BSc Thesis Applied Mathematics

Robust surgery loading using mixture distributions

Emma González Sanabria

Supervisor: Anne Zander

July, 2024

Department of Applied Mathematics
Faculty of Electrical Engineering,
Mathematics and Computer Science

UNIVERSITY OF TWENTE.

Robust surgery loading using mixture distributions

Emma González Sanabria

July, 2024

Abstract

This paper examines the impact of applying mixture distributions to surgery scheduling while addressing uncertainty in surgery durations. Slack time is introduced as an additional backup time, beyond the expected sum of surgery durations for each operating room (OR), making schedules more robust. We compare two different methods to calculate slack time: one, where surgery durations are assumed to follow a normal distribution, and another where surgeries follow a mixture distribution of normal random variables. We apply these models to three different scheduling methods to analyze their performance.

Simulations have been conducted to test the implementation of the different schedules produced, utilizing a generated waiting list from real data. We find that the schedules produced using mixture distributions generally reduce total overtime, but may lead to a less efficient OR utilization. The regret-based random sampling method applied to the mixture distribution model effectively minimizes overtime and improves scheduling outcomes in terms of overtime. However, it highlights a trade-off between operating room availability and overtime.

Keywords: surgery durations, scheduling, mixture distributions, heuristics

Contents

1 Introduction

In many healthcare systems, waiting lists have become a main issue resulting in prolonged waiting times and delayed patient access to healthcare. Additionally, there is a global ageing population, leading to an increased demand for surgical interventions [12]. For hospitals, these surgeries are of large importance, as they generate a large proportion of the profit made[13]. All these factors highlight the importance of running operating rooms efficiently.

Applications of operations research techniques have led to significant improvements in the efficiency of healthcare organizations. Many models have been developed that approximate surgery durations using unimodal distributions, most commonly, normal or log-normal distributions [1] [10]. Uncertainty in the procedure that the surgeon will follow to perform a specific surgery can make estimated surgery durations inaccurate. To address this uncertainty, the planned surgery duration can be represented by a mixture distribution of several possible procedures for a specified surgery.

Many scientific papers explore the benefits of applying a log-normal distribution [10], while others make a comparison between normal and log-normal distributions for surgery durations[1] [14]. Additionally, other factors have been investigated, such as adding more variability to the models (staff and surgery room availabilities) [2], or even addressing the uncertainty that surgery durations involve, including slack times [7].

Much focus is given to how different factors affect surgery scheduling, making comparisons of the normal and log-normal distribution, and testing how different scheduling methods can improve results. Studies have not yet explored how to model surgery durations using mixture distributions, namely, addressing the uncertainty of which surgical procedure will be chosen once the patient has been examined. We will therefore investigate whether this is a significant factor that can influence the surgery scheduling problem.

By exploring the application of mixture distributions in surgery scheduling, the main goal is to develop more robust and adaptable scheduling models that exhibit the reality of surgery procedures. This, therefore, leads us to the question, Can better surgery schedules be produced and become more realistic when surgery durations are modelled using mixture distributions?

2 Literature research

Several approaches have been proposed in literature addressing the surgery scheduling problem:

The authors of the literature review "Operating room planning and surgical case scheduling: a review of literature" [15] analyze advance scheduling; where a specific date is assigned for each operation in advance, allocation scheduling; determining the exact start time of operations and the allocation of operating room resources, the block scheduling strategy; which involves assigning OR time blocks to specialities and scheduling subsets of patients within each time block, and other heuristic methods that are used to solve the OR scheduling problem.

In the paper "Operational research in the management of the operating theatre: a survey", a structured literature review is conducted, analyzing several approaches to the surgery scheduling problem [6]. They highlight the generalization of the bin-packing problem, in addition to the portfolio effect used to minimize slack time from the paper by Hans et. al [7].

Hans et. al. [7] consider the problem of assigning surgeries with sufficient planned slack time to each operating room per day. Various constructive heuristics and local search methods are proposed, and statistical information on surgery durations is used to exploit the portfolio effect, which minimizes the required slack. This paper concludes that their approaches free a lot of operating room capacity, liberating space for additional surgeries. This paper, unlike others, claims that randomness and uncertainty have been recognized, but never addressed explicitly, as they do.

M. A. Kamran et. al. [8] propose an advanced method of surgery scheduling for patients from a patient's waiting list, taking into account several constraints. This paper considers uncertainty via stochastic surgery durations, formulating a two-stage stochastic programming, and a two-stage chance-constrained stochastic programming. In addition, it demonstrates how there is a clear variation between the deterministic and stochastic solutions when the chance-constrained approach is used.

The paper "New heuristics for planning operating rooms" [11] also addresses the decision problem of assigning an intervention date and operating room to a set of patients from a waiting list. The main goal is to minimize access time for patients with diverse clinical priority values. The authors propose a set of 83 heuristics (81 constructive heuristics, a compos- ite heuristic, and a meta-heuristic), where they compare the methods against existing heuristics. Finally, they conclude that the proposed meta-heuristic is the best for the problem under consideration, showing significant improvements in several key performance indicators. This meta-heuristic consists of the Random Extraction-Insertion algorithm (REI), involving two main steps; the destruction step, where a specified number of elements are removed from the current solution, and a construction step, where the algorithm constructs a new solution by inserting the removed elements back into the solution in a way that optimizes the objective function.

After a brief literature review, it is possible to conclude that there are many interesting approaches to solving the surgery loading problem. But, as Hans et. al. [7] describes, randomness and uncertainty are normally not addressed. Even in the aforementioned paper, we can see how the randomness within individual surgeries is not addressed, as no distinct procedures are considered. Therefore, we observe there is a gap in literature for this idea to be explored. We will make use of several proposed heuristic methods, and adapt these to an innovative way of modelling surgery durations, where uncertainty plays a central role.

3 Problem description

The general problem set-up can be described as follows: we consider a planning horizon of T days $(t = 1, ..., T)$. Every day, K parallel identical operating rooms are available for use $(k = 1, ..., K)$. From here, it follows that every operating room (OR) k on day t has a certain capacity, defined by c_{kt}. We define the pair (k, t) as the OR-plan for OR k on day t.

For this model, we assume that all ORs start operating simultaneously and have an equal capacity on each OR-plan. In addition, we consider k operating rooms to be available for only one specialty. At all times, so for each OR-plan, we assume to have sufficient personnel, including a sufficient number of surgeons and support teams; implying that each day, there are at least as many surgeons as ORs available.

We consider a list of n elective surgeries, belonging to a waiting list N of surgeries that must be scheduled $(|N| = n)$. For simplicity, we do not consider emergency patients or patients that must be planned on certain days for a medical reason. Each surgery i $(i = 1, ..., n)$ has J_i procedures that are possible to complete the surgery. This means that for each surgery i, it is uncertain which procedure j $(j = 1, ..., J_i\})$ will be selected by the surgeon. For this reason, we model the duration of surgery i using mixture distributions, each with an expected duration μ_i and expected variance σ_i^2, which will be described in further detail in Section 4.1 Each OR-plan will therefore have a set of surgeries assigned, we refer to it as N_{kt}.

To make surgery schedules more robust against overtime, a planned slack time δ_{kt} will be assigned to each OR-plan, minimizing the total overtime and preventing surgery cancellations. The amount of slack time will be influenced by the probability of overtime, chosen by management, in addition to the expected variances and durations of each surgery in the OR-plan. The probability of overtime chosen will therefore determine the total OR utilization, as a higher probability of overtime will maximize OR utilization, and a lower probability will make less usage of the ORs, but ensure a smaller risk of running into overtime.

Given a surgery allocation N_{kt} for OR k on day t, the OR-plan capacity constraint can therefore be defined as:

$$\sum_{i \in N_{kt}} \mu_i + \delta_{kt} \leq c_{kt} + O_{kt} \tag{1}$$

where $O_{kt}(O_{kt} \geq 0)$ represents the overtime occurring in OR-plan (k, t). With this constraint, we assume that surgeries may be planned on any OR-plan. In addition, we consider the best solution to this constraint the one that generates less overtime, therefore, where $\sum_{k,t} O_{kt}$ is minimized.

The situation can be analyzed as a general bin-packing problem. This is illustrated in the following figure:

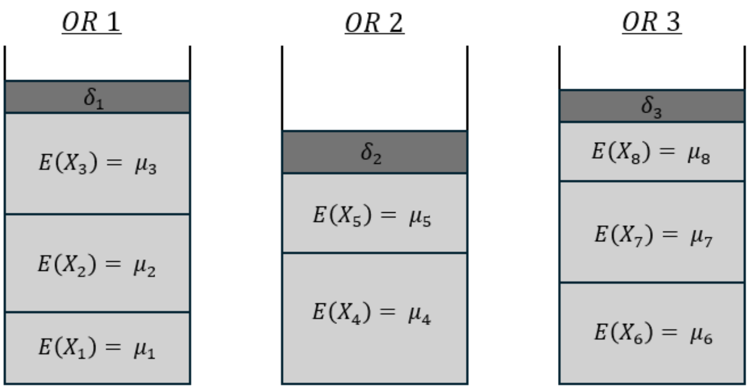

FIGURE 1: Illustration of surgeries assigned to ORs with corresponding slack times

As illustrated in Figure 1, surgeries will be scheduled into different ORs following three different scheduling methods. The following section will analyse surgery durations, slack times, and the aforementioned scheduling methods in further detail.

4 Methods

4.1 Modelling surgery durations

As described in the previous section, we will define the uncertainty of surgery durations, which enclose different procedures, using mixture distributions.

Let X_i represent the duration of surgery i, with corresponding possible procedures J_i. Assume procedures $j = 1, ..., J_i$ follow the same distribution, with parameters $(\mu_{ij}, \sigma_{ij}^2)$. Then let X_i follow a mixture distribution of random variables, with probability density function defined by:

$$f_i(x) = \sum_{j=1}^{J} w_{ij} f_{ij}(x), \tag{2}$$

where w_{ij} represent the weight/probability of following procedure j in surgery i, under the constraints of $\sum_{j=1}^{J} w_{ij} = 1$, $0 \leq w_{ij} \leq 1$. Furthermore, $f_{ij}(x)$ represents the probability density function of the individual random variables. [4]

As each surgery procedure j is independent of others, we can define the expected duration of the mixture distributed duration of surgery i as a weighted sum of the individual expected durations of procedures $j = 1, ..., J$.

$$E(X_i) = \sum_{j=1}^{J} w_{ij} \mu_{ij} = \mu_i, \tag{3}$$

Consequently, the variance of surgery i can be derived by:

$$\begin{aligned} \text{var}(X_i) &= E(X_i^2) - \mu_i^2 \\ &= \sum_{j=1}^{J_i} w_{ij} E(X_{ij}^2) - \mu_i^2. \end{aligned}$$

Then,

$$\sum_{j=1}^{J_i} w_{ij} E(X_{ij}^2) = \mathrm{var}(X_{ij}) = \sigma_{ij}^2,$$

$$\begin{aligned}\sigma_{ij}^2 &= E(X_{ij}^2) - (E(X_{ij}))^2 \\ &= E(X_{ij}^2) - \mu_{ij}^2.\end{aligned}$$

So variance of surgery i is defined as:

$$\mathrm{var}(X_i) = \sum_{j=1}^{J} w_{ij}(\sigma_{ij}^2 + \mu_{ij}^2) - \mu_i^2 = \sigma_i^2. \tag{4}$$

Specifically, we can define surgery durations following a mixture of normal random variables. In which case, we can define the probability density function $f_{ij}(x)$ in equation 2 as the probability density function of a normal random variable

$$f_{ij}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

We will compare the performance of surgery scheduling using the mixture distribution described in addition to single distributions of individual procedures. This will provide insights into the best way to model surgeries.

4.2 Defining slack time

As described in the previous section, the planned slack time will be determined by the probability of overtime, chosen by management, and by the expected durations and variances of the surgeries in each OR.

To simplify the derivation of the planned slack time δ_{kt}, consider any two surgeries, each with, for example, two procedures possible. This means we have X_1, X_2, each following a mixture distribution with parameters (μ_1, σ_1^2), (μ_2, σ_2^2) respectively. Following equation 3, we have for $i = 1, 2$,

$$\begin{aligned}\mu_i &= w_{i1}\mu_{i1} + w_{i2}\mu_{i2}, \\ \sigma_i^2 &= w_{i1}(\sigma_{i1}^2 + \mu_{i1}^2) + w_{i2}(\sigma_{i2}^2 + \mu_{i2}^2) - \mu_i^2.\end{aligned}$$

The next step is to define the probability of overtime for OR k on day t. Note that, for modelling, we assume that overtime occurs when the total surgery durations exceed the expected surgery durations plus the slack time generated. This is defined as

$$P(X > \mu_{kt} + \delta_{kt}) = p,$$

where $X = X_1 + X_2$ and $\mu_{kt} = \mu_1 + \mu_2$

As the cumulative distribution function of a mixture distribution is a weighted sum over the individual distributions, we can extend the expression above to:

$$
\begin{aligned}
P(X_1 + X_2 > \mu_{kt} + \delta_{kt}) &= 1 - P(X_1 + X_2 \leq \mu_{kt} + \delta_{kt}) \\
&= 1 - \Bigg[w_{11}w_{21}P(X_{11} + X_{21} \leq \mu_{kt} + \delta_{kt}) + w_{11}w_{22}P(X_{11} + X_{22} \leq \mu_{kt} + \delta_{kt}) \\
&+ w_{12}w_{21}P(X_{12} + X_{21} \leq \mu_{kt} + \delta_{kt}) + w_{12}w_{22}P(X_{12} + X_{22} \leq \mu_{kt} + \delta_{kt}) \Bigg] \\
&= 1 - \sum_{i=1}^{2}\sum_{j=1}^{2} w_{1i}w_{2j}P(X_{1i} + X_{2j} \leq \mu_{kt} + \delta_{kt}) \qquad (5)
\end{aligned}
$$

From here, we conclude that the unique, minimal δ_{kt} can be found by solving the equation $P(X > \mu_{kt} + \delta_{kt}) = p$. Note that this is a generalization of the case where we represent surgery durations with exactly one procedure possible, as presented by Hans et. al. with normally distributed random variables [7].

In the case where we define X_i as a random variable which follows a mixture of normally distributed random variables, we then know that

$$X_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}^2),$$

and as the sum of normal random variables is also a normal variable, then for each $i, j = 1, 2$, we know that

$$X_{1j} + X_{2j} \sim \mathcal{N}(\mu_{1i} + \mu_{2j}, \sigma_{1i}^2 + \sigma_{2j}^2).$$

This means that equation 5 can be expressed further as

$$P(X_1 + X_2 > \mu_{kt} + \delta_{kt}) = 1 - \sum_{i=1}^{2}\sum_{j=1}^{2} w_{1i}w_{2j}\Phi\left(\frac{\mu_{kt} + \delta_{kt} - (\mu_{1i} + \mu_{2j})}{\sigma_{1i} + \sigma_{2j}}\right) = p \qquad (6)$$

where $\Phi(.)$ represents the cumulative distribution function of the standard normal distribution.

So, for I surgeries, each with J_i possible procedures, equation (6) can be generalized to:

$$P(\sum_{i=1} X_i > \mu_{kt} + \delta_{kt}) = 1 - \sum_{j_1=1}^{J_1} ... \sum_{j_I=1}^{J_I} w_{ij_1}...w_{Ij_I}\Phi\left(\frac{\mu_{kt} + \delta_{kt} - (\mu_{ij_1} + ... + \mu_{Ij_I})}{\sigma_{ij_1} + ... + \sigma_{Ij_1}}\right) = p \qquad (7)$$

An approximation for δ_{kt} can be derived using the Taylor series expansion of $\Phi(.)$ around $a = \frac{\mu_{kt} - (\mu_{1i} + \mu_{2j})}{\sigma_{1i} + \sigma 2j}$ up to second order terms. Define $b = \frac{1}{\sigma_{1i} + \sigma_{2j}}$. Then the Taylor expansion of $\Phi(a + \frac{\delta_{kt}}{b})$ around a (up to second order terms) is

$$
\begin{aligned}
\Phi(a+\frac{\delta_{kt}}{b}) &= \Phi(a)+\Phi^{(1)}(a)(a+\frac{\delta_{kt}}{b}-a)+\frac{\Phi^{(2)}(a)}{2!}(a+\frac{\delta_{kt}}{b}-a)^2\\
&= \Phi(a)+\phi(a)(\frac{\delta_{kt}}{b})+\frac{\phi^{(1)}(a)}{2!}(\frac{\delta_{kt}}{b})^2\\
&= \frac{\phi^{(1)}(a)}{2b^2}\delta_{kt}^2+\frac{\phi(a)}{b}\delta_{kt}+\Phi(a)\\
&= \alpha\delta_{kt}^2+\beta\delta_{kt}+\gamma
\end{aligned}
\tag{8}
$$

This means that equation 6 can be approximated by

$$\sum_{i=1}^{2}\sum_{j=1}^{2} w_{1i}w_{2j}\left(\alpha_{1i,2j}\delta_{kt}^2+\beta_{1i,2j}\delta_{kt}+\gamma_{1i,2j}\right)=1-p$$

It is now possible to solve the quadratic equation, where we choose

$$\delta_{kt}=\min\{\delta_1,\delta_2|\delta_i>0\} \tag{9}$$

When performing simulations using real data, δ_{kt} will be solved using numerical methods. This will ensure that we obtain precise values while reducing computation time. The method that will be employed is the Newton-Raphson method.

The `fsolve` method in Python, part of the SciPy library, is used for solving systems of nonlinear equations [3]. The method relies on the underlying mathematics of the Newton-Raphson algorithm and other related techniques.

The Newton-Raphson algorithm [5] iteratively refines guesses to solve a set of nonlinear equations. In this case, we have a single nonlinear equation to solve. Therefore the general algorithm for a single nonlinear equation is as follows:

For the equation $f(x)=0$, we start with an initial guess x_0 for the root of the equation. The algorithm then calculates the derivative of the function $f(x)$ with respect to x, and updates the next approximation by $x_1=x_0-\frac{f(x_0)}{f'(x_0)}$. This process is then repeated and updates the estimated values iteratively by

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}.$$

This iteration continues until the difference between the successive approximations is below a specified tolerance level. Then, the final value obtained is an approximation of the root of the equation $f(x)=0$.

Specifically, for this problem we let

$$f(x)=1-\sum_{j_1=1}^{J_1}...\sum_{j_I=1}^{J_I} w_{ij_1}...w_{Ij_I}\Phi\left(\frac{\mu_{kt}+\delta_{kt}-(\mu_{ij_1}+...+\mu_{Ij_I})}{\sigma_{ij_1}+...+\sigma_{Ij_1}}\right)-p=0$$

4.3 Scheduling methods

The following scheduling methods have been derived from the paper by Hans et. al. [7]:

4.3.1 First Fit

The First Fit scheduling method is a straightforward approach to scheduling surgeries. As its name suggests, surgery i from a generated waiting list will be assigned to the OR where it first fits, as long as no overtime is generated, according to constraint (1). If there are no ORs where the surgery fits without generating overtime, then surgery i is scheduled into the OR where the minimum overtime is produced. This method is a standard scheduling system in several healthcare institutions, which does not take into account statistical information, such as the standard deviation of surgeries. For this reason, different heuristic methods have been proposed to analyze whether the use of this information can bring further benefits.

4.3.2 LPT

The Longest Processing Time (LPT) dispatching rule is a variant of the list-scheduling algorithm. The first step of this method is to order the tasks in non-increasing order. In this situation, the waiting list is ordered in a non-increasing order of expected durations. This approach addressed the longer surgeries early, thereby reducing potential delays that shorter surgeries might encounter if scheduled first. Once surgeries are sorted, a similar approach to the First Fit scheduling method is used. It will consider the surgeries from the beginning of the list and try to schedule them in some OR where no overtime is generated. If there are no ORs where this is possible, it will similarly aim to find the one where the least overtime is produced.

4.3.3 Regret-based random sampling method

The regret-based random sampling method is a particular sampling procedure that depends on a priority calculation and a "regret" of scheduling a certain surgery. Sampling procedures first orders surgeries in non-increasing order. From here, Z surgeries are selected for dispatching from the beginning of the list. Out of these Z surgeries, we check whether each surgery i fits a certain OR, without generating overtime, as defined in constraint (1). If this is the case, a priority v_i for surgery i is computed, and a best OR is selected from the possible options. If there are no ORs where surgery i would fit without generating overtime, we select the OR where the minimum overtime is generated, and schedule surgery i in this OR immediately.

Priorities, in this case, are based on the slack time assigned to ORs. If some OR has already had surgeries assigned to it, then it contains a slack time δ_{kt}, defined in equation (9). If surgery i was to be assigned to this OR, then the slack time would be defined as δ_{kt+i}. Define

$$\Delta_{kt} = \delta_{kt+i} - \delta_{kt}$$

as the difference in slack times in OR-plan (k, t) if surgery i was to be included. Then, the benefit of not planning surgery i into an empty OR, but into a filled OR is defined as

$$\Omega_{ikt} = \delta_i - \delta_{kt},$$

where δ_i is defined as the slack generated for surgery i only.

Using these values, we define the priority of surgery i as

$$v_i = \max_{kt}\{\Omega_{ikt}\}.$$

If there are remaining surgeries from the selected Z surgeries, we proceed as follows. Each surgery i has now an associated priority v_i and a best OR selected. Now, regret-based random sampling is applied, as this method defines how the drawing probabilities P_i are calculated for each surgery i. As its name suggests, a regret is calculated based on the priorities. The regret of surgery i can be described as the difference between its priority, and the worst of all surgery priorities:

$$w_i = v_i - \min_j\{v_j\}.$$

From here, we define the probability of surgery i being selected as

$$P_i = C \cdot (1 + w_i)^\alpha,$$

where C is the normalization constant defined as $1/\sum_j(1+w_i)^\alpha$, and α being a bias factor. Then, based on these drawing probabilities, a surgery is drawn and scheduled in its best OR selected previously.

The list of Z surgeries is then replenished and the algorithm repeats until the whole sorted waiting list has been scheduled. This procedure is carried out for several samples, from which, the best schedule based on overtime is then selected as the output.

5 Results

5.1 Data handling

We can run the derived surgery scheduling model on real data to test the effects of approximating surgery durations using mixture distributions. For this, we will make use of "Surgical case mixes and distributions of perioperative surgical process durations for German hospitals" [9]. This data set assigns a particular distribution, with its corresponding parameters, to each of the standard processes in a surgery, namely, anesthesia induction time, anesthesia emergence time, surgical lead-in, incision-to-closure time, surgical lead-out, and closure-to-incision time.

These steps contribute to a specific procedure that must be followed to perform a surgery. These procedures are described with OPS codes (Operation and Procedure Code) in the data of German hospitals. Each OPS code contains at most 6 characters, representing the most detailed level of procedure description. We get the surgical category on the right-hand side of the code, for example, 5-916. From here, the complete OPS code specifies some procedure/intervention for that specific surgery, for example, 5-916.a0, 5-916.a1, etc. These OPS codes are usually identified and assigned post-surgery and are mainly used for reimbursement purposes [9]. The reason for this is, as mentioned previously, surgeons cannot guarantee which procedure will be selected for the surgery planned.

For this problem, we are interested in the distribution of the surgery duration as a whole, as, we want to obtain the distribution of different OPS codes, that will eventually contribute to the mixture distribution of a general surgery. The way to do this is by running simulations, such that we can add up all the individual processes per surgery and then approximate some distribution for this simulated data. This procedure is done as follows:

First, we generate random samples for each process (per surgery), according to its designated distribution. In this case, we run 1000 samples for each process and surgery. Then, we can add the 5 processes per simulated data, leading to 1000 total surgery duration samples. From here, we fit this data into some distribution. For convenience of calculations, we choose surgery durations to follow a normal distribution. Using the Maximum Likelihood Estimation (MLE), it is possible to find the parameters that best fit this distribution.

In figure 2 we can see how the data of a specific surgery (surgery 5-916) with 6 possible procedures is approximated by a normal distribution:

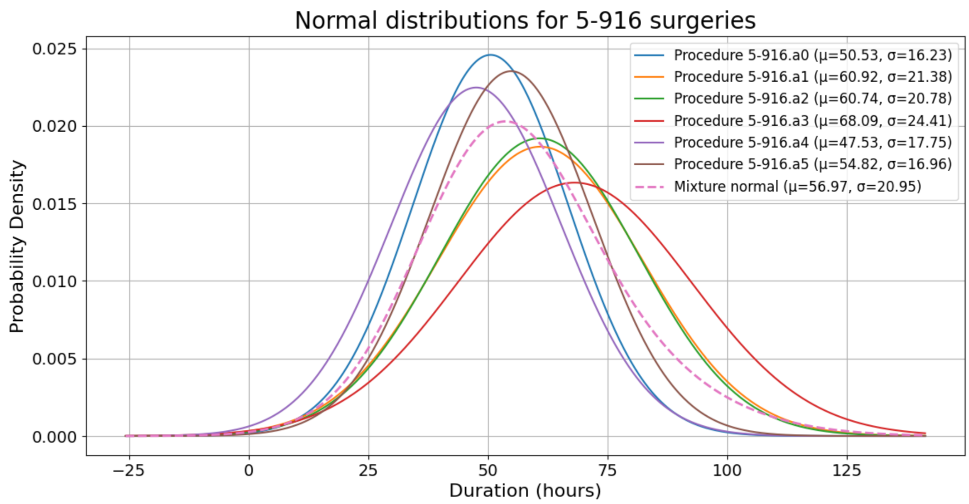

FIGURE 2: Example of mixture distribution showing unimodal curve

5.2 Data manipulation

In the dataset applied to this study, surgeries with several possible procedures have similar expected durations. For this reason, the mixture distributions of surgery durations show unimodal distributions. The data provided has been manipulated to test the effects of multimodal distributions in modelling. As seen in figure 2, surgery 5-916 generates a unimodal distribution due to the mixture of the six possible procedures. To modify this data such that a multimodal distribution become visible, we let $\mu_{ij} = \mu_{ij} + j \cdot \sigma_{ij}$. So, we add the standard deviation of the generated data by a factor j, representing the index of procedure j for surgery i. The result of this operation can be observed in figure 3:

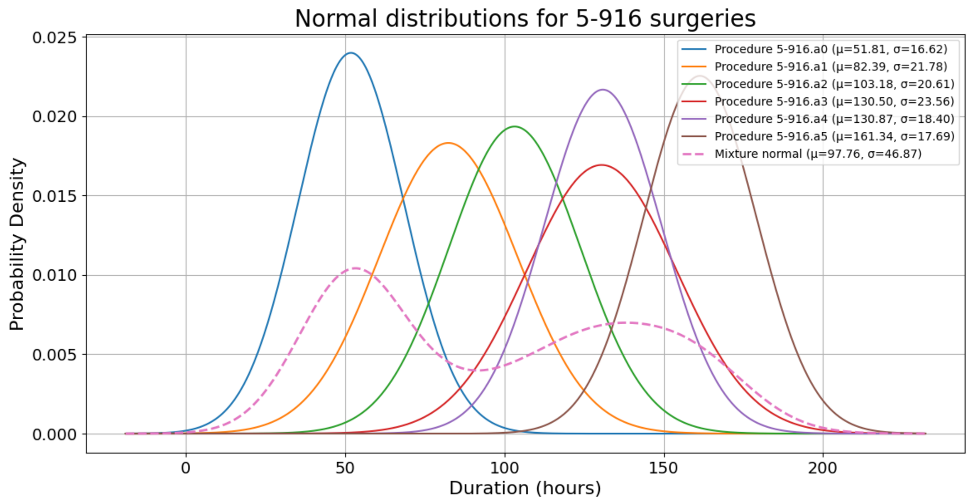

FIGURE 3: Multimodal mixture distribution as a consequence of manipulated data

As we can see, surgery 5-916 now resembles a multimodal distribution, in contrast to the unimodal distribution shown in Figure 2. The simulations in the following sections are therefore based on this processed data.

5.3 Test approach

5.3.1 Producing surgery schedules

We can test the performance of surgery scheduling using mixture distributions against single, normal distributions. For this, a waiting list will be randomly generated containing 400 surgeries, each with a corresponding expected duration and standard deviation derived from the mixture distribution of normal variables. Two methods for calculating slack time will be then tested.

The first method to calculate slack time will resemble a single normally distributed variable. This means that the parameters obtained from the mixture distribution will be used as the parameters of a normal distribution. This will lead us to a slack time that can be derived from equation (6), where there is a single contributing procedure. This means that slack time produced by surgeries $1, ..., n$, scheduled in OR-plan (k, t), will become of the form:

$$P(X_1 + ... + X_n > \mu_{kt} + \delta_{kt}) = 1 - \Phi(\frac{\delta_{kt}}{\sigma_{kt}}) = p,$$

where $\mu_{kt} = \mu_1 + ... + \mu_n$ and $\sigma_{kt} = \sqrt{\sigma_1^2 + ... + \sigma_n^2}$. Solving this further we can derive the expression defined by Hans et. al.[7], namely,

$$\delta_{kt} = \beta \sqrt{\sum_{j=1}^{n} \sigma_j^2}, \tag{10}$$

where β represents the parameter that influences the probability of overtime, defined by

$$\beta = \Phi^{-1}(1 - p).$$

The second method that we will test to produce schedules is based on the mixture distribution of the generated surgeries. This will lead to the slack time defined in equation (6). As equation (8) represents an approximation for the slack time, we can apply numerical methods to derive a precise value for slack times in the different OR-plans, as described in section 4.2.

We will test these two methods, applying the three different scheduling methods described in section 4.3. In addition, the attempt is to schedule the generated waiting list during a time span of 4 months, with 5 working days per week (with 7.5 working hours per day) Therefore, the question is, how can 1395 hours of surgeries be scheduled in a time span of 1800 hours in the most efficient way?

In addition, the probability of overtime chosen will influence the slack time assigned to ORs. We let this probability of overtime $p = 0.3$.

5.3.2 Testing surgery schedules produced

Once surgery schedules have been produced using the two different methods for slack time calculations, and the three different scheduling methods, they can be tested on a simulated data instance.

To perform these tests, we first use the generated waiting list of 400 surgeries to generate a random duration for each surgery in the list. For each surgery, a certain procedure will be selected based on the probabilities/weights of the procedures, as described in Section 4.1, which will therefore define the simulated surgery duration. Schedules produced will be tested on a simulation, representing the average values over 500 replications of the generated waiting list.

5.4 Parameter settings

For the regret-based random sampling method, several parameters have to be set. Namely, the bias factor α, the number of samples, and the surgeries Z to process in each iteration.

We have tested different combinations of α, number of samples, and Z. For this, we attempt $\alpha \in \{2, 5, 10, 25, 50, 100\}$, samples $\in \{5, 10, 15, 25, 35, 50, 100, 150\}$, and $Z \in \{2, 7, 10, 15, 25, 50, 75, 100\}$. The results obtained for these combinations can be seen in section A.3.

A summary of the selected parameters and settings to produce schedules can be seen in Table 1:

Parameter	**Value**
Surgical specialty	General
Initial guess (Newton-Raphson)	20
Number of ORs	3
Working hours	7,5 hours
Working days	80 (4 months)
Total working hours	1800 hours
Total surgery time (generated waiting list)	1395 hours
Probability of overtime (p)	0.3
α (regret-based)	100
Z (regret-based)	100
Number of samples (regret-based)	10

Table 1: Parameter settings to produce schedules

5.5 Comparison of two different methods for slack time calculations

5.5.1 Applying First Fit scheduling method

The First Fit method described in section 4.3.1 uses a simple approach to schedule surgeries. Table 2 shows the different results obtained when applying the schedules produced by First Fit scheduling when the different approaches for slack time calculations are applied:

	Using original slack	**Using slack for mixture distr.**
Total overtime (hours)	175.49	168.32
OR-plans with overtime	50	51
Overtime frequency in non-empty OR-plans (%)	25.77	25.89
Total free time (hours)	592.96	585.80
Total free OR-plans	46	43
Total free time in non-empty OR-plans (hours)	247.96	263.30

Table 2: Results of simulated data using FF

As we can see, total overtime is reduced by 7 hours when slack time for mixture distributions is applied. In addition, we see that only one extra OR-plan generates overtime. This means that, not only there is a reduction in overtime, but we also have less overtime in the OR-plans that do experience this. If we observe the free time generated, we see that the slack generated with mixture distributions has reduced the total free time, also reducing the number of free OR-plans. More meaningfully, we see that the total free time in non-empty ORs has increased. This means that the schedule was less efficient when the slack for mixture distributions was considered, as there are larger gaps in the filled OR-plans than when the original slack time model was applied.

To test the effect of the two methods applied to calculate slack time, when producing schedules using the First Fit method in a fair manner, we deviate the probability of overtime p, such that we can equate the number of free OR-plans in both models.

	Using original slack ($p = 0.25$)	**Using slack for mixture distr. ($p = 0.3$)**
Total overtime (hours)	142.71	168.32
OR-plans with overtime	55	51
Overtime frequency in non-empty OR-plans (%)	27.78	25.89
Total free time (hours)	628.18	585.80
Total free OR-plans	42	43
Total free time in non-empty OR-plans (hours)	313.18	263.30

Table 3: Results of simulated data with adjusted p (FF)

As we can observe in Table 3, we set $p = 0.25$, in order to bring the number of free OR-plans as close as possible. Here, we can better observe the effects of the two applied models. First of all, we notice that the total overtime is larger for the slack generated by mixture distributions, but, on the contrary, it occurs in less OR-plans than when the original slack is applied. Additionally, notice that the overtime frequency when using the original slack is of 25.77 %, meaning that is is surpassing the bound set by management. We observe there are also less free OR-plans for the mixture model, but, the bigger benefit is that the total free time in non-empty OR-plans reduces significantly, namely, by almost 50 hours.

5.5.2 Applying LPT scheduling method

The LPT scheduling method first orders the waiting list in non-increasing order. From here, surgeries are scheduled as described in section 4.3.2. Table 4 shows how the two methods of slack time calculations influence results.

	Using original slack	**Using slack for mixture distr.**
Total overtime (hours)	175.65	172.31
OR-plans with overtime	56	50
Overtime frequency in non-empty OR-plans (%)	29.47	25.91
Total free time (hours)	593.13	589.78
Total free OR-plans	50	47
Total free time in non-empty OR-plans (hours)	218.13	237.28

Table 4: Results of simulated data using LPT

The total overtime generated for both models is approximately equal, although there is a larger number of OR-plans with overtime when the original slack is applied. In addition, the overtime frequency is smaller when the mixture model is applied. We also notice that the total free time in non-empty OR-plans is larger when the original slack is used.

	Using original slack ($p = 0.25$)	**Using slack for mixture distr. ($p = 0.3$)**
Total overtime (hours)	162.79	172.31
OR-plans with overtime	57	50
Overtime frequency in non-empty OR-plans (%)	29.53	25.91
Total free time (hours)	602.67	589.78
Total free OR-plans	47	47
Total free time in non-empty OR-plans (hours)	250.17	237.28

Table 5: Results of simulated data with adjusted p (LPT)

When we equate the number of free OR-plans by adjusting the probability of overtime in the model using original slack to $p = 0.25$, we can better draw conclusions from these results. As we observe, the total overtime generated is greater when we apply the mixture model, although we see that this is spread over 7 fewer OR-plans than in the original slack. In addition, the total free time is reduced in the mixture model, but this also leads to a decrease in free time in non-empty OR-plans, which is desired, as we want to fill up the OR-plans as much as possible, without generating overtime.

5.5.3 Applying regret-based random sampling scheduling method

Table 17 presents the application of regret-based random sampling, with the parameter setting described in section 5.4.

	Using original slack	**Using slack for mixture distr.**
Total overtime (hours)	198.94	139.94
OR-plans with overtime	69	48
Overtime frequency in non-empty OR-plans (%)	36.32	20.25
Total free time (hours)	591.40	599.26
Total free OR-plans	50	3
Total free time in non-empty OR-plans (hours)	216.40	576.76

Table 6: Results of simulated data using regret-based random sampling

When using the regret-based random sampling method to schedule surgeries, the two methods of calculating slack time have the most significant impact. On one hand, the original model generates more overtime than the mixture model, also spread over more OR-plans, namely, 21 more OR-plans experience overtime. In addition, the original model makes the overtime frequency in non-empty OR-plans go above the established 30% bound. There is a drastic change in the number of free OR-plans when the slack for mixture distributions is applied, showing a reduction of 47 OR-plans in comparison to the original model. This can be a consequence of the reduction of overtime. As this is the main focus, then surgeries are scheduled in such a way that we obtain the least overtime, meaning that the number of free OR-plans is consequently reduced.

	Using original slack (p = 0.014)	**Using slack for mixture distr. (p = 0.3)**
Total overtime (hours)	128.17	139.94
OR-plans with overtime	35	48
Overtime frequency in non-empty OR-plans (%)	14.77	20.25
Total free time (hours)	587.79	599.26
Total free OR-plans	3	3
Total free time in non-empty OR-plans (hours)	565.29	576.76

Table 7: Results of simulated data with adjusted p (regret-based)

When equating the number of free OR-plans by setting the overtime probability to $p = 0.014$ for the original model, we observe how the original model actually produces less overtime, spread over fewer OR-plans than the mixture model. What we also see is that the overtime frequency does not satisfy the desired boundary for the original slack, which should not surpass the bound of 1.4%, and in this case, it reaches almost 15%. On the contrary, the mixture model actually produces an overtime frequency of 20%, 10% less than the established bound.

5.6 Comparison of different scheduling methods

After observing how different slack time calculations can influence results, we can now focus on the influence that different scheduling methods have on surgery scheduling, using the new, derived method for the slack time assigned.

We observe in Table 8 the three scheduling methods described previously, when the mixture model is applied to generate schedules.

	First Fit	**LPT**	**Regret-based**
Total overtime (hours)	168.32	172.31	139.94
OR-plans with overtime	51	50	48
Overtime frequency in non-empty OR-plans (%)	25.89	25.91	20.25
Total free time (hours)	585.80	589.78	599.26
Total free OR-plans	43	47	3
Total free time in non-empty OR-plans (hours)	263.30	237.28	576.76

Table 8: Comparison of different scheduling methods (simulated data, 500 samples)

As we can observe, is it clear that the regret-based random sampling method generates the least overtime, in comparison to the First Fit method, and the LPT scheduling rule. In addition, this overtime is spread over the least number of OR-plans, namely, regret-based random sampling generates overtime in 48 OR-plans. None of the schedule methods goes over the bound of OR-plans with overtime of 30%, but we notice that for regret-based random sampling, this proportion is significantly reduced, namely, by 10%. One disadvantage that the regret-based random sampling method shows is that it generates a small number of free OR-plans, in comparison to the other two scheduling methods tested. Only three OR-plans are free, and we also see it consequently produces about double the amount of free time in non-empty OR-plans. This is not desirable, as we want to fill OR-plans most efficiently without generating overtime.

	First Fit ($p = \mathbf{0.06}$)	**LPT** ($p = \mathbf{0.05}$)	**Regret-based** ($p = \mathbf{0.3}$)
Total overtime (hours)	195.07	157.96	139.94
OR-plans with overtime	41	33	48
Overtime frequency in non-empty OR-plans (%)	17.30	13.98	20.25
Total free time (hours)	595.38	554.08	599.26
Total free OR-plans	3	4	3
Total free time in non-empty OR-plans (hours)	572.88	524.08	576.76

Table 9: Comparison of scheduling methods with adjusted p

Table 9 shows a fair comparison between the three scheduling methods described. By setting $p = 0.06$ for the First Fit method and $p = 0.05$ for the LPT scheduling rule, we equate the number of free OR-plans to be able to draw conclusions on the other parameters. We see that again, the regret-based random sampling method generates the least overtime, although this occurs in more OR-plans than for the other two methods. We also notice that the overtime frequency is below the bound of 30% overtime for regret-based. The other two scheduling methods have more OR-plans with overtime than the desired amount based on the probabilities p.

6 Conclusions and further research

The experiments performed reveal insights into the effectiveness of different scheduling methods and slack time calculation models in surgical scheduling.

Comparing slack time calculation models, the use of mixture distributions generally leads to a reduction in total overtime across all scheduling methods. Consequently, it also leads to less efficient scheduling with larger gaps in filled OR-plans. This inefficiency arises as a consequence of the mixture distribution model assigning larger slack times to OR-plans. Therefore, this is an effective method for reducing overtime, although it may not fill operating rooms as compactly as the original model does.

The regret-based random sampling method in combination with slack time calculations using mixture distributions outperforms the First Fit and LPT methods in terms of minimizing overtime and the number of OR-plans with overtime. Specifically, it achieves the lowest total overtime and reduces the overtime frequency below the specified bound. However, this method also results in a considerably lower number of free OR-plans and an increase in free time within non-empty OR-plans. This indicates that there is a trade-off between reducing overtime and optimizing OR-plan utilization.

For future research, the parameter settings for the regret-based random sampling method could be optimized separately for the original model, as the experiments in this study applied the optimal settings for the mixture model to all schedules, including the original model. Also, further investigation of the initial guess for the Newton-Raphson algorithm could improve the resulting slack time calculations. The mixture model could also be expanded to incorporate lognormal distributions, aiming to better reflect the distribution of surgery durations [1]. Currently, this model only considers outpatients and surgeries from a static waiting list. Future studies could ideally include emergency patients and surgeries, as well as account for staff limitations.

References

[1] Modeling the uncertainty of surgical procedure times: comparison of log-normal and normal models - PubMed.

[2] A robust estimation model for surgery durations with temporal, operational, and surgery team effects | Health Care Management Science.

[3] scipy.optimize.fsolve — SciPy v1.13.1 Manual.

[4] Brian S. Everitt. Finite Mixture Distributions. In *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd, 2005. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0470013192.bsa231.

[5] Juan M Gers. Distributions systems analysis and automation, Chapter 3.5.2 Newton-Raphson Method.

[6] Francesca Guerriero and Rosita Guido. Operational research in the management of the operating theatre: a survey. *Health Care Management Science*, 14(1):89–114, March 2011.

[7] Erwin Hans, Gerhard Wullink, Mark van Houdenhoven, and Geert Kazemier. Robust surgery loading. *European Journal of Operational Research*, 185(3):1038–1050, March 2008.

[8] Mehdi A. Kamran, Behrooz Karimi, and Nico Dellaert. Uncertainty in advance scheduling problem in operating room planning. *Computers & Industrial Engineering*, 126:252–268, December 2018.

[9] Grigory Korzhenevich and Anne Zander. Surgical case mixes and distributions of perioperative surgical process durations for German hospitals, October 2022.

[10] Jerrold H. May, David P. Strum, and Luis G. Vargas. Fitting the Lognormal Distribution to Surgical Procedure Times*. *Decision Sciences*, 31(1):129–148, March 2000.

[11] Jose M. Molina-Pariente, Erwin W. Hans, Jose M. Framinan, and Tomas Gomez-Cia. New heuristics for planning operating rooms. *Computers & Industrial Engineering*, 90:429–443, December 2015.

[12] Samuel Negash, Endale Anberber, Blen Ayele, Zeweter Ashebir, Ananya Abate, Senait Bitew, Miliard Derbew, Thomas G. Weiser, Nichole Starr, and Tihitena Negussie Mammo. Operating room efficiency in a low resource setting: a pilot study from a large tertiary referral center in Ethiopia. *Patient Safety in Surgery*, 16(1):3, January 2022.

[13] Anne M. Schouten, Steven M. Flipse, Kim E. Van Nieuwenhuizen, Frank Willem Jansen, Anne C. Van Der Eijk, and John J. Van Den Dobbelsteen. Operating Room Performance Optimization Metrics: a Systematic Review. *Journal of Medical Systems*, 47(1):19, February 2023.

[14] David P. Strum, Jerrold H. May, Allan R. Sampson, Luis G. Vargas, and William E. Spangler. Estimating times of surgeries with two component procedures: comparison of the lognormal and normal models. *Anesthesiology*, 98(1):232–240, January 2003.

[15] Shuwan Zhu, Wenjuan Fan, Shanlin Yang, Jun Pei, and Panos M. Pardalos. Operating room planning and surgical case scheduling: a review of literature. *Journal of Combinatorial Optimization*, 37(3):757–805, April 2019.

A Appendix

A.1 QR code to source code

FIGURE 4: QR-code to python files

A.2 Pseudocodes scheduling methods

Algorithm 1 First Fit Scheduling

Initialization: generate waiting list
for surgery in waiting list **do**
    if surgery fits without generating overtime **then**
        assign surgery to first OR found
    else
        find OR where the least overtime is generated
        assign surgery to OR found
    end if
end for

Algorithm 2 LPT dispatching rule for scheduling

```
Initialization: generate waiting list
Sort waiting list in non-increasing order
for surgery in sorted waiting list do
    if surgery fits without generating overtime then
        assign surgery to first OR found
    else
        find OR where the least overtime is generated
        assign surgery to OR found
    end if
end for
```

Algorithm 3 Regret-based random sampling procedure

```
Initialization: generate waiting list
Sort waiting list in non-increasing order
for number of samples do
    select Z surgeries for dispatching,
    for surgery in Z surgeries do
        if surgery fits without generating overtime then
            calculate priority v_i for surgery i
            select the best OR for surgery i
        else
            find OR where the least overtime is generated
            assign surgery to OR found
        end if
    end for
    if some surgeries were not scheduled from Z surgeries then
        calculate the regret for each surgery
        calculate drawing probabilities
        draw a surgery
        assign the drawn surgery to its best OR selected
    end if
end for
```

A.3 Parameter setting for regret-based random sampling

(Z,α)	**(5,2)**	**(5,5)**	**(5,10)**	**(5,25)**	**(5,50)**	**(5,100)**
Total overtime (hours)	262.99	262.99	262.99	262.99	262.99	262.99
OR-plans with overtime	52	52	52	52	52	52
Overtime frequency (%)	21.67	21.67	21.67	21.67	21.67	21.67
Total free time (hours)	392.93	392.85	392.87	392.88	392.87	392.88
Total slack time (hours)	274.87	274.94	274.92	274.92	274.92	274.92
Total free OR-plans	3	4	5	4	3	6
Total running time (minutes)	5.83	5.72	5.67	5.45	5.52	5.53

TABLE 10: Comparison of results with varying α, Z = 5 (slack for mixtures)

(Z,α)	(2,100)	(7,100)	(10,100)	(15,100)	(25,100)	(50,100)	(75,100)	(100,100)
Total overtime (hours)	262.99	262.99	262.99	262.99	262.99	262.99	262.99	262.99
OR-plans with overtime	52	52	52	52	52	52	52	52
Overtime frequency (%)	21.67	21.67	21.67	21.67	21.67	21.67	21.67	21.67
Total free time (hours)	392.89	392.79	392.84	392.77	392.81	392.68	392.60	392.39
Total slack time (hours)	274.91	274.99	274.95	275.03	274.98	275.12	275.20	275.40
Total free OR-plans (days)	4	4	4	6	4	6	8	10
Total running time (minutes)	9.19	10.68	11.63	13.09	15.67	20.28	25.31	28.41

Table 11: Comparison of results with varying Z, $\alpha = 100$ (slack for mixtures)

Samples	5	10	15	25	35	50	100	150
Total overtime (hours)	262.99	262.99	262.99	262.99	262.99	262.99	262.99	262.99
OR-plans with overtime	52	52	52	52	52	52	52	52
Overtime frequency (%)	21.67	21.67	21.67	21.67	21.67	21.67	21.67	21.67
Total free time (hours)	392.49	392.32	392.51	392.40	392.39	392.54	392.59	392.58
Total planned slack (hours)	275.30	275.47	275.28	275.39	275.40	275.26	275.20	275.22
Total free OR-plans (days)	6	10	3	7	6	5	6	6
Total running time (minutes)	15.90	37.82	55.10	84.22	106.38	132.32	343.15	412.13

Table 12: Comparison of results with varying number of samples, (Z, α) = (100,100) (slack for mixtures)

A.4 Comparisons of different schedules produced

A.4.1 First Fit Schedules

	Using original slack	Using slack for mixture distr.
Total overtime (hours)	220.13	262.99
OR-plans with overtime	37	52
Overtime frequency (%)	15.42	21.67
Total free time (hours)	425.09	390.46
Total running time (minutes)	0.11	2.97
Total planned slack (hours)	199.84	277.33
Total free OR-plans	46	43

Table 13: Comparison of schedules produced for the two methods for slack time

	Using original slack	Using slack for mixture distr.
Total overtime (hours)	44.64	94.67
OR-plans with overtime	-13	1
Overtime frequency (%)	-5.41	0.42
Total free time (hours)	-167.87	-195.34
Total free OR-plans (days)	0	0

Table 14: Difference between plan and simulated data using FF

A.4.2 LPT schedules

	Using original slack	Using slack for mixture distr.
Total overtime (hours)	220.13	262.99
OR-plans with overtime	37	52
Overtime frequency (%)	15.42	21.67
Total free time (hours)	426.52	391.08
Total running time (minutes)	0.13	2.56
Total planned slack (hours)	198.42	276.72
Total free OR-plans	51	47

Table 15: Comparison of two methods for slack using LPT scheduling method

	Difference (Original Slack)	Difference (Mixture Slack)
Total overtime (hours)	44.48	90.68
OR-plans with overtime	-19	2
Overtime frequency (%)	7.91	0.84
Total free time (hours)	-166.61	-198.7
Total free OR-plans (days)	1	0

Table 16: Difference between plan and simulated data using LPT

A.4.3 Regret-based random sampling schedules

	Using original slack	Using slack for mixture distr.
Total overtime (hours)	220.13	262.99
OR-plans with overtime	37	52
Overtime frequency (%)	15.42	21.67
Total free time (hours)	426.58	392.55
Total running time (minutes)	3.52	20.29
Total planned slack (hours)	198.35	275.25
Total free OR-plans	50 days	3 days

Table 17: Comparison of two methods for slack using regret-based random sampling scheduling method

	Difference (Original Slack)	Difference (Mixture Slack)
Total overtime (hours)	44.48	91.03
OR-plans with overtime	-19	2
Overtime frequency (%)	-7.91	0.84
Total free time (hours)	-166.55	-196.89
Total free OR-plans (days)	0	0

Table 18: Difference between plan and simulated data using regret-based random sampling

A.5 Comparison of different scheduling methods

	First Fit	**LPT**	**Regret-based**
Total overtime (hours)	262.99	262.99	262.99
OR-plans with overtime	52	52	52
Overtime frequency (%)	21.67	21.67	21.67
Total free time (hours)	390.46	391.08	392.55
Total slack time (hours)	277.33	276.72	275.25
Total free OR-plans (days)	43	47	3
Total running time (minutes)	2.97	2.56	20.29

Table 19: Comparison of different scheduling methods using slack defined for mixtures

	First Fit	**LPT**	**Regret-based**
Total overtime (hours)	94.67	90.68	91.03
OR-plans with overtime	1	2	2
Overtime frequency (%)	0.42	0.84	0.84
Total free time (hours)	-195.38	-314.86	-314.52
Total free OR-plans (days)	0	0	0

Table 20: Difference between plan and simulated data for mixture slack (3 scheduling methods)