. Let $\mathbf{x}^*$ salsifies constraint $\mathbf{a}^T\,\mathbf{x} \geq \mathbf{b}_i$, when each a_{ij} obeys a symmetric distribution centered at $\bar{a}_{ij}$ and of support $[\bar{a}_{ij}-\hat{a}_{ij}, \bar{a}_{ij}+\hat{a}_{ij}]$, the probability that the constraints $\mathbf{a}^T\,\mathbf{x} \geq \mathbf{b}_i$ to be violated is at most $e^{-\frac{\Gamma_i}{2\,|J|}}$.

More generally, there are fundamental connections between distributional ambiguity, measures of risk, and uncertainty sets in robust optimization. The reader is referred to Bertsimas et al. (2010a), Chen et al. (2007) and reference therein for further details.

Dynamic Robust Optimization

Ben-Tal et al. (2004) first extended the robust optimization framework to dynamic setting. Similarly to the two-stage stochastic optimization with recourse, the decision-maker selects the *here-and-now*, or first-stage decisions, before having any knowledge of the actual value about the uncertainty. He observes then the realization of the uncertainty and after, he chooses the *wait-and-see*, or second-stage decisions according to the outcome of the uncertainty. In a short, rather than re-optimization, the decision-maker adjust his strategy to information revealed over time using policies.

Similarly to the static robust optimization, the dynamic robust optimization ensures that the solutions obtained are feasible for any realization of the uncertainty in the uncertainty set chosen. The general model of adjustable robust counterpart is as follows.

$$\begin{aligned} \min \quad & \mathbf{c}^T\,\mathbf{x} \\ \text{s.t.} \quad & \mathbf{A}\,\mathbf{x} \geq \mathbf{b}, \\ & \mathbf{T}(\omega)\,\mathbf{x} + \mathbf{W}(\omega)\,\mathbf{y}(\omega) \geq \mathbf{h}(\omega), \qquad \forall \omega \in U, \end{aligned} \tag{2.41}$$

where $\{[\mathbf{T}(\omega), \mathbf{W}(\omega), \mathbf{W}(\omega], \omega \in U\}$ is a convex uncertainty set describing the possible values taken by the uncertainty parameters. The second-stage decisions $\mathbf{y}$ are allowed to depend on the uncertainty and the first-stage decision $\mathbf{x}$ have no adaptability to the uncertainty vector ω.

Note that a problem with second-stage decision $\mathbf{y}$ in the objective can immediately be reformulate as:

$$\begin{aligned} \min \quad & Z \\ \text{s.t.} \quad & \mathbf{c}^T\,\mathbf{x} - \mathbf{d}^T\mathbf{y}(\omega) \leq Z, \qquad \forall \omega \in U, \\ & \mathbf{A}\,\mathbf{x} \geq \mathbf{b}, \\ & \mathbf{T}(\omega)\,\mathbf{x} + \mathbf{W}(\omega)\,\mathbf{y}(\omega) \geq \mathbf{h}(\omega), \qquad \forall \omega \in U, \end{aligned} \tag{2.42}$$

which has the form of Problem (2.41).

Problem (2.41) is more flexible than Problem (2.33). But this flexibility comes at the expense of tractability (mathematically, the full adaptability Problem (2.41) is NP-hard). To address this issue, Ben-Tal et al. (2004) propose the affinely adjustable robust counterpart (AARC), in which the second-stage decisions are restricted to be affinely depend on the realized data, which is called Linear Decision Rule (LDR).

In this section, we give a brief review of the AARC methodology pioneered by Ben-Tal et al. (2004). The specific from of LDR proposed by Ben-Tal et al. (2004) for the adjustable variable $\mathbf{y}$ is as follows.

$$\mathbf{y} = \mathbf{q} + \mathbf{Q}\,\omega, \tag{2.43}$$

for some $\mathbf{q}$ and $\mathbf{Q}$ to be determined. The AARC of Problem (2.41) is then

$$\begin{aligned} \min \quad & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} \quad & \mathbf{A}\,\mathbf{x} \geq \mathbf{b}, \\ & \mathbf{T}(\omega)\,\mathbf{x} + \mathbf{W}(\omega)\,(\mathbf{q} + \mathbf{Q}\,\omega) \geq \mathbf{h}(\omega), \qquad \forall \omega \in U. \end{aligned} \tag{2.44}$$

An important case of AARC is that the parameters associated with the adjustable variable in the LP are constants, independent of the uncertainty. This case is known as *fixed recourse*. Ben-Tal et al. (2004) show that AARCs with fixed recourse are computationally tractable for a wide spectrum of uncertainty set. Ben-Tal et al. (2005) employ this methodology to solve a retailer-supplier flexible commitment contract problem.

Suppose that the parameters depend affinely on uncertainty, which can be expressed as $\mathbf{T}(\omega) = \mathbf{T_0} + \mathbf{T_1}\,\omega$, $\mathbf{h}(\omega) = \mathbf{h_0} + \mathbf{h_1}\,\omega$. In what follows, we introduce the process to solve a dynamic robust optimization problem with fixed recourse. The two-stage optimization problem (2.44) can be written as

$$\begin{aligned} \min \quad & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} \quad & \mathbf{A}\,\mathbf{x} \geq \mathbf{b}, \\ & (\mathbf{T_0} + \mathbf{T_1}\,\omega)\,\mathbf{x} + \mathbf{W}\,(\mathbf{q} + \mathbf{Q}\,\omega) \geq \mathbf{h_0} + \mathbf{h_1}\,\omega, \qquad \forall\, \omega \in U. \end{aligned} \tag{2.45}$$

with the values of vectors $\mathbf{x}, \mathbf{q}$ and matix $\mathbf{Q}$ to be determined.

An important step in building the AARC formulation is the selection of the uncertainty set. Ben-Tal and Nemirovski (2000) show that if the uncertainty set is chosen to be either a polyhedral

or an ellipsoidal, the resultant AARC can be solved efficiently. We present an example by defining $\mathbf{U}$ as a polyhedral uncertainty:

$$U = \{\omega : |\omega - \bar{\omega}| \leq \rho\}. \tag{2.46}$$

The AARC formulation corresponding to the LP(2.41) is as follows.

$$\begin{aligned} \min \quad & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} \quad & \mathbf{A}\,\mathbf{x} \geq \mathbf{b}, \\ & \mathbf{h_0} - \mathbf{T_0}\,\mathbf{x} - \mathbf{W}\,\mathbf{q} + (\mathbf{h_1} - \mathbf{T_1}\,\mathbf{x} - \mathbf{W}\,\mathbf{Q})\,\omega \leq 0, \qquad \forall \omega \in U. \end{aligned} \tag{2.47}$$

For the polyhedral uncertainty set (2.46), the second inequality constraint in (2.47) is equivalent to

$$\max\{\mathbf{h_0} - \mathbf{T_0}\,\mathbf{x} - \mathbf{W}\,\mathbf{q} + (\mathbf{h_1} - \mathbf{T_1}\,\mathbf{x} - \mathbf{W}\,\mathbf{Q})\,\omega : |\omega - \bar{\omega}| \leq \rho\} \leq \mathbf{0}, \tag{2.48}$$

and its optimal solution is

$$\mathbf{h_0} - \mathbf{T_0}\,\mathbf{x} - \mathbf{W}\,\mathbf{q} + (\mathbf{h_1} - \mathbf{T_1}\,\mathbf{x} - \mathbf{W}\,\mathbf{Q})\,\bar{\omega} + |\mathbf{h_1} - \mathbf{T_1}\,\mathbf{x} - \mathbf{W}\,\mathbf{Q}|\,\rho \leq \mathbf{0}. \tag{2.49}$$

By adding a new non-negative vector $\mathbf{G}$ which has the same dimension as vector $\mathbf{b}$, the AARC formulation (2.47) can be further expressed as

$$\begin{aligned} \min \quad & \mathbf{c}^T \mathbf{x} \\ \text{s.t.} \quad & \mathbf{A}\,\mathbf{x} \geq \mathbf{b}, \\ & \mathbf{h_0} - \mathbf{T_0}\,\mathbf{x} - \mathbf{W}\,\mathbf{q} + (\mathbf{h_1} - \mathbf{T_1}\,\mathbf{x} - \mathbf{W}\,\mathbf{Q})\,\bar{\omega} + \mathbf{G}\,\rho \leq \mathbf{0}, \\ & \mathbf{G} \leq \mathbf{h_1} - \mathbf{T_1}\,\mathbf{x} - \mathbf{W}\,\mathbf{Q} \leq \mathbf{G}. \end{aligned} \tag{2.50}$$

We can see that Problem (2.50) is LP and can be solved very efficiently.

The AARC methodology uses a linear decision rule to provide the decision-maker a linear policy, even though there is no guarantee that the optimal solution (which is impossible to find) is close to a linear decision rule. However, for adaptability problems with discrete second-stage variables, the AARC methodology does not work. To the best of our knowledge, Bertsimas and Caramanis (2010) is the only work addressing the case of integer second-stage variables

within the framework of deterministic uncertainty-set. The authors propose to partition the uncertainty set into a finite number of pieces and determine a piece-wise constant recourse in each. This called finite adaptability. One important feature of this approach is that it provides a hierarchy of adaptability. Moreover, it can cooperate integer second-stage variables and non-convex uncertainty sets, while other approaches can not. Below, we present some of the results of Bertsimas and Caramanis (2010).

We consider a full adaptability Problem (2.51).

$$\begin{aligned} \min \quad & \mathbf{c}^T \mathbf{x} + \mathbf{d}^T \mathbf{y}(\omega) \\ \text{s.t.} \quad & \mathbf{A}(\omega)\,\mathbf{x} + \mathbf{B}(\omega)\,\mathbf{y}(\omega) \geq \mathbf{b}(\omega), \qquad \forall \omega \in U, \\ & \mathbf{y}(\omega) \in \mathbb{Z}^+. \end{aligned} \tag{2.51}$$

Bertsimas and Caramanis (2010) cover the uncertainty set U with a partition of K (possibly non-disjoint) pieces: $U = \sum_{k=1}^{K} U_k$, and determine K second-stage solutions as contingency plans. After observing the realization of the uncertainty, one of theses contingency plans is implemented. The optimal K-adaptability problem becomes:

$$\begin{aligned} \min \quad & \mathbf{c}^T \mathbf{x} + \max\{\mathbf{d}^T \mathbf{y_1}, ..., \mathbf{d}^T \mathbf{y_K}\} \\ \text{s.t.} \quad & \mathbf{A}(\omega)\,\mathbf{x} + \mathbf{B}(\omega)\,\mathbf{y_1} \geq \mathbf{b}(\omega), \qquad \forall \omega \in U_1, \\ & \qquad\qquad ... \\ & \mathbf{A}(\omega)\,\mathbf{x} + \mathbf{B}(\omega)\,\mathbf{y_K} \geq \mathbf{b}(\omega), \qquad \forall \omega \in U_K, \\ & \mathbf{y_1}, ..., \mathbf{y_K} \in \mathbb{Z}^+. \end{aligned} \tag{2.52}$$

The problem becomes a static robust optimization problem with K second-stage decisions which is easier to be solved than Problem (2.51). This model of adaptability also eliminates the conservativeness of the static robust formulation in the case of a two-stage optimization problem. It provides a hierarchy that bridges the cap between the static robust and fully adaptable formulation as the the level of adaptability (K) increases. Moreover, it is possible to accommodate discrete variables as second-stage decisions.

Two important issues of the finite adaptability approaches are how to partition the uncertainty set U optimally into K pieces, and the necessary conditions that any finite adaptability scheme must satisfy in order to improve the static robust solution by at least a certain quantity. The reader is referred to Bertsimas and Caramanis (2010) for detail information.

As pointed in Bertsimas and Caramanis (2010), finite adaptability is not comparable to affine

adaptability, in the sense that neither technique performs consistently better than the other.

2.3 Summary of the Current State of Art

Operations management of call centers constitutes a large stream of research. Many models in the literature address the key issue of call center staffing and scheduling under stationary parameters. The considered randomness concerns exclusively the stochastic variability of inter-arrival and service times. The impact of fluctuations in the arrival rates (and the associated flexibility issue) is ignored and the results rely on the assumption of known stationary arrival rates. However, it has become apparent that general queueing systems performance indicators are very sensitive to fluctuations of the parameters characterizing the arrival process overtime, see for example Ingolfsson et al. (2007). As a consequence, a stream of research has begun to address the problem of how call centers can better manage the capacity-demand mismatch that results from arrival rate uncertainty.

First, the pure statistical forecasting issue has been considered in several papers analyzing the probability distribution of arrival rates (see Avramidis et al. (2004); Brown et al. (2005, 2002); Weinberg et al. (2007); Shen and Huang (2008); Aldor-Noiman et al. (2009)). Various call center particularities have been pointed out in these studies.

As a second step, the analysis of performance measures of queueing systems with fluctuating arrival rates has appeared. The first setting concerns deterministic non-stationarity, i.e., some parameters evolve along time according to a known dynamics. A direct method of accommodating such time-varying parameters consists of numerically solving the complex queueing models associated to the transient system behavior, see for example Ingolfsson et al. (2007) and Yoo (1996). Another intuitive means of accommodating changes in the arrival rate is to consider piecewise stationary measures over successive intervals, while reducing the time length of the intervals over which such stationary measures could be applied. This is the essence of the point-wise stationary approximation (PSA) used in Green et al. (2007); Green and Kolesar (1991); Green et al. (2003); Ingolfsson et al. (2007). In a different setting, a few papers have considered the issue of random non-stationarity in the arrival process parameters. In Jongbloed and Koole (2001b), the authors include arrival parameter uncertainty via a Poisson mixture model for the arrival process, which permits to model the overdispersion associated with random arrival rates. They develop a generalization of the standard Erlang formula-type staffing approaches. In a different vein, in Harrison and Zeevi (2005); Whitt (2006); Robbins (2007); Steckley et al. (2004), another idea is developed. It can be summarized as estimating performance indices, by first conditioning on the random model-parameter vector, and by thereafter unconditioning to

get the effective indices. Most of these methods assume independent intervals. This would lead to inaccurate results particularly in this case of systems that are overloaded during a certain number of periods. Stolletz (2008) proposes a new approximation for time-varying queueing systems that can be overloaded. The approximation is based on the modeling of the overflow of calls between the periods. Another paper which models dependency between the periods is that of Thompson (1993). The latter does not however allow the analysis of overloaded systems.

The last issue concerns the call center staffing optimization problem under non-stationary parameters. Some models rely on a fixed staffing level methodology: there is no possible flexibility during a daily period and the staffing cannot be updated throughout the day. In Harrison and Zeevi (2005); Whitt (2006), this problem is solved via a static stochastic program using a stochastic fluid model approximation. In Jongbloed and Koole (2001b), the standard Erlang formula-type for a fixed staffing approach is generalized through a new Poisson mixture model for the arrival process.

In many situations, call centers may indeed benefit from flexible staffing, i.e., the ability to adjust staffing levels (and/or schedules) from one period to another. Such flexibility may be attained by utilizing temporary operators, in addition to the permanent operators always available to provide service. The temporary operators may be either supervisors/decision-makers or other operators who are on call. Another type of flexibility corresponds to the presence of different shifts for the operators. By combining such shifts, the operator capacity can be aligned with the time-varying average workload. A last type of flexibility consists of combining different types of calls, with different admissible delays. Some flexibility exists as less urgent calls (as e-mails or calls with a possible callback) can be kept in inventory for some time. Flexible staffing methods coupled with deterministic time-varying arrival rates has been considered in numerous papers. We refer the reader to Gans et al. (2003); Green et al. (2007) and the references therein. A stream of research has sought to use a classical rolling horizon methodology, based on deterministic arrival rate approximations, updated at each period. In Hur et al. (2004), a case study is presented in which the staffing problem under uncertain/non-stationary assumption is addressed via recoursing to a rolling horizon decision process where each step is modeled as a deterministic system. In an alternative research stream, the arrival rates are formally taken into account in the model. This approach mainly consists of generalizing the well-known fluid approximation models in order to introduce staffing level updates for the different periods coupled with available arrival rate updated forecasts. The time horizon is divided into smaller periods and deterministic forecasts for the customer arrival rates for each period are used to determine the respective staffing levels (as in Feldman et al. (2008) and Whitt (1999)).

Lastly, Robbins and Harrison (2010) consider a multi-period multiple-class call center staffing scheduling cost model, with global service constraints. The authors introduce uncertainty for parameters via a discretization of the underlying parameters probability distribution, which amounts to a scenario-based approach coupled with large scale multi-stage stochastic programs to be numerically solved. The approach has also been applied in the case of a call center with multiple call types in order to investigate the flexibility introduced by adding a proportion of cross trained workforce (see Robbins et al. (2007, 2008)). Bhandari et al. (2008) formulate, under suitable assumptions for the arrival process and service time distributions, the multi-periodic staffing problem as a Markov Decision Process with probabilistic constraints. In Bertsimas and Doan (2010), the authors develop a fluid model approximation to solve both the staffing and routing problem for large multi-class/multi-pool call centers with random arrival rates and customer abandonment. The model is solved via a robust optimization approach. Gurvich et al. (2010) propose a fluid approximation model for large-scale multi-class call centers with uncertain parameters. The optimal staffing problems is solved by a chance-constrained programming approach. Helber and Henken (2010) consider a shift scheduling problem of complex call centers with random arrival rates, skills-based routing, impatient customers and retrials. These authors propose a specific approach in which a discrete-time model captures, for a few simulated samples, the dynamics of the systems due to the time-dependent arrival rates. The associated integer program has then to be numerically solved.

Very little attention has been devoted to the issues of information update and real-time adjustment decisions. Mehrotra et al. (2010) and Gans et al. (2009) develop frameworks to make intra-day resource adjustment decisions in call centers. The former suppose that the initial schedules existed and solve the real-time agents schedule adjustment as a one-stage static problem. Gans et al. (2009) extends to include forecast updates and two-stage stochastic programs with recourse.

As for the literature that addresses methodologies issues related to optimization under uncertainty, we have already reviewed some basic knowledge of stochastic programming and robust programming in Section 2.2. Here we introduce an approach which bridges the gap between the conservatism of robust optimization and the specificity of stochastic programming. This approach optimizes the worst-case objective over a family of possible distributions, and we call this min-max stochastic optimization problem *distributionally robust*. This approach was pioneered by Žáčková (1966) and Dupačová (1987). Many other works proposed algorithms to solve min-max stochastic optimization problems: the sample-average approximation method (see Shapiro and Kleywegt (2002) and Shapiro and Ahmed (2004)), sub-gradient-based methods (see Bre-

ton and El Hachem (1995)) and cutting plane algorithms (see Riis and Andersen (2005)). This approach has numerous applications, for example, for the news-vendor problem, Scarf (1958) and Gallego and Moon (1993) derived the optimal order quantity that maximizes the worst case expected profit under the distribution that has a fixed mean and variance. As well as Yue et al. (2006) tried to minimize the worst case absolute regret for all distributions with certain mean and variance. Moreover, El Ghaoui et al. (2003) developed worst-case Value-at-Risk bounds for a robust portfolio selection problem, when only the bounds on the means of the assets and their covariance matrix are known. Natarajan et al. (2010) derived a distributionally robust model applied to portfolio optimization, where the investor maximizes his worst case expected utility over a set of ambiguous distributions described by the knowledge of the mean, covariance and support information. Calafiore and El Ghaoui (2006) considered linear optimization problems with chance constraints in which the underlying distribution is known to belong only within a given set. Erdogan and Iyengar (2006) develop a robust sampled version of ambiguous chance-constrained problems that is feasible with high probability. Chen et al. (2007) propose a tractable means of approximating distributionally robust optimization problems using directional deviation measures. Goh and Sim (2010) extend to allow for expectations of recourse variables in the constraint specifications. Bertsimas et al. (2010b) analyze two-stage min-max stochastic linear optimization problems with risk aversion, where the class of probability distributions is described by their first and second moments. Delage and Ye (2010) provide a polynomial-time algorithm for sample-driven robust stochastic programs where the mean and covariance of the primitive uncertainties are themselves subject to uncertainty. Also, some works draw connections between distributionally robust and objects that have been axiomatized and developed in the decision theory literature over the past several decades (see Xu et al. (2010) and Ben-Tal et al. (2010)).

Chapter 3

Single Shift Staffing

In this chapter, we consider a multi-period staffing problem in a single-shift call center. The call center handles inbound calls, as well as some alternative back-office jobs. The call arrival process is assumed to follow a doubly non-stationary stochastic process with a random mean arrival rate. The inbound calls have to be handled as quickly as possible, while the back-office jobs, such as answering emails, may be delayed to some extent. The staffing problem is modeled as a generalized newsboy-type model under an expected cost criterion. Two different solution approaches are considered. First, by discretization of the underlying probability distribution, we explicitly formulate the expected cost newsboy-type formulation as a stochastic program. Second, we develop a robust programming formulation. The characteristics of the two methods and the associated optimal solutions are illustrated through a numerical study based on real-life data. In particular we focus on the numerical tractability of each formulation. We also show that the alternative workload of back-office jobs offers an interesting flexibility allowing to decrease the total operating cost of the call center.

The paper version of this chapter, Liao, Koole, van Delft and Jouini Liao et al. (2010), is accepted by *OR Spectrum* for publication.

3.1 Introduction

The staffing cost is a major component in the operating costs of call centers. Unfortunately, uncertainty plaguing the arrival process and the corresponding workloads usually leads to a complex staffing problem. Traditionally, most call center models in the literature assume known and constant mean arrival rates, mainly for tractability issues. However, in addition to the usual uncertainty captured by a stochastic process modeling, real data show another uncertainty in the process parameters themselves. In this chapter, we consider the staffing problem of a single shift call center, in which we allow the mean arrival rate of calls to be uncertain. We model the arrival process of calls by a doubly non-stationary stochastic process, with random mean arrival rates. As in the traditional way, a service level constraint limits the waiting time for inbound calls. In addition to the job of calls, our call center has to process back-office jobs, such as answering emails. These additional jobs are assumed to be given at the beginning of the day and have to be processed within the same day, if necessary in overtime. We also allow the workload of back-office jobs to be random. The possibility of delaying back-office jobs introduces some flexibility to the daily workforce management. A typical example of our call center is that of a hospital, or of a government or of a public agency, where inbound calls and back-office operations are handled by agents in a single shift (during administrative hours). The agents can be, in real-time, affected to one job type or another depending on the actual workload and the operating costs.

As mentioned above, our staffing problem incorporates uncertainty in the call arrival parameters. The staffing problem is modeled as a cost optimization-based newsboy-type model. The cost criterion function includes the regular and overtime salary cost and a penalty cost for excessive waiting times for inbound calls. Our objective is to find the optimal staffing level which minimizes the total call center operating cost. We consider a multi-period single-shift call center staffing problem, with the constant staffing level as the single decision variable. We propose two solution methodologies. First, we formulate the problem as a stochastic program, by a discretization of the underlying probability distributions. The second approach relies on robust optimization theory. We prove a convexity result of the problem, which allows us to find the optimal solution via a relaxed real-valued optimization model. We then conduct a numerical study in order to illustrate the main characteristics of the two approaches and the associated optimal solutions. In the numerical illustration, we use real data gathered from a call center of a Dutch hospital handling inbound calls and emails.

We distinguish two main contributions in this chapter. The first contribution is the modeling and the analysis of the staffing problem of a call center with two types of jobs and uncertain arrival parameters: inbound calls, to be handled as quickly as possible, and back-office jobs, that

can be delayed to some extent. The second contribution is the analysis of the impact of the flexibility offered by back-office workloads. We show that combining the two types of jobs offers flexibility, partially absorbing the undesirable effects of uncertainty in the arrival parameters.

The rest of the chapter is structured as follows. In Section 3.2, we describe the call center model under consideration and formulate the associated staffing problem. In Section 3.3, we present the different solution approaches. In Section 3.4, we then conduct a numerical study to evaluate these alternative formulations. We exhibit the impact of the uncertainty of the call arrival parameter and the benefits of the flexibility offered by back-office workloads on the optimization problem. In Section 3.5, we extend the analysis to more general cases, with overflows of calls between successive periods. This chapter ends with concluding remarks and highlights some future research.

3.2 Problem Formulation

We consider a multi-period single-shift call center staffing problem. The call center handles various types of jobs: inbound calls as well as some alternative back-office jobs. The mean arrival rate of inbound calls is allowed to be uncertain. The workload of the back-office jobs is also uncertain. The inbound calls have to be handled as soon as possible, while the back-office jobs, such as emails, can be delayed to some extent within the same day. In this section, we describe the corresponding stochastic minimal cost staffing problem.

3.2.1 The Inbound Call Arrival Process

Recall the characteristics of the arrival process of calls presented in Section 2.1.2. First, it has been observed that the total daily number of calls has an overdispersion relative to the classical Poisson distribution. Second, the mean arrival rate considerably varies with the time of day. Third, there is a strong positive correlation between arrival counts during the different periods of the same day.

Assume that a given working day is divided into n distinct, and the arrival rate Λ_i is of the form

$$\Lambda_i = \Theta f_i, \text{ for } i = 1, ..., n, \tag{3.1}$$

where Θ is a positive real-valued random variable. The random variable Θ can be interpreted as the unpredictable *busyness* of a day. A large (small) outcome of Θ corresponds to a busy (not busy) day. The constants f_i model the shape of the variation of the arrival rate intensity across

the periods of the day. Formally, if a sample value in a given day of the random variable Θ is denoted by θ, the corresponding outcome of the arrival rate over period i for that day is defined by $\lambda_i = \theta f_i$. The random variable Θ is assumed to follow a discrete probability distribution, defined by the sequence of outcomes θ_l and the associated sequence of probabilities p_{θ_l}, with $l = 1, ..., L$.

We assume that service times for inbound calls are independent and exponentially distributed with rate μ. The calls arrive to a single infinite queue working under the the first come, first served (FCFS) discipline of service. Neither abandonment nor retrials are allowed.

Using the Erlang C Staffing presented in Section 2.1.3, the staffing level which guarantees the required service level is computed by

$$v_i(\theta\, f_i) \quad = \quad F^{-1}_{\theta\, i}(SL_i). \tag{3.2}$$

with the function F defined by Equation (2.3).

3.2.2 The Back-Office Workload Process

We assume that the random back-office workload arrives at the beginning of the day. As an example, one can think of a call center that stores all the emails of a given day and handles them the next day. We denote by W the number of agents required to handle this back-office workload during a single period. The random variable W is characterized by a discrete probability distribution, defined by the sequence of outcomes w_k and the associated sequence of probabilities p_{w_k}, with $k = 1, ..., K$.

3.2.3 Cost Criterion

In this chapter we consider a single-shift call center. Let us denote by y the number of agents staffed for the day. All the y agents will be therefore present all day long. We also assume that all agents are able to handle both types of jobs, calls and back-office jobs. We give priority to inbound calls as follows. For each period i, if the actual number of agents y is larger than $v_i(\theta f_i)$ (the required number of agents to handle the calls), we assign $v_i(\theta f_i)$ agents to calls and $y - v_i(\theta f_i)$ agents to back-office jobs. If $y < v_i(\theta f_i)$, all the y agents are assigned to calls. If back-office jobs are not yet finished at the end of the regular working periods in that day, they are done in overtime.

For a given period, any under-staffing situation is penalized. Under perfectly predictable arrival rates, a straightforward formulation of the optimization problem is to consider quality-of-service constraints requiring that the service level SL_i is reached in period i. However in the

presence of uncertain arrival rates as in this dissertation, the service level per period is indeed itself a random variable, depending on the outcomes of the arrival rates. A possible formulation is to adopt a chance constrained approach requiring that the quality-of-service constraints are satisfied for some pre-specified fraction of the arrival rate realizations (i.e., with some given probability). This approach has been used in the context of large skill-based routing call centers in Gurvich et al. (2010), where the authors have developed a staffing method leading to nearly optimal solutions.

More clearly, a chance constrained formulation (for a risk level α, and a stochastic parameter Θ for the arrival process) can be expressed as finding the staffing level y_α given by

$$\Pr\{y_\alpha \leq V_i(\Theta\, f_i), i = 1, .., n\} = \alpha, \tag{3.3}$$

with $V_i(\Theta\, f_i)$, the underlying random number of agents required to handle the calls in period i, in order to fulfill the required quality-of-service constraints. By choosing the risk-level α, the decision-maker may choose a trade-off between staffing costs and *safety* in terms of the likelihood with which the quality-of-service constraints are met. However, such a formulation corresponds to quite complex non-convex non-linear optimization problems requiring specific approximations and heuristics, out of the scope of this dissertation. In order to propose a solvable linear programming formulation, the risk level α is expressed via an associated under-staffing penalty cost denoted as u_α. This formulation approach has been applied for example in Robbins (2007). More concretely for each period i, a proportional under-staffing penalty u_α is paid when the actual capacity y is lower than a sample value of the required agents number $v_i(\theta f_i)$. The value of the parameter u_α can be tuned, for example, via an algorithm based on successive problem solutions and successive numerical estimations of the effective constraint violation probability α for each current staffing solution. This numerical estimation can be made through a direct computation if the probability distribution of Θ is known or, otherwise, through simulations of the arrival process. In the numerical examples presented in this chapter, this tuning procedure algorithm converged very quickly.

In our cost setting, we also assume that each agent gets a salary c per period, the overtime salary is r per agent per period. As usual, the cost parameters satisfy the ordering $c < r < u_\alpha$ for all possible values of α. The inequality $r < u_\alpha$ ensures that inbound calls have the priority over back-office jobs. The inequality $c < r$ is straightforward.

Since the time-horizon of the considered problematic is significant, the cost criterion of the formulation is the expected daily total cost associated with the staffing level y, which is expressed

as

$$C(y) = E\Big[C(y,\theta,w)\Big] = \sum_{l=1}^{L}\sum_{k=1}^{K} p_{\theta_l}\, p_{w_k}\, C(y,\theta_l,w_k), \tag{3.4}$$

with

$$C(y,\theta,w) = n\,c\,y + u_\alpha \sum_{i=1}^{n}(y - v_i(\theta f_i))^- + r\left[w - \sum_{i=1}^{n}(y - v_i(\theta f_i))^+\right]^+, \tag{3.5}$$

where $E[\cdot]$ denotes the expectation, $x^+ = max(0,x)$ and $x^- = max(0,-x)$ for $x \in \mathbb{R}$. In Equation (3.5), the first term is the salary of the agents working during regular time. The second term is the under-staffing penalty cost. The third is the overtime salary.

Under this economic framework, our objective consists of deciding on the optimal value of y which minimizes the expected daily total cost given by Equation (3.4). In the following theorem we give a convexity result for the expected daily total cost as a function of the decision variable y. All the proofs of the results in this dissertation are given in the appendix.

Theorem 3.1 *The expected daily total cost function $C(y)$ is convex in y.*

We can see from the proof that no specific assumption on the arrival rates probability distributions is required.

3.3 Solution Methodologies

The classical paradigm to solve the problem given by Equation (3.4) is to develop a deterministic approach using the expected values of the random variables Θ and W. The optimal solution under the deterministic approach might lead to a far greater cost than the actual one when the parameters take values that are different from those expected, and in particular, when the system is sensitive to data variation (for example, for a high value of u_α). This will be underlined later in the numerical study. It is thus important to take into account the effect of data uncertainties and develop better solution approaches.

In this section we develop two different approaches to solve the staffing problem given by Equation (3.4), according to the availability of the probability distributions of the random variables. These approaches are then used in the numerical study in Section 3.4. First, under the assumption that the probability distributions associated with the random variables are known exactly, a direct *stochastic programming approach* is applied to Equation (3.4), built on the discrete probability distributions characterizing Θ and W. The second approach referred to as *robust programming* consists of optimizing the staffing level with respect to (w.r.t) the worst case scenarios in a given uncertainty set.

The property given in Theorem 3.1 is directly used in the optimization procedure. The integer optimal solution is indeed known to be in the neighborhood of the real-valued relaxed optimal solution. We thus relax the integer problem and only solve the real-valued version. Then, if the optimal decision value of y is not integer as the staffing level should be, it suffices to compare the objective costs corresponding to the two nearest integers, and the optimal integer solution is that with the lower objective cost.

3.3.1 Stochastic Programming Approach

Assuming that we know the exact probability distributions associated with the random variables Θ and W, a common approach consists of expressing Equation (3.4) as a linear program via the discrete probability distributions associated with these random variables. For each sample θ_l of Θ, we use the associate sample arrival rate in each period i, $\lambda_{i,l} = \theta_l \, f_i$. The required number of agents is $v_i(\lambda_{i,l})$ and is given using Condition (2.3) as a function of $\lambda_{i,l}$.

The optimization problem from Equation (3.4) can be then formulated by the following linear program:

$$
\begin{aligned}
Min \quad & nc\, y + u_\alpha \sum_{l=1}^{L} \sum_{i=1}^{n} p_{\theta_l} \, M_{i,l}^{-} + r \sum_{k=1}^{K} \sum_{l=1}^{L} p_{\theta_l} \, p_{w_k} \, N_{k,l} && && (3.6)\\
\text{s.t.} \quad & M_{i,l} = y - v_i(\theta_l \, f_i), && \text{with} \;\; i = 1, ..., n, l = 1, ..., L, && (3.7)\\
& M_{i,l} = M_{i,l}^{+} - M_{i,l}^{-}, && \text{with} \;\; i = 1, ..., n, l = 1, ..., L, && (3.8)\\
& N_{k,l} \geq w_k - \sum_{i=1}^{n} M_{i,l}^{+}, && \text{with} \;\; l = 1, ..., L, k = 1, ..., K, && (3.9)\\
& y, M_{i,l}^{+}, M_{i,l}^{-}, N_{k,l} \geq 0, && \text{with} \;\; i = 1, ..., n, l = 1, ..., L, k = 1, ..., K. && (3.10)
\end{aligned}
$$

In this problem $M_{i,l}$ represents the difference between the staffing level and the required agent number in period i for scenario l. The positive and negative part of $M_{i,l}$ are denoted by $M_{i,l}^{+}$ and $M_{i,l}^{-}$, respectively. $M_{i,l}^{-}$ is associated to under-staffing cost in the objective function. $N_{k,l}$ is the over-time workload required in order to finish back-office jobs in scenario (k, l). This overtime induces overtime cost in the objective function. The unique decision variable in our staffing problem is the staffing level y.

In this formulation, a possible way to take into account the risk consists of bounding the conditional-value-at-risk (CVaR), see Rockafellar and Uryasev (2002). Let $0 < \beta \leq 1$ be a confident level and let $CVaR_\beta$ be the mean of the total costs belonging to the largest proportion β. Rockafellar and Uryasev (2002) proved that the minimization of $CVaR_\beta(y)$ with respect to

the decision variable y is simply given by

$$\min_{y} CVaR_{\beta}(y) = \min_{y,s}\{s + \beta^{-1}E[(C(y,\theta,w) - s)^{+}]\}. \tag{3.11}$$

Furthermore, the right-hand side of the optimization problem (3.11) is jointly convex in (y, s) if the cost function $C(y, \theta, w)$ is convex in y.

Ogryczak and Ruszczynski (2002) mention that the CVaR is a coherent risk measure which is computationally tractable in the framework of stochastic programming. For a given β, the CVaR optimization problem given by (3.11) can be formulated by the following linear program:

$$Min \quad s + \beta^{-1}\sum_{k=1}^{K}\sum_{l=1}^{L} p_{w_k}\, p_{\theta_l}\, z_{k,l} \tag{3.12}$$

$$\text{s.t.} \quad nc\, y + u_{\alpha}\sum_{i=1}^{n} M_{i,l}^{-} + r\ N_{k,l} - s \leq z_{k,l}, \quad \text{with} \quad k = 1, ..., K, l = 1, ..., L, \tag{3.13}$$

$$M_{i,l} = y - v_i(\theta_l\, f_i), \qquad \text{with} \quad i = 1, ..., n, l = 1, ..., L, \tag{3.14}$$

$$M_{i,l} = M_{i,l}^{+} - M_{i,l}^{-}, \qquad \text{with} \quad i = 1, ..., n, l = 1, ..., L, \tag{3.15}$$

$$N_{k,l} \geq w_k - \sum_{i=1}^{n} M_{i,l}^{+}, \qquad \text{with} \quad l = 1, ..., L, k = 1, ..., K, \tag{3.16}$$

$$y, M_{i,l}^{+}, M_{i,l}^{-}, N_{k,l}, z_{k,l} \geq 0, \qquad \text{with} \quad i = 1, ..., n, l = 1, ..., L, k = 1, ..., K, \tag{3.17}$$

$$s \in \mathbb{R}. \tag{3.18}$$

In Equation (3.12)-(3.18) , the variable s represents a critical threshold which is also called value-at-risk. The variable $z_{k,l}$ represents the positive gap between the total cost in scenario (k, l) and threshold s. The key point is that this conditional-value-at-risk formulation has a similar structure as that of the original formulation (3.6)-(3.10), i.e., we can again use the real-valued relaxed version in order to solve it.

3.3.2 Robust Programming Approach

Stochastic programming formulations associated with discrete distributions suffer very often from the high dimensionality of the corresponding linear programs. We refer the reader to Thenie et al. (2007) or van Delft and Vial (2004). More importantly, stochastic programming requires an accurate probabilistic description of the randomness; however, in many life applications this information is not available. An alternative, non-probabilistic approach can be implemented via a *robust optimization* formulation which adopts a *Min-Max*-type approach coupled with uncertainty sets associated to the random parameters of the problem. Robust programming based formulations are often computationally tractable even for large-scale problems and don't

require a probabilistic description of the uncertain parameters.

A main issue of the robust programming implementation is the design of an efficient uncertainty set which fixes the trade-off between robustness (i.e., protection against the worst case) and average performance (see Bertsimas and Brown (2009) and Natarajan et al. (2009) for further details). If we choose an uncertainty set covering the whole underlying sample space associated with the random parameters, implemented over sample data, this solution exhibits the best possible worst case performance, but does poorly on average (see Soyster (1973)). One can then choose an uncertainty set which does not cover the whole underlying sample space. In this case, the solution can be expected to exhibit improved average costs for sample data, however this solution will be less robust to the worst case, as some of the sample scenarios are likely to be outside of the reduced uncertainty set. By considering different sizes for the uncertainty sets, one reviews different possible trade-offs between average performance and protection against uncertainty (see Bertsimas and Sim (2004)).

We consider a robust approach associated with uncertainty sets for Θ and W. In order to analyze the above robust formulation, we first study the properties of the optimal value, denoted as $C^*(\theta, w)$, of the purely deterministic optimization problem for given outcomes θ and w,

$$
\begin{aligned}
Min \quad & nc\, y + u_\alpha \sum_{i=1}^{n} M_i^- + r\, N && && (3.19)\\
\text{s.t.} \quad & M_i = y - v_i(\theta\, f_i), && \text{with} \quad i = 1, ..., n, && (3.20)\\
& M_i = M_i^+ - M_i^-, && \text{with} \quad i = 1, ..., n, && (3.21)\\
& N \geq w - \sum_{i=1}^{n} M_i^+, && && (3.22)\\
& y, M_i^+, M_i^-, N \geq 0, && \text{with} \quad i = 1, ..., n. && (3.23)
\end{aligned}
$$

In this formulation, M_i represents the difference between the staffing level and the required agent number in period i. The positive and negative part of M_i are denoted by M_i^+ and M_i^-, respectively. M_i^- is associated to under-staffing cost in the objective function. N is the over-time workload required in order to finish back-office jobs.

In the next proposition, we exhibit some properties of $C^*(\cdot, \cdot)$, that are used in the robust programming formulation.

Proposition 3.1 *Let $C^*(\theta, w)$ be the optimal objective value of the problem defined in (3.19)-*

(3.23). For $\delta > 0$, we have the following inequalities,

$$C^*(\theta + \delta, w) \geq C^*(\theta, w), \tag{3.24}$$

$$C^*(\theta, w + \delta) \geq C^*(\theta, w). \tag{3.25}$$

Proof: See Appendix B.

Corollary 1 *For uncertainty sets defined as*

$$U = \{(\theta, w) : 0 \leq \theta \leq \overline{\theta} + \eta\,\sigma_\theta, 0 \leq w \leq \overline{w} + \eta\,\sigma_w, \eta \geq 0\}, \tag{3.26}$$

by Proposition 3.1, we have

$$\max_{(\theta,w)\in U} C^*(\theta, w) = C^*(\overline{\theta} + \eta\,\sigma_\theta, \overline{w} + \eta\,\sigma_w). \tag{3.27}$$

These results are straightforward by applying Proposition 1 and are intuitively clear: a call center with additional workload (of calls and/or back-office jobs) will require an additional cost, related to additional salary, additional under-staffing or overtime costs. The robust formulation of the staffing problem with the uncertainty set (3.26) is as follows.

$$Min \quad nc\,y + u_\alpha \sum_{i=1}^{n} M_i^- + r\,N \tag{3.28}$$

$$\text{s.t.} \quad M_i = y - v_i((\overline{\theta} + \eta\,\sigma_\theta) f_i), \qquad \text{with} \quad i = 1, ..., n, \tag{3.29}$$

$$M_i = M_i^+ - M_i^-, \qquad \text{with} \quad i = 1, ..., n, \tag{3.30}$$

$$N \geq \overline{w} + \eta\,\sigma_w - \sum_{i=1}^{n} M_i^+, \tag{3.31}$$

$$y, M_i^+, M_i^-, N \geq 0, \qquad \text{with} \quad i = 1, ..., n. \tag{3.32}$$

As in Section 3.3.1, we relax integrity constraints for the variables. The parameter $\eta \in \mathbb{R}^+$ fixes the upper bound values for the uncertain parameters Θ and W in (3.26). The decision-maker chooses to fix the trade-off between the protection level against uncertainty and the average cost performance. We note here that it is also possible to build a formulation mixing stochastic and robust programming, for example by defining an uncertainty set for Θ and a probability distribution for W (the corresponding formulation is given in Appendix A.3).

3.4 Numerical Comparison

In this section, we conduct a numerical study in order to evaluate the proposed approaches. In Section 3.4.1, we describe the numerical experiments. In Section 3.4.2, we analyze the results and give some insights.

3.4.1 Experiments

We first describe the data used in the numerical examples. We next describe the experiments and give the numerical results.

Parameter Values

Inbound calls. In the experiments we use real data from a Dutch hospital which exhibits a typical and significant workload time-of-day seasonality. Figure 3.1 displays the mean arrival rates as a function of the periods in the day. We focus on a particular day, namely Monday. With the solid line, we plot this curve for an average day, and in dashed lines we represent two samples corresponding to busy and not busy days. The mean arrival rate at the beginning and at the end of the day is quite low, exhibits a high peak in the late morning, tends to decrease around the lunch break, and finally has a second lower peak in the afternoon. Although there is a significant stochastic variability in the arrival rate from one day to another, there is a strong seasonal pattern across the periods of a given day. The day starts at 7 am, finishes at 6 pm, and is divided into $n = 11$ periods, of one hour each.

Without loss of generality, we choose $E[\Theta] = 1$. This leads to $f_i = E[\Lambda_i]$, and from a one-year-horizon data we numerically find via a standard statistical analysis, that f_1, f_2, ..., f_{11} are 3.5, 18.4, 34.4, 31.5, 29.0, 12.9, 28.4, 25.0, 17.4, 7.2, 5.3 calls per minute, respectively.

Recall that the random variable Θ describes the busyness of the day. We assume that Θ follows a discretized truncated Gaussian probability distribution. Since we normalized the realizations of Θ such that $E[\Theta] = 1$, the *busyness factor*, say θ_t, of a given day t with a total daily call number, say D_t, is $\theta_t = \frac{D_t}{\sum_{i=1}^{n} f_i}$. Collecting the θ_t values of all the days of the data, we obtain a sequence of values for which we compute a statistical standard deviation. We find $\sigma_\Theta = 0.21$. This finishes the characterization of the random variable Λ_i.

The mean service time is $\dfrac{1}{\mu} = 5$ minutes. We assume a classical service level corresponding to the well-known 80/20 rule: the probability that a call waits for less than 20 seconds has to be larger or equal to 80 percent. Using Condition (2.3), we can therefore deduce the required number of agents v_i during period i.

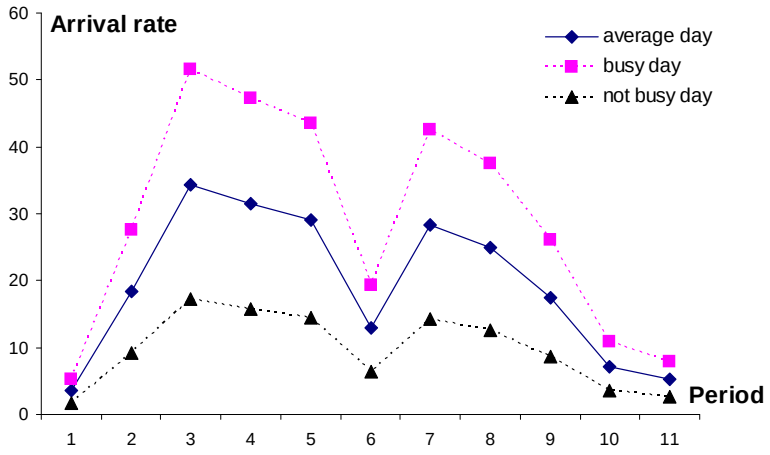

Figure 3.1: Arrival rate graph

Back-office jobs. In this real-life case, the back-office jobs correspond to the emails to be answered in a given day. The random daily workload of emails, W, is assumed to follow a truncated Gaussian distribution. We consider three scenarios with high, medium and low workload of emails, corresponding to $1000, 600, 50$ for the means and $100, 60, 5$ for the standard deviations, respectively.

Cost parameters. The salary during the regular time is $c = 15$ per agent per period. The salary during the overtime is $r = 20$ per agent per period. For each period, a penalty u_α, for being under-staffed by one agent, is incurred, within the ordering $c < r < u_\alpha$, given in Section 3.2.3. We considered three scenarios for the under-staffing probability $\alpha = 10\%$, $\alpha = 5\%$, $\alpha = 1\%$, and determined, as explained in Section 3.4, the corresponding values for the penalty cost $u_{10\%}$, $u_{5\%}$ and $u_{1\%}$. It should be mentioned that u_α depends not only on α, but also on other parameters of the call center.

Design of the Experiments

Benchmark. As an initial benchmark, we consider the simple approach based on the expected value of the random variable Θ, referred to as *average based deterministic approximation*, denoted by DA. In this case, Λ_i reduces to the single value $E[\Theta]f_i$. The required number of agents to handle the calls of period i is $v_i(E[\Theta]f_i)$ given by Condition (2.3). Here we keep the amount of emails W as a random variable, and average on all of its outcomes. The optimization problem

can be then formulated as:

$$Min \quad nc\, y + u_\alpha \sum_{i=1}^{n} M_i^- + r \sum_{k=1}^{K} p_{w_k} N_k \tag{3.33}$$

$$\text{s.t.} \quad M_i = y - v_i(E[\Theta]\, f_i), \qquad \text{with} \;\; i = 1, ..., n, \tag{3.34}$$

$$M_i = M_i^+ - M_i^-, \qquad \text{with } i = 1, ..., n, \tag{3.35}$$

$$N_k \geq w_k - \sum_{i=1}^{n} M_i^+, \qquad \text{with} \;\; k = 1, ..., K, \tag{3.36}$$

$$y, M_i^+, M_i^-, N_k \geq 0, \qquad \text{with} \;\; i = 1, ..., n. \tag{3.37}$$

In this problem, M_i represents the difference between the staffing level and the required agent number in period i for the average arrival rate $E[\Theta]\, f_i$. The positive and negative part of M_i are denoted by M_i^+ and M_i^-. N_k is the over-time workload required in order to finish back-office jobs in scenario k.

Lower bound. As a lower bound solution, we consider a perfect information model (PI) for which the value of the pair (θ, w), the actual workload, is assumed to be known before the optimization step of the variable y. For each pair (θ_l, w_k), we solve the problem (3.38)-(3.42) in order to get the optimal value of $y_{l,k}$ and its associated total cost.

$$Min \quad nc\, y_{l,k} + u_\alpha \sum_{i=1}^{n} M_{i,l}^- + r\, N_{k,l} \tag{3.38}$$

$$\text{s.t.} \quad M_{i,l} = y_{l,k} - v_i(\theta_l\, f_i), \qquad \text{with} \;\; i = 1, ..., n, l = 1, ..., L, k = 1, ..., K, \tag{3.39}$$

$$M_{i,l} = M_{i,l}^+ - M_{i,l}^-, \qquad \text{with} \;\; i = 1, ..., n, l = 1, ..., L, \tag{3.40}$$

$$N_{k,l} \geq w_k - \sum_{i=1}^{n} M_{i,l}^+, \qquad \text{with} \;\; l = 1, ..., L, k = 1, ..., K, \tag{3.41}$$

$$y_{l,k}, M_{i,l}^+, M_{i,l}^-, N_{k,l} \geq 0, \qquad \text{with} \;\; i = 1, ..., n, l = 1, ..., L, k = 1, ..., K. \tag{3.42}$$

The computation of the corresponding average total cost is then straightforward according to (3.4).

Additional Notations. We compute the optimal staffing levels given by the average based deterministic approximation (DA), the classical stochastic programming approach (SP), robust (RP) and mixed ($MxRP$) programming approaches with various robustness levels. For the robust programming approaches and the mixed robust programming approaches, the size of the uncertainty sets varies according to $\eta = 0.1$, 0.5, 1.0, 2.0 and 3.0.

Optimal policy performance simulations. In order to estimate the cost criterion probability distribution associated with the different policies, 20000 sample values are randomly generated as outcomes of (Θ, W).

Numerical Results

The results are given in Tables 3.1, A.1 and A.2, which correspond to low, medium and high volumes of emails, respectively. Tables A.1 and A.2 are given in Appendix A.4. For the value of u_α corresponding to a given estimated under-staffed probability ($\alpha = 10\%, 5\%, 1\%$) in the call center, and for a given approach, each table displays the optimal staffing level, the average total cost, the average values of the three components of the total cost, the standard deviations (STD.) of the total cost and the under-staffing cost and the over-time cost. At the end of each line, the under-staffing probability is given.

The computations have been performed using Cplex on an Intel Core Duo CPU 1.20 Ghz with 0.99 GBytes RAM. For the considered problems, the computing time of DA and RP never exceeded 0.1 seconds while for SP, this time is around 170 seconds.

3.4.2 Insights

In this section, we comment on the numerical results and derive the main insights. First, we compare the proposed approaches and show the advantage of explicitly taking into account the uncertainty in the call arrival parameters. Second, we analyze the benefits of the flexibility provided by emails on our staffing optimization problem and the number of flexible servers necessary.

Analysis of the numerical experiments

In what follows, we compare between the performance measures of the proposed approaches. First, we mention that some trade-off exists between the average cost and the associated standard deviation: Above the threshold which is the optimal staffing level of SP, the average total cost increases (see Theorem 3.1) while the associated standard deviation decreases in y. It is also obvious to see that the under-staffing probability decreases in the under-staffing penalty u_α. For large values of u_α, this probability becomes negligible. For the call centers with the same parameters as those in Tables 3.1, A.1 and A.2, we have conducted additional runs showing that when $u_\alpha = 1e + 5$, the associated under-staffing probabilities are lower than 0.015%.

Concerning the average cost, SP is as expected the most efficient. In Tables 1, 3 and 4, we observe that for a given value of the risk level α, the optimal solutions of SP of the three tables are the same. This stems from the fact that for a call center with given distributions of the busyness factor Θ and the back-office workload W, we associate an under-staffing penalty cost u_α which expresses the chance constraint (3.3). For a given period i, the distribution of the required number of agents $V_i(\Theta f_i)$ is unchanged for the three tables, since the distribution of Θ

Table 3.1: $E[\Theta] = 1$ and $\sigma_\Theta = 0.21$; $E[W] = 50$ and $\sigma_W = 5$

		Optimal staff y^*	Total Cost Average	Total Cost STD.	Salary cost	Under-staffing cost Average	Under-staffing cost STD.	Overtime cost Average	Overtime cost STD.	Constr. violation Pct.(%)
	PI	—	29764.72	5972.79	27620.88	2143.83	444.09	0.00	0.00	9.09
	DA	167	35016.90	10945.22	27555.00	7461.90	10945.22	0.00	0.00	17.22
	SP	184	34105.39	7365.34	30360.00	3745.39	7365.34	0.00	0.00	10.08
	RP									
$\alpha = 10\%$	$\eta = 0.1$	170	34710.15	10273.64	28050.00	6660.15	10273.64	0.00	0.00	15.81
u_α=145	$\eta = 0.5$	184	34105.39	7365.34	30360.00	3745.39	7365.34	0.00	0.00	10.08
	$\eta = 1.0$	201	34841.42	4532.19	33165.00	1676.42	4532.19	0.00	0.00	5.20
	$\eta = 2.0$	234	38858.72	1381.78	38610.00	248.72	1381.78	0.00	0.00	0.99
	$\eta = 3.0$	268	44239.97	278.47	44220.00	19.97	278.47	0.00	0.00	0.10
	MxRP									
	$\eta = 0.1$	170	34710.15	10273.64	28050.00	6660.15	10273.64	0.00	0.00	15.81
	$\eta = 0.5$	184	34105.39	7365.34	30360.00	3745.39	7365.34	0.00	0.00	10.08
	$\eta = 1.0$	201	34841.42	4532.19	33165.00	1676.42	4532.19	0.00	0.00	5.20
	$\eta = 2.0$	234	38858.72	1381.78	38610.00	248.72	1381.78	0.00	0.00	0.99
	$\eta = 3.0$	268	44239.97	278.47	44220.00	19.97	278.47	0.00	0.00	0.10
	PI	—	30060.42	6033.53	30060.42	0.00	0.00	0.00	0.00	0.00
	DA	182	38480.79	16040.49	30030.00	8450.79	16040.49	0.00	0.00	10.81
	SP	202	36626.72	9088.74	33330.00	3296.72	9088.74	0.00	0.00	4.98
	RP									
$\alpha = 5\%$	$\eta = 0.1$	186	37784.84	14458.39	30690.00	7094.84	14458.39	0.00	0.00	9.40
u_α=300	$\eta = 0.5$	200	36648.18	9671.77	33000.00	3648.18	9671.77	0.00	0.00	5.45
	$\eta = 1.0$	219	37436.04	5108.90	36135.00	1301.04	5108.90	0.00	0.00	2.23
	$\eta = 2.0$	255	42191.39	1116.90	42075.00	116.39	1116.90	0.00	0.00	0.26
	$\eta = 3.0$	292	48183.71	124.35	48180.00	3.71	124.35	0.00	0.00	0.01
	MxRP									
	$\eta = 0.1$	186	37784.84	14458.39	30690.00	7094.84	14458.39	0.00	0.00	9.40
	$\eta = 0.5$	200	36648.18	9671.77	33000.00	3648.18	9671.77	0.00	0.00	5.45
	$\eta = 1.0$	219	37436.04	5108.90	36135.00	1301.04	5108.90	0.00	0.00	2.23
	$\eta = 2.0$	255	42191.39	1116.90	42075.00	116.39	1116.90	0.00	0.00	0.26
	$\eta = 3.0$	292	48183.71	124.35	48180.00	3.71	124.35	0.00	0.00	0.01
	PI	—	30060.42	6033.53	30060.42	0.00	0.00	0.00	0.00	0.00
	DA	182	71579.72	78865.76	30030.00	41549.72	78865.76	0.00	0.00	10.81
	SP	233	41147.64	14645.96	38445.00	2702.64	14645.96	0.00	0.00	1.06
	RP									
$\alpha = 1\%$	$\eta = 0.1$	186	65572.94	71087.09	30690.00	34882.94	71087.09	0.00	0.00	9.40
u_α=1475	$\eta = 0.5$	200	50936.89	47552.87	33000.00	17936.89	47552.87	0.00	0.00	5.45
	$\eta = 1.0$	219	42531.78	25118.77	36135.00	6396.78	25118.77	0.00	0.00	2.23
	$\eta = 2.0$	255	42647.23	5491.41	42075.00	572.23	5491.41	0.00	0.00	0.26
	$\eta = 3.0$	292	48198.22	611.38	48180.00	18.22	611.38	0.00	0.00	0.01
	MxRP									
	$\eta = 0.1$	186	65572.94	71087.09	30690.00	34882.94	71087.09	0.00	0.00	9.40
	$\eta = 0.5$	200	50936.89	47552.87	33000.00	17936.89	47552.87	0.00	0.00	5.45
	$\eta = 1.0$	219	42531.78	25118.77	36135.00	6396.78	25118.77	0.00	0.00	2.23
	$\eta = 2.0$	255	42647.23	5491.41	42075.00	572.23	5491.41	0.00	0.00	0.26
	$\eta = 3.0$	292	48198.22	611.38	48180.00	18.22	611.38	0.00	0.00	0.01

is kept the same. Thus, the optimal staffing level y is also unchanged for the three tables. Note also that the value of the under-staffing penalty cost u_α varies with different sizes of the email workload W. We should mention that if we do not relate the chance constraint (3.3) with the under-staffing penalty cost, the optimal staffing level would change with the size of back-office workload for fixed under-staffing penalty cost (more details are given in Section 3.4.2).

The gap between the optimal staffing levels of DA and SP is remarkable, especially when the back-office workload is small. Neither DA captures the negative impact of the randomness in call arrival rates on service quality, nor on the under-staffing cost. Particularly, it can be seen that the optimal solutions of DA remains constant once the penalty cost u_α exceeds some threshold. Consequently, in the case of a high penalty cost and significant arrival rate randomness, DA should not be used. However, it can be noticed that when the back-office workload is high, the induced flexibility is quite profitable and the global performance of the DA optimal solution is in that case less affected by the randomness of the arrival process.

As described in Section 3.3.2, robust optimization relies on a worst-case-type analysis for a given uncertainty set. In order to examine different trade-offs between the average performance and the protection against risk, we have considered different η values. The higher the η value, the higher the degree of protection in the solution. An extreme case can be considered, namely $\eta = 0$, which can be viewed as equivalent to DA. By increasing the η value, the optimal RP solution includes a progressively increasing over-staffing, which eliminates under-staffing penalty costs, but at the same time increases the direct salary costs. Since RP always considers a *worst-case* setting, it is therefore important to choose an appropriate uncertainty set by taking into account the calls arrival rate variations, the target α and the flexibility offered by the back-office workload.

In Tables 3.1 and A.1, MxRP has the same optimal solution and cost performance as RP. Basically, this stems from the fact that the back-office workload uncertainty is not significant w.r.t. the arrival process randomness. The results exhibit a slight difference for increased back-office workload uncertainty (see Table A.2).

Benefits of The Flexibility Offered by Back-Office Jobs

An obvious benefit from adding back-office jobs comes from the fluctuating shape exhibited by the call arrival rate as a function of the periods of the days (see Figure 3.1). Since we are considering a single shift call center, the strongest quality-of-service constraints (corresponding to the period with the highest arrival rates), tend to force to have a typically high staffing level for the whole day. Such a level is in fact required for only some periods. Clearly, this situation

leads to over-staffing during the other periods, which can be used without any additional cost in order to handle some back-office jobs. For example, it can be seen from Tables 3.1 and A.1, that the optimal staffing levels are identical (and do not increase) while the back-office workload has been increased from 50 to 600. The savings are thus the direct salary costs corresponding to this increase (namely $600 \times c = 9000$) minus the under-staffing cost increase and minus the overtime cost increase (which are negligible w.r.t. 9000).

We note also that the variability of the calls arrival process can be smoothed by increasing back-office workload. Via Tables 3.1 and A.2, it can be observed that for a given risk level α, the associated u_α in the former (with low back-office workload) is greater than that in the latter (with high back-office workload).

In order to characterize the limits of the flexibility offered by back-office jobs, we analyze the gain function $G(w)$ associated with the flexibility offered by the back-office workload w. This function is defined as follows. Denote a given value of under-staffing penalty cost by u, for sample values θ and w, recall that the optimal total cost for the call center including the back-office jobs (see Equation (3.5)) is given by

$$C(y^*, \theta, w) = n\,c\,y^* + u \sum_{i=1}^{n} (y^* - v_i(\theta f_i))^- + r \left[w - \sum_{i=1}^{n} (y^* - v_i(\theta f_i))^+ \right]^+, \tag{3.43}$$

with y^* the optimal solution of

$$\min_{y \in \mathbb{N}} \{ n\,c\,y + u\,E[\sum_{i=1}^{n} (y - V_i(\Theta f_i))^-] + r\,E[W - E[\sum_{i=1}^{n} (y - V_i(\Theta f_i))^+]]^+ \}. \tag{3.44}$$

If the back-office workload is considered to be externally processed, for a direct cost $c\,w$, the optimal total cost for the call center without any back-office jobs is given by

$$C'(y'^*, \theta) = n\,c\,y'^* + u \sum_{i=1}^{n} (y'^* - v_i(\theta f_i))^-, \tag{3.45}$$

where y'^* is the optimal solution of

$$\min_{y' \in \mathbb{N}} \{ n\,c\,y' + u\,E[\sum_{i=1}^{n} (y' - V_i(\Theta f_i))^-] \}. \tag{3.46}$$

The gain function $G(w)$ may be therefore written as

$$G(w) \quad = \quad C'(y'^*, \theta) + c\,w - C(y^*, \theta, w). \tag{3.47}$$

If one considers an SP solving methodology, the corresponding expected profit $E[G(W)]$, given

as a function of the expected value of W, is displayed in Figure 3.2 for different penalty cost values $u = 145$, 300, and 1475. The other parameters are identical to those used in Section 3.4.1.

Figure 3.2 shows that $E[G(W)]$ is an increasing concave function of $E[W]$, asymptotically converging towards a constant level for high $E[W]$ values. We see that above a certain amount, additional back-office jobs no longer generate an additional profit. Here the thresholds are 2400, 2700 and 3000 for penalty costs respectively given by $u = 145$, 300 and 1475.

This observation brings forth consideration of the best setting up of the agents groups: the required number of flexible servers (those able to deal with both calls and back-office jobs), and that of the specialized servers. With similar parameters as in Section 3.2, we propose a model with three types of servers: single skilled servers (for calls), single skilled servers (for back-office jobs) and flexible servers. The sizes of the three groups are denoted by y_c, y_{bo}, y_{fx} respectively. In order to force the optimal solution to have a minimum number of flexible servers, while keeping unchanged the total number of servers comparing to the single type (flexible) servers model analyzed above, the flexible servers' salary, say c_{fx}, is assumed to be just slightly increased, w.r.t. the salary of single skilled servers. For our numerical example, we choose $c = 15$, and $c_{fx} = 15.0001$.

The optimization problem can be formulated by the following stochastic integer program:

$$Min \quad nc\,(y_c + y_{bo}) + nc_{fx}\,y_{fx} + u \sum_{l=1}^{L}\sum_{i=1}^{n} p_{\theta_l}\, M_{i,l}^{-} + r \sum_{k=1}^{K}\sum_{l=1}^{L} p_{\theta_l}\, p_{w_k}\, N_{k,l} \tag{3.48}$$

$$\text{s.t.} \quad M_{i,l} = y_c + y_{fx} - v_i(\theta_l\, f_i), \qquad \text{with} \quad i = 1, ..., n, l = 1, ..., L, \tag{3.49}$$

$$M_{i,l} = M_{i,l}^{+} - M_{i,l}^{-}, \qquad \text{with} \quad i = 1, ..., n, l = 1, ..., L, \tag{3.50}$$

$$R_{i,l} \le y_{fx}, \qquad \text{with} \quad i = 1, ..., n, l = 1, ..., L, \tag{3.51}$$

$$R_{i,l} \le M_{i,l}^{+}, \qquad \text{with} \quad i = 1, ..., n, l = 1, ..., L, \tag{3.52}$$

$$N_{k,l} \ge w_k - n\, y_{bo} - \sum_{i=1}^{n} R_{i,l}, \qquad \text{with} \quad l = 1, ..., L, k = 1, ..., K, \tag{3.53}$$

$$M_{i,l}^{+}, M_{i,l}^{-}, N_{k,l}, R_{i,l} \ge 0, \qquad \text{with} \quad i = 1, ..., n, l = 1, ..., L, k = 1, ..., K, \tag{3.54}$$

$$y_c, y_{fx}, y_{bo} \in \mathbb{Z}^{+}. \tag{3.55}$$

This model generalizes the original one (3.6)-(3.10), since we now allow to have three different types of agents (for calls, for emails, and for both) instead of a single type (handling both calls and emails). The variable $R_{i,l}$ represents the number of flexible agents which are assigned to handle back-office jobs in period i of scenario l.

In Figure 3.3, we plot the optimal required number of flexible servers as a function of $E[W]$, for different under-staffing penalty costs $u = 145$, 300, and 1475. Similarly to Figure 3.2, we

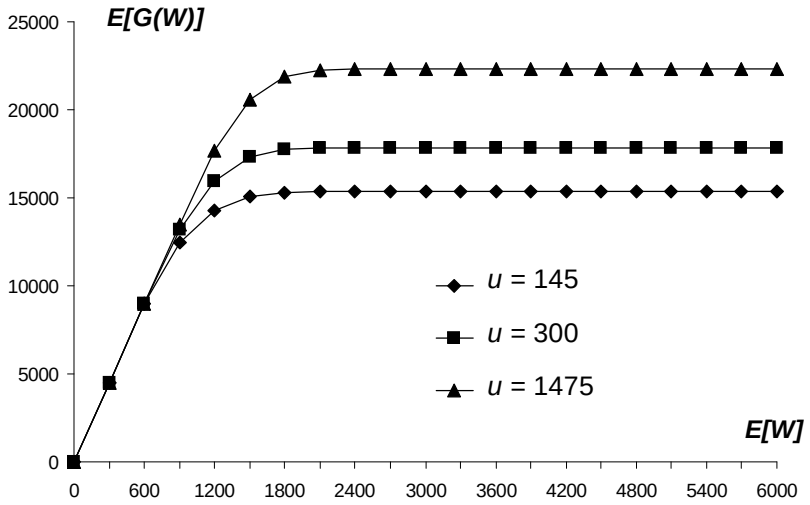

Figure 3.2: Expected gain as a function of the back-office average workload

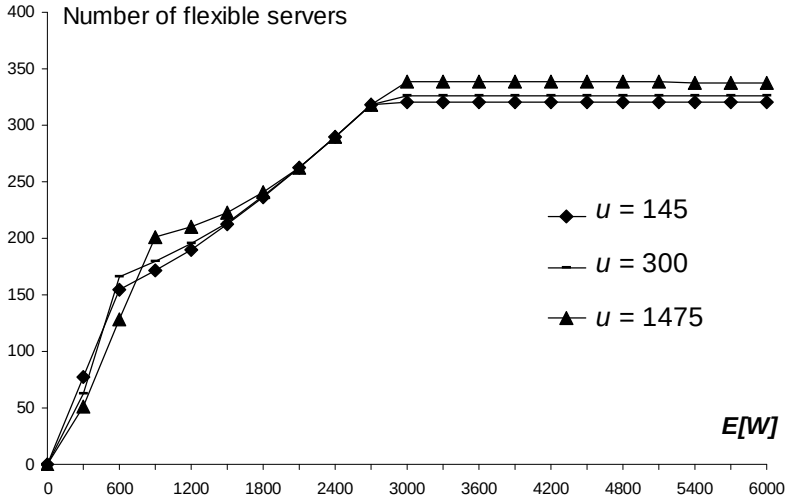

Figure 3.3: Required number of flexible servers

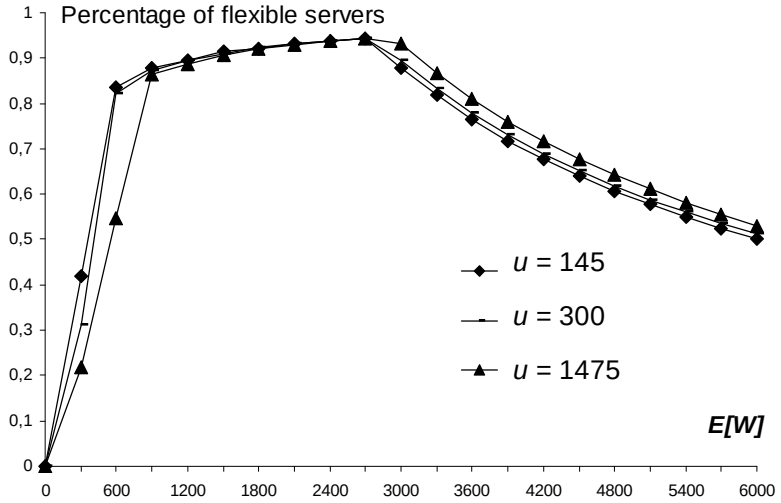

Figure 3.4: Percentage of flexible servers over the total number of servers

observe that the required number of flexible servers is increasing and concave in $E[W]$. The maximum required numbers of flexible servers are 320, 326 and 339 for $u = 145$, 300 and 1475, respectively. Call center flexibility results from the ability of the flexible servers to deal with the two types of jobs. In Figure 3.4, we plot the percentage of flexible servers from the total number of servers as a function of $E[W]$, for $u = 145$, 300, and 1475. Figure 3.4 shows that this percentage decreases after a peak around 90%. The reason is that for a given value of u, the total staffing level increases along with the back-office workload. However, above a certain amount of back-office workload, the required number of flexible servers remains constant. The ratio of flexible servers will therefore decrease.

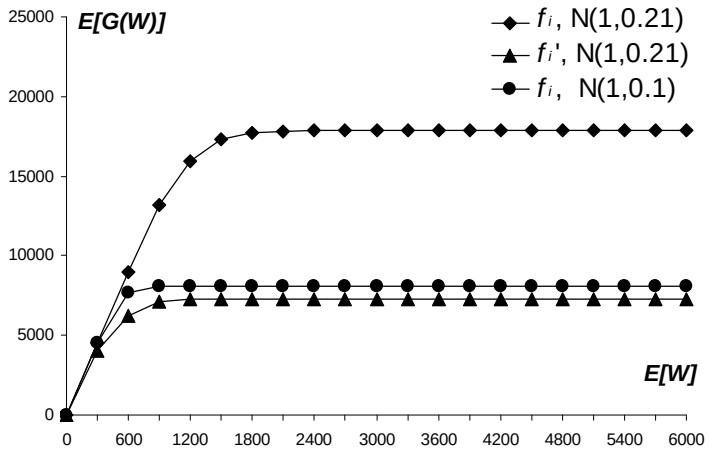

Figure 3.5: Expected gain for different seasonal pattern and busyness variance

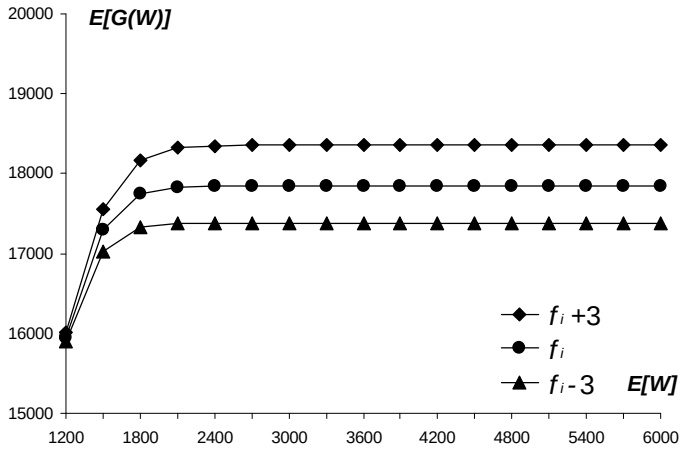

Figure 3.6: Expected gain for different call center size

The impact of the flexibility on costs performance depends also on the value of the under-staffing penalty cost and the variability of inbound calls. For small workloads of back-office jobs, Figure 3.2 shows that the gains associated with the flexibility are constant w.r.t. u. Indeed, in such situations, the over-staffed agents (for calls) can easily handle the back-office jobs without any additional cost. For large back-office workloads and high under-staffing penalty cost, the gain is larger because a higher staffing level is required, inducing more over-staffing. The next examples, illustrated by Figures 3.5 and 3.6, show the impact of variability of inbound calls on the gain associated with the flexibility offered by back-office jobs. This inbound calls variability results from the combination of the variations of the daily deterministic pattern (defined by the variations of the f_i coefficients), of the Θ random variability and of the inbound calls total average workload (defined by $\sum_{i=1}^{n} f_i$), that can be viewed as the call center size.

In Figure 3.5, we compare three cases. First, the original case is depicted, corresponding to the daily pattern f_i and the random variable Θ with a Gaussian distribution (with mean equal to 1 and standard deviation equal to 0.21). The two other cases have smoother calls workload fluctuations, but still the same global daily workload. For one example, we keep the variability of the process Θ similar, but we smooth the daily pattern by fixing f'_1, f'_2, ..., f'_{11} equal to 13.5, 18.4, 24.4, 21.5, 19, 22.9, 18.4, 25, 17.4, 17.2 and 15.3 calls per minute. It is worth noting that $\sum_{i=1}^{11} f_i = \sum_{i=1}^{11} f'_i$. In the second case, the standard deviation of the Θ random variable is decreased, from 0.21 to 0.10, but the the daily pattern is still given by the f_i parameters. As expected, it can be seen in Figure 3.5 that the benefit obtained through flexibility decreases when overall variability decreases, either because of a smoother seasonal pattern, or because of a reduction of the daily busyness variance.

In Figure 3.6, we compare three cases with similar successive periodic daily fluctuations (i.e., with similar values for the differences $f_{i+1} - f_i$) and similar Θ process. However, we vary the inbound calls total average workload (defined by $\sum_{i=1}^{n} f_i$), which can be viewed as varying the call center size. We have considered coefficients respectively given by the sequences $f_i + 3$, f_i and $f_i - 3$. The figure shows a decrease in the gain obtained from server flexibility for call center with reduced size. The underlying reason is simple: the size of the stochastic fluctuations due to Θ is reduced, for smaller f_i coefficients, and, as a consequence, the required (or useful) flexibility level is also smaller.

3.5 Extension to Models with Overflow

In this section, we extend the analysis to call centers with possible call overflows between successive periods. Some call center models which include an overflow process have been analyzed in the literature (see Thompson (1993) and Stolletz (2008)). According to these papers, we assume the outcome of the arrival rate λ_i, in period i, to be substituted by a modified auxiliary arrival rate λ_i^M, given by

$$\lambda_i^M(y) = \lambda_i + b_{i-1}(y) - b_i(y), \tag{3.56}$$

for $2 \leq i \leq n-1$, where $b_i(y)$ is the arrival rate generated by the backlog of period i. For the boundary periods $i = 1$ and $i = n$, we have $\lambda_1^M(y) = \lambda_1 - b_1(y)$ and $\lambda_n^M(y) = \lambda_n + b_{n-1}(y)$, respectively. These backlogs $b_i(y)$ are evaluated via an Erlang-loss system (see Stolletz (2008)). The overflow impacts are introduced in our cost model as follows. An overflow rate $b_{i-1}(y)$ can be viewed as associated to an under-staffing situation of $\lceil \frac{b_{i-1}(y)}{\mu} \rceil$ agents, where $\lceil x \rceil$ denotes the smallest integer not less than x. The penalty cost $u \lceil \frac{b_{i-1}(y)}{\mu} \rceil$ is then added in Equation (3.5) and we have the following new cost function expression,

$$n\,c\,y + u \sum_{i=1}^{n} (y - v_i(\theta f_i))^- + r \left[w - \sum_{i=1}^{n} (y - v_i(\theta f_i))^+ \right]^+ + u \lceil \frac{b_{i-1}(y)}{\mu} \rceil. \tag{3.57}$$

In Equation (3.57), two different kinds of under-staffing are penalized: one with respect to the agent requirement according to the service level (defined by Condition (2.3)) and another one for overflow (based on $\lceil \frac{b_{i-1}(y)}{\mu} \rceil$). The staffing level $v_i(\theta f_i)$ which guarantees the required service level is calculated based on the auxiliary arrival rate of Equation (3.56).

Since the overflow rates $b_i(y)$ are non-linear functions of the decision variable y, the non-linear optimization problem (3.57) is solved via successive iterations by updating in each iteration the values of the overflows $b_i(y)$.

The SP approach has been successively applied to the original model and to the model with overflow for three numerical examples. The parameter values of these examples are the same as in Section 3.4.1. For each example, Table 3.2 displays the optimal staffing levels for the original model without overflow, denoted by y^* and the optimal solution for the overflow model, y_M^*.

Table 3.2: Optimal staffing levels

	$E[W] = 50$, $\sigma_W = 5$	$E[W] = 600$, $\sigma_W = 60$	$E[W] = 1000$, $\sigma_W = 100$
$u_{1\%}$	1475	1475	1350
y^*	233	233	233
y_M^*	229	229	229
$u_{5\%}$	300	300	166
y^*	202	202	202
y_M^*	201	201	202
$u_{10\%}$	145	140	30
y^*	184	184	184
y_M^*	184	185	184

From Table 3.2, we see that the gap between the staffing levels for the two models is small, which tends to support the robustness of our original model. We have noticed from the numerical experiments that the algorithm with successive iterations very quickly converges after a limited number of steps.

3.6 Concluding Remarks

We have developed a single shift call center model with two types of jobs: inbound calls and back-office jobs. We focused on optimizing the staffing level w.r.t. the total operating cost of the call center.

We modeled this problem as a cost optimization-based newsboy-type model. We then proposed various approaches to solve it numerically: a classical stochastic programming approach, a robust programming approach and a mixed robust programming approach. We next conducted a numerical study in order to evaluate the performance of each approach and gain useful insights. First, by comparing with the average based deterministic approximation, we underlined the necessity of taking into account the uncertainty in the call demand parameters, which is usually not the case in the majority of existing studies. Second, we highlighted the respective advantages and drawbacks of each approach. Third, we showed to what extent the flexibility associated with storable back-office jobs helps in absorbing uncertainty in the call process.

In the succeeding chapters, we intend to extend the analysis of this chapter to a multi-shift setting, with the possibility of removing or adding agents within the same day. Another extension is to consider a global service level constraint for the whole day, instead of having a period by period constraint.

Chapter 4

Multi-shift Staffing Problem with Information Update

In this chapter we consider a multi-periodic multi-shift call center staffing problem. The call arrival process is assumed to follow a doubly non-stationary stochastic process with a random mean arrival rate. The number of agents working in each shift is decided initially before the beginning of the working day, but a real-time update is allowed within the same day. The objective is to minimize the sum of the regular salary, the update adjustment cost and the penalty cost of agents under-staffing. We focus the analysis to the case where all shifts are without break. Two different solution approaches are considered. First, by the discretization of the underlying probability distribution, we explicitly formulate the expected cost formulation as a two-stage stochastic program with recourse. Using the property of totally unimodular, we prove that the linear relaxation of this stochastic program is integral. And the large-scaled mixed integer program (MIP) can be relaxed and be solved efficiently. Second, we develop a robust programming formulation for this two-stage optimization problem. Also, we show how a property of totally unimodular can help to make integer second-stage decisions, using piece-wise linear recourse rules. The MIP becomes very easy to be solved. The efficiency and excellent performance of these two approaches are illustrated through a numerical study based on real-life data. The advantage of adding the update flexibility and the necessity of taking into account the parameter uncertainty are also demonstrated.

4.1 Introduction

As presented in the previous chapters, the arrival process of calls includes two types of uncertainty: the usual uncertainty captured by a stochastic process modeling, and the uncertainty in the process parameter themselves. In this chapter, we continue the study of the shift-scheduling problem of a call center, in which we allow the mean arrival rate of calls to be uncertain. The arrival process of calls keeps being modelled by a doubly non-stationary stochastic process, with random mean arrival rates. Different to that in Chapter 3, we consider a multi-shift call center in this chapter.

Due to the non avoidable errors on the forecasting (on calls, on the effective number of agents who will be present, etc.) which considerably affect the efficiency of the beforehand planning, and considering the existence of significant correlation between arrivals in different times intervals in the same day (see Avramidis et al. (2004)), we have recently seen in practice a new planning activity. This new activity is referred to as intra-day performance management or *trafficing*, for which decisions are made through several steps. The first decisions of staffing are made before the day of the interest, and the other decisions are taken during the day itself. The latter can be seen as corrections of the beforehand planning. As a function of the actual demand during the beginning of the day in question, trafficing consists on taking very-short horizon decisions within the same day. These decisions would correct/adjust the capacity using some available flexibility (for example adding or removing some agents) in order to efficiently handle the actual demand. Recent empirical research (Mehrotra et al. (2009)) estimates that over 70% of call centers routinely make trafficing based on largely experience and intuition. However, the literature on this subject is still quite poor with regard to a such interesting feature. In this chapter, we consider a shift-scheduling problem with trafficing.

As mentioned above, our shift-scheduling problem incorporates uncertainty in the call arrival parameters, and allows the decision-maker to adjust his initial decisions after that more call volume information have become available. The shift-scheduling problem is modeled as a cost optimization-based two-stage model. The cost criterion function includes the regular salary cost, adjustment cost and a penalty cost for under-staffing. Our objective is to find the optimal initial shift scheduling and update policy which minimize the total call center operating cost. Concerning the shifts, we assume a particular case where no breaks present at any period in the middle of the shift and each agent works consecutively until the shift ends. This leads to an important property of the period-shift matrix called totally unimodular (TU). In this chapter, we drive helpful and interesting results thanks to this totally unimodular property. One may take our results as the first step, and extend to the general case where beaks present within

shifts, using heuristic methods (Gans et al. (2003)). We propose two solution methodologies. First, we formulate the problem as a two-stage stochastic program with recourse, by discretizing the probability distribution and constructing the associated event-trees and scenarios. Using the totally unimodular property, we prove that the large scale Mixed Integer Problem (MIP) can be solved efficiently by just relaxing it to a linear problem (LP). The second approach relies on adjustable robust optimization theory. Again by taking advantage of the totally unimodular property, we make some modification on the Affinely Adjustable Robust Counterpart (AARC) methodology which is supposed to be applied in linear problems. Then we solve the MIP very efficiently and get piecewise linear update policies. A numerical study is conducted in order to illustrate the efficiency of the two approaches and the advantage to provide update flexibility. In the numerical illustration, we use real data gathered from a call center of a Dutch hospital.

We distinguish two main contributions in this chapter. The first contribution is, under the assumption that the period-shift matrix has the totally unimodular property, we propose two approaches to modeling and to solve efficiently a large scale two-stage call center shift-scheduling problem with uncertain arrival parameters. The second contribution is the analysis of the added advantage of using dynamic adjustment (update). We show that the update action reduces the operational cost and the under-staffing probability.

The most related work to ours which develops frameworks to make intra-day resource adjustment decisions in call centers is that of Mehrotra et al. (2010) and Gans et al. (2009). However, the former suppose that the initial schedules existed and solve the real-time agents schedule adjustment as a one-stage static problem. Gans et al. (2009) extends to include forecast updates and two-stage stochastic programs with recourse. But they use neither the totally unimodular property nor the adjustable robust approach.

The rest of the chapter is structured as follows. In Section 4.2, we describe the call center model under consideration and formulate the associated shift-scheduling problem. In Section 4.3 and 4.4, we present the different solution approaches and analyze their interesting properties. In Section 4.5, we then conduct a numerical study to evaluate these alternative approaches. We also exhibit the advantage of doing trafficing. The chapter ends with concluding remarks.

4.2 Problem Formulation

We consider a multi-period multi-shift call center with a single type of inbound calls. The inbound call mean arrival rate in each period is allowed to be uncertain. The periods of a day are assumed to be divided into two time horizons. During the early time horizon, the decision-maker implements the first-stage (here-and-now) decisions without information on the actual

requirements. After the uncertainty in the early time horizon has realized, the decision-maker has a better estimation of the uncertainty in the later time horizon. Then he chooses the second-stage (wait-and-see) decisions for the later time horizon according to his observation. In this section, we describe the corresponding two-stage recourse workforce shift-scheduling problem.

4.2.1 The Inbound Call Arrival Process

Due to the characteristics of the arrival process of calls presented in Section 2.1.2 (Chapter 2), in order to address uncertain and time-varying mean arrival rates coupled with significant correlations, we model the inbound call arrival process by a doubly stochastic Poisson process (see Avramidis et al. (2004); Harrison and Zeevi (2005), and Whitt (1999)) as follows. We assume that a given working day is divided into t distinct, equal periods of length T, so that the overall horizon is of length tT. The period length in practice is often 15 or 30 minutes.

The inbound calls arrive following a stochastic process with a random arrival rate in each period i, denoted by Λ_i. Furthermore, using the modeling in Avramidis et al. (2004) and in Whitt (1999), we assume that the arrival rate Λ_i is of the form

$$\Lambda_i = \Theta f_i, \text{ for } i = 1, ..., t, \tag{4.1}$$

where Θ is a positive real-valued random variable. The random variable Θ can be interpreted as the unpredictable busyness of a day: A large (small) outcome of Θ corresponds to a busy (not busy) day. The constants f_i model the intra day seasonality, e.x., the shape of the variation of the arrival rate intensity across the periods of the day, and they are assumed to be known. Formally, if a sample value in a given day of the random variable Θ is denoted by θ, the corresponding outcome of the arrival rate over period i for that day is defined by $\lambda_i = \theta f_i$.

We assume that service times for inbound calls are independent and exponentially distributed with rate μ. The calls arrive to a single infinite queue working under the the first come, first served (FCFS) discipline of service. Neither abandonment nor retrials are allowed. The staffing level which guarantees the required service level is then computed by

$$n_i(\theta) \quad = \quad F^{-1}_{\theta f_i}(SL_i). \tag{4.2}$$

with the function F defined by Equation (2.3). We denote the required staffing level by $n_i(\theta)$ because f_i is constant for each i.

4.2.2 Shifts Setting

We denote the period sets of the overall daily horizon by I. In this chapter, we consider a multi-shift call center. Two catalogs of shifts are considered: the initial decisions (before update) and the update decisions.

Let J be the set of all the feasible work schedules, each of which dictates if an agent answers calls in period $i \in I$. We define the $|I| \times |J|$ matrix $\mathbf{A} = [a_{ij}]$, where for $i \in I$ and $j \in J$,

$$a_{ij} = \begin{cases} 1, & \text{if agents in schedule } j \text{ answer calls during period } i, \\ 0, & \text{otherwise.} \end{cases}$$

We divide the overall horizon into the early horizon and late horizon, denoted by I_1 and I_2, respectively. After observing the call volumes in the early horizon, a real-time update of staff capacity is allowed at the beginning of the late horizon. We define also the $|I| \times |J|$ matrix $\mathbf{A}' = [a'_{ij}]$ with

$$a'_{ij} = \begin{cases} 1, & \text{if agents in new (after updating) schedule } j \text{ answer calls during period } i, \\ 0, & \text{otherwise.} \end{cases}$$

Note that here the first $|I_1|$ lines of matrix $\mathbf{A}'$ are all zeros.

4.2.3 Cost Criterion

Let the first-stage (here-and-now) decision variables X_j, $j \in J$, represent the numbers of agents assigned to the various schedules implemented before the start of the overall planning horizon. And the second-stage (wait-and-see) decision variables Y_j and Z_j, $j \in J$, denote respectively the numbers of agents added to and removed from the schedule j after the observation of uncertainty of the early horizon. And $n_i(\theta)$ represents the required agents number for period i related to the busyness factor realization θ.

Each agent initially assigned to shift j gets a salary c_j for the day. For the recourse actions, the cost of adding an agent to schedule j is d_j, the cost saving by removing an agent from scheduling j is r_j, $j \in J$. In each period, a per person under-staffing penalty u is penalized if the number of agents assigned to this period is less than the required agents number. As usual, the cost parameters satisfy the ordering $r_j < c_j < d_j$ for $j \in J$, which ensures that it costs more to modify the staffing level by second-stage decisions than to determiner it by the initial scheduling. The setting that u is larger than periodic (adding agents or regular) salary is straightforward.

Primarily, we use a small numeric example to illustrate the advantage of making trafficing.

Example:

Suppose that a call center is divided into three period: morning, noon and afternoon, ($i = 1, 2, 3$). The busyness of a day has three scenarios ($l = 1, 2, 3$): busy, average, not busy day, with L as the set of scenarios. Consequently a subscript l relating to scenario is added for n_i, Y_j and Z_j and the notations become n_{il}, Y_{jl}, Z_{jl}. The required agents numbers are $\mathbf{n} = [n_{il}]$: $[15, 20, 25; 45, 60, 75; 30, 40, 50]$. The probabilities of the three busyness scenarios (busy, average, not busy) are $p_1 = 0.3$, $p_2 = 0.4$ and $p_3 = 0.3$ respectively. Suppose that there are two feasible shifts ($j = 1, 2$), with the definition of the matrix $\mathbf{A}$ as $[1, 1; 0, 1; 1, 1]$. The update is done after the first period, and we define the matrix $\mathbf{A}'$ as $[0, 0; 0, 1; 1, 1]$. The regular salary cost $c_1 = 2, c_2 = 3$, and the adjustment update cost is defined as $d_1 = 1.2, d_2 = 2.4$ and $r_1 = 0.8, c_2 = 1.6$. The under staffing penalty is $u = 5$.

We assume that the decision-maker knows exactly which scenario the day belongs to after his observation during the first period. We solve the following two shift-scheduling problems: Problem (4.3) without update and Problem (4.4) allowing update, and compare their total costs.

$$
\begin{aligned}
Min \quad & \sum_{j \in J} c_j \, X_j + \sum_{i \in I} u \, M_{il} \\
\text{s.t.} \quad & \sum_{j \in J} a_{ij} X_j + M_{il} \geq n_{i,l}, i \in I, l \in L, \\
& M_{il} \geq 0, i \in I, l \in L, \\
& X_j \in \mathbb{Z}^+, j \in J.
\end{aligned}
\tag{4.3}
$$

$$
\begin{aligned}
Min \quad & \sum_{j \in J} c_j \, X_j + \sum_{j \in J, l \in L} p_l \, (d_j \, Y_{jl} - r_j \, Z_{jl}) + \sum_{i \in I} u \, M_{il} \\
\text{s.t.} \quad & \sum_{j \in J} a_{ij} X_j + M_{il} \geq n_{i,l}, i \in I_1, l \in L, \\
& \sum_{j \in J} a_{ij} X_j + \sum_{j \in J} a'_{ij} (Y_{jl} - Z_{jl}) + M_{il} \geq n_{i,l}, i \in I_2, l \in L, \\
& X_j \geq Z_{jl}, j \in J, l \in L, \\
& M_{il} \geq 0, i \in I, l \in L, \\
& X_j, Y_{jl}, Z_{jl} \in \mathbb{Z}^+, j \in J, l \in L.
\end{aligned}
\tag{4.4}
$$

In the two equations above, it is obvious that Y_{jl}, Z_{jl} and $n_{i,l}$ are related to the realized busyness factor θ_l, and we let M_{il} present the agents number shortfall in period i of scenario l. We find out that the optimal costs of Problems (4.3) and (4.4) are 202.5 and 181.5 respectively. This small example shows the interest on cost saving of using available information to adjust the

staffing capacity in real-time.

4.2.4 Totally Unimodular Matrix

In this chapter we consider a particular case where each agent is to work consecutive periods, without breaks. Then every column of both matrix $\mathbf{A}$ and $\mathbf{A}'$ has contiguous ones and this kind of matrix is totally unimodular. Totally unimodular matrices are of extreme importance in polyhedral combinatorics and combinatorial optimization. It is well known that if matrix $\mathbf{A}$ is totally unimodular and vector $\mathbf{b}$ is integral, every extreme point of the feasible region $\{\mathbf{x} \mid \mathbf{Ax} \geq \mathbf{b}\}$ is integral and thus the feasible region is an integral polyhedron. This gives a quick way to verify that a linear program is integral (has an integral optimum, when any optimum exists). The property of totally unimodular facilitates a lot the solving process for our solution methodologies presented in Section 4.3 and 4.4.

4.2.5 Information Update

As mentioned above, Θ is an random variable which can be interpreted as the unpredictable busyness factor. Since the decision-maker has an opportunity to adjust the agents capacity during the day, he may would like to know better the realized busyness factor θ before doing the adjustment. The reason is simply that θ is the crucial and unique unknown parameter to calculate the required gents number. Unfortunately, in real life, the decision-maker can not observe directly the value of θ, but only use the observed call volumes in the early horizon periods to get an estimate value $\tilde{\theta}$. In this chapter, we consider this realistic situation.

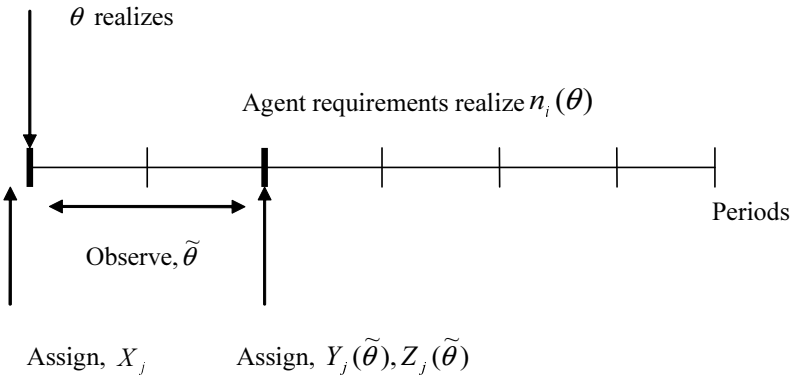

Figure 4.1: Two-stage staffing process

Figure 4.1 shows the staffing process we consider in this problem. At the beginning of the

day, the initial staffing policy $X_j, j \in J$, is applied. The true busyness factor θ realizes but the decision-make observes and get only an estimate busyness factor $\tilde{\theta}$. He then chooses the second-stage (wait-and-see) decision variables Y_j and Z_j, $j \in J$, according to the estimated uncertain parameter value $\tilde{\theta}$.

4.2.6 Problem Setting

For any given fist-stage decisions $X_j, j \in J$, a true busyness factor θ, and an observed busyness factor $\tilde{\theta}$, the second-stage cost of the call center, associated with a second-stage decision variables Y_j, Z_j, is

$$\pi'(X_j, \tilde{\theta}) := \sum_{j \in J} d_j\, Y_j - \sum_{j \in J} r_j\, Z_j + \sum_{i \in I_2} u\,(n_i(\theta) - \sum_{j \in J} a_{ij} X_j - \sum_{j \in J} a'_{ij}(Y_j - Z_j))^+. \tag{4.5}$$

The variables Y_j, Z_j will be optimized for this given X_j and observed busyness factor $\tilde{\theta}$. In Equation (4.5), the first and second terms correspond to the staffing capacity adjustment cost, and the last term corresponds to the agents shortfall penalty for the late horizon periods. The optimization problem at the second-stage is to determine the optimal set of variables $Y_j, Z_j, j \in J$ to minimize the second-stage cost $\pi'(X_j, \tilde{\theta})$. In order to keep being logic, we ask that the number of agents removed is less than the number of agents initially assigned, $X_j \geq Z_j$, for each shift $j \in J$.

In order to formulate the problem as an integer linear program, we replace x^+ by $y \geq x, y \geq 0$. The two-stage recourse shift-scheduling problem can be formulated as the following mixed integer program(MIP):

$$\begin{aligned}
Min \quad & \sum_{j \in J} c_j\, X_j + \sum_{j \in J} d_j\, Y_j - \sum_{j \in J} r_j\, Z_j + \sum_{i \in I_1} u\, M_i + \sum_{i \in I_2} u\, M'_i \\
\text{s.t.} \quad & \sum_{j \in J} a_{ij} X_j + M_i \geq n_i(\theta), i \in I_1, \\
& \sum_{j \in J} a_{ij} X_j + \sum_{j \in J} a'_{ij}(Y_j - Z_j) + M'_i \geq n_i(\theta), i \in I_2, \\
& X_j \geq Z_j, j \in J, \\
& M_i \geq 0, i \in I_1, \\
& M'_i \geq 0, i \in I_2, \\
& X_j, Y_j, Z_j \in \mathbb{Z}^+, j \in J.
\end{aligned} \tag{4.6}$$

The objective of this model is to minimize the total cost of initial scheduling, recourse decisions and the under-staffing penalty cost associated with failure to satisfy all staffing require-

ments. The first and second sets of constraints calculate the agents number shortfall M_i ($i \in I_1$) and M_i' ($i \in I_2$), which depend on θ and $(\theta, \tilde{\theta})$ respectively. The third set of constraints ensures that the number of agents removed is less than the number of agents initially assigned, for each shift $j \in J$. The last three sets of constraints define the non-negativity and integer conditions for program variables.

In the following two sections we develop two methods to solve the shift-scheduling problem given by Equation (4.6). These approaches are then demonstrated by the numerical study in Section 4.5. First, under the assumption that the probability distributions associated with the random variables are known exactly, a *two-stage stochastic programming approach* is applied to problem (4.6), built on the discrete probability distributions characterizing uncertain parameters. The second approach refereed to as *adjustable robust programming* consists of optimizing the staffing level with respect to the worst case scenarios in a given uncertainty set. In what follows, we describe these approaches.

4.3 Two-Stage Stochastic Programming Approach

The traditional way to take into account parameters uncertainty consists of using a stochastic programming formulation, which minimizes the expected cost by assuming that the parameters obey a known probability distribution.

Suppose that the true busyness factor θ has the probability density function $p(\theta)$, the standard way (as that in Gans et al. (2009)) to estimate the probability density function of the observed busyness factor $\tilde{\theta}$ is as follows. Assume that the joint probability density function of the true and estimated busyness factor is $\psi(\theta, \tilde{\theta})$ and $p(\theta) > 0$ for each θ. After the beginning of the day that a true busyness factor θ is realized, the conditional probability density function of the estimated busyness factor $\tilde{\theta}$, given that $\Theta = \theta$, is given by $pb(\tilde{\theta}|\theta) = \psi(\theta, \tilde{\theta})/p(\theta)$.

In this chapter, we assume the random variable Θ to follow a discretized probability distribution, defined by the sequence of outcomes θ_l and the associated sequence of probabilities p_l, $l \in L$, with L as the set of scenarios of possible true busyness factor. We have $\sum_{l \in L} p_l = 1, p_l \geq 0$. For a given θ_l, the decision-maker estimates it as $\tilde{\theta}_k$ with probability $pb_{l,k}$, $k \in K$, with K as set of scenarios of possible estimated busyness factor. For each $l \in L$, we have $\sum_{k \in K} pb_{l,k} = 1, pb_{l,k} \geq 0$. To better understand the probability relation between θ_l and $\tilde{\theta}_k$ please see Figure 4.2.

Let $X_j, j \in J$, represent initial scheduling decisions. Late-horizon recourse decisions vary by the estimated value of $\tilde{\theta}_k$. The decision variables $Y_{j,k}, Z_{j,k}$ with $j \in J, k \in K$ represent the full set of the recourse actions.

As shown in Figure 4.3, before the day begins, the initial scheduling decisions $X_j, j \in J$, are

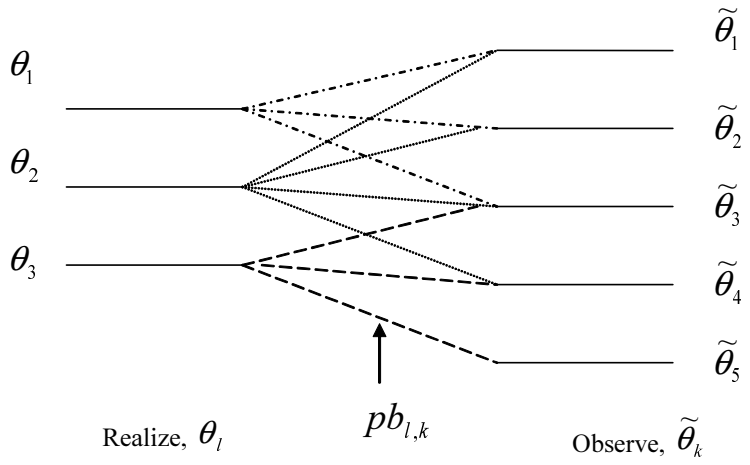

Figure 4.2: Probability relation between θ_l and $\tilde{\theta}_k$

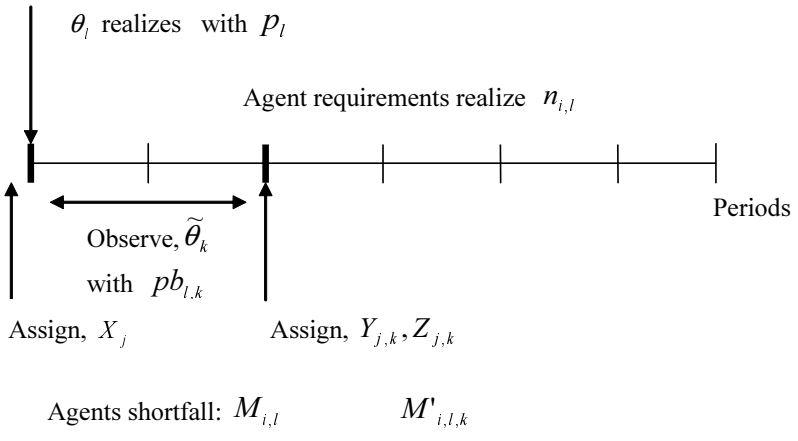

Figure 4.3: Two-stage staffing process in stochastic setting

assigned. When the day begins, a scenario l may realize with the true busyness factor θ_l, the periodic required numbers of agents also realize, we dote them by $n_{i,l}$ for $i \in I$. The number of agents shortfall for each early horizon periods is denoted by $M_{i,l}$ for $i \in I_1$. After the observation, the true busyness factor θ_l is estimated to be $\tilde{\theta}_k$, $k \in K$. According to the estimated $\tilde{\theta}_k$, the decision-maker then implements the corresponding second-stage policy $Y_{j,k}, Z_{j,k}$ with $j \in J$. In this case, the number of agents shortfall for each later horizon period is denoted by $M'_{i,l,k}$, for $i \in I_2$.

The objective the two-stage stochastic integer program is to optimize the expected total cost of initial scheduling, recourse actions and under-staffing penalty. It determines the optimal set of first-stage decisions and the set of recourse actions for each possible estimated value $\tilde{\theta}_k$. Mathematically, the two-stage stochastic counterpart of problem (4.6) can be formulated as the following MIP:

$$
\begin{aligned}
Min \quad & \sum_{j\in J} c_j\, X_j + \sum_{l\in L}\sum_{i\in I_1} p_l\, u\, M_{i,l} + \sum_{l\in L}\sum_{k\in K} pb_{l,k}\,[\sum_{j\in J} d_j\, Y_{j,k} - \sum_{j\in J} r_j\, Z_{j,k} + \sum_{i\in I_2} u\, M'_{i,l,k}] \\
\text{s.t.} \quad & \sum_{j\in J} a_{ij} X_j + M_{i,l} \geq n_{i,l}, i \in I_1, l \in L, \qquad (4.7) \\
& \sum_{j\in J} a_{ij} X_j + \sum_{j\in J} a'_{ij}(Y_{j,k} - Z_{j,k}) + M'_{i,l,k} \geq n_{i,l}, i \in I_2, l \in L, k \in K, \\
& X_j \geq Z_{j,k}, j \in J, k \in K, \\
& M_{i,l} \geq 0, i \in I_1, l \in L, \\
& M'_{i,l,k} \geq 0, i \in I_2, l \in L, k \in K, \\
& X_j, Y_{j,k}, Z_{j,k} \in \mathbb{Z}^+, j \in J, k \in K.
\end{aligned}
$$

The size of this model is impacted mainly by the size of shift schedules set J, the sizes of scenarios sets L and K. The number of integer variables is equal to $|J| + 2 \times |J| \times |K|$, while the number of continuous variables is equal to $|I_1| \times |L| + |I_2| \times |L| \times |K|$.

Example:

We consider a day with 25 periods where the first 5 periods are considered as early horizon and the rest are later horizon. The number of possible shifts is 162. Set L contains 200 scenarios, and set K contains 21 scenarios. This implies the requirement to solve models with 6966 integer variables and 85 000 continuous variables. As analyzed in van Delft and Vial (2004), this kind of scenario tree leads to a large scale problem. Fortunately, we discover an interesting and important property of the MIP (4.7) as presented in the following theorem.

Theorem 4.1 *The relaxed linear programs of MIPs (4.7) is integral (has an integral optimum, when any optimum exists).*

The proofs is given in the appendix.

The property given in Theorem 4.1 is directly used in the optimization procedure. We relax the MIP (4.7) and solve a linear problem.

4.4 Adjustable Robust Approach

In contrast with the stochastic programming framework which explicitly requires a probability description of the uncertainty, robust optimization models uncertain parameters using uncertainty sets. The objective is then to minimize the worst-case cost in those sets. Suppose that the uncertainty realizations of the parameters θ lie in the uncertainty set U. If we ignore the fact that the decision variables Y_j and Z_j could be determined after observation and getting more information about the uncertain parameter θ, the static one-stage robust counterpart of problem (4.6) can be formulated as:

$$\begin{aligned}
\min_{X_j, Y_j, Z_j} \quad & \max_{\theta \in U} \sum_{j \in J} c_j X_j + \sum_{j \in J} d_j Y_j - \sum_{j \in J} r_j Z_j + \sum_{i \in I} u M_i \\
\text{s.t.} \quad & \sum_{j \in J} a_{ij} X_j + \sum_{j \in J} a'_{ij} (Y_j - Z_j) + M_i \geq n_i(\theta), i \in I, \\
& X_j \geq Z_j, j \in J, \\
& M_i \geq 0, i \in I, \\
& X_j, Y_j, Z_j \in \mathbb{Z}^+, j \in J.
\end{aligned} \tag{4.8}$$

Note that in the formulation above, M_i with $i \in I$, are the state variables which describe the staffing capacity shortfall, and they depend on the value of θ. The decision variables are X_j, Y_j and $Z_j, j \in J$. The optimal solution (X_j^*, Y_j^*, Z_j^*) of (4.8) satisfies the constraints for all possible realizations θ, and guarantees a worst case objective cost. Thus it is called robust. But this worst case objective cost would be grossly overestimated since the static one-stage robust counterpart losses the flexibility that (Y_j, Z_j) can be decided after getting a better information about the uncertainty.

Ben-Tal et al. (2004) first extended the robust optimization framework to dynamic setting. Similar to the two-stage stochastic optimization with recourse, the decision-maker selects the here-and-now, or first-stage decisions, before having any knowledge of the actual value about the uncertainty. He observes then the realization of the uncertainty and after, and he chooses the wait-and-see, or second-stage decisions according to the outcome of the uncertainty. In a short, rather than re-optimization, the decision-maker adjusts his strategy according to information revealed over time using policies. Recall that in our problem setting, the second-stage (wait-and-

see) decisions do not have to be determined a priori, the staffing adjustment strategy Y_j and Z_j could be adjusted according to the revealed information θ.

We introduce as follows the dynamic formulation by assuming that the decision-maker could estimate the true uncertainty parameter θ after observation. We can reformulate the Min-Max adjustable robust counterpart of problem (4.6) as follows.

$$\begin{aligned}
\min_{X_j, Y_j, Z_j} \quad & \sum_{j\in J} c_j\, X_j + \max_{\theta\in U}\{\sum_{j\in J} d_j\, Y_j(\theta) - \sum_{j\in J} r_j\, Z_j(\theta) + \sum_{i\in I} u\, M_i\} \\
\text{s.t.} \quad & \sum_{j\in J} a_{ij} X_j + M_i \geq n_i(\theta), i \in I_1, \\
& \sum_{j\in J} a_{ij} X_j + \sum_{j\in J} a'_{ij}(Y_j(\theta) - Z_j(\theta)) + M_i \geq n_i(\theta), i \in I_2, \\
& X_j \geq Z_j(\theta), j \in J, \\
& M_i \geq 0, i \in I, \\
& X_j, Y_j(\theta), Z_j(\theta) \in \mathbb{Z}^+, j \in J.
\end{aligned} \tag{4.9}$$

Problem (4.9) is more flexible than Problem(4.8), the notations $Y_j(\theta)$ and $Z_j(\theta)$ indicate that adjustable strategy Y_j and Z_j depend on the revealed information θ. But this flexibility comes at the expense of tractability. Mathematically, even by relaxing the integer constraints in Problem (4.9) and assuming that parameters $n_i(\theta)$ are linear in θ, the full adaptability is NP-hard.

To address this issue, in Ben-Tal et al. (2004), the authors suggested an approximation to the linear adjustable robust counterpart which they call the *Affinely Adjustable Robust Counterpart*(AARC). They proposed a so called *linear decision rule* (LDR) which restricts that the future decisions as affine functions of the revealed uncertainty. An important case of AARC is that the parameters associated with the adjustable variable in the linear problem are constants, independent of the uncertainty. This case is known as *fixed recourse*. Ben-Tal et al. (2004) show that AARCs with fixed recourse are computationally tractable for a wide spectrum of uncertainty set. A brief review of the AARC methodology is given in Section 2.2.2 in Chapter 2.

In the problem we consider in this chapter, the parameters associated with the adjustable variable are constants (fixed recourse), however, there exists three major difficulties:

1. The decision-maker can not observe the true busyness factor θ. He can only observe the call volume and get an estimate busyness factor $\tilde{\theta}$, which contains some error on estimation.

2. Parameters $n_i(\theta)$ are not affine, but non-linear increasing concave functions of random parameter θ, we don't have an affinely adjustable robust counterpart.

3. The second-stage decisions Y_j and Z_j are integer variables, consequently, the linear decision

rule can not be applied directly to the second-stage decision variables.

In the following three subsections, we explain our methods to treat these difficulties. In Section 4.4.1 we introduce the robust adjustable model with error on estimating the uncertainty parameter θ. In Section 4.4.2 we use piecewise linear approximation to approach the non-linear function $n_i(\theta)$. Section 4.4.3 presents the way to treat the second-stage integer decision variables: We apply the LDR on the periodic staffing level $\sum_{j\in J} a'_{ij}(Y_j - Z_j)$ instead of on shift staffing level Y_j and Z_j. With this modification, on the one hand, we can determine the parameters of the linear decision rule just like what is done in Ben-Tal et al. (2004). One the other hand, we can take advantage of the totally unimodularity of matrix $\mathbf{A}'$, and solve the mixed integer problem efficiently.

4.4.1 The Robust Adjustable Model

In our problem setting, the busyness factor $\tilde{\theta}$ estimated by the decision-maker is not exactly the true busyness factor θ, but contains some error on estimation: $\tilde{\theta} = \theta + \epsilon$ with $-\tau \leq \epsilon \leq \tau, \tau > 0$. Consequently, once the decision-maker gets an estimate value $\tilde{\theta}$, he has a better information on the true busyness factor θ by knowing that it falls in a reduced interval $[\tilde{\theta}-\tau, \tilde{\theta}+\tau]$. Following the main idea of robust optimization, he makes the second-stage decisions in order to minimize the worst cost associated with all possible true parameter values θ within this reduced interval, which is the cost associated with $\theta = \tilde{\theta} + \tau$ (see Appendix B.2 for the proof). The relation between the realized θ, the observed value $\tilde{\theta}$ and the speculated interval of possible true busyness factor according to the observation, is presented in Figure 4.4.

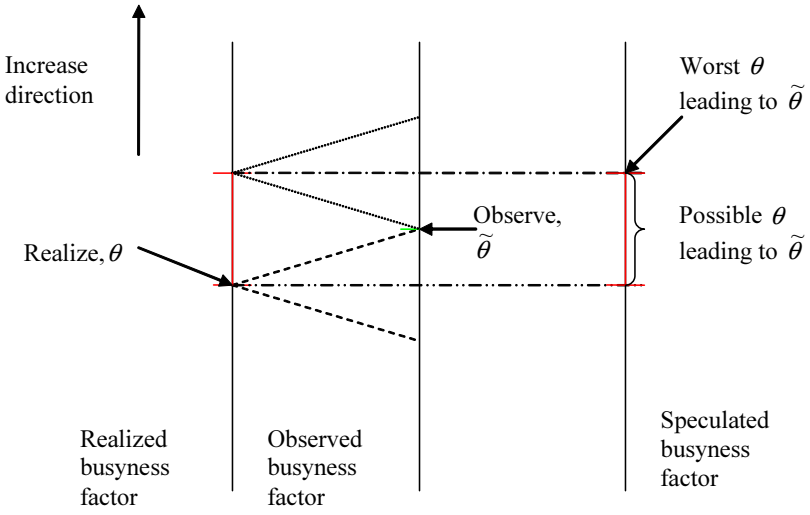

Figure 4.4: Relation between θ and $\tilde{\theta}$ in robust setting

The staffing process is as follows (as presented in Figure 4.5). Before the beginning of the day, the initial scheduling decisions $X_j, j \in J$, are assigned. When the day begins, the true busyness factor θ realizes, consequently the agent equipments $n_i(\theta)$ realize. Unfortunately, during the early horizon periods, the decision-maker can not find the true busyness factor θ. He can only observe the realized call volumes and get an estimate busyness factor $\tilde{\theta}$. This estimate busyness factor $\tilde{\theta}$ contains some error of estimation: $\tilde{\theta} = \theta + \epsilon$. With this estimate $\tilde{\theta}$, the decision-maker speculates that the true busyness factor may fall in a reduced interval $[\tilde{\theta} - \tau, \tilde{\theta} + \tau]$ with $\tau > 0$. At the end of the early horizon, the decision-maker then implements the second-stage decisions Y_j and Z_j in order to protect against the worst case busyness factor $\tilde{\theta} + \tau$.

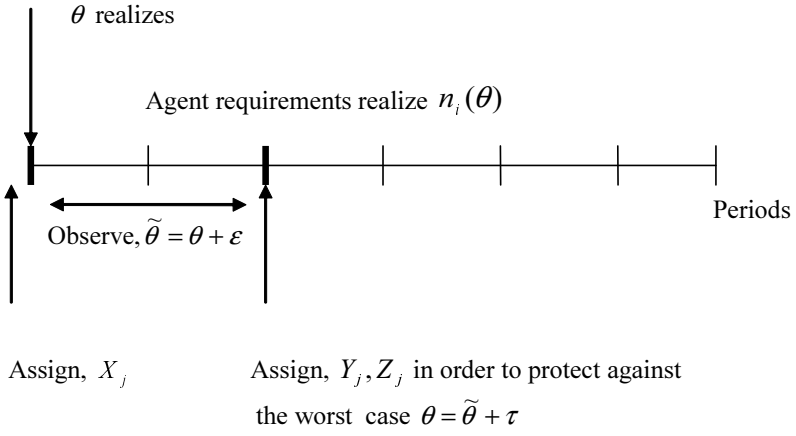

Figure 4.5: Two-stage staffing process in robust setting

We can then reformulate the Min-Max adjustable robust counterpart of problem (4.6) as follow.

$$
\begin{aligned}
\min_{X_j, Y_j, Z_j} \quad & \max_{\theta \in U, \tilde{\theta} = \theta + \epsilon} \sum_{j \in J} c_j X_j + \sum_{j \in J} d_j Y_j - \sum_{j \in J} r_j Z_j + \sum_{i \in I_1} u M_i + \sum_{i \in I_2} u M_i' \\
\text{s.t.} \quad & \sum_{j \in J} a_{ij} X_j + M_i \geq n_i(\theta), i \in I_1, \\
& \sum_{j \in J} a_{ij} X_j + \sum_{j \in J} a_{ij}'(Y_j - Z_j)(\tilde{\theta} + \tau) + M_i' \geq n_i(\theta), i \in I_2, \\
& X_j \geq Z_j, j \in J, \\
& M_i \geq 0, i \in I_1, \\
& M_i' \geq 0, i \in I_2, \\
& X_j, Y_j, Z_j \in \mathbb{Z}^+, j \in J.
\end{aligned}
\tag{4.10}
$$

Note that M_i and M_i' are the state variables which denote the staffing capacity shortfall for the periods in early and late horizon respectively. We use the notation $\sum_{j\in J} a'_{ij}(Y_j - Z_j)(\tilde{\theta}+\tau)$ in order to emphasize the fact that we will take the periodic staffing level adjustment as an entirety to define the adjustable strategies. Similar to the model (4.9), due to the unknown functional relations between $\sum_{j\in J} a'_{ij}(Y_j - Z_j)(\tilde{\theta}+\tau)$ and $\tilde{\theta}+\tau$, it is a common result that the two-stage full adaptive optimization problem is often computational intractable (NP-hard) even for linear problem with simple uncertainty sets(see Ben-Tal et al. (2004)). Model (4.10) is again more complex as the second-stage decisions variables are integers, and the parameters $n_i(\theta)$ are not affinely affected by uncertainty θ.

4.4.2 Piecewise Linear Approximation

To the best of our knowledge, the only work addressing the case of integer second-stage variables within the framework of deterministic set-based uncertainty is that of Bertsimas and Caramanis (2010). The author propose the *finite adaptability* where the uncertainty set is partitioned into several pieces and a best recourse in each is determined. There, the second-stage variables are piecewise constant functions of the uncertainty, and the decision-maker commits to one of them only after observation of the uncertainty realization. With this method, the dynamic integer problem is approximated by several static one-stage integer problem, and the computational complexity increases along with the number of partitioned pieces. Nevertheless the original intention of Bertsimas and Caramanis (2010) to partition the uncertainty set is to address the issue of over-conservatism, we use this idea for the purpose of partitioning non-linear functions into piecewise linear functions. Denote the function which defines the required agents number $n_i(\theta)$ by $\varphi_i(\theta)$, we have $n_i(\theta) = \lceil \varphi_i(\theta) \rceil$, for $i \in I$, where $\lceil \cdot \rceil$ denotes the ceiling of a continuous variable. In what follows, we consider a general case where $\varphi_i(\theta)$ is non-linear function of θ. We adapt to our setting the principle of k-adaptable robust counterpart proposed by Bertsimas and Caramanis (2010), which covers the uncertainty set with a partition of k pieces, and selects a contingency plan for each subset.

Firstly, we approximate the function $\varphi_i(\theta)$ by a set of piecewise linear functions, denote that set as K. Since the uncertain variable θ is of one-dimension, corresponding to the projection of these piecewise linear functions, the uncertainty set U is partitioned into $|K|$ disjoint regions: $U = U_1 \cup ... \cup U_{|K|}$. Recall that we assume that the observed $\tilde{\theta}$ contains small error on estimating the true value of θ: $\tilde{\theta} = \theta + \epsilon$ with $-\tau \leq \epsilon \leq \tau$. For all $\theta \in U_k$, the uncertainty set which contains the associated $\tilde{\theta}$ is denoted as $U'_k(\theta)$. We have the whole set of possible $\tilde{\theta}$ values: $U'(\theta) = U'_1(\theta) \cup ... \cup U'_{|K|}(\theta)$. In order to partition the uncertainty set U' into disjoint regions,

we let the joint part of two adjacent uncertainty sets belong only to the one with smaller indices value. Mathematically, we denote:

$$U_k = \{\theta : |\theta - \bar{\theta_k}| \leq \rho_k\}, k \in K, \tag{4.11}$$

$$U'_k(\theta) = \{\tilde{\theta} : \tilde{\theta} = \theta + \epsilon,\ \theta \in U_k, -\tau \leq \epsilon \leq \tau\} \setminus U'_{k-1}, k \in K, \tag{4.12}$$

where $\rho_k > 0$, and the mean value of each partitioned uncertainty set U_k is denoted by $\bar{\theta_k}$. $\bar{\theta_k}$ increases along with the indices k. In order to better explain the idea, we plot the definition of the relation between set U_k and U'_k by Figure 4.6.

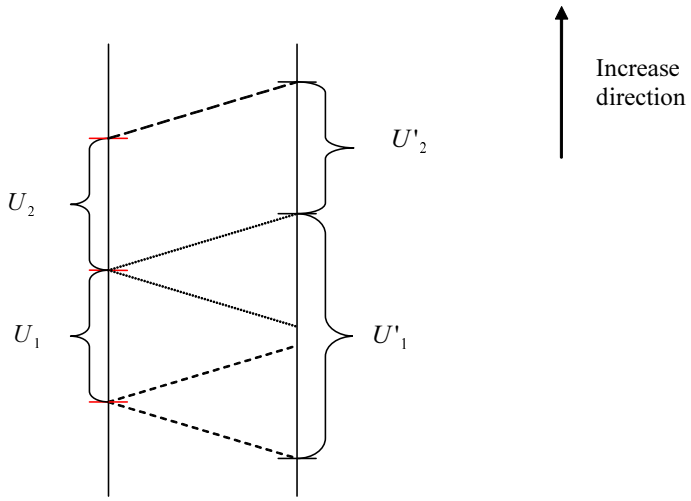

Figure 4.6: Relation between sets U_k and U'_k

We formulate $\varphi_i(\theta)$ by the piecewise linear functions approximation:

$$\varphi_i(\theta) = \varphi_1^{ik} + \varphi_2^{ik}\,\theta, \qquad i \in I, \theta \in U_k, k \in K. \tag{4.13}$$

The optimal piecewise adaptability problem can be formulated as Equation (4.14)-(4.21).

$$\min_{X_j} \quad \sum_{j\in J} c_j\, X_j + \max_{k\in K}\{\min_{Y_{jk},Z_{jk}} \sum_{j\in J} d_j\, Y_{jk} - \sum_{j\in J} r_j\, Z_{jk} + \sum_{i\in I_1} u\, M_{ik} + \sum_{i\in I_2} u\, M'_{ik}\} \tag{4.14}$$

$$\text{s.t.} \quad \sum_{j\in J} a_{ij} X_j + M_{ik} \geq N_{ik}(\theta), \qquad i \in I_1, \theta \in U_k, k \in K, \tag{4.15}$$

$$\sum_{j\in J} a_{ij} X_j + \sum_{j\in J} a'_{ij}(Y_{jk} - Z_{jk})(\tilde{\theta} + \tau) + M'_{ik} \geq N_{ik}(\theta),$$

$$i \in I_2, \theta \in U_k, \tilde{\theta} \in [\theta - \tau, \theta + \tau] \cap U'_k(\theta), k \in K, \tag{4.16}$$

$$N_{ik}(\theta) \geq \varphi_1^{ik} + \varphi_2^{ik}\, \theta, \qquad i \in I, \theta \in U_k, k \in K, \tag{4.17}$$

$$X_j \geq Z_{jk}, \quad j \in J, k \in K, \tag{4.18}$$

$$M_{ik} \geq 0, \qquad i \in I_1, k \in K, \tag{4.19}$$

$$M'_{ik} \geq 0, \qquad i \in I_2, k \in K, \tag{4.20}$$

$$X_j, Y_{jk}, Z_{jk} \in \mathbb{Z}^+, \qquad j \in J, k \in K. \tag{4.21}$$

In this equation, for each couple of partitioned uncertainty sets $(U_k, U'_k(\theta))$ with $k \in K$, the required number of agents are denoted by variable $N_{ik}(\theta)$ with $i \in I$, and the total cost is defined as the sum of the initial shift scheduling salary (which is common for all couples of uncertainty sets), the recours salary and the agents shortfall penalty. The objective function is to minimize the worst one among all these total costs. The first-stage decision variables are the initial shift scheduling $X_j, j \in J$. And for each $k \in K$, a contingency plan $(Y_{jk}, Z_{jk}, j \in J)$ is selected.

In the next theorem, we exhibit some properties of Equation (4.14)-(4.21), that are used in the robust programming formulation.

Theorem 4.2 *If there exists a value of θ, denoted by $\hat{\theta}$, for which $N_i(\hat{\theta}) \geq N_{ik}(\theta)$ with $i \in I$ for any $\theta \in U_k, k \in K$, then Equation (4.14)-(4.21) can be simplified as Equation (4.22)-(4.29).*

Proof: see Appendix B.3.

Let $\hat{\theta} + \epsilon$ denote the observed value of $\hat{\theta}$, which is slightly different from $\hat{\theta}$ and affects the

second-stage decisions.

$$\min_{X_j,Y_j,Z_j} \quad \sum_{j\in J} c_j\, X_j + \sum_{j\in J} d_j\, Y_j - \sum_{j\in J} r_j\, Z_j + \sum_{i\in I_1} u\, M_i + \sum_{i\in I_2} u\, M_i' \tag{4.22}$$

$$\text{s.t.} \quad \sum_{j\in J} a_{ij} X_j + M_i \geq N_i(\hat{\theta}), i \in I_1, \tag{4.23}$$

$$N_i(\hat{\theta}) \geq \varphi_1^{ik} + \varphi_2^{ik}\, \theta, i \in I, \theta \in U_k, k \in K, \tag{4.24}$$

$$\sum_{j\in J} a_{ij} X_j + \sum_{j\in J} a'_{ij}(Y_j - Z_j)(\hat{\theta} + \epsilon + \tau) + M_i' \geq N_i(\hat{\theta}), i \in I_2, \tag{4.25}$$

$$X_j \geq Z_j, j \in J, \tag{4.26}$$

$$M_i \geq 0, i \in I_1, \tag{4.27}$$

$$M_i' \geq 0, i \in I_2, \tag{4.28}$$

$$X_j, Y_j, Z_j, N_i(\hat{\theta}) \in \mathbb{Z}^+, j \in J, i \in I. \tag{4.29}$$

Similar to Equation (4.10), Problem (4.22)-(4.29) with full adaptive second-stage decisions is NP-hard. Specially, the second-stage decisions are integers. In what follows, we analysis how to solve Problem (4.22)-(4.29) efficiently.

In model (4.22)-(4.29), the matrix A' associated with the adjustable variables is constant, this is the case of fixed recourse. But the second-stage decisions Y_j and Z_j are subject to be integers while the LDR is a continuous function. This ends that LDR can not be applied directly on the second-stage decision variables Y_j and Z_j.

In Section 4.4.3, we provide the technique how we relate the discrete second-stage variables with the LDR. Thanks to the totally unimodular property of the period-shift matrix, the problem is solved without computational difficulty.

4.4.3 Relating Discrete Seconde-Stage Variables with Affine Adaptability

As mentioned above, a continuous decision rule can not be implemented directly on the discrete second-stage decision variables Y_j and Z_j for $j \in J$. We thus propose to implement the decision rule on a group of auxiliary state variables $Q_i, i \in I_2$, rather than directly on the second-stage decision variables. This is the major difference between our method and the traditional AARC methodology.

Similar to AARC methodology, we define the linear decision rules as

$$G_{ik}(\tilde{\theta}) = G_1^{ik} + G_2^{ik}\,(\tilde{\theta} + \tau), \quad i \in I_2, k \in K. \tag{4.30}$$

In these linear decision rules, parameters G_1^{ik} and G_2^{ik} are to be determined by solving model

(4.31)-(4.42).

Let the auxiliary integer state variables Q_i denote the quantity of staff capacity adjustment for each period in the late horizon I_2. The optimal piecewise linear adaptability problem can be formulated as follows.

$$\min_{X_j, Y_j, Z_j} \quad \sum_{j\in J} c_j\, X_j + \sum_{j\in J} d_j\, Y_j - \sum_{j\in J} r_j\, Z_j + \sum_{i\in I_1} u\, M_i + \sum_{i\in I_2} u\, M_i' \tag{4.31}$$

$$\text{s.t.} \quad \sum_{j\in J} a_{ij} X_j + M_i \geq N_i, \qquad i \in I_1, \tag{4.32}$$

$$N_i \geq \varphi_1^{ik} + \varphi_2^{ik}\,\theta, \qquad i \in I_1, \theta \in U_k, k \in K, \tag{4.33}$$

$$\sum_{j\in J} a'_{ij}(Y_j - Z_j) \geq Q_i, \qquad i \in I_2, \tag{4.34}$$

$$Q_i \geq G_1^{ik} + G_2^{ik}\,(\tilde{\theta} + \tau),$$

$$i \in I_2, k \in K, \theta \in U_k, \tilde{\theta} \in [\theta - \tau, \theta + \tau] \cap U'_k(\theta), \tag{4.35}$$

$$\sum_{j\in J} a_{ij} X_j + M_i' \geq O_i, \qquad i \in I_2, \tag{4.36}$$

$$O_i \geq \varphi_1^{ik} + \varphi_2^{ik}\,\theta - G_1^{ik} - G_2^{ik}\,(\tilde{\theta} + \tau),$$

$$i \in I_2, k \in K, \theta \in U_k, \tilde{\theta} \in [\theta - \tau, \theta + \tau] \cap U'_k(\theta), \tag{4.37}$$

$$X_j \geq Z_j,\ j \in J, \tag{4.38}$$

$$M_i \geq 0,\ N_i \in \mathbb{Z}^+, \quad i \in I_1, \tag{4.39}$$

$$M_i' \geq 0,\ Q_i \in \mathbb{Z}, O_i \in \mathbb{Z}^+, i \in I_2, \tag{4.40}$$

$$X_j, Y_j, Z_j \geq 0, j \in J, \tag{4.41}$$

$$G_1^{ik}, G_2^{ik} \in \mathbb{R}, i \in I_2, k \in K. \tag{4.42}$$

In order to simplify the presentation, in this equation and those follows, we replace $N_i(\theta)$ by N_i, similar omissions are done for all other variables depending on θ and/or $\tilde{\theta}$. In this equation, the first-stage decision variables are the initial shift scheduling $X_j, j \in J$. And for each $k \in K$, the parameters of a linear decision rule, G_1^{ik}, G_2^{ik} with $i \in I_2$, are determined.

The objective (4.31), Constraints (4.32) and (4.33) are similar to (4.22), (4.23) and (4.24), respectively. Constraints (4.34)-(4.37) replace constraints set (4.25). Constraint (4.34) restricts that the adding or removing actions should grantee that at least Q_i^+ agents are added, or at most Q_i^- agents are removed. We use inequality instead of equality constraints in order that the problem has feasible solutions. For $i \in I_2$, Q_i are required to be integer, and they are approximated by the maximum value determined by all decisions rules G_{ik} with $k \in K$, as shown in Constraint (4.35). The variable O_i in Constraints (4.36)-(4.37) equals to $N_i - Q_i$, for

$i \in I_2$. Constraint (4.38) restricts that the agents removed are less than that assigned initially. Constraints (4.39)-(4.42) define the non-negative and integer conditions for programs variables.

We underline Constraint (4.41) in which the integer constraints on $X_j, Y_j, Z_j, j \in J$, are relaxed to be linear. The reason is that by solving Problem (4.31)-(4.42), the solutions X_j, Y_j and Z_j are integer. The explanation is as follows. The feasible region of $\{X_j, Y_j, Z_j\}$ is constructed by Constraints (4.32), (4.34), (4.36), (4.38),

$$\begin{array}{rcll} \sum_{j\in J} a_{ij}X_j + M_i & \geq & N_i, & i \in I_1, \\ \sum_{j\in J} a'_{ij}(Y_j - Z_j) & \geq & Q_i, & i \in I_2, \\ \sum_{j\in J} a_{ij}X_j + M'_i & \geq & O_i, & i \in I_2, \\ X_j & \geq & Z_j, & j \in J. \end{array}$$

Since the matrices are $(\mathbf{A}\,\mathbf{I})$ and $\mathbf{A}'$ are totally unimodular and variables N_i, Q_i and O_i are integer, the results in Appendix B.1 show that the feasible region of $\{X_j, Y_j, Z_j\}$ is an integral polyhedron. Thus the solutions X_j, Y_j and Z_j are automatically integer. This is the main interest of our technique which relates discrete seconde-stage variables with affine adaptability.

Similar to the AARC methodology, with the uncertainty set (4.11) and (4.12), the inequality constraints (4.33), (4.35) and (4.37) are equivalent to

$$\begin{aligned} &N_i \geq \varphi_1^{ik} + \varphi_2^{ik}\,\bar{\theta_k} + |\varphi_2^{ik}|\,\rho_k, \qquad i \in I_1, k \in K, \\ &Q_i \geq G_1^{ik} + G_2^{ik}\,(\bar{\theta_k} + \tau) + |G_2^{ik}|\,(\rho_k + \tau), \qquad i \in I_2, k \in K, \\ &O_i \geq \varphi_1^{ik} - G_1^{ik} + (\varphi_2^{ik} - G_2^{ik})\,\bar{\theta_k} + |\varphi_2^{ik} - G_2^{ik}|\,\rho_k + \tau\,| - G_2^{ik}| - \tau\,G_2^{ik}, \qquad i \in I_2, k \in K. \end{aligned}$$

By adding non-negative variables $P_{ik}, R_{ik}, i \in I_2, k \in K$, Equation(4.31)-(4.42) can be further

expressed as

$$\min_{X_j,Y_j,Z_j} \quad \sum_{j\in J} c_j\, X_j + \sum_{j\in J} d_j\, Y_j - \sum_{j\in J} r_j\, Z_j + \sum_{i\in I_1} u\, M_i + \sum_{i\in I_2} u\, M_i' \tag{4.43}$$

$$\text{s.t.} \quad \sum_{j\in J} a_{ij} X_j + M_i \geq N_i, i \in I_1, \tag{4.44}$$

$$\sum_{j\in J} a'_{ij}(Y_j - Z_j) \geq Q_i, i \in I_2, \tag{4.45}$$

$$\sum_{j\in J} a_{ij} X_j + M_i' \geq O_i, i \in I_2, \tag{4.46}$$

$$N_i \geq \varphi_1^{ik} + \varphi_2^{ik}\,\bar{\theta_k} + |\varphi_2^{ik}|\,\rho_k, i \in I_1, k \in K, \tag{4.47}$$

$$Q_i \geq G_1^{ik} + G_2^{ik}\,(\bar{\theta_k} + \tau) + R_{ik}\,(\rho_k + \tau), i \in I_2, k \in K, \tag{4.48}$$

$$O_i \geq \varphi_1^{ik} - G_1^{ik} + (\varphi_2^{ik} - G_2^{ik})\,\bar{\theta_k} - \tau\, G_2^{ik} + P_{ik}\,\rho_k + R_{ik}\,\tau, i \in I_2, k \in K, \tag{4.49}$$

$$-P_{ik} \leq \varphi_2^{ik} - G_2^{ik} \leq P_{ik}, i \in I_2, k \in K, \tag{4.50}$$

$$-R_{ik} \leq G_2^{ik} \leq R_{ik}, i \in I_2, k \in K, \tag{4.51}$$

$$X_j \geq Z_j, j \in J, \tag{4.52}$$

$$M_i \geq 0,\ N_i \in \mathbb{Z}^+, \quad i \in I_1, \tag{4.53}$$

$$M_i' \geq 0,\ Q_i \in \mathbb{Z}, O_i \in \mathbb{Z}^+, i \in I_2, \tag{4.54}$$

$$X_j, Y_j, Z_j \geq 0, j \in J, \tag{4.55}$$

$$G_1^{ik}, G_2^{ik} \in \mathbb{R},\ P_{ik}, R_{ik} \geq 0, i \in I_2, k \in K. \tag{4.56}$$

The MIP (4.43)-(4.56) contains $|I| + |I_2|$ integer and $4 \times |I_2| \times |K| + |I| + 3 \times |J|$ continuous variables.

Example:

Consider the same structure of call center as described in Section 4.3 (a day with 25 periods and 162 possible shifts, where the first 5 periods are considered as early horizon), if we partition the uncertainty set into 11 pieces, this implies to solve a model with 45 integer variables and 1391 continuous variables. This Equation (4.43)-(4.56) is very well structured. Firstly, the number of integer variables keeps the same even the shifts number and partitioned sets number increase. Secondly, and more importantly, the integer constraints are only to find the ceilings of some continuous values, which is quite a simply requirement. Consequently, the computational difficulty of solving Equation (4.43)-(4.56) is almost as that of a linear problem.

In conclusion, we consider the two-stage shift scheduling problem, in which both first and second-stage decisions are discrete variables, and the parameters are not affinely affected by the

data uncertainty. Firstly, for the non-affine affection of data uncertainty on the parameter functions, we approximate the non-linear function by several piecewise linear functions, equivalently partition the one dimensional uncertainty set into several pieces, in order to determiner a decision rule for each sub uncertainty set. Then we propose a method to relate the discrete variables and the affine decision rule, the approximate MIP has a very good structure thanks to the totally unimodular property. The approximate problem can be solved with no computational difficulty.

4.5 Numerical Experiments and Results

In this section, we conduct a numerical study in order to evaluate and compare between the proposed approaches. In Section 4.5.1, we describe the numerical experiments. In Section 4.5.2, we analyze the results and derive various insights.

4.5.1 Experiments

We describe in this section the data used in the numerical examples first, and then the design of experiments.

Parameter Values

Inbound calls. In the experiments, we use real data from a Dutch hospital which exhibits a typical and significant workload time-of-day seasonality. In Figure 4.7, we plot the curve of the mean arrival rates as a function of the periods of the day. We focus on a particular day, namely Monday. In solid line, we plot this curve for an average day, and in dashed lines we plot two examples of not busy and busy days. The mean arrival rate at the beginning and at the end of the day is quite low. It has a high peak in the late morning and tends to decrease around the lunch break, but a second lower peak occurs also in the afternoon. Although there is a significant stochastic variability in the arrival rate from one day to another, there is a strong seasonal pattern across the periods of a given day. The day starts at 8:00 am, finishes at 8:30 pm, and is divided into $n = 25$ periods of half hour each. The first 5 periods are considered as early horizon, and the other periods construct the later horizon.

Without loss of generality, we choose $E[\Theta] = 1$. This leads to $f_i = E[\Lambda_i]$, and from a one-year-horizon data we numerically find via a standard statistical analysis, that f_1, f_2, ..., f_{25} are 98.8, 148, 200, 226.4, 237.6, 236.4, 242.8, 231.6, 218.8, 219.2, 226.8, 215.6, 216.8, 210, 213.6, 226.8, 232.8, 212, 174.4, 158, 136, 122.4, 102, 92.8 and 74.8 calls per minute, respectively.

Recall that the random variable Θ describes the busyness of the day. We assume that Θ follows a discretized truncated Gamma probability distribution. From the data, we estimate the

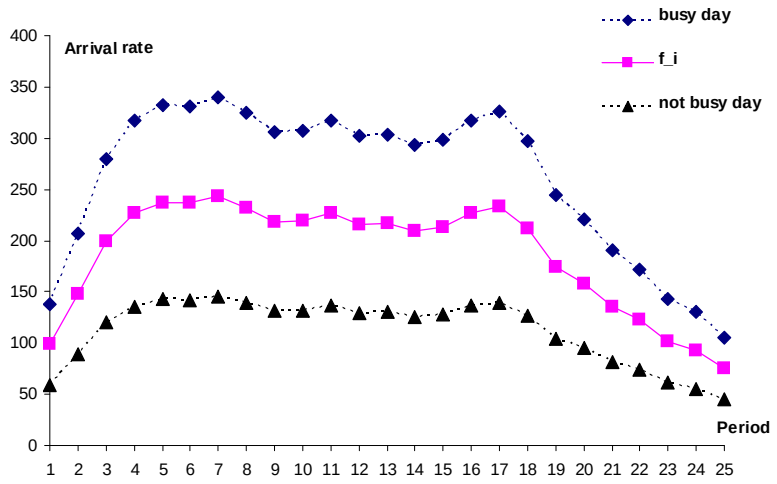

Figure 4.7: Arrival rate graph

distribution as Gamma$(25, 0.04)$ with mean 1 and variance 0.04. This finishes the characterization of the random variable Λ_i. The mean service time is $\frac{1}{\mu} = 5$ minutes. We assume a classical service level corresponding to the well-known 80/20 rule: the probability that a call waits for less than 20 seconds has to be larger or equal to 80 percent. Using Condition (2.3), we can therefore deduce the required number of agents n_i during period i. For a given period i, the relation between n_i and θ can be approximated by piecewise linear functions.

Information Update Method. In what follows, we present how the decision-maker uses the actual call volumes early in a day to get an estimated busyness factor $\tilde{\theta}$. First, we follow the work of Brown et al. (2005), and use the variance-stabilizing transformation for Poisson data. If V is Poisson(λ), then $\sqrt{V + \frac{1}{4}}$ has approximately mean $\sqrt{\lambda}$ and variance $\frac{1}{4}$. In addition, $\sqrt{V + \frac{1}{4}}$ is asymptotically normal (as $\lambda \to \infty$). Using this approximation and Equation (4.1), we obtain the following model for a given value of θ:

$$\sqrt{V_i + \frac{1}{4}} = \sqrt{\theta\, f_i} + \varepsilon_i, \text{ with } \varepsilon_i \sim Gaussian((0, 1/4). \tag{4.57}$$

Next, using the model (4.57), we estimate the busyness factor from the observed call volumes V_i, $i \in I_1$. Operationally, it exist a trade-off between having more information about call volumes and adjusting the agents capacity earlier in the day. One achieves a better approximation of the busyness factor by having more collected information, but the decision-maker then has to wait longer to implement the second-stage decisions. Since it requires many observed periodical call volumes to estimate θ using the variance of ε_i, we use its mean instead. Equation (4.57) shows

that ε_i for $i \in I_1$ has approximately mean 0. We then have

$$\sum_{i \in I_1} \left(\sqrt{V_i + \frac{1}{4}} - \sqrt{\theta\, f_i} \right) \approx 0, \tag{4.58}$$

and we get an estimated $\tilde{\theta}$ by

$$\tilde{\theta} = \left(\frac{\sum_{i \in I_1} \sqrt{V_i + \frac{1}{4}}}{\sum_{i \in I_1} \sqrt{f_i}} \right)^2 . \tag{4.59}$$

As mentioned in Section 4.3, for a given θ_l (associated with probability p_l, $l \in L$), the decision maker estimates it as $\tilde{\theta}_k$ with probability $pb_{l,k}, k \in K$. The way we obtain the set K and parameters $pb_{l,k}$ is as follows. Given θ_l, the call volume V_{il} is Poisson $(\lambda_{i,l})$, $i \in I_1$. The probability that a particular combination of early periods call volumes occurs, is p_l multiples the product of the poisson probabilities of these call volumes. We calculate the estimated $\tilde{\theta}$ associated as explained above. The sum of the probabilities associated with combinations which lead to the given $\tilde{\theta}_k$, is defined as probability $pb_{l,k}$. Set K assembles all possible values of $\tilde{\theta}_k$. We find that for a given θ_l, its estimated value falls highly probably not far from its own value.

Cost parameters. Agents work between 7 with 18 half-hour periods a day, without intermediate break. A shift can start at the beginning of any period, while guaranteeing the length of between 7 with 18 half-hour periods before the day ends. Enumeration shows that there are 162 feasible schedules.

We define the shifts with a length less than 11 half-hour periods as part-time shifts, and the others as full-time shifts. Without loss of generality, we use a normalized cost of 1 for each half-hour an agent works at regular full time shifts. In order to avoid that first-stage decisions $X_j, j \in J$ takes too many part-time agents which are usually less professional, we defined the regular part time shifts unit cost as 1.1 per half-hour. Therefore $c_j = \sum_{i \in I} a_{ij}$ for full-time shifts and $c_j = 1.1 \sum_{i \in I} a_{ij}$ for part-time ones. The temporarily added shift should be payed more expensively than the regular one, we then define $d_j = 1.2 \times \sum_{i \in I} a'_{ij}$. Removing an agent from shift j makes some cost saving, but this cost saving should be less than c_j. We choose $r_j = 0.5 \times \sum_{i \in I} a'_{ij}$. The value of under-staffing penalty parameter is taken as $u = 25$, which is chosen via successive trials/corrections based on stochastic two-stage recourse programming approach (SRA), and estimates that the optimal solution of SRA leads to about 3% under-staffing. This way of choosing of under-staffing penalty value u is similar to that in Liao et al. (2010).

Design of the Experiments

Lower bound. As a lower bound solution, we consider a perfect information model (PI) for which the value of θ, the actual workload, consequently the required agents number n_i, is assumed to be known before the optimization step of the variable $X_j, Y_j, Z_j, j \in J$. For each sample $\theta_s, s \in S$, with S as the set of samples, we solve the problem (4.60) in order to get the optimal value of $X_{sj}, Y_{sj}, Z_{sj}, j \in J$ and its associated total cost. The computation of the corresponding average total cost is then straightforward.

$$
\begin{aligned}
Min \quad & \sum_{j\in J} c_j\, X_{sj} + \sum_{j\in J} d_j\, Y_{sj} - \sum_{j\in J} r_j\, Z_{sj} + \sum_{i\in I} u\, M_{is} \\
\text{s.t.} \quad & \sum_{j\in J} a_{ij} X_{sj} + M_{is} \geq n_i(\theta_s\, f_i), \quad i \in I_1, \\
& \sum_{j\in J} a_{ij} X_{sj} + \sum_{j\in J} a'_{ij}(Y_{sj} - Z_{sj}) + M_{is} \geq n_i(\theta_s\, f_i), \qquad i \in I_2, \\
& X_{sj} \geq Z_{sj}, \qquad j \in J, \\
& X_{sj}, Y_{sj}, Z_{sj} \in \mathbb{Z}^+, j \in J, \\
& M_{is} \geq 0, \quad i \in I.
\end{aligned}
\tag{4.60}
$$

Benchmark1. As an initial benchmark, we consider the simple approach based on the expected value of the random variable Θ, referred to as average based deterministic approximation, denoted by DA. In this case, Λ_i reduces to the single value $E[\Theta]\, f_i$. The required number of agents to handle the calls of period i is $n_i(E[\Theta]\, f_i)$. The optimization problem can be then formulated as:

$$
\begin{aligned}
Min \quad & \sum_{j\in J} c_j\, X_j + \sum_{j\in J} d_j\, Y_j - \sum_{j\in J} r_j\, Z_j + \sum_{i\in I} u\, M_i \\
\text{s.t.} \quad & \sum_{j\in J} a_{ij} X_j + M_i \geq n_i(E[\Theta]\, f_i), \qquad i \in I_1, \\
& \sum_{j\in J} a_{ij} X_j + \sum_{j\in J} a'_{ij}(Y_j - Z_j) + M_i \geq n_i(E[\Theta]\, f_i), \qquad i \in I_2, \\
& X_j \geq Z_j, \qquad j \in J, \\
& X_j, Y_j, Z_j \in \mathbb{Z}^+, \qquad j \in J, \\
& M_i \geq 0, \quad i \in I.
\end{aligned}
\tag{4.61}
$$

Benchmark2. Another benchmark we consider is the approach where the decision maker takes into account of the randomness of variable Θ, but the second-stage variables have no dependence on $\tilde{\theta}$, referred to as static one-stage stochastic approximation. The static one-stage stochastic optimization problem (4.62) is built on the discrete probability distributions characterizing Θ.

The difference between (4.7) and (4.62) is that in the latter all decisions variables $X_j, Y_j, Z_j, j \in J$ are decided before the day begins.

$$
\begin{aligned}
Min \quad & \sum_{j\in J}(c_j\, X_j + d_j\, Y_j - r_j\, Z_j) + \sum_{l\in L}\sum_{i\in I} p_l\, u\, M_{i,l} \\
\text{s.t.} \quad & \sum_{j\in J} a_{ij} X_j + M_{i,l} \geq n_{i,l} && i \in I_1, l \in L, \\
& \sum_{j\in J} a_{ij} X_j + \sum_{j\in J} a'_{ij}(Y_j - Z_j) + M_{i,l} \geq n_{i,l}, && i \in I_2, l \in L \\
& X_j \geq Z_j, && j \in J, \\
& M_{i,l} \geq 0, && i \in I, l \in L, \\
& X_j, Y_j, Z_j \in \mathbb{Z}^+, && j \in J.
\end{aligned}
\tag{4.62}
$$

The relaxed linear programs of all MIPs (4.60), (4.61) and (4.62) are integral (see Appendix B.1). We thus relax these MIPs and solve linear problems.

Additional Notations. We compute the optimal staffing levels given by the average based deterministic approximation (DA), the static one-stage stochastic programming approach (SSA), the stochastic two-stage recourse programming approach (SRA) and piecewise linear adjustable robust approach(ARA). For the piecewise linear adjustable robust approach, we partitioned the uncertainty set $U = \{\theta : 0.3 \leq \theta \leq 1.4\}$ into 11 uncertainty set with harmonious size. Then we have $U_k = \{\theta : |\theta - 0.1 * (k + 2.5)| \leq 0.05\}$, for $k = 1, ..., 11$. The corresponding uncertainty sets for $\tilde{\theta}$ are defined as in (4.12), and the value of τ is determined to be 0.01.

Optimal policy performance simulations. In order to estimated the cost criterion associated with the different policies, 10000 samples values are randomly generated as outcomes of Θ. For each sample, the arrival rates and the required numbers of agents of the 25 periods are calculated using the sample value $\theta_s, s \in S$. And the call volumes of the first five periods $V_i (i = 1, .., 5)$, are randomly given around the corresponding mean arrival rates. The estimated $\tilde{\theta}$ is calculated based on these call volumes.

In order to evaluate the stochastic two-stage recourse programming, the decision-maker implements the initial scheduling decisions and choose the second-stage decisions from its recourse actions set according to the estimated $\tilde{\theta}$. Similarly, in order to evaluate the adjustable robust approach, the initial scheduling decisions are implemented first, and if $\tilde{\theta}$ falls in the uncertainty set $\hat{k}$, then the decision maker get the capacity adjustment value $G_{i\hat{k}}, i \in I_2$ according to decision rule (4.30), with parameters determined by (4.43)-(4.56). Since the uncertainty set we considered may not cover all possible values of $\tilde{\theta}$, in this case, we apply the decision rule of the sub uncertainty set which is the most close to $\tilde{\theta}$. A simple transforms is need to get the recours

actions Y_j, Z_j:

$$
\begin{aligned}
Min \quad & \sum_{i \in I_2} Gap_i \\
& Gap_i \geq \sum_{j \in J} a'_{ij}(Y_j - Z_j) - Q_i, \;\; i \in I_2, \\
& Q_i \geq G_1^{i\hat{k}} + G_2^{i\hat{k}}\,(\tilde{\theta} + \tau), \;\; i \in I_2 \\
& Gap_i \geq 0, Q_i \in \mathbb{Z}, \;\; i \in I_2, \\
& Y_j, Z_j \geq 0, \;\; j \in J.
\end{aligned} \tag{4.63}
$$

The optimal solutions of the two benchmark approaches are implemented directly. The average total cost, probability of under-staffing and computing time associated with all the approaches are reported in the next section.

4.5.2 Insights

In this section, we comment on the numerical results and derive the main insights. First, we report the computing time of each solution approach. Second,we compare the average total costs and under-staffing probabilities between the approaches. We show the necessity of explicitly taking into account the uncertainty in the call arrival parameters. And the advantage of the flexibility provided by information update is also analyzed.

Computing time report

For the SRA and SSA, We discretize Gamma distribution to 200 scenarios. The set K contains 21 scenarios for SRA. And for ARA, the considered uncertainty sets are partitioned into 11 pieces. The size of the problems and their computing time are reported in Table 4.1. The computations have been performed using Cplex on an Intel Core Duo CPU 1.20 Ghz with 0.99 GBytes RAM.

Table 4.1: Computing time and problem size

Approache	Computing time (Sec.) total time/ solving time	variable numbers
SRA	194s/72.80 s	91 966 continous
ARA	0.17s/0.03 s	45 integers, 1391 continous
SSA	1.48s/0.84 s	5 486continous
DA	0.17s/0.02s	511 continous

It is shown that both the large size LP of SRA and MIP of ARA require very little time to be

solved. This is one of the most important contributions in this chapter. We construct the model of a two-stage call center staffing-scheduling recourse problem, of which the solving process is supposed to be quite time consuming, but we solve it very efficiently by two approaches.

Cost Comparison

Table 4.2 displays for each approach (SRA, ARA, SSA, DA and PI), the average total cost, the average values of the three components of the total cost, the standard deviations (Std.) of the total cost, the update cost and the under-staffing cost. At the end of each line, the under-staffing probability is given. The PI solution is clearly the ultimate lower bound on the minimal cost of all the other solution method.

Table 4.2: $\Theta \sim$ Gamma(25,0.04)

Approach	Total Cost		Regular	Update cost		Under-staff. cost		Under-staff.
	Average	Std.	salary	Average	Std.	Average	Std.	probability
PI	24844.60	4869.61	24844.6	0.00	0.00	0.00	0.00	0.00%
SRA	30534.32	6483.75	27007.4	2939.10	3930.04	587.82	3556.61	0.87%
ARA	37013.14	19529.28	34519.3	683.92	168.92	1809.92	19525.48	3.02%
SSA	36335.84	16908.21	33801.6	57.60	0.00	2476.64	16908.21	3.89%
DA	72813.81	75804.59	24860.4	0.00	0.00	47953.41	75804.59	47.01%

We can see that SRA for which the probability distribution is available, performs better than ARA in terms of the average total cost, the total cost standard deviation and the probability of under-staffing. We underline here that if no reliable knowledge of the probability distributions of the uncertain parameter is available, it is reasonable to apply methods such as the ARA to use a min-max objective function. Otherwise, SRA may be more appropriate and will result better on average total cost, and even on other criteria.

The main idea of the (adjustable) robust approach is that it allows an adjustment of the level of robustness of the solution in order to trade off between performance and protection against uncertainty (see Bertsimas et al. (2004)). The key point in line with practical efficiency of robust programs consists of the design of the uncertainty sets. In addition, for the adjustable robust approach, Ben-Tal et al. (2005) argue that there is no guarantee that the optimal solution is close to a linear decision rule, although AARC proposes the decision rule to be linear. We declare a similar property for the ARA in this chapter.

Comparing the two approaches SRA and SSA for both of which the information of probability distribution is available, we can see that the SSA has larger average total cost and bigger under-staffing probability than that of SRA, this emphasizes the advantage of adding the update

flexibility.

The gaps between DA and other approaches on average total cost, total cost standard derivation and under-staffing probability are quite large. It is thus very important to take into account the effect of data uncertainties and develop better solution approaches.

4.6 Conclusion

We have developed a multi-shift call center model allowing a real-time staff capacity adjustment. We focused on optimizing the initial staffing level and the real-time adjustable policy w.r.t. the total operating cost of the call center.

We modeled this problem as a cost optimization-based two-stage model. We then proposed two approaches to solve it numerically: a stochastic two-stage recourse programming approach and an adjustable robust programming approach. These two approaches can both solve the mixed integer problem of large size without computational difficulties. We next conducted a numerical study in order to evaluate the performance of each approach and gain useful insights. First, computing time and problem sizes illustrate the efficiency of the approaches we proposed. Second, by comparing the average total cost and under-staffing probability between these two approaches, the static one-stage stochastic approach and the average based deterministic approximation, we underlined the advantage of adding the update flexibility and the necessity of taking into account the uncertainty in the call demand parameters.

This chapter consider only the case where the shifts are without breaks. One extension of this work is to take our results as the fist step, and place breaks within shifts using heuristical methods.

Chapter 5

Multi-shift Staffing Problem with Distributionally Robust Optimization

This chapter continues to deal with the call center scheduling aims to set-up the workforce so as to meet target service levels, in a multi-periodic multi-shift setting. The service level depends on the mean rate of arrival calls, which fluctuates during the day and from day to day. The staff scheduling must adjust the workforce period per period during the day, but the flexibility in so doing is limited by the workforce organization by shifts. The challenge is to balance salary costs and possible failures to meet service levels. We consider uncertain arrival rates, that vary according to an intra-day seasonality and a global busyness factor. Both factors (seasonal and global) are estimated from past data and are subject to errors. We propose an approach combining stochastic programming and distributionally robust optimization to minimize the total salary costs under service level constraints. The performance of the robust solution is simulated via Monte-Carlo techniques and compared to the pure stochastic programming.

5.1 Introduction

In this chapter, we continue to allow the mean arrival rate of calls to be uncertain and to follow some periodical fluctuation pattern. We model the arrival process of calls by a doubly non-stationary stochastic process, with random mean arrival rates which are related to a random parameter called busyness factor of the day. In each period, one can estimate the theoretical number of agents required to efficiently handle the inbound calls. The assigned agents number is allowed to be less than that required (under-staffing), but the total expect under-staffing in the whole day should not exceed a certain limit. The staff-scheduling problem is modeled as a cost optimization-based model with constraints. The cost criterion function is the agents salary cost. Our objective is to find the optimal shift scheduling which minimizes the salary cost under condition of respecting the total expect under-staffing limit.

There are many different approaches to modeling uncertainty in the context of such optimization problem. The traditional way to take into account the parameters uncertainty is using stochastic programming techniques, which minimizes the expected cost by assuming that the parameters obey a known probability distribution (see Birge and Louveaux (1997); Ruszczynski and Shapiro (2003); Shapiro et al. (2009)). Even though this can give us a complete picture of randomness, as a typical critique of stochastic approaches, the exact probability distributions are however often unknown in practice and can be computationally unwieldy (see Shapiro and Nemirovski (2005)). In contrast with this stochastic programming framework which explicitly incorporates a probability description of the uncertainty, robust optimization models uncertain parameters using uncertainty sets. The objective is then to minimize the worst-case cost in that sets. For a summary of the state-of- the-art in robust optimization, the reader is referred to Ben-Tal et al. (2009). Robust programming based formulations are often computationally tractable even for large-scale problems. The main disadvantage of the robust programming is that the solution tends to be conservative, since the approach fundamentally implements a worst-case analysis in some given uncertainty set. As a consequence, the uncertainty sets used in the formulations have to be carefully designed in order to solve efficiently the trade off between performance and protection against uncertainty (see for instance Bertsimas and Sim (2004); Bertsimas and Brown (2009)).

It is then natural to consider an approach which bridges the gap between the conservatism of robust optimization and the specificity of stochastic programming. This approach optimizes the worst-case expected objective function over a family of possible probability distributions. This class of min-max stochastic optimization problems is known as *distributionally robust*. Pioneering papers along this line are Scarf (1958), with a closed form solution for a distributionally robust

newsvendor model, and Dupačová (1987); Žáčková (1966) who considered general distributionally robust stochastic linear programs.

A key issue in this approach is the structure of the family of possible probability distributions. In a first stream of research, the possible distributions are described only using properties such as their support and/or moments Bertsimas et al. (2010b); Calafiore and El Ghaoui (2006); Delage and Ye (2010); El Ghaoui et al. (2003); Natarajan et al. (2010).

Under situations when samples data of the uncertainty are available, it is however a pity to ignore full information of sample data but to use only part of available information to describe the family of possible distributions. In this chapter, we fully exploits the sample data, and consider a set of discrete probability distributions within a given range from the observed sample distribution. The most similar work to ours is Wang et al. (2009), in which they authors construct a model to minimizes the worst case cost under the set of all distributions that maintain a certain level of likelihood of the observed data.

For each structure of family of probability distributions, the corresponding min-max optimization models potentially exhibit specific linearity/convexity properties and associated complexity. As a consequence various solution methodologies and algorithms have been developed (see for instance Bertsimas et al. (2010b); Breton and El Hachem (1995); Calafiore and El Ghaoui (2006); Delage and Ye (2010); Erdogan and Iyengar (2006); El Ghaoui et al. (2003); Natarajan et al. (2010); Riis and Andersen (2005); Shapiro and Kleywegt (2002); Shapiro and Ahmed (2004)).

The rest of the chapter is structured as follows. In Section 5.2, we describe the call center model under consideration and formulate the associated staff-scheduling problem. The classic stochastic programming model of this problem is given. In Section 5.3, we introduce the distributionally robust model of the staff-scheduling problem. In Section 5.4, we conduct a numerical study to evaluate these alternative formulations. We exhibit the impact of the uncertainty of the distributional probability.

5.2 Problem Formulation

We consider a call center with a single type of inbound calls in a multi-period multi-shift setting. The service level depends on the current workforce (number of servers) and of the inbound call arrival process. The latter is of the Poisson type; it essentially depends on the mean arrival rate, which vary during the day and according to the random busyness of the day. To account for these variations, Liao et al. (2010) proposed a stochastic programming formulation of the single shift problem. We present here a closely related formulation with an extension to the multi-shift problem. A main difference with the paper quoted above concerns the handling of understaffing.

In the present chapter, we put the constraint that understaffing does not exceed a fraction of the required staff, while in Liao et al. (2010) understaffing was simply part of the objective with a penalty factor.

5.2.1 The Inbound Call Arrival Process

As described in Chapter 2, several characteristics of the arrival process of calls have been underlined in the recent call centers literature. First, it has been observed that the total daily number of calls has an over-dispersion relative to the classical Poisson distribution. Second, the mean arrival rate considerably varies with the time of day. Third, there is a strong positive correlation between arrival counts during the different periods of the same day.

In order to address uncertain and time-varying mean arrival rates coupled with significant correlations, we model the inbound call arrival process by a doubly stochastic Poisson process (see Avramidis et al. (2004); Harrison and Zeevi (2005); Whitt (1999)) as follows. We assume that a given working day is divided into n distinct periods of equal length T, so that the overall horizon is of length nT. The period length is 15 or 30 minutes in practice.

The inbound calls arrive following a stochastic process with a random arrival rate in each period i, denoted by Λ_i. Furthermore, using the modeling in Avramidis et al. (2004); Whitt (1999), we assume that the arrival rate Λ_i is of the form

$$\Lambda_i = \Theta f_i, \text{ for } i = 1, ..., n, \tag{5.1}$$

where Θ is a positive real-valued random variable. The random variable Θ can be interpreted as the unpredictable level of busyness of a day. A large (small) outcome of Θ corresponds to a busy (not busy) day. The constants f_i model the intra-day seasonality, i.e. the shape of the variation of the arrival rate intensity across the periods of the day, and they are assumed to be known. Formally, if a sample value in a given day of the random variable Θ is denoted by θ, the corresponding outcome of the arrival rate over period i for that day is defined by $\lambda_i = \theta f_i$.

We assume that service times for inbound calls are independent and exponentially distributed with rate μ. The calls arrive to a single infinite queue working under the the first come, first served (FCFS) discipline of service. Neither abandonment nor retrials are allowed. The staffing level which guarantees the required service level is then computed by

$$n_i(\theta\, f_i) \quad = \quad F^{-1}_{\theta\, f_i}(SL_i). \tag{5.2}$$

with the function F defined by Equation (2.3).

5.2.2 Shifts Setting

We denote the period sets of the day by I. Let J be the set of all the feasible work schedules, each of which dictates if an agent answers calls in period $i \in I$. For $i \in I$ and $j \in J$, we define the $|I| \times |J|$ matrix $\mathbf{A} = [a_{ij}]$, where

$$a_{ij} = \begin{cases} 1, & \text{if agents in schedule } j \text{ answer calls during period } i, \\ 0, & \text{otherwise.} \end{cases}$$

We furthermore assume that each agent works during consecutive periods, without breaks. Under this assumption, it is direct to see that every column of both matrix $\mathbf{A}$ has contiguous ones and this kind of matrix is totally unimodular, i.e., for any integral vector $\mathbf{b}$, every extreme point of the feasible region $\{\mathbf{x} \mid \mathbf{Ax} \geq \mathbf{b}\}$ is integral and thus the feasible region is an integral polyhedron.

5.2.3 Stochastic Programming Models for An Optimal Staffing

Model with seasonality factors f known with certainty

We assume first that the f_i are certain and that Θ follows a discrete probability distribution, defined by the sequence of outcomes θ_l, $l \in L$ with L as the set of outcomes set. The assumed probability distribution is presented by q_l, with constraint $\sum_{l \in L} q_l = 1$, $q_l \geq 0$. For period $i \in I$, the parameters $N_{il} = F^{-1}_{i,\theta_l f_i}(SL_i)$, estimated via (5.2), represent the required number of agents in period i associated with a particular busyness factor value θ_l.

Let x_j, $j \in J$, be the decision variables representing the numbers of agents assigned to the various schedules implemented before the start of the day. Each agent assigned to shift j gets a salary c_j for the day. In order to optimize the call center operational cost, Liao et al. (2010) proposed the following stochastic programming model

$$\begin{aligned} \min \quad & \sum_{j \in J} c_j x_j \\ \text{s.t.} \quad & \sum_{l \in L} \sum_{i \in I} q_l M_{il} \leq \bar{M} \\ & \sum_{j \in J} a_{ij} x_j + M_{il} \geq N_{il},\ i \in I,\ l \in L \\ & x_j \in \mathbb{Z}^+, j \in J \\ & M_{il} \geq 0,\ i \in I,\ l \in L. \end{aligned} \tag{5.3}$$

The objective of Problem (5.3) is to minimize the agents salary cost. The variables M_{il} represent the amount of under-staffing at period i in event l. The first constraint states that the total

expected under-staffing should not exceed the prescribed limit $\bar{M}$. The second constraint bounds from below the understaffing amount at each period of day and for each level of the busyness factor. When the first constraint on the expected understaffing is active at the optimum, the second constraint will also be active and is equivalent to defining the understaffing as $M_{il} = \max\{0, N_{il} - \sum_{j\in J} a_{ij}x_j\}$. The last two sets of constraint defines the non-negativity and integer conditions for program variables.

It is possible to take advantage of the totally unimodular structure of matrix $A = (a_{ij})$ and make Problem (5.3) computationally much easier by adding auxiliary variables ($y_i \in \mathbb{Z}^+, i \in I$) to represent the available work force $\sum a_{ij}x_j$. Indeed, the variable x appears in equation $\sum a_{ij}x_j = y_i$ and in the objective, but nowhere else. Hence, the integrality condition on y is sufficient to enforce integrality of the x in any solution produced by the Simplex algorithm. The new formulation is

$$\begin{aligned}
\min \quad & \sum_{j\in J} c_j x_j \\
\text{s.t.} \quad & \sum_{l\in L}\sum_{i\in I} q_l M_{il} \le \bar{M} \\
& \sum_{j\in J} a_{ij}x_j = y_i,\ i \in I \\
& y_i + M_{il} \ge N_{il},\ i \in I,\ l \in L \\
& y_i \in \mathbb{Z}^+,\ i \in I \\
& x_j \ge 0,\ j \in J \\
& M_{il} \ge 0,\ i \in I,\ l \in L.
\end{aligned} \tag{5.4}$$

Clearly, Problem (5.4) is equivalent of Problem (5.3). Notice that the integer constraints on x_j are relaxed, Problem (5.4) contains $|I|$ integer variables and $|J|+|I|\times|L|$ continuous variables while Problem (5.3) contains $|J|$ integer variables and $|I|\times|L|$ continuous variables. The integer constraints in Problem (5.4) are only to find the ceilings of some continuous values, they are thus less computationally consuming than that in Problem (5.3).

Model with uncertain seasonality factors f

The seasonality factors may not be known with certainty. Their value is usually estimated through some statistical scheme, and their true value may differ from the estimated one. The difference between the true f_i and its estimator is taken to be a white noise ϵ_i in period i :

$$f_i = \hat{f}_i + \epsilon_i.$$

We assume that θ and the noises ϵ_i are independent. The theoretical staff size that is required to meet the desired service level in period i also depends on the random noise ϵ_i. We now replace the continuous distribution of the ϵ_i by a discrete one, or equivalently a discrete distribution of the f_i. Let f_{ik}, $k \in K_i$ be the set of discrete values and let π_{ik}, with $\sum_{k \in K_i} \pi_{ik} = 1$, be the associated probabilities. For period $i \in I$, the parameters $N_{ikl} = F^{-1}_{i,\theta_l f_{ik}}(SL_i)$, estimated via (5.2), represent the required number of agents associated with a particular busyness factor value θ_l and seasonality factor f_{ik}. We can now formulate an extension of the base model of Liao et al. (2010) to account for the stochastic variability of the seasonality factors f_i:

$$
\begin{aligned}
\min \quad & \sum_{j \in J} c_j x_j \\
\text{s.t.} \quad & \sum_{l \in L} q_l \sum_{i \in I} \sum_{k \in K_i} \pi_{ik} M_{ikl} \leq \bar{M} \\
& \sum_{j \in J} a_{ij} x_j = y_i, \ i \in I \\
& y_i + M_{ikl} \geq N_{ikl}, \ i \in I, \ k \in K_i, \ l \in L \\
& y_i \in \mathbb{Z}^+, \ i \in I \\
& x_j \geq 0, \ j \in J \\
& M_{ikl} \geq 0, \ i \in I, \ k \in K_i, \ l \in L.
\end{aligned}
\tag{5.5}
$$

5.3 Distributionally Robust Model

In the above stochastic programming formulation, the true distribution of θ was assumed to be known, and as a consequence the different constraints of the models are satisfied for any outcome θ_l associated with this distribution.

At the end of the previous section we proposed an extension of Liao et al. (2010) to account for the stochastic variability of the seasonality factors. We now turn our attention to the busyness factor θ. The same argument as for the seasonality factors holds concerning the imperfect knowledge on the true distribution of θ. To make the solutions of models (5.4) and/or (5.5) robust with respect to this imperfect knowledge, we substitute to the estimated probabilities of the busyness values θ a family of alternative probabilities distributions compatible with the observed values of θ. The distributionally robust solution is such that it solves the stochastic programming staffing problem against the worst probability distribution in the class of alternative distributions for θ.

A standard question in such an approach is size of the probability distribution set. It is well known that too large sets, i.e., in our case, sets including all potential probability distributions,

can be extremely conservative in the sense that the robust solution has an objective function value much worse than the objective function value of the solution of the nominal distribution.

It is thus necessary to consider partial uncertainty sets, in the sense that some potential distributions are not included. The idea consists then of introducing, by tuning the size of the uncertainty set, efficient tradeoffs between the probability of constraint violation and the objective function value. Our approach allows thus the modeler to vary the level of conservatism of the robust solutions in terms of probabilistic bounds of constraint violations. Clearly, in such a process, theoretical bounds linking uncertain sets size and constraints violation probabilities are required.

5.3.1 Uncertainty Set Based on A Statistical Dispersion Model

The true probability distribution of the random factor Θ is not known. It must be estimated by some statistical mean. For instance, we can imagine that a set $(\hat{\theta}_1, \ldots, \hat{\theta}_N)$ of historical data is available. The maximum likelihood estimator of the true probability p_l is the observed frequency $q_l = n_l/N$. Moreover, the classical Pearson's test of goodness of fit is based on the quantity

$$X^2 = \sum_l \frac{(n_l - Np_l)^2}{Np_l} = \sum_l N\frac{(q_l - p_l)^2}{p_l}.$$

Asymptotically X^2 follows a χ^2 distribution with $|L| - 1$ degrees of freedom. This asymptotic distribution probability makes it possible to define a first confidence region around q for the true probability p. To this end, we define the dispersion measure $\sum_l N\frac{(q_l-p_l)^2}{q_l}$ and introduce the set of alternative probabilities

$$H_\alpha = \{p \geq 0 : \sum_l N\frac{(q_l - p_l)^2}{q_l} \leq \alpha, \sum_l p_l = 1\} \tag{5.6}$$

that are somehow compatible with the observed frequencies q_l.

The goal of the present analysis would be to incorporate this formulation into the stochastic programming formulation (5.3). Namely, we shall try to solve (5.3) for the worst possible distribution of p in the confidence region (5.6). The formal implementation of this idea consists of replacing the constraint $\sum_{l\in L}\sum_{i\in I} q_l M_{il} \leq \bar{M}$ in (5.3) with its robust counterpart

$$\sum_{l\in L}\sum_{i\in I} p_l M_{il} \leq \bar{M}, \text{ for all } p \in H_\alpha. \tag{5.7}$$

Note that (5.7) is equivalent to

$$\max_{p\in H_\alpha}\left\{\sum_{l\in L}\sum_{i\in I} p_l M_{il}\right\} \leq \bar{M}.$$

However, it can be shown that this infinite dimensional robust counterpart has an equivalent formulation as a conic quadratic constraint. Due to the presence of integer variables x, the equivalent robust counterpart leads to a nonlinear mixed integer problem, possibly a difficult one to solve. We shall not use this test in our analysis, but we shall be inspired by it to define a kind of confidence level set for the true probability p. We shall see that we can replace (5.7) by a more restrictive constraint that is equivalent to a set of linear inequalities. In this way we remain in the realm of linear programming with integer variables.

In order to remain in the realm of mixed integer linear programming for which powerful commercial solvers exist, we shall replace the maximization over the confidence region H_α, by the maximization over a larger, but linear, set

$$\mathcal{P}_\beta = \{p \geq 0 : \sum_{l\in L} p_l = 1,\ \sum_{l\in L} \frac{|p_l - q_l|}{\sqrt{q}_l} \leq \beta\}. \tag{5.8}$$

The larger β, the larger the admissible dispersion and the higher is the protection against the unfavorable probability distributions. Clearly, in order to be coherent with (5.6), the set size factor β has to be chosen to enforce $H_\alpha \subset \mathcal{P}_\beta$. From the simple inequality on norms, we have for any $\theta \in \mathbb{R}^{|L|}$

$$\sum_{l\in L} |\theta_l| \leq \sqrt{|L|}\sqrt{\sum_{l\in L}\theta_l^2}.$$

It follows that for $\beta_\alpha = \sqrt{|L|}\sqrt{\alpha/N}$ the set P_{β_α} contains the set H_α. Therefore, one has

$$\beta_\alpha = \sqrt{|L|}\sqrt{\frac{\alpha}{N}} \Rightarrow H_\alpha \subset \mathcal{P}_{\beta_\alpha}.$$

Hence

$$\max_{p\in H_\alpha}\left\{\sum_{l\in L}\sum_{i\in I} p_l M_{il}\right\} \leq \max_{p\in P_{\beta_\alpha}}\left\{\sum_{l\in L}\sum_{i\in I} p_l M_{il}\right\}$$

and (5.8) implies (5.7). Let us now derive the equivalent counterpart of (5.8). Let

$$\begin{aligned} F &= \max_{p\in\mathcal{P}_\beta} \sum_{l\in L} p_l \sum_{i\in I} M_{il} \\ &= \max_p \{\sum_{l\in L} p_l \sum_{i\in I} M_{il} : \sum_{l\in L} \frac{|p_l - q_l|}{\sqrt{q_l}} \leq \beta,\ \sum_{l\in L} p_l = 1,\ p_l \geq 0, \forall l \in L\} \end{aligned} \tag{5.9}$$

We shall now explicit problem (5.9) as a linear programming problem. Define the new variables

$$\delta_l = p_l - q_l,$$

the problem becomes

$$\begin{aligned} \max_{\delta} \quad & \sum_{l\in L} q_l \sum_{i\in I} M_{il} + \sum_{l\in L}\sum_{i\in I} M_{il}\delta_l \\ \text{s.t.} \quad & \sum_{i\in L} \frac{|\delta_l|}{\sqrt{q_l}} \leq \beta \\ & \sum_{l\in L} \delta_l = 0 \\ & \delta_l \geq -q_l,\ l \in L. \end{aligned} \tag{5.10}$$

We consider the dual of Problem (5.10),

$$\begin{aligned} \min_{v,w,z} \quad & \sum_{l\in L} q_l \sum_{i\in I} M_{il} + \sum_{l\in L} q_l w_l + \beta z \\ \text{s.t.} \quad & z \geq \sqrt{q_l}\left[\sum_{i\in I} M_{il} + v + w_l\right],\ l \in L \\ & z \geq -\sqrt{q_l}\left[\sum_{i\in I} M_{il} + v + w_l\right],\ l \in L \\ & w_l \geq 0,\ l \in L. \end{aligned} \tag{5.11}$$

By strong duality, since Problem (5.10) is feasible and bounded, then the dual Problem (5.11) is also feasible and bounded and their objective values coincide.

Back to the global formulation of the staffing problem with uncertain busyness daily factors, we obtain the following mixed integer linear programming problem in the original variables

(x, M) and the auxiliary variables (v, w, z)

$$
\begin{aligned}
\min \quad & \sum_{j \in J} c_j x_j \\
\text{s.t.} \quad & \textstyle\sum_{l \in L} q_l (\sum_{i \in I} M_{il}) + \sum_{l \in L} q_l w_l + \beta z \leq \bar{M} \\
& -z \leq \sqrt{q_l} \left[\sum_{i \in I} M_{il} + v + w_l \right] \leq z, \quad \forall l \in L \\
& \sum_{j \in J} a_{ij} x_j + M_{il} \geq N_{il}, \; i \in I, \; l \in L \\
& x_j \in \mathbb{Z}^+, \; j \in J \\
& M_{il} \geq 0, \; i \in I, \; l \in L \\
& w_l \geq 0, \; l \in L.
\end{aligned}
\tag{5.12}
$$

Problem (5.12) is the equivalent robust counterpart of the robust version of Problem (5.3) with uncertainty set (5.8) for the underlying business factor probability distribution. It is worth elaborating on the first constraint in Problem (5.12). The first term on the left-hand side is the expected under-staffing taken with respect to the reference, or nominal, probability distribution q. The other two components are safety factors the extra under-staffing that could occur when the true probability distribution is the worst possible in the uncertainty set. Note that the safety term βz is proportional to the *immunization* factor β. The larger β, the larger the admissible dispersion and the higher is the protection against the risk of incurring an extra under-staffing if the distance between the true distribution p and the nominal distribution q increases.

Similar to that we proposed in Problem (5.4), a possible way to make Problem (5.12) easier to be solved is to add some auxiliary variables ($y_i \in \mathbb{Z}^+, i \in I$), Problem (5.12) can then be

reformulated as

$$
\begin{aligned}
\min \quad & \sum_{j\in J} c_j x_j \\
\text{s.t.} \quad & \sum_{l\in L}\sum_{i\in I} q_l M_{il} + \sum_{l\in L} q_l w_l + \beta z \le \bar{M} \\
& -z \le \sqrt{q_l}\left[\sum_{i\in I} M_{il} + v + w_l\right] \le z, \quad \forall l \in L \\
& \sum_{j\in J} a_{ij} x_j = y_i,\ i \in I \\
& y_i + M_{il} \ge N_{il},\ i \in I,\ l \in L \\
& y_i \in \mathbb{Z}^+, i \in I \\
& x_j \ge 0, j \in J \\
& M_{il} \ge 0,\ i \in I,\ l \in L \\
& w_l \ge 0,\ l \in L.
\end{aligned}
\tag{5.13}
$$

Problem (5.13) is equivalent to Problem (5.12). Notice that the integer constraints on x_j are relaxed, since y_i are restricted to be integers, thanks to the total unimodularity property of matrix $\mathbf{A}$, x_j are automatically integers. Problem (5.13) contains $|I|$ integer variables and $|J| + |I| \times |L| + |L| + 2$ continuous variables while Problem (5.12) contains $|J|$ integer variables and $|I| \times |L| + |L| + 2$ continuous variables.

Model (5.13) is easily extended to the case with uncertain seasonality factors as it was done in Section 5.2.3. The equivalent robust counterpart is then

$$
\begin{aligned}
\min \quad & \sum_{j\in J} c_j x_j \\
\text{s.t.} \quad & \sum_{l\in L}\sum_{i\in I} q_l \sum_{k\in K_i} \pi_{ik} M_{ikl} + \sum_{l\in L} q_l w_l + \beta z \le \bar{M} \\
& -z \le \sqrt{q_l}\left[\sum_{i\in I}\sum_{k\in K_i} \pi_{ik} M_{ikl} + v + w_l\right] \le z, \quad \forall l \in L \\
& \sum_{j\in J} a_{ij} x_j = y_i,\ i \in I \\
& y_i + M_{ikl} \ge N_{ikl},\ i \in I,\ k \in K_i,\ l \in L \\
& y_i \in \mathbb{Z}^+, i \in I \\
& x_j \ge 0, j \in J \\
& M_{ikl} \ge 0,\ i \in I,\ k \in K_i,\ l \in L \\
& w_l \ge 0,\ l \in L.
\end{aligned}
\tag{5.14}
$$

5.3.2 Standard Uncertainty Set: An Alternative Formulation

A statistical dispersion measure, like Pearson's, is a sensible choice for the design of an efficient uncertainty set. Unfortunately, it does not seem possible, via such a measure, to compute a reasonable estimate of the probability that the robust solution satisfies the uncertain constraint. In this subsection, we propose an alternative uncertainty set formulation enabling such calculation for constraint violation probability. The derivation is based on the equivalence

$$\begin{cases} \sum_{l\in L} p_l M_l \le \bar{M} \\ \sum_{l\in L} p_l = 1,\ p \ge 0 \end{cases} \Leftrightarrow \begin{cases} \sum_{l\in L} p'_l (M_l - \bar{M}) \le 0 \\ p' \ge 0, \end{cases} \tag{5.15}$$

which holds in the following sense: if the left part holds for some p, the right part holds for $p' = p$; if the right part holds for $p' \neq 0$, the left part holds for $p = p' / \sum_{l\in L} p'_l$). This naturally leads to the following uncertainty model

$$\begin{cases} p'_l = q_l(1+\xi_l),\ \forall l \in L \\ \xi_l \in [-1,1],\ \forall l \in L \\ p = p' / \sum_{l\in L} p'_l. \end{cases} \tag{5.16}$$

The definition is meaningful if $\max_{l\in L} q_l \le 0.5$ and $p'_l \neq 0$. A sufficient condition for the latter is $\xi_l > -1$ for all $l \in L$.

With this model of probability, the condition on the uncertain constraint (5.15) becomes

$$\sum_{l\in L} p'_l (M_l - \bar{M}) = \sum_{l\in L} q_l (M_l - \bar{M}) + \sum_{l\in L} \xi_l q_l (M_l - \bar{M}) \le 0.$$

Define the uncertainty set

$$\Xi = \{\xi : ||\xi||_\infty \le 1,\ ||\xi||_2 \le k\}.$$

The robust counterpart of the uncertain constraint is thus

$$\sum_{l\in L} q_l M_l + \sum_{l\in L} \xi_l q_l (M_l - \bar{M}) \le \bar{M},\ \forall \xi \in \Xi.$$

The equivalent robust counterpart (see Babonneau et al. (2010)) is the inequality

$$\sum_{i\in I} q_l M_l + k||Q(M - \bar{M}) + w||_2 + ||w||_1 \le \bar{M}, \quad \text{for some } u, \tag{5.17}$$

where Q is a diagonal matrix with main diagonal $(q_l)_{l\in L}$.

The bound on the probability of constraint satisfaction is given by the following theorem (see Ben-Tal et al. (2009))

Theorem 5.1 *Assume ξ_l, $l \in L$ are independent random variables with range $[-1, 1]$ and common expectation $E(\xi_l) = 0$. Then, for any $z \in \mathbb{R}^{|L|}$*

$$Prob(\sum_{l \in L} z_l \xi_l \geq k||z||_2) \leq e^{-\frac{k^2}{2}}.$$

The theorem directly applies to a formulation with the ellipsoidal uncertainty set $\{\xi : ||\xi||_2 \leq k\}$. Because the theorem holds under the hypothesis $||\xi||_\infty \leq 1$, we can replace the ellipsoidal uncertainty set by Ξ, which is the intersection of the two balls in the l_2 and l_∞ norms. We thus have

Corollary 5.1 *Assume ξ_l, $l \in L$ are independent random variables with range $[-1, 1]$ and common expectation $E(\xi_l) = 0$. Then for any solution to the equivalent robust counterpart (5.17)*

$$Prob(\sum_{l \in L} p_l M_l \geq \bar{M}) \leq e^{-\frac{k^2}{2}}.$$

Because our problem involves integer variables, it is computationally more efficient (for the time being) to replace the ellipsoidal uncertainty set by on in the l_1-norm. Because the following inequalities hold for any $a \in \mathbb{R}^{|L|}$

$$\frac{1}{\sqrt{|L|}}||a||_1 \leq ||a||_2 \leq \sqrt{|L|}||a||_\infty$$

we can replace Ξ by the larger uncertainty set

$$\{\xi : ||\xi||_\infty \leq 1,\ ||\xi||_1 \leq k\sqrt{|L|}\} \supseteq \Xi$$

and the equivalent robust counterpart (5.17) by the stricter inequality

$$\sum_{l \in I} q_l M_l + k\sqrt{|L|}\,||Q(M - \bar{M}) + w||_\infty + ||w||_1 \leq \bar{M}, \quad \text{for some } w.$$

Finally, as shown in Proposition 1 of Babonneau et al. (2010), the above inequality is equivalent

to the set of inequalities

$$\begin{aligned}\sum_{l\in I} q_l M_l + k\sqrt{|L|}z + \sum w_l &\leq \bar{M}\\ z + w_l &\geq q_l(M_l - \bar{M}),\ l \in L\\ z + w_l &\geq q_l(\bar{M} - M_l),\ l \in L\\ w \geq 0,\ z \geq 0,&\end{aligned}$$

where $w \in \mathbb{R}^{|L|}$ and $z \in \mathbb{R}$ are auxiliary variables.

In order to have a model associated with a theoretical bound for the constraint violation probability, we plug this inequalities in our distributionally robust call center model, we obtain a new model, similar to (5.14). Namely

$$\begin{aligned}\min\quad & \sum_{j\in J} c_j x_j\\ \text{s.t.}\quad & \textstyle\sum_{l\in I} q_l(\sum_{i\in I}\sum_{k\in K_i}\pi_{ik}M_{ikl}) + k\sqrt{|L|}z + \sum w_l \leq \bar{M}\\ & z + w_l \geq q_l(\sum_{i\in I}\sum_{k\in K_i}\pi_{ik}M_{ikl} - \bar{M}), \quad \forall l \in L\\ & z + w_l \geq q_l(\bar{M} - \sum_{i\in I}\sum_{k\in K_i}\pi_{ik}M_{ikl}), \quad \forall l \in L\\ & \sum_{j\in J} a_{ij}x_j = y_i,\ i \in I\\ & y_i + M_{ikl} \geq N_{ikl},\ i \in I,\ k \in K_i,\ l \in L\\ & y_i \in \mathbb{Z}^+, i \in I\\ & x_j \geq 0, j \in J\\ & M_{ikl} \geq 0,\ i \in I,\ k \in K_i,\ l \in L\\ & z \geq 0,\ w_l \geq 0,\ l \in L.\end{aligned} \tag{5.18}$$

We conclude this subsection by showing that the uncertainty set Ξ could be viewed as a form of dispersion measure. Namely, we define the set of probability distributions

$$\mathcal{P}_k = \left\{p : p = \frac{p'}{\sum_{l\in L} p'_l},\ \sum_{l\in L}\left(\frac{p'-q}{q}\right)^2 \leq k^2, p' \geq 0\right\}. \tag{5.19}$$

This definition is compatible with an assumption of independence of the variables p'_l. It lead us to assume that the quantities $(p'_l - q_l)/q_l$ are independent random variables with range $[-1, 1]$. Note that it does not imply that the p_l are independent. Thanks to the independence assumption, we

have been able to compute a bound on the probability of satisfaction of the uncertain constraint. The alternative formulation (5.19) bypasses the difficulty we've met with H_α. There, p only enters the definition and the condition $\sum_{l\in L} p_l = 1$ creates an explicit dependence among the variables.

5.4 Numerical Experiments and Results

The numerical results reported in this section aim at assessing empirically the merit of the distributionally robust approach as compared with the plain stochastic programming approach. A robust, or stochastic programming, solution consists in a set of shifts x. The behavior of this solution is analyzed on large samples of daily operations scenarios.

In this section, we conduct a numerical study in order to evaluate and compare between the classic stochastic programming approach and the distributionally robust programming approach. In Section 5.4.1, we describe the numerical experiments. In Section 5.4.2, we analyze the results and derive various insights.

5.4.1 Setting of the Experiments

We describe in this section the data used in the numerical examples first, and then the design of experiments.

Parameter values

Inbound calls. In the experiments, we use real data from a Dutch hospital which exhibits a typical and significant workload time-of-day seasonality. To give an idea of the pattern of the mean arrival rate, we consider three days, a normal one, a busy one and a not so busy one. The solid line in Figure 5.1, represents arrival in a normal day, while the dashed lines represent the two other cases. Clearly the three lines have a similar pattern, with low values at the beginning and at the end of the day, with a two peaks one in late morning and one in the afternoon, and a relative decrease in-between during the lunch break. This illustrates the choice of the model, with (almost) fixed seasonality factors and a multiplicative busyness factor. The day starts at 8:00 am, finishes at 8:30 pm, and is divided into $|I| = 50$ periods of 15 minutes each.

From this observation, we construct an illustrative example as follows. The average rate of arrivals at each period of the day is supposed to have been estimated by statistical analysis on a record of $n = 400$ working days. The estimated seasonal factors are given in Table 5.1. Note that the seasonal factors could have been normalized because the true arrival rate is obtained by multiplying those values by the busyness day factor.

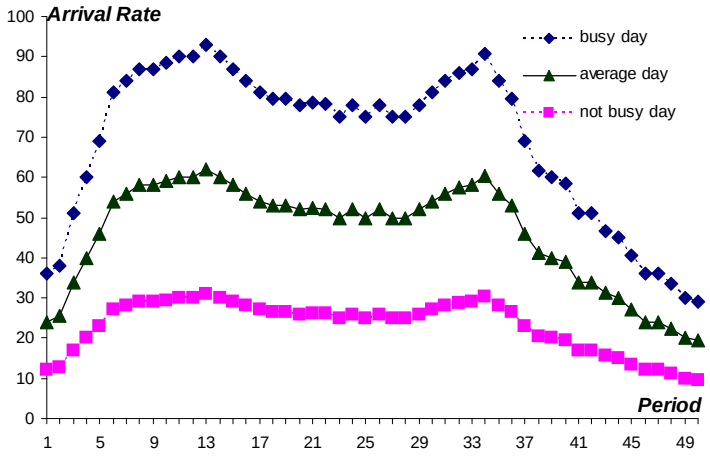

Figure 5.1: Arrival rate graph

Table 5.1: Average seasonality factors estimated from a sample of $n = 400$ working days

f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}
6	6.35	8.5	10	11.5	13.5	14	14.5	14.5	14.75	15	15	15.5

f_{14}	f_{15}	f_{16}	f_{17}	f_{18}	f_{19}	f_{20}	f_{21}	f_{22}	f_{23}	f_{24}	f_{25}	f_{26}
15	14.5	14	13.5	13.25	13.25	13	13.1	13.05	12.5	13	12.5	13

f_{27}	f_{28}	f_{29}	f_{30}	f_{31}	f_{32}	f_{33}	f_{34}	f_{35}	f_{36}	f_{37}	f_{38}	f_{39}
12.5	12.5	13	13.5	14	14.35	14.5	15.1	14	13.25	11.5	10.3	10

f_{40}	f_{41}	f_{42}	f_{43}	f_{44}	f_{45}	f_{46}	f_{47}	f_{48}	f_{49}	f_{50}
9.75	8.5	8.5	7.8	7.5	6.75	6	6	5.6	5	4.85

The uncertain environment of the problem is built as follows. First we consider that each individual seasonal factor is subject to an independent noise. For the sake of the illustration, we selected a discrete distribution for each seasonal factor with three outcomes $f_i - f_i/10,\ f_i,\ f_i + f_i/10$, with respective probabilities $0.25,\ 0.5,\ 0.25$. This choice is arbitrary, but can be easily replaced by an alternative one. In the numeric experiments, we analyze the general case with uncertain seasonal factors. The seasonal factors known with certainty can be considered as a special case of uncertain seasonal factors.

The second element that introduces uncertainty is the distribution of Θ, the random busyness factor. It is estimated by comparing the record of the mean arrival rate of each working day with the average of all these means. We assume that the distribution of Θ has been estimated from past records by a discrete distribution with $|L| = 41$ outcomes θ_l and probabilities q_l. To construct a plausible distribution, we choose to discretize a continuous distribution. In Avramidis et al. (2004), the authors postulate in their Model 1 that Θ follows a gamma distribution with shape parameter $\gamma > 0$ and scale parameter 1. In this chapter, we assume that Θ can take values from interval $[0.00, 12.00]$, we take 41 equidistant points including the two endpoints 0.00 and 12.00, which gives $|L| = 41$ possible values of θ_l. And we consider 3 types of estimate probability distributions q: distributions A, B and C, which are discretized from a gamma distribution with scale parameter 1 and shape parameter γ (consequentially the mean $E[\Theta]$) as $2, 4$ and 6, respectively. For each type of estimate probability distributions q, we have $\sum_{l \in L} q_l = 1, q_l \geq 0$. Figure 5.2 shows the three probability density functions.

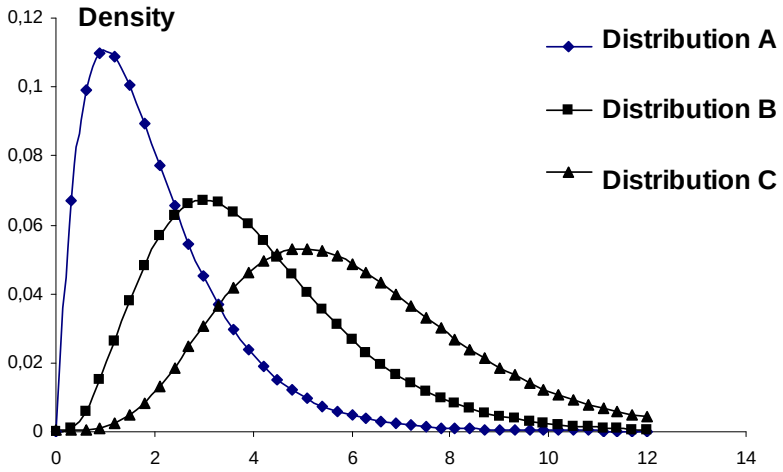

Figure 5.2: Some probability density functions

Finally, for each value of the arrival rate at period i with busyness factor θ_l and seasonal factor (given by one of the three values $f_i - f_i/10,\ f_i,\ f_i + f_i/10$) we compute the staffing

requirement that is needed to meet the service level. To this end, we start with the assumption that mean service time is $1/\mu = 5$ minutes. We use the classical service level corresponding to the well-known 80/20 rule: the probability that a call waits for less than 20 seconds has to be larger or equal to 80 percent. Using Condition (5.2) and Definition (5.1), we deduce the required number of agents N_{ikl} during period i, associated to the values θ_l and f_{ik}.

The understaffing bound $\bar{M}$. The quantity is user dependent. We chose it as follows. We compute the average or the size $N = \sum_{i,k,l} q_l \pi_{ik} N_{ikl}$ of ideal staff. We consider three values for $\bar{M}$: 0, $1\% \times N$ and $2\% \times N$. Note that the value $\bar{M}$ imposes that all $M_{ikl} \geq 0$ in the constraint

$$\sum_{l \in L} \sum_{i \in I} \sum_{k \in K_i} p_l \pi_{ik} M_{ikl} \leq \bar{M} = 0$$

are zero. This case corresponds to the conservative position of 100% protection.

Cost parameters. Agents work 4 or 8-hour days, with neither break nor overtime. Full-time shifts (8-hour) start at the hour or the hour and a half. Part-time shifts (4-hour) start at the hour between 8 am to 2 pm. There are 17 part-time and full-time feasible schedules. Without loss of generality, we use a normalized cost of 1 for each period an agent works at full-time shifts. The part-time shifts unit cost, assumed to be larger, is equal to 1.4 per period. Therefore the agent salary is $c_j = \sum_{i \in I} a_{ij}$ for full-time shifts and $c_j = 1.4 \sum_{i \in I} a_{ij}$ for part-time ones.

The distributionally robust solution

The DR solution is obtained from Problem (5.14) with parameters values as described above. In this formulation, one critical factor remains to be determined, namely the immunization factor β or, in other words, the size of the uncertainty set considered in the distributionally robust model. Ideally, its value would be chosen so that the robust solution ensures that the constraint

$$\sum_{l \in L} p_l \sum_{i \in I} \sum_{k \in K_i} \pi_{ik} M_{ikl} \quad \leq \quad \bar{M} \tag{5.20}$$

is satisfied with a given probability α, say 95%. A seemingly natural approach would be to use the probability that the true distribution belongs to the uncertainty set

$$\mathcal{P}_\beta = \{p \geq 0 : \sum_{l \in L} p_l = 1, \ \sum_{l \in L} \frac{|p_l - q_l|}{\sqrt{q_l}} \leq \beta\}$$

as a lower bound of the probability of satisfaction of the constraint by the robust solution. Indeed, suppose the p's are obtained by making a Monte-Carlo of n from the probabilities q. One

could argue the Pearson indicator $\sum_l n\frac{(q_l-p_l)^2}{q_l}$ is asymptotically approximated by a χ-square distribution with $|L|-1$ degrees of freedom. Using the property that the confidence region H_α is included in P_β we compute a value β, ensuring that H_α has a large enough probability. Unfortunately this path leads to a gross over estimation of β. Indeed, the values of β computed in this way are much too large and the robust solution is overly conservative. This phenomenon is well-known. The fact is clear when one consider the critical value $\beta = 0$ implying an uncertainty set reduced to a singleton with probability zero. The DR solution with $\beta = 0$ still enforces the constraint (5.20) half of the time. This just says that only one half of the possible distributions p, close to or far from q are harmful. This phenomenon is discussed in Babonneau et al. (2010).

The literature proposes much stronger approximations of chance programming (see, e.g., chapter 2 of the book by Ben-Tal et al. (2009)). Those approximations strongly rely on the assumption that the random coefficients in the uncertain equation, namely the p's in the constraint (5.20) are independent random variables. This is not the case here, because the condition $\sum_l p_l$ make them dependent. It is not clear to us that the known techniques can be extended to handle our case.

Our approach to determine the β will be purely empirical. We shall let β vary from 0 to 1 and observe the behavior of the robust solution on simulations, as described in the next subsection. This approach is quite common in robust optimization. We conclude this discussion by pointing out that the DR solution with $\beta = 0$ is nothing else than the SP solution. A similar approach can be considered with the uncertainty set (5.19) and the parameter k. It is worth noting that the bound appears to be loose in our setting for most numerical applications due to the monotonic structure of the under-staffing process with respect to the θ_l values.

Simulations

The idea of simulation is to create K scenarios of day operations. To this end, we first draw by Monte-Carlo sampling, a value for p. This is done as follows. We perform n independent random trials with respect to the probability distribution q. For each θ_l we record the frequency of occurrence of θ_l; this frequency defines p_l. Next we draw a value for each seasonal factor among the 3 possibilities with respect to the given probabilities (here, 0.25, 0.5 and 0.25). Given the day operation conditions, we can compute the understaffiing of the DR for that day. We have thus K realizations of the understaffing of the DR solution.

We compute three types of statistics

1. The proportion of times the constraint on understaffing is violated, i.e., the expected understaffing $M = \sum_{l\in L}\sum_{k\in K_i}\sum_{i\in I} p_l\pi_{ik}M_{ilk}$ exceeds $\bar{M}$.

2. The conditional expectation value of $(M - \bar{M})$ conditionally to $M - \bar{M} > 0$.

3. The worst case for $(M - \bar{M})$.

5.4.2 Analysis of the Numerical Results

In this section, we comment on the numerical results and derive the main insights. Four criterions are considered in order to evaluate the performance of both SP and DR methods: The salary cost, the probability of violation of the constraint $M \leq \bar{M}$, the conditional expectation value of $(M - \bar{M})$ for M that exceeds $\bar{M}$, and the maximum $(M - \bar{M})$ among all the $K = 10000$ trials. We compare the performance between SP and DR with different sizes of uncertainty sets $\mathcal{P}_\beta$ (defined by (5.8)) and $\mathcal{P}_k$ (defined by (5.19)), for different under-staffing bound $\bar{M}$. We analyse the trade-off between salary cost and the other three criterions, and show the necessity of taking into account the uncertainty in the probability distribution. These comparison are done based on the 3 types of estimate probability distributions presented previously.

For the 3 types of estimate probability distributions, the value of the under-staffing bound $\bar{M}$, defined as 1% of the total required workforce is $64.97, 120.77$ and 179.19 respectively. That defined as 2% of the total required workforce is $\bar{M} = 129.94, 241.54$ and 358.38. For the models with uncertain seasonal factor f_i, uncertainty set $\mathcal{P}_\beta$, and $\bar{M}$ as 1% (2%)of the total required workforce, Table 5.2 (Table 5.3) displays for each type of estimate probability distribution, the four evolutional criterions mentioned above. Table 5.4 (Table 5.5) has similar structure, but it is related to models with uncertainty set $\mathcal{P}_k$.

In order to examine the trade-off between the salary cost and the protection against risk, we consider for DR different values of β (or k), which correspond to uncertainty sets with different sizes. The higher the β (or k) value, the higher the degree of protection against the uncertainty in probability distribution. An extreme case can be considered, namely $\beta = 0$ (or $k = 0$), which can be viewed as equivalent to SP. For information, given the uncertainty set $\mathcal{P}_\beta$ ($\mathcal{P}_k$) chosen in the following tables, we have observed the percentage that the sampled true probability distribution p falls outside the uncertainty set. We find that almost 100% of p falls outside the uncertainty set $\mathcal{P}_\beta$ ($\mathcal{P}_k$), but numeric results show that not all of them lead to constraint violation.

From Table 5.2 to 5.5, we can observe a trade-off between the salary cost and the other three criterions which present the protection against risk. By increasing the $\beta(ork)$ value, which increases the uncertainty set size, the constraint violation percentage, the conditional expectation of $(M - \bar{M})$, and the max case $(M - \bar{M})$ are eliminated progressively, with an increase in salary cost. Figure 5.3 shows the trade-off between salary cost and the constraint violation percentage. And Figure 5.4 shows the decreasing tendency of the other two criterions in total cost. As

expected, SP has the lowest salary cost. However, the constraint violation percentage for the method SP is remarkable. For the 3 types of estimate probability distribution, the solutions of SP tend to violate the constraints by about half chance. The performance of DR is quite nice. For example, given $\beta = 0.2$, both Table 5.2 and 5.3 show that, for the estimate probability distribution A, B and C, DR reduces the constraint violation percentage more than 33%, 34% and 40% by only increasing about 11%, 4% and 3% the salary cost, respectively. Similar remarks can be found from the results in Table 5.4 and 5.5.

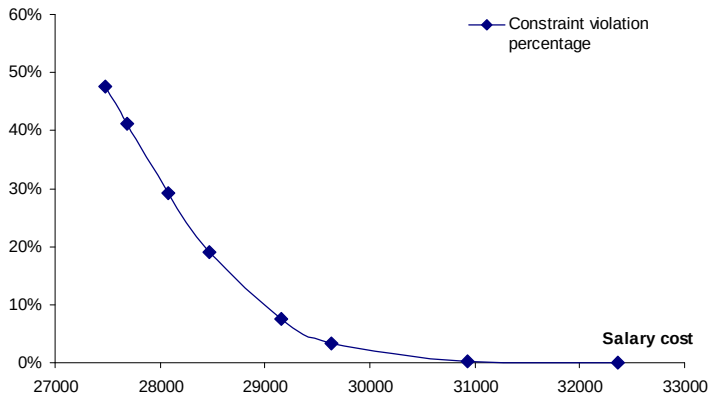

Figure 5.3: Trade-off between the salary cost and constraint violation percentage

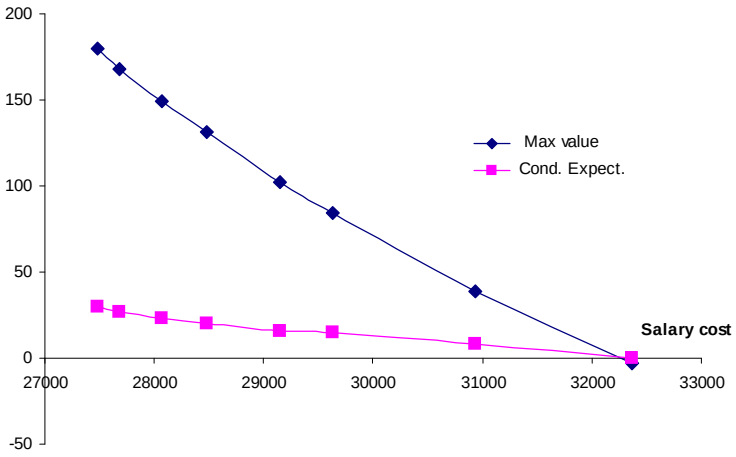

Figure 5.4: Trade-off of the max and conditional expected $(M - \bar{M})$ with the salary cost

In general, the method SP which does not take into account the uncertainty on the probability distribution, leads to violation of constraint (infeasibility) with a quite high proportion. While

the DR method we proposed avoids this trouble, by only paying a relatively small increase on the salary cost. This illustrates the necessity of taking into account the uncertainty on the probability distribution.

For both $\bar{M}$ equals to 1% and 2% of the required total workforce, we find similar performance for both SP and DR, as presented above. An extreme value of $\bar{M}$ is 0, with all M_{ikl} of both SP and DR are zero. Consequently, SP and DR behave the same. The salary cost is the upper bound for all further results. As $\bar{M}$ grows, it is likely that SP and DR diverge more and more. For both models with uncertainty set $\mathcal{P}_\beta$ and $\mathcal{P}_k$, Table 5.6 displays the upper bound salary cost for the 3 types of estimate probability. We observe that given the model with uncertain f_i, the upper bound costs are the same for the 3 types of probability distribution. The reason is simply that in our numeric example, the random variable Θ takes values from the same interval $[0.00, 12.00]$, and all f_{ik} are defined by the same way, then the largest required agents number N_{ikl} are the same for the 3 types of probability distribution.

5.5 Conclusion

We consider a staffing-scheduling problem of a multi-shift call center, of which the probability distribution of random parameter is ambiguous. We introduce firstly the classic stochastic programming formulation of this problem. Then we fully exploit the sample data, and construct a linear distributionally robust model of the staffing-scheduling problem. We next conduct a numerical study in order to evaluate the performance of these two methods and gain useful insights. The necessity of taking into account the uncertainty on probability distribution is exhibited.

In terms of formulation, two factors are worth mentioning. First, we have considered the uncertainty set of probability distribution of the busyness factor Θ. It may be of interest to consider additionally un uncertainty set of probability distribution on uncertain seasonal factors f_i. Second, we have treated the agent shortfall in each period in a similar way. However, for periods with different numbers of required agents, the same quantity of agent shortfall may lead the customers to experience different additional waiting time. Consequently, a given agent shortfall in different period may of different importance. Extended formulation could introduce weights per period that depend on the values of required agent number.

β	Salary cost	Constr. violation (%)	Expectation (M - $\bar{M}$|$M > \bar{M}$)	Worst case $M - \bar{M}$
Set A of probabilities q, $\bar{M} = 64.97$				
0	21500.8	45.46	28.96	168.88
0.01	21660.8	43.13	27.82	164.58
0.05	22313.6	33.55	24.16	147.87
0.1	23164.8	22.73	20.47	126.14
0.2	24896.0	8.77	14.96	88.56
0.5	29372.8	0.11	7.62	21.54
0.8	32361.6	0	NaN	-12.08
1	33673.6	0	NaN	-23.40
Set B of probabilities q, $\bar{M} = 120.77$				
0	27481.6	47.19	29.73	162.76
0.01	27555.2	44.76	28.57	157.71
0.05	27849.6	35.95	24.72	142.81
0.1	28220.8	25.11	21.42	126.20
0.2	28953.6	10.25	16.03	94.16
0.5	30998.4	0.13	7.11	24.62
0.8	32713.6	0	NaN	-20.22
1	33667.2	0	NaN	-39.90
Set C of probabilities q, $\bar{M} = 179.19$				
0	32752.0	48.97	28.1	150.00
0.01	32809.6	46	26.81	145.28
0.05	33030.4	34.44	22.64	129.64
0.1	33302.4	21.56	18.56	110.80
0.2	33814.4	6.09	12.82	77.82
0.5	35171.2	0.01	4.29	4.29
0.8	36243.2	0	NaN	-41.17
1	36841.6	0	NaN	-62.68

Table 5.2: Models with uncertain f_i, uncertainty set $\mathcal{P}_\beta$ and $\bar{M}$ is 1% of total required workforce

β	Salary cost	Constr. violation (%)	Expectation (M - $\bar{M}|M > \bar{M}$)	Worst case $M - \bar{M}$
Set A of probabilities q, $\bar{M} = 129.94$				
0	18134	45.91	40.39	271.99
0.01	18227	43.62	39.58	268.08
0.05	18605	35.2	36.19	249.23
0.1	19098	26.03	31.9	224.26
0.2	20157	12.33	25.66	178
0.5	23578	0.4	13.11	62.66
0.8	26611	0	NaN	-3.36
1	28342	0	NaN	-36.14
Set B of probabilities q, $\bar{M} = 241.54$				
0	24438	47.42	44.8	236.19
0.01	24493	45.66	43.31	231.43
0.05	24704	37.55	38.82	212.71
0.1	24973	28.1	34.18	190.61
0.2	25504	13.4	27.41	148.93
0.5	27059	0.38	13.22	49.83
0.8	28550	0	NaN	-20.39
1	29466	0	NaN	-56.78
Set C of probabilities q, $\bar{M} = 358.38$				
0	29891	48.66	46.11	223.46
0.01	29939	46	44.43	217.43
0.05	30125	35.34	39.02	197.46
0.1	30352	23.59	34.13	173.81
0.2	30800	8.44	26.49	128.4
0.5	32035	0.05	11.72	20.55
0.8	33120	0	NaN	-59.46
1	33766	0	NaN	-99.46

Table 5.3: Models with uncertain f_i, uncertainty set $\mathcal{P}_\beta$ and $\bar{M}$ is 2% of total required workforce

k	Salary cost	Constr. violation (%)	Expectation (M - $\bar{M}$|$M > \bar{M}$)	Worst case $M - \bar{M}$
Set A of probabilities q, $\bar{M} = 64.97$				
0	21500.8	44.03	29.27	161.16
0.10	21865.6	38.96	26.81	152.16
0.30	22678.4	27.73	23.04	133.95
0.50	23574.4	18.28	19.25	114.24
0.80	25152.0	7.38	14.80	84.10
1.00	26246.4	3.43	12.43	65.16
1.50	29456.0	0.13	4.65	17.01
2.00	32656.0	0.00	NaN	-18.91
Set B of probabilities q, $\bar{M} = 120.77$				
0	27481.6	47.67	29.62	180.18
0.10	27686.4	41.20	26.73	168.26
0.30	28073.6	29.20	22.93	149.06
0.50	28473.6	19.05	19.83	131.70
0.80	29152.0	7.56	15.82	102.42
1.00	29628.8	3.37	14.75	84.59
1.50	30934.4	0.26	8.15	38.89
2.00	32361.6	0.00	NaN	-2.81
Set C of probabilities q, $\bar{M} = 179.19$				
0	32752.0	48.27	27.69	157.67
0.10	33040.0	32.83	22.35	135.80
0.30	33577.6	10.82	15.71	97.75
0.50	34041.6	2.75	12.39	68.12
0.80	34556.8	0.41	8.59	38.26
1.00	34819.2	0.10	10.03	24.44
1.50	35561.6	0.00	NaN	-10.70
2.00	36403.2	0.00	NaN	-45.40

Table 5.4: Models with uncertain f_i, uncertainty set $\mathcal{P}_k$ and $\bar{M}$ is 1% of total required workforce

k	Salary cost	Constr. violation (%)	Expectation (M - $\bar{M}$|$M > \bar{M}$)	Worst case $M - \bar{M}$
Set A of probabilities q, $\bar{M} = 129.94$				
0	18134.4	45.88	40.41	272.00
0.10	18483.2	37.71	37.22	254.79
0.30	19264.0	23.10	31.14	217.21
0.50	20131.2	12.73	26.21	181.70
0.80	21673.6	3.50	18.56	122.38
1.00	22764.8	1.10	15.82	86.56
1.50	26073.6	0.01	1.32	1.32
2.00	29593.6	0.00	NaN	-53.38
Set B of probabilities q, $\bar{M} = 241.54$				
0	24438.4	47.41	44.78	236.15
0.10	24640.0	39.95	40.03	218.17
0.30	25059.2	25.20	33.18	184.33
0.50	25513.6	13.28	27.51	149.05
0.80	26249.6	3.38	19.22	97.65
1.00	26771.2	0.90	15.76	67.33
1.50	28195.2	0.00	NaN	-6.78
2.00	29760.0	0.00	NaN	-69.16
Set C of probabilities q, $\bar{M} = 358.38$				
0	29891.2	48.43	45.44	255.78
0.10	30137.6	34.78	37.37	226.47
0.30	30608.0	13.31	27.59	173.67
0.50	31043.2	3.86	22.42	128.40
0.80	31619.2	0.45	15.53	72.91
1.00	31932.8	0.10	19.82	44.92
1.50	32803.2	0.00	NaN	-26.15
2.00	33808.0	0.00	NaN	-96.50

Table 5.5: Models with uncertain f_i, uncertainty set $\mathcal{P}_k$ and $\bar{M}$ is 2% of total required workforce

Table 5.6: $\bar{M} = 0$, the upper bound for the salary cost

Estimate prob. distribution	uncertain f_i	
	DR Salary	SP Salary
A	48956.8	48956.8
B	48956.8	48956.8
C	48956.8	48956.8

Chapter 6

Workforce Optimization of a Call Center under a Global Service Level Constraint and Information Update

In this chapter, we address a multi-periodic multi-shift call center staffing problem. We consider a global service level constraint, and allow the staffing level to be updated. The call arrival process is assumed to follow a doubly non-stationary stochastic process with a random mean arrival rate. To the contrary to all the problems treated in the previous chapters, we combine the staffing step which determines the number of required agents for each period, and the shift-scheduling step which determines the agents number working in each shift. The staffing-scheduling policy is initially decided before the beginning of the working day and a real-time update is allowed within the same day. The objective is to minimize the sum of the regular salary and the update adjustment cost, with respect to a global service level constraint. We construct two models using two-stage stochastic program with recourse to describe this problem. Through numerical experiments, we illustrate the excellent performance of these two approaches. The advantages of adding the update flexibility and those of using a global service level, instead of period to period ones, are also shown.

6.1 Introduction

In all of the previous chapters, we considered period to period service level (SL) constraints. This type of SL constraints are referred to as *hard constraints*. As a first step, we made use of the Erlang formula in order to determine for each period the required staffing level, and as a second step, we optimized the scheduling of the shifts. Another optimization approach in practice consists on considering a global SL constraint, referred to as *soft constraint*. The soft constraint requires to meet an SL objective for the whole planning horizon (several periods) which might be one day, one week or even longer. This means that the call center could reach low SLs during some periods (or intervals) and high ones during other periods. Roughly speaking, since shifts span multiple intervals, a model with hard constraints would lead to overcapacity in certain periods, and the actual global SL would be much higher than required. A model with a soft constraint avoids this inconvenience, however, solving the optimization problem would be more complex.

In the context of call centers, Koole and van der Sluis (2003) are the first to deal with a global SL constraint. The authors prove the multimodularity property of the shift scheduling problem. Then, they develop a heuristic based on a local search algorithm to solve the problem. Robbins (2007) constructs a stochastic model with a global SL constraint that takes into account the uncertainty in arrival parameters. It is however a static single-stage model. He determines the staffing level once for the beginning of the day, and can not adjust it later on, if needed. The author proposes a piecewise linear approximation in order to describe the nonlinear Telephone Service Factor (TSF) curve. In this chapter, we extend his results by allowing recourse actions.

Similarly to the previous chapters, we again consider the uncertainty in arrival rates. Recall that there is a strong positive correlation between arrival rates during the different periods of the same day. Also, we allow the decision-maker to modify the staffing level after observing the actual busyness of day. If, the call center is actually over-staffed (under-staffed), then she has the possibility to reduce (increase) the staffing level. To the best of our knowledge, the only work which considers a combined staffing-scheduling problem with recourse actions and a global service constraint in a stochastic setting is Gans et al. (2009). The authors construct a two-stage stochastic model, which is an extension of the model of Robbins (2007). These two models are smartly built, but there exist a disadvantage: they force the service level for each period in each scenario to be higher than a certain value. This leads to the problem as follows. For the periods before updating, the staffing level is common for all scenarios. If a scenario with large mean arrival rates exists, since the SL for each scenario should exceed a certain level, the decision-maker has to assign in that case a too high staffing level not necessary for other

less busier scenarios. Using this approach, the staffing policy would be somewhat influenced by scenarios with large arrival rates. Such scenarios occur however with negligible probabilities.

In this chapter, we consider a multi-period multi-shift staffing-scheduling problem with staffing adjustment (update) and randomness of the call arrival rate parameters. The objective is to determine the optimal schedules that minimize the operational costs of the call center while achieving a predefined global service level. The global service level is defined as the average of the SLs achieved over all periods, weighted by the intensity of arrival rates. Our approach is based on a the two-stage stochastic program with recourse.

The achieved SL in a given period is a function of the value of the call arrival rate. Thus, the parameters associated with the staffing adjustable variables are not constant, but depend on the uncertainty of the call arrival rate. This was not the case for the fixed recourse as presented in Chapter 4. Hence, even if we construct an adjustable robust model for this staffing-scheduling problem, we can not use the linear decision rule to solve it while keeping a computationally tractable model. For this reason, we only consider here a two-stage stochastic programming optimization problem.

We build two stochastic models: the first uses the piecewise linear approximation to approach the TSF curve, as that proposed in Robbins (2007). The second uses a simple linear function to approach that curve. The difference and advantages of our models comparing to that in Gans et al. (2009), is that we don't give any restriction on the SL in any period or scenario. The risk mentioned above for the models of Robbins (2007) and Gans et al. (2009) is then avoided. We then conduct a comparison study between the different models: the above two models, the model with hard constraints and staffing update, and the static one with global service level constraint and without staffing update. The comparison shows the advantages of adding the update flexibility, and points out the impact of having a global service level constraint.

We distinguish two major contributions in this chapter. The first contribution is the proposition of two new models for the staffing-scheduling problem with global SL constraint, taking into account the call arrival rate uncertainty and allowing the update of the staffing level. Even though the first one is time consuming, the numerical experiments show that both models are good approaches for the current staffing problem. The second contribution is the analysis of the impact of the flexibility offered by the recourse action and the global SL constraint. We show that allowing staffing level adjustment reduces total cost, and period to period SL constraints lead to unnecessary over-staffed situations.

The rest of this chapter is structured as follows. In Section 6.2, we describe the call center model under consideration and formulate the first model associated with the staffing-scheduling

problem, using piecewise linear functions to approach the TSF curve. In Section 6.3.2, we present a simple model associated with the same problem, by using a single linear function to approach the TSF curve. In Section 6.4, we conduct a numerical study with a small size problem in order to evaluate the two new models. We exhibit the benefit offered by staffing level adjustment on the optimization problem, and the over capacity caused by period to period SL constraints. In Section 6.5, we extend the analysis to a larger size problem. This chapter ends with concluding remarks.

6.2 Problem Formulation

We consider a single class call center. The problem is similar to that in Section 4.2.1 in Chapter 4. The new feature is that the target SL is no longer required for each period, but only for the whole day. Consequently, the required number of agents for each period is not determined through Equation (4.2). It is now a variable related to the global SL target.

6.2.1 The inbound call arrival process

In order to address uncertain and time-varying mean arrival rates coupled with significant correlations, the same as the previous chapters, we model the inbound call arrival process by a doubly stochastic Poisson process as follows.

We denote the set of periods of the day of interest by I. The mean arrival rate Λ_i for period i, $i \in I$, is assumed to be of the form

$$\Lambda_i = \Theta f_i, \tag{6.1}$$

where Θ is a positive real-valued random variable. The random variable Θ can be interpreted as the unpredictable busyness of the day. The constants f_i model the shape of the variation of the mean arrival rate intensity across periods, and they are assumed to be known. If a sample value in a given day of the random variable Θ is denoted by θ, the corresponding outcome of the arrival rate over period i for that day is $\lambda_i = \theta f_i$. The random variable Θ is assumed to follow a discretized probability distribution, defined by the sequence of outcomes θ_l, for $l \in L$, with L as the set of the possible scenarios of the busyness factors. An outcome θ_l occurs with probability p_l, with $\sum_{l \in L} p_l = 1$.

We assume that service times for inbound calls are independent and exponentially distributed with rate μ. The calls arrive to a single infinite queue working under the the first come, first served (FCFS) discipline of service. Neither abandonment nor retrials are allowed.

6.2.2 Shifts Setting

Let J be the set of the feasible work schedules, each of which dictates whether an agent answers calls during period i, or not. For $i \in I$ and $j \in J$, we define the $|I| \times |J|$ matrix $\mathbf{A} = [a_{ij}]$, with

$$a_{ij} = \begin{cases} 1, & \text{if agents in schedule } j \text{ answer calls during period } i, \\ 0, & \text{otherwise.} \end{cases}$$

We divide the overall horizon into the early horizon and late horizon, denoted by the sets of periods I_1 and I_2, respectively. After observing the call volumes in the early horizon, a real-time update of staff capacity is allowed at the beginning of the late horizon. We define also the $|I| \times |J|$ matrix $\mathbf{B} = [b_{ij}]$ with

$$b_{ij} = \begin{cases} 1, & \text{if agents in the new schedule } j \text{ answer calls during period } i, \\ 0, & \text{otherwise.} \end{cases}$$

Note here that all the terms of the first $|I_1|$ lines of matrix $\mathbf{B}$ are all zeros.

We furthermore assume that the schedules have no breaks in the middle. Then, the binary matrix $\mathbf{A}$ and $\mathbf{B}$ have contiguous 1 terms. Thus, $\mathbf{A}$ and $\mathbf{B}$ are totally unimodular.

6.2.3 TSF Curve and Global Service Level

The SL archived during period i is counted in the global SL through the weight $\frac{f_i}{\sum_i f_i}$. For a given outcome of the arrival rate in a given period, the service level, also called the telephone service factor (TSF) (Gans et al. (2003)), is defined based on the Erlang C model, presented in Section 2.1.2 in Chapter 2.

We consider an example of a call center with a mean call arrival rate $\lambda = 6.88$ during a given period, and mean service time $\frac{1}{\mu} = 5$. Figure 6.1 shows the TSF curve which corresponds to the percentage of customers served within 20 seconds as a function of the number of agents.

Figure 6.1 reveals that the TSF curve is non-linear in the number of agents. Point A is the demarcation point from which the SL starts to be non-zero and the TSF curve becomes concave. This non-linear character of the TSF curve was not an issue for the analysis in the previous chapters. The reason is the SL target per period determines the minimum number of agents. The latter is then used as an input in the mathematical program of optimization. Unfortunately, this is no longer possible for problem with a global SL target, since the SL per period is a variable itself. For tractability and in the usual way as in Robbins (2007) for example, we need to use a linear approximation model relating the TSF with the number of agents.

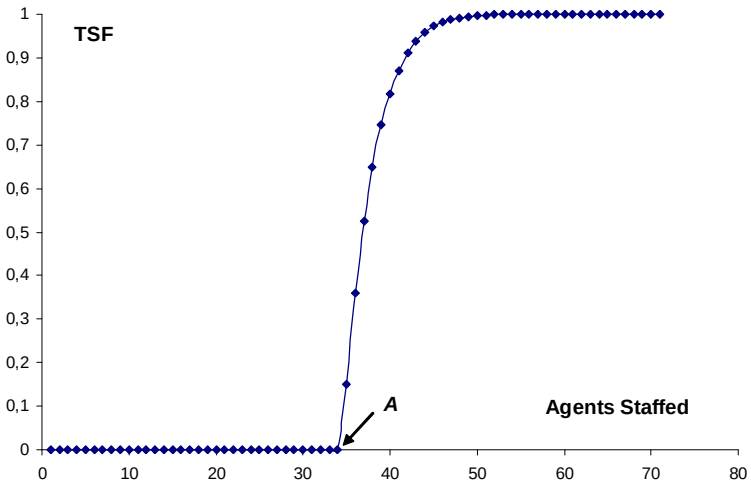

Figure 6.1: Example of a TSF curve

6.3 Two-Stage Stochastic Programming Models

The models we consider in this chapter attempt to optimize the staffing and scheduling costs subject to satisfying a predefined global service level. Given uncertainty in the mean arrival rates, we allow the achieved global SL to be less than the target (shortfall), but we do oblige the expected shortfall to not exceed a certain limit. We use two methods to approximate the non-liner TSF curves: A piece-wise linear approximation and a simple linear approximation. The related two models are presented in Section 6.3.1 and 6.3.2, respectively.

Both models are formulated as two-stage mixed integer stochastic programs. At the beginning of the day, the initial (first-stage) staffing policy X_j, for $j \in J$, is applied. After that the true busyness factor has been revealed as θ_l with probability p_l, for $l \in L$, the decision-maker then chooses the associated adjustable staffing (second-stage) decisions Y_{jl} and Z_{jl}, for $j \in J$. In Figure 6.2, we explain this two-stage staffing process.

6.3.1 Piece-Wise Linear Approximation Model

In this section, we describe the model with piece-wise linear approximation of the TSF curves. As shown in Figure 6.3, we choose the service level in the left side of point A to be zero, and the right side concave curve is approximated by piecewise linear functions. This approximation is similar to but more accurate than that in Robbins (2007). In Figure 6.3, the straight lines represent the individual linear functions. The extreme case with an infinite number of linear functions leads to an exact model.

In what follows, we describe the notations of the sets, the parameters and the variables.

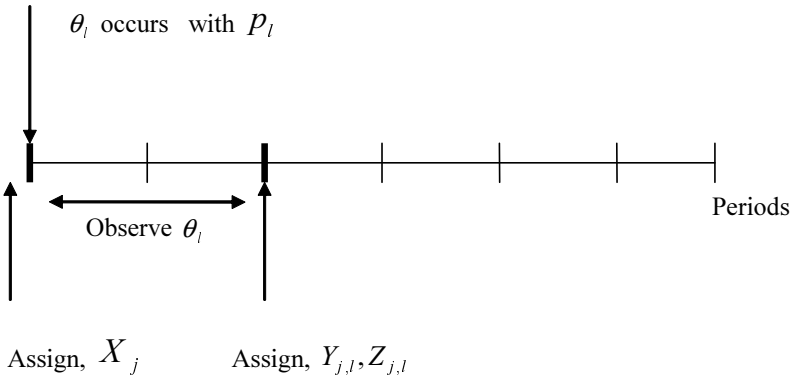

Figure 6.2: Two-stage staffing process

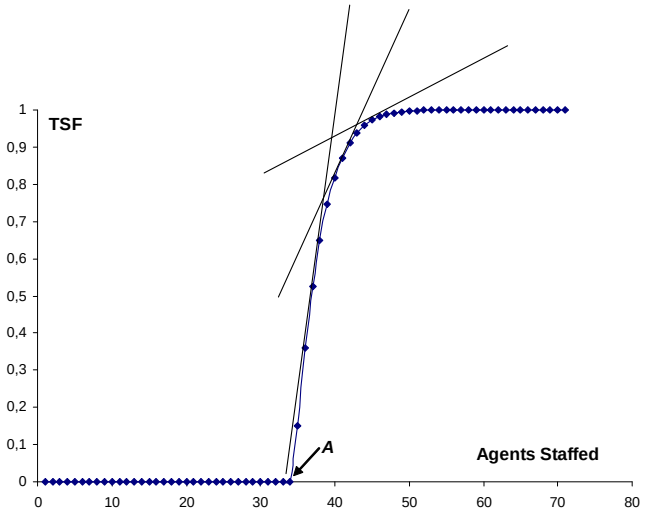

Figure 6.3: Piecewise approximation of TSF

Sets

I: set of time periods in the whole horizon (one day),

I_1: set of the early horizon periods, which contains the time periods before updating,

J: set of schedules,

L: set of sample scenarios.

Decision variables

x_j: number of agents assigned to schedule j, for $j \in J$,

y_{jl}: number of agents added to schedule j after updating in scenario l, for $j \in J$ and $l \in L$,

z_{jl}: number of agents reduced from schedule j after updating in scenario l, for $j \in J$ and $l \in L$.

Deterministic parameters

c_j: cost of schedule j, for $j \in J$,

d_j: the cost of adding an agent to schedule j, for $j \in J$,

r_j: the cost saving by removing an agent from scheduling j, for $j \in J$,

a_{ij}: indicates if schedules j is staffed at period i or not, for $i \in I$ and $j \in J$,

b_{ij}: indicates if schedules j is staffed at period i or not after updating, for $i \in I$ and $j \in J$. Note that $b_{ij} = 0$ for $i \in I_1$ and $j \in J$,

ρ_i: weight of period i, $\rho_i = \dfrac{f_i}{\sum_{i \in I} f_i}$, for $i \in I$,

g: global SL target for the whole day,

p_l: probability associated with scenario l, for $l \in L$,

m_{ilh}: slope of piecewise TSF approximation h at period i in scenario l, for $i \in I, l \in L$ and $h \in H$,

e_{ilh}: intercept of piecewise TSF approximation h at period i in scenario l, for $i \in I, l \in L$ and $h \in H$,

Δ_{il}: the minimum number of agents which leads to non-zero SL, at period i in scenario l, for $i \in I$ and $l \in L$,

State variables

s_l: global SL shortfall in scenario l, for $l \in L$,

Y_{il}: number of employees treating calls at period i in scenario l, for $i \in I$ and $l \in L$,

B_{il}: is equal to 1, if the number of employees treating calls Y_{il} exceeds Δ_{il}, at period i in scenario l, $i \in I, l \in L$. If not, it is equal to 0.

V_{il}: the positive part of $Y_{il} - \Delta_{il}$, at period i of scenario l, for $i \in I$ and $l \in L$,

ν_{il}: percentage of customers served within the acceptable waiting time (example 20 seconds). In another words, it is the achieved SL at period i in scenario l, for $i \in I$ and $l \in L$.

Based on the above notations, our problem can be formulated as

$$\min \quad \sum_{j\in J} c_j\, x_j + \sum_{j\in J, l\in L} p_l\,(d_j\, y_{jl} + r_j\, z_{jl}) \tag{6.2}$$

s.t.

$$\forall\, i,l, \quad Y_{il} = \sum_{j\in J} a_{ij}\, x_j + \sum_{j\in J} b_{ij}\,(y_{jl} - z_{jl}), \tag{6.3}$$

$$\forall\, i,l, \quad Y_{il} \geq \Delta_{il}\, B_{il} + V_{il}, \tag{6.4}$$

$$\forall\, i,l, \quad V_{il} \leq 1000\, B_{il}, \tag{6.5}$$

$$\forall\, i,l, \quad \nu_{il} \leq B_{il}, \tag{6.6}$$

$$\forall\, i,l,h \quad \nu_{il} \leq m_{ilh}\, V_{il} + e_{ilh}, \tag{6.7}$$

$$\forall\, l, \quad \sum_{i\in I} \rho_i\, \nu_{il} + s_l \geq g, \tag{6.8}$$

$$\sum_{l\in L} p_l\, s_l \leq \bar{s}, \tag{6.9}$$

$$\forall\, j,l, \quad x_j \geq z_{jl}, \tag{6.10}$$

$$\forall\, j,l, \quad x_j, y_{jl}, z_{jl} \geq 0, \tag{6.11}$$

$$\forall\, i,l, \quad Y_{il} \in \mathbb{Z}^+, B_{il} \in \{0,1\}, V_{il} \geq 0, \nu_{il} \geq 0. \tag{6.12}$$

The objective of this model is to minimize the total cost of staffing: the initial scheduling and the update adjustment. The optimization occurs over a set of $|L|$ sample realizations of mean call arrival rates. These samples are called scenarios. Constraints (6.3) define the state variables Y_{il} as the number of agents assigned to treating calls at period i in scenario l. Constraints (6.4) and (6.5) define two important state variables, B_{il} and V_{il}. They are related to the staffing level Y_{il} and the parameter Δ_{il} which indicates the smallest staffing level leading to non-zero SL. These two constraints define the binary variables B_{il} to indicate whether $Y_{il} \geq \Delta_{il}$ or not, and the variable V_{il} which is $(Y_{il} - \Delta_{il})^+$. Since all parameters m_{ilh} and e_{ilh} are positive, in order to maximize the periodic SL ν_{il}, the variable V_{il} tends to take the biggest possible value which is less than or equal to $(Y_{il} - \Delta_{il})^+$. Constraints (6.6) and (6.7) define the variable ν_{il} as the periodic SL at period i in scenario l. Constraints (6.8) calculate the global SL shortfall, and Constraint (6.9) limits the expect value of the global service shortfall to $\bar{s}$. Constraints (6.10) ensure that the number of agents reduced from schedule j is less than that assigned initially. Constraints (6.11) and (6.12) define the non-negativity and integer conditions for program variables.

This model is similar to those in the previous chapters, but with some extensions. First, the staffing and scheduling steps are combined into only one optimization problem. Second, the model uses a piecewise linear approximation for the TSF curve derived from an Erlang C model.

One important issue that should be discussed for this model is the required computation effort to solve it. Since we consider in this chapter the special case that each shift contains no breaks, the period-shift matrix $\{a_{ij}\}$ and $\{b_{ij}\}$ are totally unimodular. This simplifies the structure of the problem, and the integer variables x_j, y_{jl} and z_{jl} are relaxed to real ones. The problem hence contains $|I| \times |L|$ integer and $|I| \times |L|$ binary variables. However, the structure of Constraints (6.4)-(6.6) makes the solving of the model time consuming. For the numerical experiments, we are then forced to consider only small size problems with restricted numbers of shifts and scenarios.

6.3.2 Linear Approximated Model

>From the one hand, the model presented in the previous section is a very good approximation. From the other hand, solving it is very time consuming. In this section, we propose another appropriate alternative that allows to quickly solve the optimization problem even for large systems, by choosing a simpler approximation for the TSF function.

As shown in Figure 6.4, we draw a line starting from point $B(0,0)$ and ending at point C which is the first point with the maximum service level 1. This gives a simple linear function $y = m'\,x$ with m' as the slope. Similarly to the model (6.2)-(6.12), we consider the SL in the left side of A to be zero, and the SL in right side part of C to be 1. In contrast to Figure 6.3 which uses several linear functions to approximate the concave part of the TSF curve, we use here a single linear function. Its curve, as shown in Figure 6.4, is the solid part of the drawn line (from point D to point C). As one may see, this approximation is rougher than the piecewise linear approximation, but it allows to construct a much simpler model, with much less constraints.

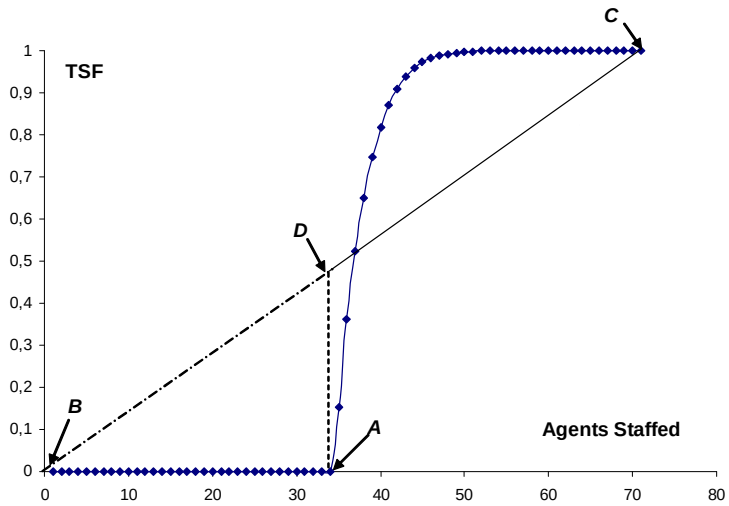

Figure 6.4: Linear approximation of TSF

We denote the slope of the linear function (point B to C in Figure 6.4) by m'_{il} for period i of scenario l. The linear approximation model can be written as

$$\min \quad \sum_{j\in J} c_j\, x_j + \sum_{j\in J, l\in L} p_l\,(d_j\, y_{jl} + r_j\, z_{jl}) \tag{6.13}$$

s.t.

$$\forall\, i,l, \quad Y_{il} = \sum_{j\in J} a_{ij}\, x_j + \sum_{j\in J} b_{ij}\,(y_{jl} - z_{jl}), \tag{6.14}$$

$$\forall\, i,l, \quad Y_{il} \geq \Delta_{il}\, B_{il}, \tag{6.15}$$

$$\forall\, i,l, \quad \nu_{il} \leq B_{il}, \tag{6.16}$$

$$\forall\, i,l, \quad \nu_{il} \leq m'_{il}\, Y_{il}, \tag{6.17}$$

$$\forall\, l, \quad \sum_{i\in I} \rho_i\, \nu_{il} + s_l \geq g, \tag{6.18}$$

$$\sum_{l\in L} p_l\, s_l \leq \bar{s}, \tag{6.19}$$

$$\forall\, j,l, \quad x_j \geq z_{jl}, \tag{6.20}$$

$$\forall\, j,l, \quad x_j, y_{jl}, z_{jl} \geq 0, \tag{6.21}$$

$$\forall\, i,l, \quad Y_{il} \in \mathbb{Z}^+, B_{il} \in \{0,1\}, \nu_{il} \geq 0. \tag{6.22}$$

This model substitutes the previous constraints (6.4-6.7) in the previous model (6.2-6.12) by constraints (6.15-6.17), and keeps all the other constraints and the objective function the same. Constraint (6.15) defines the binary variables which identify whether $Y_{il} \geq \Delta_{il}$, or not. For period i of scenario l, if $Y_{il} < \Delta_{il}$, Constraint (6.16) defines the associated SL to be zero. Otherwise, Constraint (6.16) calculates the approximated SL. Even though the number of integer and binary variables do not decrease in this model compared to the previous one, the structure is much simpler and it has much less constraints related to the integer and binary variables.

6.4 Numerical Implementation for Small Problems

In this section, we conduct a numerical study for small size problems in order to evaluate the two models. In Sections 6.4.1 and 6.4.2, we describe the numerical experiments. In Section 6.4.3, we analyze the results and give some insights.

6.4.1 Parameter Values

Inbound calls: In the experiments we analyze a small size call center. The day starts at 7:00 am, finishes at 6:00 pm, and is divided into $n = 5$ periods of two hours each. The first 2 periods are considered as early horizon. Similarly to the characteristics of the arrival rates described in

the previous chapters, although there is a significant stochastic variability in the arrival rates from one day to another, a strong seasonal pattern across the periods of a given day is preserved.

We set the seasonal factors f_i as $5.475, 16.475, 10.475, 13.35$ and 6.315 calls per minutes. Concerning the random busyness factor Θ, in order to construct a plausible distribution, we choose to discretize a continuous distribution. Similarly to Chapter 5, we assume that Θ follows a gamma distribution with shape parameter $\gamma = 2$ and scale parameter 1, and can take values in $[0.2, 4.0]$. We consider 20 equidistant points including the two endpoints 0.20 and 4.0, which gives $|L| = 20$ possible values of θ_l.

For the piecewise linear approximation model, for the TSF curve of each period and each scenario, we use a linear function to approach each point which achieves an SL higher than zero but less than 1. We use for each period and each scenario $|H| = 62$ constraints.

The mean service time is assumed to be $\frac{1}{\mu} = 5$ minutes, and the acceptable waiting time is 20 seconds. The target global service level g is determined as 0.8, which indicates that 80% of calls during the day is replied with 20 seconds. The upper bound of the expect shortfall of global service level $\bar{s}$ is 0.1.

Cost parameters: An agent can work either 3, 4 or 5 periods during the day (with no breaks). A shift can start at the beginning of any period, enumeration shows that there are 6 feasible schedules. For the salary c_j, d_j and r_j, without loss of generality, we use a normalized cost of 1 for each period an agents works in regular shifts, therefore $c_j = \sum_{i \in I} a_{ij}$. And the temporarily added shift should be payed more expensively than the regular one, we then define $d_j = 1.4 \times \sum_{i \in I} b_{ij}$. Removing an agent from shift j makes some cost saving, but this cost saving should be less than c_j. We choose $r_j = 0.5 \times \sum_{i \in I} b_{ij}$.

6.4.2 Design of the Experiments

Benchmark1: As a first benchmark, we consider the static single-stage model which contains only the initial staffing-scheduling decision variables. The static single-stage stochastic optimization problem is also built on the discrete probability distributions characterizing Θ. The difference between this static single-stage model and the model (6.2-6.12) is that in the static single-stage model, only the initial decision x_j determines the staffing level, and the decision-maker does not have the flexibility to update the staffing level later on during the day. Conse-

quently the staffing level Y_{il} for each scenario is the same.

$$\min \quad \sum_{j\in J} c_j\, x_j \tag{6.23}$$

s.t.

$$\forall\, i,l, \quad Y_{il} = \sum_{j\in J} a_{ij}\, x_j, \tag{6.24}$$

$$\forall\, i,l, \quad Y_{il} \geq \Delta_{il}\, B_{il} + V_{il}, \tag{6.25}$$

$$\forall\, i,l, \quad V_{il} \leq 1000\, B_{il}, \tag{6.26}$$

$$\forall\, i,l, \quad \nu_{il} \leq B_{il}, \tag{6.27}$$

$$\forall\, i,l,h \quad \nu_{il} \leq m_{ilh}\, Y_{il} + e_{ilh}, \tag{6.28}$$

$$\forall\, l, \quad \sum_{i\in I} \rho_i\, \nu_{il} + s_l \geq g, \tag{6.29}$$

$$\sum_{l\in L} p_l\, s_l \leq \bar{s}, \tag{6.30}$$

$$\forall\, j, \quad x_j \geq 0, \tag{6.31}$$

$$\forall\, i,l, \quad Y_{il} \in \mathbb{Z}^+, B_{il} \in \{0,1\}, V_{il} \geq 0. \tag{6.32}$$

All the parameters and variables keep the same definitions as those in the model (6.2-6.12). We consider this static single-stage model in order to compare its total cost and SL performance to those of the dynamic model (6.2-6.12), and analyze the benefits of the flexibility offered by update.

Benchmark2: Another benchmark we consider is the model which is similar to that in Chapter 4, where a target service level is required for each period, instead of a global one for the whole day. This model keeps the flexibility of update, which leads to a two-stage stochastic programming optimization problem. We also build this problem on the discrete probability distribution characterizing Θ. For period i of scenario l, N_{il} denotes the required number of agents in order to achieve that target SL, calculated through the Erlang C formula. A shortfall of agents M_{il} is allowed for each period and scenario, but the expected sum of shortfall during the whole day should not exceed the limit $\bar{M}$.

In order to be comparable to the models presented above, N_{il} is calculated with a target SL as 0.8 for each period. $\bar{M}$ is defined as the difference between $\sum_{i\in I}\sum_{l\in L} p_l\, N_{il}$ with target SL

as 0.8 and that with target SL as 0.7. The value of $\bar{M}$ in this small problem is 9.

$$\min \quad \sum_{j\in J} c_j\, x_j + \sum_{j\in J, l\in L} p_l\,(d_j\, y_{jl} + r_j\, z_{jl}) \tag{6.33}$$

s.t.

$$\forall\, i,l, \quad Y_{il} = \sum_{j\in J} a_{ij}\, x_j + \sum_{j\in J} b_{ij}\,(y_{jl} - z_{jl}), \tag{6.34}$$

$$\forall\, i,l, \quad Y_{il} + M_{il} \geq N_{il}, \tag{6.35}$$

$$\sum_{l\in L, i\in I} p_l\, M_{il} \leq \bar{M}, \tag{6.36}$$

$$\forall\, j,l, \quad x_j \geq z_{jl}, \tag{6.37}$$

$$\forall\, j,l, \quad x_j, y_{jl}, z_{jl} \geq 0, \tag{6.38}$$

$$\forall\, i,l, \quad Y_{il} \in \mathbb{Z}^+, M_{il} \geq 0. \tag{6.39}$$

In this model, the objective and the first constraints are similar to those in the model (6.2)-(6.12). Constraints (6.35) calculate the agents number shortfall for each period and scenario, and Constraint (6.36) limits the expect sum of this shortfall. Constraints (6.37) ensure that the number of agents removed from shift j is less than that is initially assigned, and Constraints (6.38)-(6.39) define the non-negativity and integer conditions for program variables.

We compare this model with the model (6.2-6.12) in order to identify the advantage of considering a global service level.

Additional Notations: We compute the optimal staffing levels given by the dynamic piecewise linear approximation model with global service level constraint(DPLG), the dynamic linear approximation model with global service level constraint(DLG), the static single-stage piecewise linear approximation model with global service level constraint(SPLG) and the dynamic model with periodic service level constraint (DP).

Optimal policy performance simulations: In order to estimate the cost criterion associated with the different policies, $10,000$ samples values are randomly generated as outcomes of Θ.

For the three dynamic models (DPLG, DLG and DP), the corresponding optimal policy is selected according to each of the $10,000$ samples. And for the static single-stage model SPLG, each sample implements the same staffing level. We then calculate the expected global SL, and the expected global SL shortfall comparing to the target SL $g = 0.8$, associated with the optimal policies of these four models.

6.4.3 Insights

First of all, we give a report of the size of the problems and their computing time in Table 6.1. The computations have been performed using Cplex on an Intel Core Duo CPU 1.20 Ghz with 0.99 GBytes RAM.

Table 6.1: Computing time and problem size

Approach	Computation time (seconds)	Number of variables
DPLG	73.17	100 integers, 100 binary, 446 continuous
DLG	2.97	100 integers, 100 binary, 346 continuous
SPLG	321	100 integers, 100 binary, 206 continuous
DP	0.06	100 integers, 346 continuous

It is shown that even though the number of integer and binary variables in models DPLG and DLG is unchanged, model DLG requires much less computation time than the former. Both the models DPLG and SPLG are time consuming, due to the complicate constraint structure. The model DP requires little time, the reason is related to the total unimodularity of matrix $\{a_{ij}\}$ and $\{b_{ij}\}$, as discussed in Chapter 5.

Cost and Global SL Comparison

Table 6.2 provides for each model (DPLG, DLG, SPLG and DP), the total cost, the average values of the global SL and the shortfall comparing target global SL $g = 0.8$. Since we use as many as needed piecewise linear functions to approximate the TSF curve, the optimal policy solved by DPLG is the real optimal solution for this small size problem. The DPLG solution is clearly the lower bound on the minimal cost of all the three models with a global service level constraint. And the average global SL and average shortfall on global SL achieved by DPLG is almost what the decision-maker expects to obtain.

Table 6.2: Total cost and SL archived

Approach	Total cost	Expected Global SL (%)	Expected Shortfall (%)
DPLG	591.63	73.53	15.86
DLG	644.22	72.31	13.55
SPLG	735.00	74.10	18.83
DP	769.50	85.36	8.49

Firstly, it is exciting to find out that the performance measures of DLG and DPLG are similar. DLG costs only slightly higher than DPLG on the total cost. And the expected global SLs of

both models fall in the target global SL interval $[0.7, 0.8]$. Even more, the expected shortfall on global SL of DLG is even slightly smaller than that of DPLG, and approaches more to the limit $\bar{s} = 0.1$. Sum up the performance and the computational complexity, we conclude that the model DLG is a good approach to the exact model DPLG, the former can be used to replace the latter in order to solve larger size problems.

The gap between the total cost of DPLG and SPLG is remarkable. SPLG gets an expected global SL similar to that of DPLG, by paying 24% additional total cost. Moreover, the expected shortfall on global SL of SPLG is much bigger than the limit $\bar{s} = 0.1$. Recall that the difference between the models DPLG and SPLG is that the former takes advantages of the flexibility of the recourse action. Numerical results show that the flexibility allowing staffing level adjustment during the day is favorable.

Lastly, we compare between the performance measures of model DPLG and DP, which considers the soft and hard service level constraints, respectively. We find out that the total cost of model DP is 30% higher than that of DPLG, and the expected global service level is much higher than required. This shows that the model with hard service level constraints will lead to staffing level overcapacity, the service level is higher than required, and the total cost is much higher than necessary. If the decision-maker evaluates the performance of a call center by a global view, it is then more favorable to use models with global service constraint to optimize staffing and scheduling policies.

6.5 Numerical Implementation for Large Problems

In this section, we conduct a numerical study with larger problem sizes, in the the same order as those analyzed in the previous chapters.

6.5.1 Experiments

Inbound calls: We consider a case where the day starts at 8:00 am, finishes at 9:00 pm, and is divided into $|I| = 11$ periods of one hour each. The first 3 periods are considered as early horizon and the update actions take place at the beginning of the 4^{th} period.

The seasonal factors f_i are 3.5, 18.4, 34.4, 31.5, 29,12.9, 28.4, 25, 17.4, 7.2 and 5.3 calls per minute. For the random busyness factor Θ, we also discretize the gamma distribution with shape parameter $\gamma = 2$ and scale parameter 1. We assume that Θ can take values in $[0.2, 6.0]$. We consider 30 equidistant points including the two endpoints 0.20 and 6.0, which give $|L| = 30$ possible values for θ_l. The mean service time is again 5 minutes, and the acceptable waiting time is 20 seconds. In order to observe the performance associated with varying global service level,

here we set the target global service level g to 0.9. The upper bound of the expect shortfall of global service level $\bar{s}$ is 0.1.

For the model DP, $\bar{M}$ is defined as the difference between $\sum_{i\in I}\sum_{l\in L} p_l\, N_{il}$ with target SL = 0.9 and that with target SL = 0.8. The value of $\bar{M}$ in this large size problem is 41.

Cost parameters: An agent can work between 7 or 8 hour periods in a day, without intermediate breaks. A shift can start at the beginning of any period. We have 9 feasible schedules. The definition of the salary costs c_j, d_j and r_j are the same as those in Section 6.4: $c_j = \sum_{i\in I} a_{ij}$, $d_j = 1.4 \times \sum_{i\in I} b_{ij}$, and $r_j = 0.5 \times \sum_{i\in I} b_{ij}$.

Design of the Experiments: This larger problem size contains $|L| = 30$ scenarios and $|I| = 11$. The models DPLG and SPLG become intractable due to computational complexity. After the comparison between the performance of the models DPLG and DLG in Section 6.4.3, we find that the model DLG can be a good approximation to the model DPLG, and gains the feasibility for larger size problems.

During this numeric implementation, we thus compare between the performance of the models DLG, DP, and the static single-stage version of model DLG:

$$\min \quad \sum_{j\in J} c_j\, x_j \tag{6.40}$$

s.t.

$$\forall\, i,l, \quad Y_{il} = \sum_{j\in J} a_{ij}\, x_j, \tag{6.41}$$

$$\forall\, i,l, \quad Y_{il} \geq \Delta_{il}\, B_{il}, \tag{6.42}$$

$$\forall\, i,l, \quad \nu_{il} \leq B_{il}, \tag{6.43}$$

$$\forall\, i,l, \quad \nu_{il} \leq m'_{il}\, Y_{il}, \tag{6.44}$$

$$\forall\, l, \quad \sum_{i\in I} \rho_i\, \nu_{il} + s_l \geq g, \tag{6.45}$$

$$\sum_{l\in L} p_l\, s_l \leq \bar{s}, \tag{6.46}$$

$$\forall\, j,l, \quad x_j \geq 0, \tag{6.47}$$

$$\forall\, i,l, \quad Y_{il} \in \mathbb{Z}^+, B_{il} \in \{0,1\}, \nu_{il} \geq 0. \tag{6.48}$$

This model differs from the model DLG only by the fact that it is static without update and does not allow the real time update. It is referred to as SLG. We do the same experiments as those in Section 6.4 in order to obtain further insights on the performance of optimal policies of different models.

6.5.2 Insights

This problem is larger than that in Section 6.4. It contains $|I| = 11$ periods, $|L| = 30$ scenarios and $|J| = 9$ shifts. We list the problem size of the models DLG, SLG and DP in Table 6.3.

Table 6.3: Computing time and problem size

Approach	Computation time (seconds)	Number of variables
DLG	127	330 integers, 330 binary, 879 continuous
SLG	6711.58	330 integers, 330 binary, 339 continuous
DP	0.30	330 integers, 879 continuous

It is shown that all the three models could be solved within acceptable time duration. Table 6.4 shows for each model (DLG, SLG and DP), the total cost, the average values of the global service level and the shortfall from the target global service level $g = 0.9$. Recall that the allowed

Table 6.4: Total cost and SL archived

Approach	Total cost	Expected Global SL (%)	Expected Shortfall (%)
DLG	2746.40	81.23	10.65
SLG	3914.00	80.94	11.08
DP	3682.78	94.36	2.97

expected global SL shortfall is limited to $\bar{s} = 0.1$. It is shown in Table 6.4 that the expected shortfall of the global SL of DLG approaches very closely this limit. The optimal solution of the model DLG also achieves an average global SL as it is expected. This confirms the good quality of model DLG as an approximation of model DPLG.

The expected value and the expected shortfall of global SL associated with model SLG are very close to those of DLG. The only difference is that SLG costs much more than DLG. This again confirms the benefit of the flexibility of allowing staffing adjustment. Comparing between the performance of the models DLG and DP, we obtain similar remarks and conclusions as those in Section 6.4: the hard constraints formulation leads to overcapacity on staffing level, global service level higher than required and unnecessary total cost.

6.6 Concluding Remarks

We have developed a multi-period multi-shift call center problem with staffing level update and global service level. We focused on optimizing the staffing level w.r.t. the operating cost of the call center, under a global service level constraint.

We modeled our problem as a cost optimization-based model. We proposed two models. The first used piecewise linear approximations to approach the non-linear TSF curve. The second model used a single linear function to roughly approach the TSF curve. The first model is limited by the size of the problems. However, the second model could solve problems with large sizes and gives at the same time appropriate results.

We compare between a model with staffing level update and global service level constraint, a static single-stage model with global service level constraint, and a dynamic model with period by period service level constraints. The advantages of considering a global service level constraint and allowing staffing level dynamic adjustments are numerically analyzed.

Chapter 7

Conclusion and Perspectives

In this chapter, we give general concluding remarks and present directions for future research. For further details, we refer the reader to the concluding sections of the previous chapters.

7.1 Conclusions

A call center, or in general a contact center, is defined as a service system in which agents serve customers, over telephone, fax, email, etc. Over the past few years, call centers have emerged as an essential component of the customer relationship management strategy for many large companies. This customer service has become a key issue to attracting and maintaining companies market shares. As a consequence, call centers performance indices, as typical customer waiting time, are considered now as important assets to be optimized, in particular through efficient workforce management of skilled operators. This thesis deals with the operation management of call centers.

For this service sector, the staffing cost is a major component in the operating costs. Call center scheduling aims to set-up the workforce so as to meet target service levels, and minimize the staffing cost at the same time. The service level depends on the mean rate of arrival calls, which fluctuates during the day and from day to day. This thesis focuses on the optimization of staffing-scheduling problem of call centers in a stochastic setting, where the mean rates of arrival calls are assumed to be uncertain. This subject is at the same time scientifically interesting and practically relevant.

We investigate the impact of uncertainty on the capacity management decision and develop models that explicitly incorporate uncertainty in the staffing planning process. The arrival process of calls is modelled by a doubly non-stationary stochastic process, with random mean arrival rates related to a random business factor of the day. We have considered the staffing-scheduling problem based on four cost optimization-based models:

1. Blending Single Shift Scheduling Model, in which there exists some flexibility to modify in real-time (within the same day) the affectation of capacity between calls and emails.

2. Multi Shift Scheduling Model with Recourse which decides an initial schedule before the beginning of the working day and allows for real-time recourse actions to adjust the initially scheduled staffing levels in reaction to the realized deviations from arrival-rate forecasts.

3. Distributionally Robust Optimization Model, where the probability distribution of the random parameter is ambiguous and belongs to some probability distribution set. We propose an approach combining stochastic programming and distributionally robust optimization to construct this model and solve the problem.

4. Global Service Level Model, with possibility to adjust the staffing level during the day, and the target is a global service level for the whole day (instead of a target per period).

We have proposed some approaches such as stochastic programming (with recourse), (two-stage) robust programming, distributionally robust programming, in order to efficiently address the above problems and gain useful insights.

7.2 Future Research

The research in this dissertation can of course be extended and expanded. As detailed in the conclusion remarks in this chapter, we addressed specific extensions of each model. We point out some of the key areas for potential future research in what follows.

More General Shifts: This dissertation considers only the case where the shifts are without breaks. One extension of this work is to take our results as the fist step, and place breaks within shifts using heuristical methods.

Multi-types of Calls: In this dissertation we consider only single type of calls. In practice a call centre is often a multi-skill environment which often contains multi-types of calls and multi-skilled agents. The workflows are often very complex dues to skills based routing. The models in this dissertation cold be extended to address multi-skilled call routing.

Abandonment: We use the model Erlang C which does not consider the abandonment of calls to calculate the required agent number and construct the TSF curve in this dissertation. Further development to incorporate customer impatience (abandonment) as that in the Erlang A system is worth considering.

Uncertainty on Absenteeism: The uncertainty we take into account in this dissertation concerns only on the randomness of the arrival rates. In practice, another source of uncertainty is the absenteeism of agents which highly affects the efficiency of the before-hand planned agents in order to meet the quality of service constraints. It is of interest to take into account this type of uncertainty in the extended model.

Uncertainty on Seasonal Factors: For the distributionally robust model we developed, we have considered the uncertainty set of probability distribution of the busyness factor Θ. It may be of interest to consider additionally un uncertainty set of probability distribution on uncertain seasonal factors f_i.

Agent shortfall penalty with weights: We have treated the agent shortfall in each period in a similar way. However, for periods with different numbers of required agents, the same quantity of agent shortfall may lead the customers to experience different additional waiting time. Consequently, a given agent shortfall in different period may of different importance. Extended formulation could introduce weights per period that depend on the values of required agent number.

Variable Time Horizons: The total time horizon we consider in this dissertation is basically one day, and the SL requirement is based on one period of several hours or on a single day as global SL. An extension of this work would analyze the situation where the service level is evaluated over a longer time such as a week or a month.

Appendix A

Appendix of Chapter 3

This appendix deals with the analysis of Chapter 3. In Appendix A.1 and A.2, we give the proof of the Theorem 3.1 and Proposition 3.1 in Section 3.2.3 and Section 3.3.2, respectively. We present in Appendix A.3 the mixed robust model mentioned in Section 3.3.2. Finally, Appendix A.4 provides numeric supports for the analysis in Section 3.4.1

A.1 Proof of Theorem 3.1

Recall the definition of Theorem 3.1 as follows: *The expected daily total cost function $C(y)$ is convex in y.*

We assume that $C(y)$ is a continuous function over $y \in \mathbb{R}^+$, Θ and W are continuous random variables. It is clear that proving the convexity in the continuous case implies proving it for the original discrete case. We denote by $f_\Theta(.)$ and $f_W(.)$ ($F_\Theta(.)$ and $F_W(.)$) the probability density functions (the cumulative probability distribution functions) of the random variables Θ and W, respectively. For a given outcome of Θ, denoted by θ, we use $v_i(\theta)$ to denote the required number of agents to handle the calls in period i. And V_i denotes the underlying random number of agents required to handle calls in period i. The continuous version of the total cost, given in Equation (3.4) becomes

$$C(y) = n\,c\,y + u_\alpha \sum_{i=1}^{n} \int_{\theta_i^*(y)}^{\infty} (v_i(\theta) - y) f_\Theta(\theta)\, d\theta + r \int_{Q(y)}^{\infty} (x - Q(y)) f_W(x)\, dx, \tag{A.1}$$

where

$$Q(y) = E[\sum_{i=1}^{n} (y - V_i)^+] = \sum_{i=1}^{n} \int_0^{\theta_i^*(y)} (y - v_i(\theta)) f_\Theta(\theta)\, d\theta, \tag{A.2}$$

$$\theta_i^*(y) = \min\{\theta : v_i(\theta) \geq y\}, \quad i = 1, .., n. \tag{A.3}$$

Proving the convexity of the $C(\cdot)$ function is equivalent to prove that $\frac{d^2C(y)}{dy^2} \geq 0$ for $y \in \mathbb{R}^+$. Applying Leibniz formula, we have

$$\frac{dQ(y)}{dy} = \sum_{i=1}^{n} \int_0^{\theta_i^*(y)} f_\Theta(\theta)\, d\theta = \sum_{i=1}^{n} F_\Theta(\theta_i^*(y)). \tag{A.4}$$

Combining now Equations (A.1) and (A.4), we obtain

$$\begin{aligned}
\frac{dC(y)}{dy} &= nc - u_\alpha \sum_{i=1}^{n} \int_{\theta_i^*(y)}^{\infty} f_\Theta(\theta) d\theta + r \int_{Q(y)}^{\infty} \frac{\partial\Big(x - Q(y)\Big)}{\partial y} f_W(x) dx & \text{(A.5)}\\
&= nc + u_\alpha \left(\sum_{i=1}^{n} F_\Theta(\theta_i^*(y)) - n \right) - r \int_{Q(y)}^{\infty} \sum_{i=1}^{n} F_\Theta(\theta_i^*(y))\, f_W(x) dx & \text{(A.6)}\\
&= n(c - u_\alpha) + \Big(u_\alpha - r\left(1 - F_W(Q(y))\right)\Big) \sum_{i=1}^{n} F_\Theta(\theta_i^*(y)). & \text{(A.7)}
\end{aligned}$$

We have

$$\frac{d^2C(y)}{dy^2} = \frac{d}{dy}\left(\Big(u_\alpha - r\left(1 - F_W(Q(y))\right)\Big) \cdot \sum_{i=1}^{n} F_\Theta(\theta_i^*(y)) \right). \tag{A.8}$$

Since for $i = 1, ..., n$, $F_\Theta(\cdot) \geq 0$ and $F'_\Theta(\cdot) \geq 0$, $F_W(\cdot) \geq 0$ and $F'_W(\cdot) \geq 0$ and by assumption $r < u_\alpha$ (see Section 3.2), we have

$$u_\alpha - r\Big(1 - F_W(Q(y))\Big) \geq 0, \tag{A.9}$$

and

$$\frac{d}{dy}\Big(u_\alpha - r\left(1 - F_W(Q(y))\right)\Big) = r\frac{dF_W(Q(y))}{dy} \geq 0. \tag{A.10}$$

We thereafter conclude that $\frac{d^2C(y)}{dy^2} \geq 0$, which finishes the proof of the theorem.

□

A.2 Proof of Proposition 3.1

We recall the Proposition 3.1 as follows: *Let $C^*(\theta, w)$ be the optimal objective value of the problem defined in (3.19)-(3.23). For $\delta > 0$, we have the following inequalities,*

$$C^*(\theta + \delta, w) \geq C^*(\theta, w), \tag{A.11}$$

$$C^*(\theta, w + \delta) \geq C^*(\theta, w). \tag{A.12}$$

Recall that $C(y, \theta, w)$ is the cost associated with a given staffing level y, a business level θ and a back-office workload w as defined in Equation (3.5). For given sample values θ and w, we denote the optimal solution of problem (3.19)-(3.23) as $y^*_{\theta,w}$. Furthermore, for a given staffing level y, and sample values θ and w, we denote the corresponding variables $M_{\theta,\,w,\,i}(y)$, $M^-_{\theta,\,w,\,i}(y)$, $M^+_{\theta,\,w,\,i}(y)$ and $N_{\theta,\,w}(y)$. We now prove that for $\delta \geq 0$, we have $C(y^*_{\theta+\delta,\,w}, \theta, w) \leq C(y^*_{\theta+\delta,\,w}, \theta + \delta, w)$.

It is straight forward to see that the variables $M^-_{\theta,\,w,\,i}(y)$ and $M^+_{\theta,\,w,\,i}(y)$ can not take strictly positive values, simultaneously. In case of over-staffing (under-staffing) at period i, we have $M^+_{\theta,\,w,\,i}(y) > 0$ and $M^-_{\theta,\,w,\,i}(y) = 0$ ($M^+_{\theta,\,w,\,i}(y) = 0$ and $M^-_{\theta,\,w,\,i}(y) > 0$).

Furthermore, using the Erlang C formula, we have for each period i, $v_i(\theta\, f_i) \leq v_i((\theta + \delta)\, f_i)$. For the staffing level $y^*_{\theta+\delta,\,w}$, we have $M_{\theta,\,w,\,i}(y^*_{\theta+\delta,\,w}) \geq M_{\theta+\delta,\,w,\,i}(y^*_{\theta+\delta,\,w})$, which means $M^+_{\theta,\,w,\,i}(y^*_{\theta+\delta,\,w}) \geq M^+_{\theta+\delta,\,w,\,i}(y^*_{\theta+\delta,\,w})$ and $M^-_{\theta,\,w,\,i}(y^*_{\theta+\delta,\,w}) \leq M^-_{\theta+\delta,\,w,\,i}(y^*_{\theta+\delta,\,w})$. Furthermore, by constraint (3.22), we have $N_{\theta,\,w}(y^*_{\theta+\delta,\,w}) \leq N_{\theta+\delta,\,w}(y^*_{\theta+\delta,\,w})$. As $n, c, u_\alpha, r > 0$, we easily get $C(y^*_{\theta+\delta,\,w}, \theta, w) \leq C(y^*_{\theta+\delta,\,w}, \theta + \delta, w)$. Since $C^*(\theta, w) = min_{y\in\mathbb{N}} C(y, \theta, w)$, we have $C^*(\theta, w) \leq C(y^*_{\theta+\delta,\,w}, \theta, w)$. Therefore $C^*(\theta, w) \leq C^*(\theta + \delta, w)$, which gives (A.11).

Consider now sample values θ and $w + \delta$, with $\delta > 0$. For a given staffing level y, by constraint (3.20), we have $M_{\theta,\,w,\,i}(y) = M_{\theta,\,(w+\delta),\,i}(y)$, and $N_{\theta,\,w}(y) \leq N_{\theta,\,w+\delta}(y)$. This leads to $C(y^*_{\theta,\,w+\delta}, \theta, w) \leq C(y^*_{\theta,\,w+\delta}, \theta, w + \delta)$. Since $C^*(\theta, w) \leq C(y^*_{\theta,\,w+\delta}, \theta, w)$, we obtain $C^*(\theta, w) \leq C^*(\theta, w + \delta)$, which gives (A.12). □

A.3 Mixed Robust Programming Formulation

Here we give a formulation mixing stochastic and robust programming (see the end of Section 3.3.2). With an uncertainty set for Θ defined as

$$U' = \{\theta : 0 \leq \theta \leq \overline{\theta} + \eta\, \sigma_\theta, \text{ with } \eta \geq 0\} \tag{A.13}$$

and with the random back-office workload process described as in Section 3.2.2, a mixed robust programming formulation can be given as follows.

$$Min \quad nc\,y + u_\alpha \sum_{i=1}^{n} M_i^- + r \sum_{k=1}^{K} p_{w_k} N_k \tag{A.14}$$

$$\text{s.t.} \quad M_i = y - v_i((\theta + \eta\,\sigma_\theta) f_i), \qquad \text{with} \;\; i = 1, ..., n, \tag{A.15}$$

$$M_i = M_i^+ - M_i^-, \qquad \text{with} \;\; i = 1, ..., n, \tag{A.16}$$

$$N_k \geq w_k - \sum_{i=1}^{n} M_{i,\,l}^+, \qquad \text{with} \;\; k = 1, ..., K, \tag{A.17}$$

$$y, M_i^+, M_i^-, N_k \geq 0, \qquad \text{with} \;\; i = 1, ..., n, k = 1, ..., K. \tag{A.18}$$

In this problem, M_i represents the difference between the staffing level and the required agent number in period i for the highest arrival rate in the considered uncertainty set, $(\theta + \eta\,\sigma_\theta) f_i$. N_k is the over-time workload required in order to finish back-office jobs in scenario k.

A.4 Additional Numerical Results

In Tables A.1 and A.2, we give additional support for the numerical analysis of Section 3.4.1.

Table A.1: $E[\Theta] = 1$ and $\sigma_\Theta = 0.21$; $E[W] = 600$ and $\sigma_W = 60$

		Optimal staff y^*	Total Cost Average	Total Cost STD.	Salary cost	Under-staffing cost Average	Under-staffing cost STD.	Overtime cost Average	Overtime cost STD.	Constr. violation Pct.
	PI	—	29842.15	5698.79	28006.24	1830.28	819.21	5.64	23.82	8.23
	DA	167	35096.13	11161.08	27555.00	7204.60	10567.79	336.54	781.65	17.22
	SP	184	34016.92	7248.73	30360.00	3616.24	7111.36	40.69	244.55	10.08
	RP									
$\alpha = 10\%$	$\eta = 0.1$	170	34726.28	10404.54	28050.00	6430.49	9919.38	245.79	660.63	15.81
$u_\alpha = 140$	$\eta = 0.5$	184	34016.92	7248.73	30360.00	3616.24	7111.36	40.69	244.55	10.08
	$\eta = 1.0$	201	34785.34	4386.80	33165.00	1618.62	4375.90	1.72	40.21	5.20
	$\eta = 2.0$	234	38850.14	1334.13	38610.00	240.14	1334.13	0.00	0.00	0.99
	$\eta = 3.0$	268	44239.28	268.87	44220.00	19.28	268.87	0.00	0.00	0.10
	MxRP									
	$\eta = 0.1$	170	34726.28	10404.54	28050.00	6430.49	9919.38	245.79	660.63	15.81
	$\eta = 0.5$	184	34016.92	7248.73	30360.00	3616.24	7111.36	40.69	244.55	10.08
	$\eta = 1.0$	201	34785.34	4386.80	33165.00	1618.62	4375.90	1.72	40.21	5.20
	$\eta = 2.0$	234	38850.14	1334.13	38610.00	240.14	1334.13	0.00	0.00	0.99
	$\eta = 3.0$	268	44239.28	268.87	44220.00	19.28	268.87	0.00	0.00	0.10
	PI	—	30169.06	5827.46	30159.40	0.00	0.00	9.66	36.60	0.00
	DA	182	38535.26	16209.41	30030.00	8450.79	16040.49	54.47	288.14	10.81
	SP	202	36628.06	9097.57	33330.00	3296.72	9088.74	1.34	35.04	4.98
	RP									
$\alpha = 5\%$	$\eta = 0.1$	186	37814.84	14566.28	30690.00	7094.84	14458.39	30.01	205.77	9.40
$u_\alpha = 300$	$\eta = 0.5$	200	36650.35	9684.86	33000.00	3648.18	9671.77	2.17	45.95	5.45
	$\eta = 1.0$	219	37436.04	5108.90	36135.00	1301.04	5108.90	0.00	0.00	2.23
	$\eta = 2.0$	255	42191.39	1116.90	42075.00	116.39	1116.90	0.00	0.00	0.26
	$\eta = 3.0$	292	48183.71	124.35	48180.00	3.71	124.35	0.00	0.00	0.01
	MxRP									
	$\eta = 0.1$	186	37814.84	14566.28	30690.00	7094.84	14458.39	30.01	205.77	9.40
	$\eta = 0.5$	200	36650.35	9684.86	33000.00	3648.18	9671.77	2.17	45.95	5.45
	$\eta = 1.0$	219	37436.04	5108.90	36135.00	1301.04	5108.90	0.00	0.00	2.23
	$\eta = 2.0$	255	42191.39	1116.90	42075.00	116.39	1116.90	0.00	0.00	0.26
	$\eta = 3.0$	292	48183.71	124.35	48180.00	3.71	124.35	0.00	0.00	0.01
	PI	—	30169.06	5827.46	30159.40	0.00	0.00	9.66	36.60	0.00
	DA	182	71634.19	79033.33	30030.00	41549.72	78865.76	54.47	288.14	10.81
	SP	233	41147.64	14645.96	38445.00	2702.64	14645.96	0.00	0.00	1.06
	RP									
$\alpha = 1\%$	$\eta = 0.1$	186	65602.95	71194.13	30690.00	34882.94	71087.09	30.01	205.77	9.40
$u_\alpha = 1475$	$\eta = 0.5$	200	50939.05	47565.89	33000.00	17936.89	47552.87	2.17	45.95	5.45
	$\eta = 1.0$	219	42531.78	25118.77	36135.00	6396.78	25118.77	0.00	0.00	2.23
	$\eta = 2.0$	255	42647.23	5491.41	42075.00	572.23	5491.41	0.00	0.00	0.26
	$\eta = 3.0$	292	48198.22	611.38	48180.00	18.22	611.38	0.00	0.00	0.01
	MxRP									
	$\eta = 0.1$	186	65602.95	71194.13	30690.00	34882.94	71087.09	30.01	205.77	9.40
	$\eta = 0.5$	200	50939.05	47565.89	33000.00	17936.89	47552.87	2.17	45.95	5.45
	$\eta = 1.0$	219	42531.78	25118.77	36135.00	6396.78	25118.77	0.00	0.00	2.23
	$\eta = 2.0$	255	42647.23	5491.41	42075.00	572.23	5491.41	0.00	0.00	0.26
	$\eta = 3.0$	292	48198.22	611.38	48180.00	18.22	611.38	0.00	0.00	0.01

Table A.2: $E[\Theta] = 1$ and $\sigma_\Theta = 0.21$; $E[W] = 1000$ and $\sigma_W = 100$

			Total Cost		Salary cost	Under-staffing cost		Overtime cost		Constr. violation
		Optimal staff y^*	Average	STD.		Average	STD.	Average	STD.	Pct.
	PI	—	32380.01	3803.42	32229.51	93.24	269.54	57.26	52.52	2.47
	DA	195	34421.47	3306.33	32175.00	467.01	1124.20	1779.45	2434.60	6.72
	SP	184	34179.90	4181.90	30360.00	774.91	1523.86	3044.99	3007.52	10.08
	RP									
$\alpha = 10\%$	$\eta = 0.1$	198	34573.05	3055.65	32670.00	403.20	1028.04	1499.85	2256.24	5.94
$u_\alpha = 30$	$\eta = 0.5$	210	35543.77	2081.96	34650.00	216.09	700.19	677.68	1529.68	3.49
	$\eta = 1.0$	225	37409.53	1101.63	37125.00	91.08	408.53	193.45	775.65	1.64
	$\eta = 2.0$	255	42090.97	183.05	42075.00	11.64	111.69	4.33	97.67	0.26
	$\eta = 3.0$	284	46860.91	22.09	46860.00	0.90	21.93	0.01	1.05	0.03
	MxRP									
	$\eta = 0.1$	192	34306.34	3554.89	31680.00	538.98	1225.90	2087.36	2606.67	7.56
	$\eta = 0.5$	200	34695.45	2888.18	33000.00	364.82	967.18	1330.63	2134.48	5.45
	$\eta = 1.0$	212	35756.59	1932.22	34980.00	193.57	654.10	583.02	1415.12	3.17
	$\eta = 2.0$	233	38585.79	728.44	38445.00	54.97	297.88	85.82	490.39	1.06
	$\eta = 3.0$	250	41276.60	257.09	41250.00	16.97	141.93	9.63	146.71	0.37
	PI	—	32725.57	4252.19	32677.93	0.00	0.00	47.64	54.13	0.00
	DA	195	36538.59	8082.15	32175.00	2584.14	6220.55	1779.45	2434.60	6.72
	SP	202	36328.43	6585.24	33330.00	1824.18	5029.10	1174.25	2011.94	4.98
	RP									
$\alpha = 5\%$	$\eta = 0.1$	198	36400.88	7422.82	32670.00	2231.03	5688.50	1499.85	2256.24	5.94
$u_\alpha = 166$	$\eta = 0.5$	210	36523.37	5067.04	34650.00	1195.69	3874.39	677.68	1529.68	3.49
	$\eta = 1.0$	225	37822.42	2859.08	37125.00	503.98	2260.54	193.45	775.65	1.64
	$\eta = 2.0$	255	42143.73	674.59	42075.00	64.40	618.02	4.33	97.67	0.26
	$\eta = 3.0$	292	48182.05	68.81	48180.00	2.05	68.81	0.00	0.00	0.01
	MxRP									
	$\eta = 0.1$	192	36749.70	8764.76	31680.00	2982.34	6783.30	2087.36	2606.67	7.56
	$\eta = 0.5$	200	36349.29	6997.70	33000.00	2018.66	5351.71	1330.63	2134.48	5.45
	$\eta = 1.0$	219	37186.11	3641.13	36135.00	719.91	2826.93	331.21	1045.39	2.23
	$\eta = 2.0$	255	42143.73	674.59	42075.00	64.40	618.02	4.33	97.67	0.26
	$\eta = 3.0$	292	48182.05	68.81	48180.00	2.05	68.81	0.00	0.00	0.01
	PI	—	32725.57	4252.19	32677.93	0.00	0.00	47.64	54.13	0.00
	DA	195	54970.04	52282.73	32175.00	21015.59	50588.78	1779.45	2434.60	6.72
	SP	233	41004.42	13747.43	38445.00	2473.61	13404.77	85.82	490	1.06
	RP									
$\alpha = 1\%$	$\eta = 0.1$	198	52313.78	47841.24	32670.00	18143.93	46261.93	1499.85	2256.24	5.94
$u_\alpha = 1350$	$\eta = 0.5$	210	45051.66	32601.04	34650.00	9723.98	31508.61	677.68	1529.68	3.49
	$\eta = 1.0$	225	41417.05	18936.65	37125.00	4098.60	18383.90	193.45	775.65	1.64
	$\eta = 2.0$	255	42603.06	5078.16	42075.00	523.73	5026.03	4.33	97.67	0.26
	$\eta = 3.0$	292	48196.67	559.57	48180.00	16.67	559.57	0.00	0.00	0.01
	MxRP									
	$\eta = 0.1$	192	58021.33	56967.56	31680.00	24253.97	55165.38	2087.36	2606.67	7.56
	$\eta = 0.5$	200	50747.44	45022.91	33000.00	16416.81	43522.97	1330.63	2134.48	5.45
	$\eta = 1.0$	219	42320.89	23739.77	36135.00	5854.68	22990.06	331.21	1045.39	2.23
	$\eta = 2.0$	255	42603.06	5078.16	42075.00	523.73	5026.03	4.33	97.67	0.26
	$\eta = 3.0$	292	48196.67	559.57	48180.00	16.67	559.57	0.00	0.00	0.01

Appendix B

Appendix of Chapter 4

B.1 Proof of Theorem 4.1

In this Appendix we give the proof of Theorem 4.1: *The relaxed linear programs of MIP (4.7) is integral (has an integral optimum, when any optimum exists).*

Recall the formulation of the MIP (4.7):

$$
\begin{aligned}
Min \quad & \sum_{j\in J} c_j\, X_j + \sum_{l\in L}\sum_{i\in I_1} p_l\, u\, M_{i,l} + \sum_{l\in L}\sum_{k\in K} pb_{l,k} \left[\sum_{j\in J} d_j\, Y_{j,k} - \sum_{j\in J} r_j\, Z_{j,k} + \sum_{i\in I_2} u\, M'_{i,l,k}\right] \\
\text{s.t.} \quad & \sum_{j\in J} a_{ij} X_j + M_{i,l} \geq n_{i,l}, i \in I_1, l \in L, \\
& \sum_{j\in J} a_{ij} X_j + \sum_{j\in J} a'_{ij}(Y_{j,k} - Z_{j,k}) + M'_{i,l,k} \geq n_{i,l}, i \in I_2, l \in L, k \in K, \\
& X_j \geq Z_{j,k}, j \in J, k \in K, \\
& M_{i,l} \geq 0, i \in I_1, l \in L, \\
& M'_{i,l,k} \geq 0, i \in I_2, l \in L, k \in K, \\
& X_j, Y_{j,k}, Z_{j,k} \in \mathbb{Z}^+, j \in J, k \in K.
\end{aligned}
$$

Lemma B.1 *Let* $\mathbf{H}$, $\mathbf{W}$ *be vectors of* m *dimension with each element being integer, infinite or infinitesimal, we prove that the set* E *defined in (B.1) is an integer polyhedron.*

$$
\begin{aligned}
W_i \geq e_i \geq H_i, \quad & i = 1, .., m, \\
e_i \geq e_j, \qquad & i = 1, .., m, j = 1, .., m, i \neq j.
\end{aligned}
\tag{B.1}
$$

Proof: We use the reductio ad absurdum method to prove this. Recall that in mathematics,

an extreme point of a convex set E in a real vector space is a point in E which does not lie in any open line segment joining two points of E. Suppose the polyhedron E contains an extreme point $\mathbf{e}'$ of which all the elements are not integers. Let S denotes the ensemble index where the element of $\mathbf{e}'$ is non-integer, $S := \{i : e'_i \notin \mathbb{Z}, i = 1, .., m\}$. And let ϕ be the minimum of the gaps between these non-integer elements and their own closest integers. This minimum is the largest value satisfying

$$\phi \leq \lceil e'_i \rceil - e'_i, i \in S, \tag{B.2}$$

$$\phi \leq e'_i - \lfloor e'_i \rfloor, i \in S. \tag{B.3}$$

We have thereafter $0 < \phi < 1$. As an extreme point of set E, then we have $\mathbf{e}' \in E$, consequently $\mathbf{e}'$ satisfies the inequalities in (B.1). Let Φ be the vector of m dimension defined as $\Phi = \{\Phi_i = \phi, \text{for} i \in S; \Phi_i = 0, \text{otherwise}\}$. It is easy seen that both $\mathbf{e}' + \mathbf{\Phi}$ and $\mathbf{e}' - \mathbf{\Phi}$ satisfy Inequalities (B.1), the non-integer point $\mathbf{e}'$ can not be an extreme point of polyhedron E.

Lemma B.2 *The following linear problem (B.4-B.8) is integral.*

$$Min \quad \sum_{j\in J} c_j X_j + \sum_{j\in J} d_j Y_j - \sum_{j\in J} r_j Z_j + \sum_{i\in I} u M_i \tag{B.4}$$

$$\text{s.t.} \quad \sum_{j\in J} a_{ij} X_j + \sum_{j\in J} a'_{ij}(Y_j - Z_j) + M_i \geq n_i, \ \ i \in I, \tag{B.5}$$

$$X_j \geq Z_j, \qquad j \in J, \tag{B.6}$$

$$M_i \geq 0, \qquad i \in I, \tag{B.7}$$

$$X_j, Y_j, Z_j \geq 0, \qquad j \in J. \tag{B.8}$$

Proof: Problem (B.4-B.8) is a linear function, the feasible solution set is convex and the optimal solution is an extreme point. We show that all extreme points are integers for this feasible solution set.

Recall the definition of the matrix $\mathbf{A}$ and $\mathbf{A}'$ in Section 4.2.6, denote vector $\mathbf{x} = (X_j | j \in J)$, vector $\mathbf{y} = (Y_j | j \in J)$, vector $\mathbf{z} = (Z_j | j \in J)$, vector $\mathbf{M} = (M_i | i \in I)$ and vector $\mathbf{n} = (n_i | i \in I)$. Constraint (B.5) can be rewritten as:

$$\begin{pmatrix} \mathbf{A} & \mathbf{A}' & \mathbf{A}' & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \\ -\mathbf{z} \\ \mathbf{M} \end{pmatrix} = \begin{pmatrix} \mathbf{n} \end{pmatrix}.$$

Recall that both matrix $\mathbf{A}$ and $\mathbf{A'}$ are totally unimodular. Hence every column of the left side matrix ($\mathbf{A}$ $\mathbf{A'}$ $\mathbf{A'}$ $\mathbf{I}$) would have continuous ones, this matrix is also totally unimodular. As the required agents number n_i is integral, Constraint (B.5) defines a polyhedron all of whose extreme points are integer valued. This integer polyhedron can be rewritten as:

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \\ -\mathbf{z} \\ \mathbf{M} \end{pmatrix} \geq \begin{pmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{pmatrix},$$

where $\alpha, \beta, \gamma, \delta$ are integer vectors. Combine with Constraints (B.6)-(B.8), the feasible solution set can be formulated as:

$$X_j \geq Z_j, j \in J, \tag{B.9}$$

$$\begin{pmatrix} \infty \\ \infty \\ (-\gamma)^+ \\ \infty \end{pmatrix} \geq \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{z} \\ \mathbf{M} \end{pmatrix} \geq \begin{pmatrix} \alpha^+ \\ \beta^+ \\ \mathbf{0} \\ \delta^+ \end{pmatrix}.$$

It is obvious that Equation (B.9) has the same structure as Equation (B.1). Therefore, the feasible set of Equation (B.4)-(B.8) is an integer polyhedron consequently the LP (B.4)-(B.8) is integral.

Theorem B.1 *The relaxed linear programs of MIPs (4.7) is integral.*

Using the results of Lemmas B.1 and B.2, we prove this theorem. For demonstration convenience, we consider the number of scenarios in MIP (4.7) as $|L| \times |K|$ for each period $i \in I$ and reformulate the relaxed linear program as in Equation (B.10)-(B.17). As parameters, for a given period i and scenario l, we define that the $n_{i,l,k} = n_{i,l}$ for $k \in K$. Then the linear relaxed

version of (4.7) can be written as

$$Min \quad \sum_{j\in J} c_j\, X_j + \sum_{l\in L}\sum_{k\in K} pb_{l,k}\left(\sum_{j\in J} d_j\, Y_{j,l,k} - r_j\, Z_{j,l,k} + \sum_{i\in I} u\, M'_{i,l,k}\right), \tag{B.10}$$

$$\text{s.t.} \quad \sum_{j\in J} a_{ij} X_j + \sum_{j\in J} a'_{ij}(Y_{j,l,k} - Z_{j,l,k}) + M'_{i,l,k} \geq n_{i,l,k}, i \in I, l \in L, k \in K, \tag{B.11}$$

$$X_j \geq Z_{j,1,k}, j \in J, k \in K, \tag{B.12}$$

$$Y_{j,1,k} = Y_{j,l,k}, j \in J, l \in L, k \in K, \tag{B.13}$$

$$Z_{j,1,k} = Z_{j,l,k}, j \in J, l \in L, k \in K, \tag{B.14}$$

$$M'_{i,l,1} = M'_{i,l,k}, i \in I_1, l \in L, k \in K, \tag{B.15}$$

$$M'_{i,l,k} \geq 0, i \in I, l \in L, k \in K, \tag{B.16}$$

$$X_j, Y_{j,l,k}, Z_{j,l,k} \geq 0, j \in J, l \in L, k \in K. \tag{B.17}$$

Equalities (B.13)-(B.14) guarantee that all scenarios which share the estimated $\tilde{\theta}_k$, implement the same second-stage decisions. Equality (B.15) concerning the agents shortfall in the early horizon periods are straight forward. Denote vector $\mathbf{x} = (X_j | j \in J)$. For $l \in L, k \in K$, we denote the vectors as follow. $\mathbf{M'_{l,k}} = (M'_{i,l,k} | i \in I)$, $\mathbf{y'_{l,k}} = (Y_{j,l,k} | j \in J)$, $\mathbf{z'_{l,k}} = (Z_{j,l,k} | j \in J)$, and $\mathbf{n'_{l,k}} = (n_{i,l,k} | i \in I)$.

Constraint (B.11) can be presented as:

$$\begin{pmatrix} A & A' & A' & I & & 0 & & 0 & & .. & & .. & & 0 & \\ A & & 0 & & A' & A' & I & 0 & & .. & & .. & & 0 & \\ .. & & .. & & & .. & & .. & & .. & & .. & & .. & \\ A & & 0 & & & 0 & & .. & A' & A' & I & .. & & 0 & \\ .. & & .. & & & .. & & .. & & .. & & .. & & .. & \\ A & & 0 & & & 0 & & .. & & 0 & & .. & A' & A' & I \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y_{1,1}} \\ \mathbf{-z_{1,1}} \\ \mathbf{M_{1,1}} \\ \mathbf{y_{1,2}} \\ \mathbf{-z_{1,2}} \\ \mathbf{M_{1,2}} \\ .. \\ \mathbf{y_{l,k}} \\ \mathbf{-z_{l,k}} \\ \mathbf{M_{l,k}} \\ .. \\ \mathbf{y_{|L|,|K|}} \\ \mathbf{-z_{|L|,|K|}} \\ \mathbf{M_{|L|,|K|}} \end{pmatrix} = \begin{pmatrix} \mathbf{n_{1,1}} \\ \mathbf{n_{1,2}} \\ .. \\ \mathbf{n_{l,k}} \\ .. \\ \mathbf{n_{|L|,|K|}} \end{pmatrix}.$$

And the feasible solution set of the Equation (B.10)-(B.17) can be presented as:

$$\begin{aligned} &X_j \geq Z_{j,1,k}, j \in J, k \in K, \\ &Y_{j,1,K} = Y_{j,l,k}, j \in J, l \in L, k \in K, \\ &Z_{j,1,K} = Z_{j,l,k}, j \in J, l \in L, k \in K, \\ &M'_{i,l,1} = M'_{i,l,k}, i \in I_1, l \in L, k \in K, \end{aligned} \tag{B.18}$$

,

$$\begin{pmatrix} \infty \\ \infty \\ (-\gamma')^+ \\ \infty \end{pmatrix} \geq \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \\ \mathbf{Z} \\ \mathbf{M''} \end{pmatrix} \geq \begin{pmatrix} \alpha'^+ \\ \beta'^+ \\ \mathbf{0} \\ \delta'^+ \end{pmatrix},$$

where we denote the vector $\mathbf{X} = (X_j | j \in J)$, vector $\mathbf{Y} = (Y_{j,l,k} | j \in J, l \in L, k \in K)$, vector $\mathbf{Z} = (Z_{j,l,k} | j \in J, l \in L, k \in K)$, vector $\mathbf{M''} = (M'_{i,l,k} | i \in I, l \in L, k \in K)$. And the vectors $\alpha', \beta', \gamma', \delta'$ have the corresponding dimensions. We can see that Equation (B.18) also has the same structure as Equation (B.1). Therefore, the relaxed linear program of the MIP (4.7) is integral. This finishes the proof. □

B.2 Proof of the worst case θ value within a given interval

Given the fist-stage decision variables X_j, and the estimate business factor $\tilde{\theta}$, by the following we prove that, among all the true business factor θ falling in the interval $[\tilde{\theta}-\tau, \tilde{\theta}+\tau]$, the one associated with the worst cost is $\theta = \tilde{\theta}+\tau$.

Consider the following Problem (B.19). It is obvious that the optimal solution of this problem with $\theta = \tilde{\theta}+\tau$ is feasible for the same problem with other θ values in $[\tilde{\theta}-\tau, \tilde{\theta}+\tau]$, since $n_i(\theta)$ is an increasing concave function in θ. This indicates that the optimal cost associated with $\theta = \tilde{\theta}+\tau$ is bigger or equal than that associated with other values of θ in this interval.

$$
\begin{aligned}
\min_{X_j, Y_j, Z_j} \quad & \sum_{j\in J} c_j\, X_j + \sum_{j\in J} d_j\, Y_j(\theta) - \sum_{j\in J} r_j\, Z_j(\theta) + \sum_{i\in I} u\, M_i \\
\text{s.t.} \quad & \sum_{j\in J} a_{ij} X_j + M_i \geq n_i(\theta), i \in I_1, \\
& \sum_{j\in J} a_{ij} X_j + \sum_{j\in J} a'_{ij}(Y_j(\theta) - Z_j(\theta)) + M_i \geq n_i(\theta), i \in I_2, \\
& X_j \geq Z_j(\theta), j \in J, \\
& M_i \geq 0, i \in I, \\
& X_j, Y_j(\theta), Z_j(\theta) \in \mathbb{Z}^+, j \in J.
\end{aligned}
\tag{B.19}
$$

B.3 Proof of Theorem 4.2

In this section we give the proof of Theorem 4.2: *If there exists a value of θ, denoted by $\hat{\theta}$, for which $N_i(\hat{\theta}) \geq N_{ik}(\theta)$ with $i \in I$, for any $\theta \in U_k, k \in K$, then Equation (4.14)-(4.21) can be simplified as Equation (4.22)-(4.29).*

Given values of $X_j, j \in J$, let $C_k(\theta, Y_{jk}, Z_{jk}, M_{ik}, M'_{ik})$ denote the cost associated with a given solution $(Y_{jk}, Z_{jk}, M_{ik}, M'_{ik})$.

Let $C^*_k(\theta), k \in K$, denote the optimal cost of the optimization problem (B.20)-(B.27), with optimal solution $(Y^*_{jk}, Z^*_{jk}, M^*_{ik}, M'^*_{ik})$.

$$\min_{Y_{jk},Z_{jk}} \quad \sum_{j\in J} d_j\, Y_{jk} - \sum_{j\in J} r_j\, Z_{jk} + \sum_{i\in I_1} u\, M_{ik} + \sum_{i\in I_2} u\, M'_{ik} \tag{B.20}$$

$$\text{s.t.} \quad \sum_{j\in J} a_{ij}X_j + M_{ik} \geq N_{ik}(\theta), i \in I_1, \theta \in U_k, \tag{B.21}$$

$$\sum_{j\in J} a_{ij}X_j + \sum_{j\in J} a'_{ij}(Y_{jk} - Z_{jk})(\tilde{\theta} + \tau) + M'_{ik} \geq N_{ik}(\theta),$$

$$i \in I_2, \theta \in U_k, \tilde{\theta} \in [\theta - \tau, \theta + \tau] \cap U'_k(\theta) \tag{B.22}$$

$$N_{ik}(\theta) \geq \varphi_1^{ik} + \varphi_2^{ik}\,\theta, i \in I, \theta \in U_k, \tag{B.23}$$

$$X_j \geq Z_{jk}, j \in J, k \in K, \tag{B.24}$$

$$M_{ik} \geq 0, i \in I_1, \tag{B.25}$$

$$M'_{ik} \geq 0, i \in I_2, \tag{B.26}$$

$$Y_{jk}, Z_{jk} \in \mathbb{Z}^+, j \in J. \tag{B.27}$$

Let also $C^*(\hat{\theta})$ be the optimal cost of the optimization problem (B.28)-(B.34), with optimal solution $(\hat{Y}_j, \hat{Z}_j, \hat{M}_i, \hat{M}'_i)$.

$$\min_{Y_j,Z_j} \quad \sum_{j\in J} d_j\, Y_j - \sum_{j\in J} r_j\, Z_j + \sum_{i\in I_1} u\, M_i + \sum_{i\in I_2} u\, M'_i, \tag{B.28}$$

$$\text{s.t.} \quad \sum_{j\in J} a_{ij}X_j + M_i \geq N_i(\hat{\theta}), \qquad i \in I_1, \tag{B.29}$$

$$\sum_{j\in J} a_{ij}X_j + \sum_{j\in J} a'_{ij}(Y_j - Z_j)(\hat{\theta} + \epsilon + \tau) + M'_i \geq N_i(\hat{\theta}), \qquad i \in I_2, \tag{B.30}$$

$$X_j \geq Z_j, j \in J, \tag{B.31}$$

$$M_i \geq 0, i \in I_1, \tag{B.32}$$

$$M'_i \geq 0, i \in I_2, \tag{B.33}$$

$$Y_j, Z_j \in \mathbb{Z}^+, j \in J. \tag{B.34}$$

We prove by following that $C^*(\hat{\theta}) \geq C_k^*(\theta)$ for $k \in K$. According to the assumption in Theorem 4.2 that $N_i(\hat{\theta}) \geq N_i(\theta)$ with $i \in I$ for any $\theta \in U_k, k \in K$, it is easy to see that $(\hat{Y}_j, \hat{Z}_j, \hat{M}_i, \hat{M}'_i)$ is a feasible solution for Problem (B.20)-(B.27). Then we have $C_k(\theta, \hat{Y}_j, \hat{Z}_j, \hat{M}_i, \hat{M}'_i) = C^*(\hat{\theta})$ for $k \in K$. Since $C_k^*(\theta) \leq C_k(\theta, \hat{Y}_j, \hat{Z}_j, \hat{M}_i, \hat{M}'_i)$, we have $C^*(\hat{\theta}) \geq C_k^*(\theta)$ for $k \in K$. This ends the proof of Theorem 4.2. □

Bibliography

Aksin, Z., Armony, M., and Mehrotra, V. (2007). The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research. *Production and Operations Management*, 16:665–688.

Aldor-Noiman, S., Feigin, P., and Mandelbaum, A. (2009). Workload Forcasting for a Call Center: Methodology and a Case Study. *Annals of Applied Statistics*, 3:1403–1447.

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

Babonneau, F., Vial, J.-P., and Apparigliato, R. (2010). *Robust Optimization for Environmental and Energy Planning.* Springer Verlag. In Filar, J. and Haurie, A., editors, Handbook on "Uncertainty and Environmental Decision Making", International Series in Operations Research and Management Science, pages 79–126.

Ben-Tal, A., Bertsimas, D., and Brown, D. (2010). A Soft Robust Model for Optimization Under Ambiguity. *Operations Research*, 58:1220–1234.

Ben-Tal, A., El Ghaoui, L., and Nemirovski, A. (2009). *Robust Optimization.* Princeton University Press.

Ben-Tal, A., Golany, B., Nemirovski, A., and Vial, J. (2005). Retailer-Supplier Flexible Commitments Contracts: A Robust Optimization Approach. *MSOM*, 7:248–271.

Ben-Tal, A., Goryashko, A., Guslitzer, E., and Nemirovski, A. (2004). Adjustable Robust Solutions of Uncertain Linear Programs. *Math. Programming*, 99:351–376.

Ben-Tal, A. and Nemirovski, A. (1998). Robust Convex Optimization. *Math. Oper. Res.*, 23:769–805.

Ben-Tal, A. and Nemirovski, A. (1999). Robust solutions of uncertain linear programs. *Operations Research Letters*, 25:1 – 13.

Ben-Tal, A. and Nemirovski, A. (2000). Robust Solutions of Linear Programming Problems Contaminated with Uncertain Data. *Mathematical Programming*, 88:411–424.

Benders, J. F. (1962). Partitioning Procedures for Solving Mixed Variables Programming Problems. *Numerische Mathematik*, 4:238–252.

Bertsimas, D. and Brown, D. (2009). Constructing Uncertainty Sets for Robust Linear Opitmization. *Operations Research*, 57:1486–1495.

Bertsimas, D., Brown, D. B., and Caramanis, C. (2010a). Theory and Applications of Robust Optimization. To appear in SIAM Review.

Bertsimas, D. and Caramanis, C. (2010). Finite Adaptability in Multistage Linear Optimization. *IEEE Transactions on Automatic Control*, 55:2751–2766.

Bertsimas, D. and Doan, X. (2010). Robust and Data-Driven Approaches to Call Centers. *European Journal of Operational Research*, 207:1072–1085.

Bertsimas, D., Doan, X., Natarajan, K., and Teo, C. (2010b). Models for Minimax Stochastic Linear Optimization Problems with Risk Aversion. *Mathematics of Operations Research*, 35:580–602.

Bertsimas, D., Pachamanova, D., and Sim, M. (2004). Robust Linear Optimization under General Norms. *Operations Research Letters*, 32:510–516.

Bertsimas, D. and Sim, M. (2003). Robust Discrete Optimization and Network Flows. *Mathematical Programming*, 98:49–71.

Bertsimas, D. and Sim, M. (2004). The Price of Robustness. *Operations Research*, 52:35–53.

Bertsimas, D. and Thiele, A. (2006). Modern Decision-making under Uncertainty: Robust and Data-driven Optimization. INFORMS Annual Meeting, Pittsburgh, Pennsylvania.

Bhandari, A., Scheller-Wolf, A., and Harchol-Balter, M. (2008). An Exact and Efficient Algorithm for the Constrained Dynamic Operator Staffing Problem for Call Centers. *Management Science*, 54:339–353.

Birge, J. R. and Louveaux, F. (1997). *Introduction to Stochastic Programming*. Springer.

Borst, S., Mandelbaum, A., and Reiman, M. (2004). Dimensioning Large Call Centers. *Operations Research*, 52:17–34.

Breton, M. and El Hachem, S. (1995). Algorithms for the Solution of Stochastic Dynamic Minimax Problems. *Computational Optimization and Applications*, 4:317–345.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. *J. Amer. Statist. Assoc.*, 100:36–50.

Brown, L., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2002). Multifactor Poisson and Gamma-Poisson Models for Call Center Arrival Times. Technical report, University of Pennsylvania.

Cai, X., McKinney, D., Lasdon, L., and Watkins Jr, D. (2001). Solving Large Nonconvex Water Resources Management Models using Generalized Benders Decomposition. *Operations Research*, 49:235–245.

Calafiore, G. and El Ghaoui, L. (2006). On Distributionally Robust Chanceconstrained Linear Programs with Applications. *Journal of Optimization Theory and Applications*, 130:1–22.

Charnes, A. and Cooper, W. (1959a). Uncertain Convex Programs: Randomize Solutions and Confidence Level. *Management Science*, 6:73–79.

Charnes, A. and Cooper, W. W. (1959b). Chance-Constrained Programming. *Management Science*, 6:73–79.

Chen, W., Sim, M., Sun, J., and Teo, C. P. (2010). From Cvar to Uncertainty Set: Implications in Joint Chance-Constrained Optimization. *Operations Research*, 58:470 – 485.

Chen, X., Sim, M., and Sun, P. (2007). A Robust Optimization Perspective to Stochastic Programming. *Operations Research*, 55:1058–1071.

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

Dantzig, G. B. (1955). Linear Programming under Uncertainty. *Management Science*, 1:197–206.

Delage, E. and Ye, Y. (2010). Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems. *Operations Research*, 58:595–612.

Dupačová, J. (1987). The Minimax Approach to Stochastic Programming and an Illustrative Application. *Stochastics*, 20:73–88.

El Ghaoui, L., Oks, M., and Oustry, F. (2003). Worst-case Value-at-Risk and Robust Portfolio Optimization: A Conic Programming Approach. *Operations Research*, 51:543–556.

Erdogan, E. and Iyengar, G. (2006). Ambiguous Chance Constrained Problems and Robust Optimization. *Mathematical Programming*, 107:37–61.

Ernst, A. T., Jiang, H., Krishnamoorthy, M., Owens, B., and Sier, D. (2004). An Annotated Bibliography of Personnel Scheduling and Rostering. *Annals of Operations Research*, 127:21–144.

Feldman, A., Mandelbaum, A., Massey, W., and Whitt, W. (2008). Staffing of Time-Varying Queues to Achieve Time-Stable Performance. *Management Science*, 54:324–338.

Gallego, G. and Moon, I. (1993). The Distribution Free Newsboy Problem: Review and Extensions. *The Journal of the Operational Research Society*, 44:825–834.

Gans, N., Koole, G., and Mandelbaum, A. (2003). Telephone Calls Centers: A Tutorial and Literature Review and Research Prospects. *Manufacturing and Service Operations Management*, 5:79–141.

Gans, N., Shen, H., and Zhou, Y. (2009). Parametric Stochastic Programming Models for Call-Center Workforce Scheduling. working paper.

Ghaoui, L. E. and Lebret, H. (1997). Robust Solutions to Least-Squares Problems with Uncertain Data. *SIAM J. Matrix Anal. Appl.*, 18:1035–1064.

Ghaoui, L. E., Oustry, F., and Lebret, H. (1998). Robust Solutions to Uncertain Semidefinite Programs. *SIAM J. on Optimization*, 9:33–52.

Goh, G. and Sim, M. (2010). Distributionally Robust Optimization and Its Tractable Approximations. *Operations Research*, 58:902–917.

Green, L. and Kolesar, P. (1991). The Pointwise Stationary Approximation for Queues with Nonstationary Arrivals. *Management Science*, 37:84–97.

Green, L., Kolesar, P., and Soares, J. (2003). An Improved Heuristic for Staffing Telephone Call Centers with Limited Operating Hours. *Production and Operations Management*, 12:46–61.

Green, L., Kolesar, P., and Whitt, W. (2007). Coping with Time-Varying Demand When Setting Staffing Requirements for a Service System. *Production and Operations Management*, 16:13–39.

Gross, D. and Harris, C. (1998). *Fundamentals of Queueing Theory*. Wiley Series in Probability and Mathematical Statistics. 3rd Edition.

Gurvich, I., Luedtke, J., and Tezcan, T. (2010). Staffing Call-Centers with Uncertain Demand Forecasts: A Chance-Constraints Approach. *Management Science*, 56:1093–1115.

Halfin, S. and Whitt, W. (1981). Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research*, 29:567–588.

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

Helber, S. and Henken, K. (2010). Profit-Oriented Shift Scheduling of Inbound Contact Centers with Skills-Based Routing, Impatient Customers, and Retrials. *OR Spectrum*, 32:109–134.

Higle, J. (2005). Stochastic Programming: Optimization when Uncertainty Matters. Tutorials in OR, INFORMS.

Hur, D., Mabert, V., and Bretthauer, K. (2004). Real-Time Work Schedule Adjustment Decisions: An Investigation and Evaluation. *Production and Operations Management*, 13:322–339.

Ingolfsson, A., Akhmetshina, A., Budge, S., Li, Y., and Wu, X. (2007). Survey and Experimental Comparison of Service Level Approximation Methods for Nonstationary M(t)/M/s(t) Queueing Systems. *INFORMS Journal on Computing*, 19:201–214.

Jongbloed, G. and Koole, G. (2001a). Managing Uncertainty in Call Centers Using Poisson Mixtures. *Applied Stochastic Models in Business and Industry*, 17:307–318.

Jongbloed, G. and Koole, G. (2001b). Managing Uncertainty in Call Centers Using Poisson Mixtures. *Applied Stochastic Models in Business and Industry*, 17:307–318.

Kall, P. and Mayer, J. (2005). *Stochastic Linear Programming: Models, Theory, and Computation*. Springer.

Kall, P. and Wallace, S. (1994). *Stochastic Programming*. John Wiley and Sons.

Koole, G. and Mandelbaum, A. (2002). Queueing Models of Call Centers: An Introduction. *Annals of Operations Research*, 113:41–59.

Koole, G. and van der Sluis, E. (2003). Optimal Shift Scheduling with A Global Service Level Constraint. *IIE Transactions*, 35:1049–1055.

Liao, S., van Delft, C., Koole, G., and Jouini, O. (2010). Staffing a Call Center with Uncertain Non-Stationary Arrival Rate and Flexibility. To appear in OR Spectrum.

Mehrotra, V., Ozluk, O., and Saltzman, R. (2010). Intelligent Procedures for Intra-Day Updating of Call Center Agent Schedules. *Production and Operations Management*, 19:353–367.

Mehrotra, V., Wright, C., and Patel, S. (2009). An Investigation into the Business Processes, Job Descriptions, and Human Resource Factors Associated with Successful Workforce Management Practice. Working Paper, San Francisco State University, San Francisco, CA.

Natarajan, K., Pachamanova, D., and Sim, M. (2009). Constructing Risk Measures from Uncertainty Sets. *Operations Research*, 57:1129–1141.

Natarajan, K., Sim, M., and Uichanco, J. (2010). Tractable Robust Expected Utility and Risk Models for Portfolio Optimization. *Mathematical Finance*, 20:695–731.

Nemirovski, A. and Shapiro, A. (2006). Convex Approximations of Chance Constrained Programs. *SIAM Journal on Optimization*, 17:969–996.

Ogryczak, W. and Ruszczynski, A. (2002). Dual Stochastic Dominance and Related Mean-Risk Models. *SIAM Journal on Optimization*, 13:60–78.

Riis, M. and Andersen, K. A. (2005). Applying the Minimax Criterion in Stochastic Recourse Programs. *European Journal of Operational Research*, 165:569–584.

Robbins, T. (2007). Managing Service Capacity under Uncertainty. Ph.D. Dissertation, Penn State University.

Robbins, T. and Harrison, T. (2010). A Stochastic Programming Model for Scheduling Call Centers with Global Service Level Agreements. *European Journal of Operational Research*, 207:1608–1619.

Robbins, T., Medeiros, D., and Harrison, T. (2007). Partial Cross Training in Call Centers with Uncertrain Arrivals and Global Service Level Agreements. In *Proceedings of the 2007 Winter Simulation Conference*.

Robbins, T., Medeiros, D., and Harrison, T. (2008). Optimal Cross Training in Call Centers with Uncertain Arrivals and Global Service Level Agreements. Working paper. Pennsylvania State University.

Rockafellar, T. and Uryasev, S. (2002). Conditional Value-at-Risk for General Loss Distributions. *Journal of Banking and Finance*, 26:1443–1471.

Ruszczynski, A. and Shapiro, A. (2003). *Stochastic Programming.* Elsevier. Handbooks in Operations Research and Management Science, Vol. 10.

Scarf, H. (1958). A Min-Max Solution of an Inventory Problem. *Studies in The Mathematical Theory of Inventory and Production*, pages 201–209.

Shapiro, A. (2008). Stochastic Programming Approach to Optimization under Uncertainty. *Mathematical Programming*, 112:183–220.

Shapiro, A. and Ahmed, S. (2004). On a Class of Minimax Stochastic Programs. *SIAM Journal on Optimization*, 14:1237–1249.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. (2009). *Lectures on Stochastic Programming: Modeling and Theory.* MPS/SIAM Series on Optimization.

Shapiro, A. and Kleywegt, A. (2002). Minimax Analysis of Stochastic Programs. *Optimization Methods and Software*, 17:523–542.

Shapiro, A. and Nemirovski, A. (2005). *On Complexity of Stochastic Programming Problems.* Springer-Verlag New York, Inc. In V. Jeyakumar and A. M. Rubinov (Eds.), Continuous optimization: Current trends and applications (pp. 111-144).

Shen, H. and Huang, J. (2008). Interday Forecasting and Intraday Updating of Call Center Arrivals. *Manufacturing and Service Operations Management*, 10:391–410.

Soyster, A. L. (1973). Convex Programming with Set-Inclusive Constraints and Applications to Inexact Linear Programming. *Operations Research*, 21:1154–1157.

Steckley, S., Henderson, S., and Mehrotra, V. (2004). Service System Planning in the Presence of a Random Arrival Rate. Technical report, Cornell University, Ithaca, NY.

Stolletz, R. (2008). Approximation of the Non-Stationary M(t)/M(t)/c(t)-Queue using Stationary Queueing Models: The Stationary Backlog-Carryover Approach. *European Journal of Operational Research*, 190:478–493.

Thenie, J., van Delft, C., and Vial, J. (2007). Automatic Formulation of Stochastic Programs Via an Algebraic Modeling Language. *Computational Management Science*, 4:17–40.

Thompson, G. (1993). Accounting for The Multi-Period Impact of Service When Determining Employee Requirements for Labor Scheduling. *Journal of Operations Management*, 11:269–287.

van Delft, C. and Vial, J. (2004). A Practical Implementation of Stochastic Programming: An Applications to the Evaluation of Option Contracts in Supply Chains. *Automatica*, 40:743–756.

Žáčková, J. (1966). On Minimax Solution of Stochastic Linear Programming Problems. *Časopis pro Peštováaní*, 91:423–430.

Wang, Z., Glynn, P., and Ye, Y. (2009). Likelihood Robust Optimization for Data-Driven Newsvendor Problem. Working Paper.

Weinberg, J., Brown, L., and Stroud, J. (2007). Bayesian Forecasting of an Inhomogeneous Poisson Process with Applications to Call Center Data. *Journal of the American Statistical Association*, 102:1186–1199.

Whitt, W. (1999). Dynamic Staffng in a Telephone Call Center Aiming to Immediately Answer All Calls. *Operations Research Letters*, 24:205–212.

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

Xu, H., Caramanis, C., and Mannor, S. (2010). A Distributional Interpretation of Robust Optimization. Proceedings of the Allerton Conference on Communication, Control and Computing.

Yoo, J. (1996). Queueing Models for Staffing Service Operations. Ph.D. Dissertation, University of Maryland.

Yue, J., Chen, B., and Wang, M. (2006). Expected Value of Distribution Information for the Newsvendor Problem. *Operations Research*, 54:1128–113.

Zhao, G. (2001). A Log-Barrier Method with Benders Decomposition for Solving Two-Stage Stochastic Linear Programs. *Mathematical Programming*, 90:507–536.