NTNU
Norwegian University of
Science and Technology

# A Stochastic Programming Approach to Daily Surgery Scheduling Under Uncertainty at a Norwegian Hospital

**Marthe Siren Anvik**
**Jon Erik Medhus**
**Mikkel Treu Os**

Industrial Economics and Technology Management
Submission date: June 2016
Supervisor: Asgeir Tomasgard, IØT

Norwegian University of Science and Technology
Department of Industrial Economics and Technology Management

# Preface

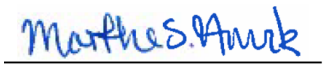
The submission of this thesis completes our Master of Science degree in Industrial Economics and Technology Management at The Norwegian University of Science and Technology. The thesis is motivated by Lars Hellemo, research scientist at SINTEF, in collaboration with representatives from St. Olavs Hospital, but the final problem formulation was developed independently by the authors.

Several individuals deserve special acknowledgements for their contribution to the completion of this thesis. First and foremost, we want to extend our gratitude towards our supervisor, Asgeir Tomasgard, who has greatly guided our learning process while at the same time granting us valuable autonomy. Secondly, we sincerely appreciate the inspiration and assistance from Lars Hellemo, Kjetil Trovik Midthun, and Michal Kaut at SINTEF Technology and Society.

The helpfulness of St. Olavs Hospital's personnel has also been essential throughout the duration of the project. In particular, we would like to thank Liv-Inger Stenstad and Stian Saur for facilitating our research. The opportunity to gain insights into practical applications of operations research has been highly rewarding.

We are forever grateful for the unwavering support and motivation of our families, friends, and classmates, through this intense journey culminating in our academic climax.

*Trondheim, 2016*



Marthe Sirén Anvik

Jon Erik Medhus

Mikkel Treu Os

# Abstract

This thesis proposes two mathematical stochastic optimisation models handling two different aspects of uncertainty in the daily problem of deciding start times for a set of surgeries in a single operating room. The uncertainty related to surgery durations are modelled by scenarios generated using moment-matching on a statistical basis of detailed data on almost 90 000 past surgeries at St. Olavs Hospital.

Both problems have the objective of minimising the expected cost of waiting time, idle time and overtime. The first is based on the hypothesis that, for some surgeries, the uncertainty in duration depends on the start time, a hypothesis that is tested using a two-sample Kolmogorov-Smirnov test of independence. The model formulated to solve this problem is a mixed integer program including decision-dependent uncertainty, something that complicates the problem considerably on a computational level. The second problem includes both stochastic surgery durations and stochastic arrival of emergency patients, a combination that, to the authors' best knowledge, has not yet been covered by existing literature.

The model formulations are tightened by the introduction of several valid inequalities, the most effective of which is a cut strengthening the link between two types of sequencing variables. On average this reduces execution time by roughly 40%. To further overcome the computational challenges posed by decision-dependent uncertainty, and in order to investigate alternative solution methods, we test the heuristics most commonly used in surgery scheduling literature. These tests conclude that the best performance is by a heuristic sorting surgeries by increasing variance, while the popular Bailey-Welch heuristic shows poor performance.

Based on insights gained from our practical analysis, and with foundation in literature, we propose a decision rule stating that you should sequence the surgeries by ascending variance, and set start times with intervals equal to each surgery's mean duration. We proceed to show that this rule captures large parts of the total potential gain from solving the stochastic models using optimisation. Averaged over our problem instances, the model results signify that the case hospital can reduce waiting time, idle time and overtime by 160, 22, and 16 minutes per day, respectively. The contribution of this thesis is thus both on a practical and a theoretical level.

# Sammendrag

Denne masteroppgaven presenterer to stokastiske optimeringsmodeller som håndterer to ulike aspekter av det daglige operasjonsplanleggingsproblemet som går ut på å bestemme starttidspunkter for en mengde operasjoner på ett operasjonsrom. Usikkerheten forbundet med operasjoners varighet er modellert ved hjelp av scenarioer generert ved bruk av moment-matching på et statistisk grunnlag bestående av data fra nesten 90 000 operasjoner utført ved St. Olavs Hospital.

Målfunksjonen i begge problemene minimerer forventet kostnad relatert til ventetid, dødtid og overtid. Det første problemet er basert på en hypotese om at usikkerheten i varigheten til noen operasjoner avhenger av starttidspunktet, en hypotese som undersøkes ved hjelp av en Kolmogorov-Smirnov-test. Modellen som løser dette problemet formuleres som et blandet heltallsprogram, og inkluderer beslutningsavhengig usikkerhet som øker den beregningsmessige kompleksiteten. Det andre problemet hensyntar usikker ankomst av akuttpasienter i tillegg til usikker varighet, en kombinasjon som, så vidt oss bekjent, ikke har blitt dekket i eksisterende litteratur.

Modellenes mulighetsområder reduseres ved hjelp av flere gyldige ulikheter, hvor det mest effektive kuttet styrker forbindelsen mellom to typer sekvensvariabler. I gjennomsnitt reduserer dette kjøretiden med omtrent 40%. For å overkomme de beregningsmessige utfordringene som oppstår som følge av beslutningsavhengig usikkerhet, og for å undersøke alternative løsningsmetoder, tester vi de mest brukte heuristikkene fra litteraturen. Fra disse testene konkluderes det med at heuristikken som sorterer operasjoner etter økende varians gir best ytelse, mens den populære Bailey-Welch heuristikken gir dårlige resultater.

Basert på innsikt hentet fra praktiske analyser, og med forankring i relevant litteratur, foreslår vi en ny beslutningsregel om å ordne operasjonene etter økende varians og sette starttider med intervaller tilsvarende hver operasjons gjennomsnittlige varighet. Vi viser at denne regelen fanger opp store deler av den totale potensielle forbedringen man kan oppnå ved å løse de stokastiske modellene ved hjelp av modellering. I snitt for våre instanser, viser modellresultatene at sykehuset kan redusere ventetid, dødtid og overtid med henholdsvis 160, 22 og 16 minutter per dag. Bidraget til denne masteroppgaven er derfor både på et praktisk og et teoretisk nivå.

# Contents

# Abbreviations

| | |
|---|---|
| ASP | Advance scheduling problem |
| ALSP | Allocation scheduling problem |
| CV | Coefficient of variance |
| EVPI | Expected value of perfect information |
| EVPP | Expected value of planning with phases |
| EEV | Expectation of the expected value problem |
| I.i.d. | Independent and identically distributed |
| MIP | Mixed integer linear program |
| SA | Simulated annealing |
| SP | Objective value of the stochastic program |
| SPNP | Value of stochastic solution without phases |
| VSS | Value of stochastic solution |
| WS | Value of wait-and-see problem |

# List of Figures

# List of Tables

# Part I

# Background Information

# Chapter 1

# Introduction

Health care services are a major cost driver in national budgets, and the sector is continuously challenged to provide high quality treatment with limited resources [1]. The governmental funding of the somatic specialist health care services in Norway consists of a basic part and an activity-based part. In 2015, each of these constitute about 50 percent of the total funding. The activity-based funding makes the budgets depend on the number of patients and what type of treatment they receive [2]. Surgeries constitute, in this way, a critical part of the funding, as well as the most expensive activity in many hospitals. The high costs are first of all a result of expensive resources used in surgery. Secondly, surgery execution affects many other activities in the hospital [1]. Careful planning and scheduling is therefore crucial to be able to manage human material resources efficiently and to provide appropriate treatment [1, 3]. Research indicates that inadequate schedules may be one of the main factors contributing to the inefficiencies in the health sector [3].

The inherently uncertain environment of hospitals makes surgery scheduling a complex task. Not only do the uncertain surgery durations make the schedules prone to disruptions, but the unpredictable arrival of emergency patients also poses major challenges on the process of making good plans.

The overall objective of this thesis is to enhance the understanding of how surgery scheduling should account for uncertainty. Specifically, the thesis will investigate a daily problem of scheduling a given set of surgeries in a single operating room. This will be studied in the environment of the Department of Orthopaedic Surgery at St. Olavs Hospital, whence we analyse an extensive collection of surgery data from the past decade.

By applying operations research, we will formulate two different stochastic mathematical programs with the objective of finding the optimal scheduling strategy when surgery durations are uncertain. One of these will incorporate decision-dependent uncertainty, a field that is yet to be given much attention in literature. The other will combine uncertain

surgery durations with uncertain arrival of emergency patients in a multi-stage model. The performance of several heuristics from literature will be evaluated in order to explore alternative solution methods. In addition, we aim to extract from our results a simple decision rule that may contribute to establishing a best practice for surgery scheduling.

This thesis is structured as follows: In Chapter 2, the problems we investigate are put into context by providing background information about the hospital and their current practice in terms of surgery scheduling. Chapter 3 provides a review of the relevant literature in the field of surgery scheduling and classifies our contribution in an academic context. Also, the chapter describes and explains some of the most relevant theoretical topics and solution methods used in the thesis. Then, a generalised problem description is presented in Chapter 4, along with an explanation of the most important assumptions used. Chapter 5 explains and analyses the statistical basis on which the input data to the models is generated. The two models proposed in this thesis are presented in Chapter 6, along with suggestions for strengthening the formulations and heuristics that can help guide the solution process. Then, Chapter 7 provides an extensive computational study, which starts with a discussion of performance measures before it checks the stability of the stochastic models we use. The chapter also analyses our results, both on a computational and on a practical level, and quantifies the value of accounting for uncertainty when scheduling surgeries. Finally, the conclusion of the thesis is found in Chapter 8.

# Chapter 2

# Background

The purpose of this chapter is to provide a context for the problems studied in this thesis, and a description of the case hospital including its current practice in terms of surgery scheduling.

## 2.1 St. Olavs Hospital and the Department of Orthopaedic Surgery

St. Olavs Hospital is located in Trondheim, Norway, with additional services situated in Orkdal, Røros, and Hysnes, and it is integrated with the Norwegian University of Science and Technology (NTNU). It is owned by the Central Norway Regional Health Authority [4], and its activities consist of specialist health care services in both somantic and mental health care. The hospital functions as a local hospital for the population of Sør-Trøndelag, in addition to having both regional and national responsibilities for the population of the three counties of Møre og Romsdal, Sør-Trøndelag and Nord-Trøndelag, performing patient treatment, education and research. Moreover, it treats complex surgical cases referred from other Norwegian hospitals.

The Department of Orthopaedic Surgery performs surgical procedures related to conditions concerning the musculoskeletal system, i.e. diseases and injuries in bones, joints, tendons and muscles [5]. The department employs eleven operating rooms in Trondheim, five in Orkdal, and two in Røros [6]. When referring to a specific physical hospital in this thesis, we will refer to it by its location (e.g. the Røros Hospital), while St. Olavs refers to the hospital organisation as a whole, including all locations. Correspondingly, when referring to the Department of Orthopaedic Surgery this includes the department's activities at all three physical hospitals.

## 2.2 Surgery scheduling at St. Olavs Hospital

The process of surgery scheduling is often split into two separate processes. The first one comprises the allocation of a date, an operating room and a surgical team to all pending surgeries. This is what the literature refers to as the advance scheduling problem (ASP). The second problem is referred to as the allocation scheduling problem (ALSP), and determines the sequence and start times of all surgeries that are to be performed each day on each operating room. The focus of this thesis is on the latter of the two but, in order to provide a context of the problem, the following part describes both of these processes.

### 2.2.1 The current advance scheduling process

Surgery scheduling at the Department of Orthopaedic Surgery is performed by a dedicated team of patient coordinators where each coordinator manages different surgical groups (hand surgeries, prostheses, arthroscopy, etc.). The scheduling is a manual process based mainly on guidelines and agreements within the hospital, resulting from several interdependent planning processes. Within the boundaries of these agreements, the scheduling strategy depends on rules of thumb and private knowledge of each patient coordinator. The scheduling strategy used for one surgery group may therefore differ from the strategy used for another.

Two other schedules pose restrictions on the surgery schedule. Firstly, the work schedule of the surgeons determines when each surgeon is available for surgery as opposed to being occupied with other tasks such as research or lectures. Secondly, the master surgery schedule shows, for every day of the week, which surgery groups are given priority at which operating rooms. Figure 2.1 displays the current master surgery schedule for eight of the operating rooms employed by the Department of Orthopaedic Surgery in Trondheim. Together with the surgeons' work schedule, this sets guidelines for the surgery schedules. The planners also take into account the preference of surgeons to avoid moving between different operating rooms on the same day, and they coordinate resource usage between departments and operating rooms such that all equipment, anaesthesia, etc. are available when needed.

Elective patients are organised in a waiting list, to which they are added upon referral from a general practitioner and an assessment by an orthopaedics specialist. The assessment determines the appropriate surgical procedure, a deadline before which it must be performed, and often which surgical team is to perform it. The waiting list is ordered based on the imminence of each surgery's deadline, and when allocating operating rooms and days, the patient coordinators make sure this complies with the work schedule of the surgeons as well as the master surgery schedule. They generally pick from the top of the list, but they also try to accommodate special preferences that patients may have.

| | Monday | Tuesday | Wednesday | Thursday | Friday |
|---|---|---|---|---|---|
| 1 | Local | Closed | Hand | Local | Closed |
| 2 | Reconstruction | Reconstruction | Plastic | Plastic | Reconstruction |
| 3 | Plastic | Plastic | Plastic | Plastic | Closed |
| 4 | Hand | Plastic | Arthroscopy | Arthroscopy | Arthroscopy |
| 5 | Arthroscopy | Arthroscopy | Arthroscopy | Arthroscopy | Arthroscopy |
| 6 | Back | Back | Back | Hand | Closed |
| 7 | Prosthesis | Prosthesis | Prosthesis | Prosthesis | Misc. |
| 8 | Prosthesis | Prosthesis | Prosthesis | Closed | Closed |

Figure 2.1: Master surgery schedule for eight of the operating rooms at the Department of Orthopaedic Surgery

Table 2.1: Excerpt from the time matrix, durations specified in minutes

| **Surgery type** | **Procedure code** | **Pre-time** | **Knife-time** | **Post-time** |
|---|---|---|---|---|
| Athroscopy | NBK13 | 60 | 40 | 15 |
| Athroscopy | NHL49 | 45 | 60 | 15 |
| Plastic | HAD30 | 45 | 60 | 15 |

### 2.2.2 The current allocation scheduling process

Once it has been determined what surgeries to perform at a given operating room on a given day, we get to the problem of this thesis: deciding the start time for each surgery. A central part of this scheduling process is to determine how much time to allocate. Because the surgery durations are uncertain, the problem of setting appointment times is not a trivial one, as stated by Robinson and Chen [7]. Estimates of the duration of all surgical procedures are given by a time matrix that the hospital has developed based on averages of historical surgery durations. This specifies the expected time required for preparation (pre-time), the surgery itself (knife-time), and wrap-up (post-time), all of which take place in the operating room and therefore cannot be performed in parallel. In addition, the room must be cleaned between surgeries. An excerpt from the time matrix is given in Table 2.1. According to the coordinators, the time matrix is primarily used for pre- and post-times, although they sometimes adjust for patient characteristics such as age, medical conditions, etc. Knife-time durations, on the other hand, differ from surgeon to surgeon to the extent that the coordinators rarely adhere to the time matrix, but instead base the estimates on experience. Naturally, this is especially the case for common surgeries that have been performed several times by the same surgeon.

The current practice in terms of setting start times for surgeries differs from hospital to hospital. In Trondheim, all surgeries are planned back to back, such that the preparation phase of one surgery is set to start exactly when the cleaning after the previous patient is set to finish. In the ideal and completely unrealistic case of all surgeries lasting exactly as

long as predicted, this results in zero idle time and zero waiting time. However, knowing that the durations are unpredictable, this policy is very prone to schedule disruptions, and without the slack needed to absorb potential delays, it is likely to lead to high waiting time. Of course, given that the estimated duration actually provides a realistic average of the durations, you will also have the case of some surgeries lasting shorter than predicted. This may, to some extent, offset the effects of delays but, since patients and equipment are not necessarily ready before their scheduled surgery time, the former effect is expected to dominate. At the Røros hospital, the planners quite often try to eliminate idle time by scheduling the two first surgeries to start at the same time. The second patient simply has to wait for the first one to be completed, guaranteeing a delay as they immediately get behind schedule. This is also likely to lead to high waiting times for both patients and staff. There is no consistent practice, but with both of the mentioned scheduling policies allowing waiting time to avoid idle time, there seems to be a certain understanding that waiting time is preferred to idle time. From the hospital's point of view, this makes sense, but the question of exactly how to trade these off against each other is not trivial. Chapter 3 provides a discussion on how to set these weights, and Chapter 7 analyses how different weight combinations affect the solution of the models we propose.

In addition to the uncertainty in the surgery durations, there is uncertainty related to the arrival of emergency patients requiring surgery on short notice. Since orthopaedic surgeries do not concern vital organs, emergency surgeries within this field usually do not require immediate action but must be treated within one or a few days. The patient coordinators are given a list of emergency patients that they are required to schedule within a deadline that is at least twelve hours ahead. This means that, in the beginning of any given day, they know how many and which emergency patients to schedule, and they can adjust the planned schedule to accommodate for this. During certain events or seasons (e.g. during Easter, when skiing accidents peak) there is time allocated for emergency surgeries but, apart from this, there is no slack or reserved time for such patients. Quite often, the arrival of an emergency patient therefore leads to either overtime or an elective patient being rescheduled to another day.

Another uncertain element that would be expected to affect the scheduling are patients failing to meet up for their scheduled surgery, or patients that arrive late. No-shows and late arrivals lead to high idle time unless there are patients on stand-by that are ready on short notice. According to the hospital, this does not happen often enough for it to be a substantial problem and, when it does, they can often summon an inpatient waiting for surgery to keep the operating room employed. Therefore, no-shows and late arrivals are kept out of the analysis presented in this thesis.

### 2.2.3 Other considerations

In general, when it comes to scheduling the surgeries on a given day at a given operating room, the coordinators can choose freely how to sequence the surgeries. However, for certain patients and surgery types there are extra considerations that need to be accounted for:

1. *Travel distance* - Patients with a long travel distance are usually not scheduled early in the morning or late in the afternoon, with regards to the comfort of the patients.

2. *Preparations* - To avoid operating room idle time, patients that require medical tests, blood samples or other lenghty preparations prior to surgery are not scheduled as the first surgery of the day.

3. *Health conditions* - Patients who, due to their age or medical condition, have problems fasting a long period of time are scheduled early in the morning to make the experience as pleasant as possible.

4. *Surgery complexity* - There is a policy of postponing surgeries to a later day if delays in prior surgeries mean that they cannot be carried out without the incurrence of overtime. Since complex surgeries require more planning and coordination in terms of staffing and equipment, the coordinators reduce the risk of such surgeries being postponed by scheduling them early in the day.

The majority of the surgeries to be scheduled are not subject to any of the considerations above, so the mathematical models proposed in this thesis leave these considerations out. This is important because it facilitates an analysis focusing on the uncertainty aspects we want to investigate, and it accommodates a pertinent computational study. The reader should thus note that the considerations listed above are stated only to provide extra insights to the complexity of the planning process.

## 2.3 The motivation for the two problems

Based on discussions and interviews with planners and other staff at the hostpital, we formulated the hypothesis that the probability distribution for the duration of a given surgery performed by a given surgeon is dependent on the time of the day. For instance, according to the planners, some surgeons might be less efficient in the afternoon, due to weariness and lack of concentration, while other surgeons might perform better later in the day. Especially for complex surgeries, and surgeries that require high precision, the hypothesis is that these effects become considerable and may affect surgery durations. If this hypothesis holds, it could play a significant role when determining surgery schedules because it provides additional information to the planners, who should take this factor into account when determining the schedules. This is the motivation for the first problem

proposed in this thesis. The Phase Model, in Section 6.1, is designed to solve this.

The second problem is motivated by what the medical staff claims to be their biggest challenge in surgery scheduling: stochastic arrival of emergency patients. If time is reserved for potential emergency patients and no one arrives, idle time will occur. Conversely, if no available time is reserved and an emergency patient arrives, waiting time or cancellations are likely to occur. The trade-off between these is difficult to balance, and this is the objective of the Emergency Model, in Section 6.2. We will study the arrival rate of the emergency patients and try to determine an optimal scheduling strategy taking this into account.

# Chapter 3

# Literature review

This chapter will review the most relevant literature and theory related to our problem. We will first give a brief introduction to stochastic programming, in order to establish a theoretical basis for the models we propose. Further, we will describe different formulations of the ALSP, to give ideas of how the problem can be formulated, including typical assumptions used by other researchers. Apart from a few special cases, the problem is too intractable to be solved to optimality [7]. We will therefore present various solution techniques utilised in literature. Also, in order to evaluate surgery schedules, we will give a detailed discussion on different performance measures that can be applied. The next part will describe how uncertainty is modelled, before the last two sections contextualise the two models we present by introducing decision-dependent uncertainty and stochastic arrival of emergency patients.

## 3.1 Introduction to stochastic programming

In deterministic programming, all information is assumed to be known with certainty in when making decisions [8]. However, since elements in mathematical programming may be uncertain, they might be more appropriately represented by random variables. These problems often involve decisions that must be made before important information is available [9]. Stochastic programming incorporates this uncertainty and information structure explicitly in the model formulation [10]. Different outcomes of the uncertain elements are considered collectively and the impact of different scenarios are balanced against each other [10]. Most common in literature are problems where the decisions do not impact the uncertainty of the problem [11].

The timing of the decisions relative to the resolution of the uncertainty must be specified in a stochastic model [10]. Decisions that can be delayed until after disclosure of information

offers an opportunity to adjust or adapt to the received information [10]. A stochastic formulation values this flexibility, while a deterministic model do not capture the possibility of responding to new information [8].

Various solution methods are used to solve mathematical problems with uncertain elements. If the number of possible outcomes or scenarios in the stochastic program is sufficiently small, deterministic solution approaches may be appropriate with the use of the deterministic equivalent of the stochastic formulation. This is often not the case and solution methods that exploit the structure of the stochastic model must be applied [10]. One common approach is decomposition, e.g. the L-shaped method or Benders' decomposition, which decompose the stochastic problem by stages [10]. Another approach is to use statistically based methods such as sample average approximation.

## 3.2 Formulations of the allocation scheduling problem

The ALSP has a close resemblance to machine scheduling problems, about which there exists extensive research. However, an important distinction should be pointed out. Once appointments are set in surgery scheduling, patients are not available prior to the scheduled start time, even if the server, i.e. operating room, is idle. In general, there is a lot more flexibility related to shuffling jobs around in machine scheduling. Thus, this review will focus on research related to surgery scheduling, but will to some degree find inspiration from other more generalised scheduling literature, such as job shop scheduling and flow shop scheduling.

Pham and Klinkert [12] address the deterministic ALSP by extending the job shop problem. As in the classical job shop problem there are $n$ jobs to be processed on $m$ machines, where a common objective is to minimise the makespan. Each job consists of several activities, which is the processing of the job on a given resource for a known duration. The set of resources needed for an activity is called a mode, and there may be several possible modes for a given activity. Once a mode is chosen, the resources of the mode are occupied for the entire processing duration. In this framework, Pham and Klinkert [12] formulate the scheduling problem as a mixed integer linear program (MIP) that assigns a mode to each activity and determinesx5 the start- and end time of the modes. The authors argue that this formulation gives a lot of flexibility and adaptability.

Charnetski [13] looks at the stochastic version of the ALSP, determines the start times for a given sequence of surgeries, and proposes a simulation procedure to model it. He utilises a two-stage Monte Carlo sampling plan, which generates both the surgery type and duration based on empirical data. The purpose of this study is to determine a relationship between the scheduled time for a surgery and the average waiting- and idle time. The paper uses a heuristic to determine the amount of time scheduled for surgery $i$, given by $d_i(h) = \mu_i + h\sigma_i$, where $h$ is a constant. The goal is to find approximate functions for the waiting time and

idle time given by $\mu, \sigma$ and $h$, and the value of $h$ that minimises the weighted sum of these.

Batun et al. [14] look at daily decisions that include the number of operating rooms to open, assignment of surgeries to operating rooms, and the sequence within each operating room and start time for each surgeon. They formulate a two-stage stochastic program to evaluate the effect of pooling operating rooms as a shared resource and utilise parallel surgery processing.

Wang [15] studies the problem of scheduling $n$ jobs to a single-server system. He considers both the static and the dynamic case, where revisions of the scheduled arrivals throughout the day are only included in the latter. The author assumes exponential service times where the goal is to minimise the weighted sum of customer flow time and system completion time.

Wang [16] extends his previous study to any service time distribution that can be approximated with a phase-type distribution, using the assumption of independent and identically distributed (i.i.d.) service durations. As pointed out by Mancilla and Storer [17], the sequencing of jobs are irrelevant in cases like this, because durations are assumed i.i.d. and the costs of waiting are equal for all jobs. However, even with these assumptions, the optimal arrival time intervals for the individual jobs are not equal, but dome-shaped. That is, more time is allotted to patients in the middle of the day.

In contrast to the paper by Wang [16], Denton and Gupta [18] assume that the job sequence is fixed, and address the problem of determining the start time and job allowance for each job. The goal is to minimise the weighted sum of expected waiting times, idle times and tardiness. The jobs have uncertain durations, which can be drawn from different distributions, relaxing the assumption of i.i.d. durations. The authors model the ALSP as a two-stage stochastic program. The first-stage decisions are the job allowances, while the waiting times, idle times, tardiness and earliness are second-stage decisions. They perform some experiments, indicating that the first two moments are sufficient to compute well performing job allowances when idle costs are high relative to waiting costs. To solve the model, they recognise that it exhibits a block-diagonal structure and exploit this in an adaption of the standard L-shaped method.

Denton et al. [19] further extend the model proposed by Denton and Gupta [18] by introducing the sequence of surgeries as decisions. They also provide a two stage stochastic program, where first stage binary variables represent the job allowances and the sequencing decisions. The sequencing decisions make the model considerably more complex than that of Denton and Gupta [18]. Thus, heuristic approaches are proposed, tested and evaluated. To find the optimal solution when total enumeration gives unacceptable computation times, a pairwise interchange heuristic, with similar steps as the L-shaped method, is used.

Mancilla and Storer [17] address the same problem as Denton et al. [19], but formulate a slightly different model. Instead of letting the first stage variables determine the job allowances and surgery precedence, they define binary variables to be equal to 1 when a surgery $j$ is in position $i$ and a continuous variable for the scheduled surgery start time. These variable definitions are useful when they use Benders' decomposition to solve the problem. The rest of the model, however, is equivalent to that of Denton et al. [19]. The authors propose a heuristic based on Benders' decomposition, where they evaluate three different algorithms to improve the heuristic.

### 3.3 Solution methods

The ALSP is considered a computationally difficult problem [20], which makes solution methods often include heuristics. Literature often studies decision rules based on statistical measures of the uncertain duration of a surgery.

Among the first papers on the topic, Bailey [21] and Welch and Bailey [22] propose a scheduling rule where $k$ patients are scheduled to arrive at the beginning of a session, while the subsequent patients are scheduled at intervals equal to the average surgery duration. They conclude that using $k$ = 2 provides the best trade-off between patient waiting time and idle time for the surgical team. Ho and Lau [23] evaluate nine scheduling rules using simulation. These include heuristics found in literature, in addition to some that are original for their paper. After testing several rules and variations of parameters, the authors are unable to dislodge the simple Bailey-Welch rules, and reveal that they are surprisingly robust.

Weiss [24] proves that the optimal sequence of two surgeries are in order of increasing variance of duration for certain distributions of durations. He also shows that this sequencing rule does not guarantee optimality when the number of jobs increases. According to Wang [16], the optimal sequence of surgeries is in order of the mean surgery durations if the durations are exponential distributed, the horizon is zero, and the goal is to minimise a convex combination of waiting- and idle time.

Three heuristics are proposed by Denton et al. [19]. The first sequences surgeries in order of increasing mean of durations, and the second in order of increasing variance of durations. The third sequences in order of increasing coefficient of variation (CV) of durations. The authors conclude that the second heuristic dominates the other two in nearly all tests and that the effects of optimal sequencing depend on the relative weight of performance measures. Dexter and Marcon [25] analyse the impact of several sequencing rules on staffing. They find that the commonly used rule of sequencing the longest surgeries first performs poorly from a staffing perspective, while sequencing the shortest surgeries first is more efficient.

Sicking and Kolisch [26] provide a generalisation of the Bailey-Welch rules. This is used

to find an initial schedule for a neighbourhood search heuristic. They divide a surgery day into $n$ service slots. The core of this heuristic is the definition of the neighbourhood, which comprises all schedules that can be found by increasing the number of patients in slot $i$ by one, while decreasing the number of patients of a slot $j \neq i$ by one.

Bosch and Dietz [27] claim there are no easy sequencing rules based on patient characteristics. Instead, they use a local search heuristic suggested by Bosch [28], that on average gives a cost 0.02% higher than the optimum. The heuristic first determines the cost of an initial sequence. Then, for each possible pairwise swap on this sequence, the cost of the optimal schedule is determined. If the best of these swaps result in an improved sequence, the heuristic goes back to the step of pairwise swaps, otherwise the current sequence is accepted as the optimal one. Robinson and Chen [7] use Monte Carlo-based techniques to compare the performance of several heuristics to determine patient appointment times. They use a cost based formulation, where the value of patient waiting time is expressed as a fraction of the value of surgeon idle time.

Kaandorp and Koole [29] derive a local search procedure, where the goal is to minimise a weighted average of expected waiting time, idle time and tardiness. It is also possible to include no-shows of patients. They prove that the scheduling algorithm converges to the global optimum by showing that their objective is multimodular. They also report that the appointment intervals in the optimal solutions are dome-shaped, equal to the observations made by Wang [15], Robinson and Chen [7] and Denton and Gupta [18].

Robinson and Chen [7] notice that almost all published heuristics are tested only against other heuristics, and not against the optimal policy. Regarding uncertain surgery durations, the authors also report that most papers do not even recognise that the means of the surgery durations can be eliminated from the formulation, leading to arbitrarily poor performance. This supports the findings of Bosch [28] some years earlier, who realises that the optimal sequence of surgeries places patients with identical characteristics at very different places in the schedule.

Batun et al. [14] have trouble solving their two-stage stochastic model for realistically sized instances. They use the L-shaped method to decompose the problem, but fails to solve even small problems within a reasonable amount of time. This is because the $\theta$ they define carries only limited information between the two stages of the model. To speed up the convergence time of the L-shaped method, they strengthen the formulation using lower bounding valid inequalities for $\theta$, based on Jensen's inequality [30]. Similarly, Laporte et al. [31] derive two lower bounds for $\theta$, used in optimality cuts in the L-shaped method.

Surgery scheduling is a specialisation of job shop scheduling and usually has additional constraints. However, solution methods proposed for machine scheduling can be useful in constructing procedures for surgery scheduling and should be considered. Applegate and Cook [32] describe a cutting-plane method for obtaining lower bounds on job-shop

problems and look at several cuts for this problem. They describe eight different inequalities from other job shop literature, such as Balas [33] and Dyer and Wolsey [34]. The basic cuts they consider impose restrictions on the possible sequences of jobs on a given machine, which is dependent on the earliest possible start time on that machine. Their results show bounds superior to the standard methods, but requires a greater computational effort. They also argue that finding classes of valid inequalities that will close the large optimality gap within a reasonable amount of computation time, remains a research challenge.

When constructing search algorithms for generalised job shop and surgery scheduling problems, similar same trade-offs must be considered. Common considerations are regarding the neighbourhood function and size, and whether to search the neighbourhood using first or best improvement. Vaessens et al. [35] state that the solution representation is a crucial ingredient of a local search algorithm together with the neighbourhood function. They discuss several neighbourhoods used in literature, including neighbourhood functions based on interchanges, swaps and reinsertions.

## 3.4 Performance measures

There are a variety of performance measures used in literature to evaluate surgery planning and scheduling procedures. Cardoen et al. [36] mention several widely used metrics: utilisation, makespan, levelling, throughput, patient deferrals, financial measures, and preferences. The utilisation should be maximised as unused resources are wasteful. Conversely, high utilisation often results in dense schedules, which implies solutions that are sensitive to changes. A similar measure is the minimisation of makespan which is the length of time required to complete all operations [37]. Decreasing makespan often involves a more dense schedule, i.e. higher utilisation, and will therefore have challenges similar to maximisation of utilisation. Levelling may reduce capacity problems by avoiding peaks of resource usage. Throughput of patients is the number of patients treated within a certain time period, which is a common metric. Moreover, minimisation of patient deferrals is another, making sure that the patients are treated within adequate time. Cardoen et al. [36] argue that the financial measure is the most general of all, as all of the measures can be represented by costs.

In the ALSP, performance is often measured using waiting time, idle time and overtime [38]. The interpretation and valuation of these measures varies. Waiting time may be valued as the patients' or surgical teams' waiting time. Idle time may be understood as the cost of not using the operating room or a surgical team. Overtime is sometimes interpreted as the cost of having a surgical team working after hours, and sometimes as a penalty for exceeding your estimates. The trade-off between these measures must be addressed by the decision maker. A common approach is to assign relative weights, as opposed to monetary values, and minimise the expected total cost of the system [38]. Fries

and Marathe [39] point out that the determination of these weights may be difficult. They write that the costs related to idle time are often available from standard cost accounting, but that assigning costs to waiting time requires the inclusion of intanglible aspects such as the effect on goodwill and social welfare. In addition, costs may differ across surgeries, further complicating the decision.

Denton and Gupta [18] minimise the expected cost of patients' waiting, operating room idling and overtime penalty when the session lasts longer than expected. They assume that the cost of all three performance measures are equal across surgeries. Values between 1 and 9 are tested for all measures, and these values are used by Kong et al. [40]. They also test costs of waiting time and overtime of 1 and 1-40, respectively.

Denton et al. [19] estimate the value of their costs based on consultations with the hospital staff. The waiting cost is set to 3 when a surgical team performs consecutive surgeries. When the surgical team is changed between surgeries, the cost of waiting time is set to 8 to include the cost of both patient and surgical team waiting. Operating room idle time is set to cost 8 and overtime cost to 4. Overtime is viewed as a penalty for late completion rather than a precise overtime cost. In addition, they perform a cost sensitivity analysis where they use 1 and 3 for waiting and overtime cost, respectively, and set idling cost to 0. They conclude that the relative importance of optimising sequence and start time depends on the choice of weights.

Cayirli et al. [41] minimise average patients' waiting time, surgeons' idle time per patient and surgeons' overtime per patient. Different ratios between idle time and waiting time are calculated depending on the scheduling scheme. The overtime cost is set to 1.5 times the idle cost. The same cost combinations are used by Cayirli et al. [42] with idling cost normalised to 1 and the overtime set to either 1.5 or 3. Zacharias and Pinedo [43] interpret the costs the same way as Cayirli et al. [41] and similarly normalise the idle time cost to 1 and overtime cost to 1.5. The patients' waiting cost is tested for several values between 0 and 1. Cayirli et al. [42] state that the best scheduling strategy is related to the choice of costs.

As opposed to the previous mentioned papers, Mancilla and Storer [17] use test cases with both equal and different cost across surgeries. For example, the idle cost of one surgery may be different from the idle cost of another. In all cases, waiting and idle costs are independently drawn from a uniform (20,150) distribution, and these are tested both with and without overtime. When overtime is included, this is set to 1.5 times the average waiting cost.

Some papers use monetary terms for the cost of waiting time, idle time and overtime. Batun et al. [14] estimate the cost parameters based on historical data from St. Marys Hospital in Rochester, MN. The overtime is estimated to \$12.37 per minute, which is 50% higher than the regular operating room cost. Due to difficulties in estimating the exact cost of surgeon idle time, they define both a low and high idle time costs, which are

calculated as fractions of the daily fixed cost of opening an operating room. The low and high idle time costs they use are $17.75 and $88.74 per minute, respectively. The authors do not consider patient waiting time. Keller and Laughhunn [44] operate similarly and estimate the cost of idle time by dividing the annual surgeon salary by the number of hours worked per year and use the minimum wage as the opportunity cost of the patients' waiting time.

All the previously reviewed literature assumes a linear relationship between waiting time, idle time and overtime. Klassen and Rohleder [45], however, point out that this relationship may actually be non-linear, because one patient waiting 40 minutes may have a different cost than 20 patients waiting 2 minutes each. Some papers include other measures to test whether their solution is biased. F8or example, Cayirli et al. [42] include a fairness measure which is measured as the standard deviation of the patients waiting time.

## 3.5 Distributions used for modelling surgery durations

A variety of probability distributions are chosen in papers addressing uncertain surgery durations. Some suggest distributions based on empirical data from clinics [27, 46], while other analytical studies assume the durations are drawn from distributions that make their models more tractable [38]. According to Cayirli and Veral [38], the majority of studies use i.i.d. surgery durations for all patients. Other papers divide the patients into unique patient classes where the surgery durations are i.i.d. within each class. Charnetski [13] notices that different types of procedures have different service time distributions. This is revisited by Gupta and Denton [47], who assume that surgery durations are normally distributed.

The CV is commonly used as a measure for the variability of surgery durations. Denton and Gupta [18] find that optimal solutions are mostly dependent on mean and variance, but may exhibit some dependence on higher moments like skewness. On the other hand, May et al. [48] found the skewness to be important, while the CV to have little impact when selecting which type of lognormal distribution to use. According to May et al. [48], literature suggests that normal and lognormal distributions are the only two viable candidate distributions to consider. They seek to find the distribution that gives the best overall fit to data by using an appropriate statistical test.

Based on previous studies, Soriano [49] defines a gamma distribution for the surgery durations. The author also performs a chi-square goodness of fit test to show that the fitted distribution is satisfactory at a significance level of 0.05. According to Yang et al. [50], Ho and Lau [51] show that the exact shape of a surgery duration distribution is not important. For this reason, Yang et al. [50] choose to use the gamma distribution to replicate the surgery durations.

O'Keefe [52] finds empirically that the higher moments of surgery duration distributions can be of significance, such that the distributions are not sufficiently described by simple two-parameter density functions. For this reason, he uses a lognormal distribution that includes skewness and kurtosis. Similarly, Hancock et al. [53] observe that surgery duration plots usually reveal a truncation on the left side and a tail on the right side, which they argue might be better represented by a two-parameter lognormal distribution. Robb and Silver [54], on the other hand, use a three-parameter lognormal distribution.

Different from the papers reviewed so far, Jansson [55] and Fries and Marathe [39] assume that surgery durations are exponentially distributed to make an analytical solution approach tractable. Liao et al. [56] consider a dynamic arrival problem where surgery durations are Erlang distributed. Bosch [28] extends and formalises Simeoni's [57] approach that also assumes that surgery durations follow an Erlang distribution. Bosch [28] justifies the chosen distribution, as he claims there is good evidence in his case that the optimal schedule is relatively insensitive to the third and higher moments. Lastly, papers like Liu and Liu [58] compare a simulation scheme for multiple types of distributions, including uniform, exponential, and Weibull.

### 3.6 Decision-dependent uncertainty

Stochastic problems may be classified as exogenous or endogenous. Exogenous uncertainty is widely studied in literature and includes problems where the uncertainty is independent of the decisions. In endogenous stochastic problems, the decisions have an impact on the uncertainty, either by changing the information structure or the probability of different outcomes [11]. Literature addressing this type of stochastic problems is far more sparse, as they are significantly more difficult to solve [59]. Decision-dependent probability problems include both problems where the decisions affect the probability of different outcomes and the parameters in the problem. To the best of our knowledge, no publications regarding allocation scheduling incorporate endogenous uncertainty. To provide insights into literature on endogenous uncertainty, papers of subjects outside allocation scheduling will be reviewed in this section. The main focus is on decision-dependent probabilities, but papers addressing decision-dependent information structure will be mentioned.

A general stochastic program without decision-dependent uncertainty may be formulated as

$$min \quad F(x;P) \coloneqq E_p\big[f(x,\omega)\big] \quad on \quad X \tag{1}$$

where $P$ denotes the probability distribution of the possible outcomes $\omega \in \Omega$ and $X$ is a closed non-empty subset of a Euclidian space. The decision $x$ is made before the disclosure of $\omega$. The cost of the decision is quantified by a real-valued function $f(x,\omega)$ [60].

In contrast, a stochastic problem *with* decision-dependent uncertainty may be formulated as

$$min \quad F(x) \coloneqq \int_{\Omega} f(x,\omega)P(x;\omega) \quad on \quad X \tag{2}$$

which differs from equation (1) by making the probability distribution dependent on the decisions [60]. Even though $f(x,\omega)$ may be convex, this property may be lost for $F(x)$. This makes the problem far more complex and puts limitation on the number of available efficient optimisation techniques.

Ahmed [61] was the first to address decision-dependent probabilities according to Hellemo et al. [11]. He presents several problems incorporating decision-dependent probabilities, which are related to network design, server selection and facility location. Ahmed [61] formulates these problems as MIPs and shows that these can be solved with linear programming-based branch and bound methods.

Viswanath et al. [62] formulate a problem that is categorised as a decision-dependent distribution selection problem by Hellemo et al. [11] and Goel and Grossmann [63]. They introduce a shortest path problem between a predefined origin and destination in a network. The network consists of links subject to disruptive events. The links have different probabilities of survival. To strengthen the weak elements and increase the probability of survival, investments in the links can be made at a cost. The problem is formulated as a two-stage stochastic problem. The first-stage decisions are whether or not to invest in each link without any knowledge of of how the network will survive a disruptive event. The underlying probability distributions of the random variables are dependent on these decisions. The second-stage decisions are made after the event have occurred and consist of finding the shortest path from origin to destination. A deterministic equivalent to the the stochastic formulation is presented and structural results are derived. Viswanath et al. [62] propose approximate solution procedures solving the problem, which are tested with numerical experiments and prove to give good results for small problem instances.

Hellemo et al. [11] introduce an extended taxonomy of stochastic problems with decision-dependent uncertainty. They present relevant models and applications, and classify papers and formulations within decision-dependent uncertainty. Starting with an initial formulation of a two-stage stochastic program with decision-dependent uncertainty and recourse, they show how direct and indirect manipulation of the probability distributions can be incorporated. Four different manipulations are presented, two of which are indirect manipulations. In these, transformations of the probability distributions are performed, either by linear scaling or a convex combination of the distributions. The two last formulations include direct manipulation by changing the parameters of the distribution, either in a Kumaraswamy or approximated normal distribution. To test and compare the different formulations, Hellemo et al. [11] look at capacity expansions of power generation where

an investor seeks to minimise the cost of meeting a stochastic demand. The formulations introduce many non-linear terms and consequently non-convex programs.

Decision-dependent information structure is more widely addressed in literature than decision-dependent probabilities. Jonsbråten et al. [59] were among the first to introduce this type of problem. They propose an enumeration algorithm for stochastic programs with decision-dependent information structure and two decision stages. Goel and Grossmann [64] propose a model to facilitate decision-making in investment and operational planning of gas field development under uncertainty. The resolution of uncertainty depends on the investment decision. They formulate a stochastic mathematical model and use a decomposition-based approximation algorithm to solve it. The same authors [65] extend this model by introducing theoretical properties satisfied by any feasible solution to reduce the size of the model. In addition they present a Lagrangean duality-based branch and bound algorithm which is guaranteed to find the optimal solution and reduces the model size significantly. This work is further extended by Tarhan et al. [66] where the resolution of uncertainty is gradual over time instead of immediately. More recent improvements of the works by Goel and Grossmann [63,64] are made by Gupta and Grossmann [67]. In this publication, they try to introduce a more compact representation of the nonanticipativity constraints. Moreover, they propose three solution procedures that are tested on two process network problems.

## 3.7 Stochastic arrival of emergency patients

In addition to uncertain surgery durations, another aspect of uncertainty is related to the arrival of emergency patients, as pointed out in Chapter 2. A common assumption in literature is that elective and emergency patients consume different resources and are handled by different personnel. Among the exceptions from this assumption is the paper by Gerchak et al. [68], including both types of patients. In their work, the number of emergency patients is modelled as a random variable. However, they address the ASP rather than the ALSP. The ASP is also addressed both by Lamiri et al. [69] and Lamiri et al. [70], who use a random variable to represent the capacity used by emergency patients with a given distribution, which can easily be estimated from historical data. In a third paper, Lamiri et al. [71] use a column generation approach to solve the stochastic ASP with uncertain surgery durations and uncertain demand for emergency capacity. Sickinger and Kolisch [26] provide an interesting perspective of emergency patients as stochastic downtime of the resources they occupy.

On a higher level of consideration, a common approach to dealing with emergency surgeries is to reserve operating room capacity, which is believed to increase responsiveness [72]. However, Wullink et al. [72] find that performing emergency surgeries in elective operating rooms is more efficient than having designated operating rooms for emergency patients. The authors do not consider how sequencing of emergency patients among elective patients

is handled most efficiently. To the best of our knowledge, there are no literature addressing the combination of stochastic surgery durations and stochastic arrival of emergency patients for the ALSP.

# Chapter 4

# Problem description

This chapter formulates the two problems of this thesis, in a generic manner, and includes elaborations on the assumptions made and the implications of these.

## 4.1 Problem statement

The problems consider a single day on a single operating room, and decides how to set start times for, and thus decide the sequence of, a predetermined set of surgeries with stochastic duration. The objective is to minimise the expected weighted sum of waiting time, idle time, and overtime, with the implicit assumption that the quality of a schedule can be adequately measured using only these three performance measures. Waiting time is defined as the difference between a scheduled surgery start time and the actual start time, idle time is defined as the time between the cleaning after one surgery until the start of the next one, and overtime is the amount of time by which the end of the last surgery exceeds the end of the regular working day.

The three main assumptions of the problems are listed in Table 4.1. Firstly, it is assumed that the patients arrive exactly in time for their scheduled start time (Assumption 1). This means that no-shows and late arrivals are not a problem and, at the same time, it implies that a surgery can never commence before its scheduled start. Moreover, the problems assume that surgical team and all necessary resources are available as required (Assumption 2), so that the only thing that can prevent a surgery from starting at the scheduled start is if the previous surgery does not finish on time. Finally, it is assumed that all surgeries must be performed on the day when they are planned (Assumption 3), meaning that postponing a surgery to another day is not an option.

Two separate problems are formulated to investigate two different aspects of uncertainty. The first problem addresses the aspect of uncertainty in surgery durations being dependent

Table 4.1: The main problem assumptions

| | |
|---|---|
| *Assumption 1* | Patients arrive exactly at their scheduled time |
| *Assumption 2* | No other resources represent bottlenecks |
| *Assumption 3* | All scheduled surgeries must be performed |

on the surgeries' start time, meaning that the decision of when a surgery is scheduled affects the uncertainty in its duration. The second problem addresses the aspect of uncertain arrival of emergency patients. The focus is then how to schedule the elective surgeries knowing the probabilities that given numbers of emergency patients arrive, and how to adjust the schedule once you know how many and which emergency patients arrive.

## 4.2 Explanations and implications of assumptions

The assumption of no-shows and late arrivals not being a problem reflects the real situation, as explained in Section 2.2.2. As has already been stated, Assumption 1 also implies that surgeries cannot be started before their scheduled start. In reality, if a patient is already at the hospital, or can be asked to arrive early, it might be possible to start the surgery before schedule, but without having information on what patients are on stand by this cannot be modelled in a realistic way. When evaluating the solutions we get, we therefore note that in reality the hospital possibly could have avoided some of the idle time by starting some surgeries ahead of schedule.

With operating room and surgeon idle time being considered very expensive, the coordinators are normally able to make sure that other resources are available when needed, making Assumption 2 a realistic assumption.

Assumption 3 follows mainly from the fact that the decision of which surgeries can and cannot be postponed is based on a subjective medical assessment that we are unqualified to make. In reality, it sometimes happens that surgeries are postponed to another day if they cannot be performed without overtime incurring. According to the coordinators, performing surgeries after regular working hours is unfavourable and should be avoided if possible. This is both with respect to the working hours of the surgical staff, because overtime is expensive for the hospital, and because potential complications are handled more easily if more staff are at work and available. The policy of what to do when facing the prospect of overtime varies across the three different hospitals. Coordinators at the Trondheim hospital tend to postpone surgeries to another day if they see that performing them is likely to result in overtime, whereas at the Røros hospital they normally finish all planned surgeries even though overtime incurs. In any case, penalising overtime in the objective accounts for the inclination towards avoiding it. Also, since the inconvenience of postponing a surgery depends on a lot of different factors, Assumption 3 lets us evaluate the quality of the schedules we generate and appropriately compare these to the current practice at St. Olavs.

# Part II

# Analyses and Discussions

# Chapter 5

# Data analysis

This chapter will explain the analyses performed on the data set, the selection of which instances to use when evaluating the models we propose, and the choice of scenario generation technique.

## 5.1 About the data

The data on which the following analyses are based was extracted from the hospital's "OpPlan" system and provided by the hospital's analytics manager in February, 2016. It contains extensive descriptive information on all surgeries performed by the hospital's Department of Orthopaedic Surgery from 1 January 2006 to 31 December 2015, across Trondheim, Orkdal and Røros. For each surgery there is information about

(a) the patient, including demographics and diagnosis/diagnoses

(b) the surgical team, including surgeon(s), nurses and anaesthesia personnel

(c) the surgical procedure, including procedure type, urgency, operating room, and realised times and durations related to the surgery

Surgical procedures are classified according to the NOMESCO Classification of Surgical Procedures (NCSP), developed by the Nordic Medico-Statistical Committee in 1996 [73]. A surgery may include multiple surgical procedures and thus be given a set of surgery procedure codes. The data has been anonymised to exclude details related to patient and staff identity.

The data set was pre-processed to exclude surgeries whose inputs were obviously erroneous, e.g. surgeries whose durations were below or equal to zero minutes and surgeries where data was missing. For instance, one surgery had a reported surgery duration of roughly

-1 000 000 minutes, heavily distorting the statistical properties of the data set. The validation process removed 25 of the surgeries, leaving the data set with a total of 87 100 surgeries dating back to 2006.

## 5.2 Analysing the data

The following part will first describe the segmentation of the data and how it has been adjusted for trends. Next, it will investigate whether the effect of the hypothesis described in 2.3 is statistically apparent. Note that since only knife-time is considered stochastic, as explained in Section 2.2.2, the analysis considers only this component of the total surgery durations.

### 5.2.1 Pre-analysis

The overall objective of analysing the statistical properties of the data is being able to make more accurate estimates on surgery durations, in order to diminish the negative impacts of the underlying uncertainty. When determining the daily schedule within the scope of this thesis, it is given which procedures are going to be performed by which surgeons at a given operating room. In order to make as accurate estimates as possible, the data has been segmented in order to isolate the effect of the relevant uncertainties, while keeping the segments large enough for them to provide statistical significance. In particular, if we want to make predictions about the duration of a specific procedure performed by a specific surgeon, we want to base this on a sample containing data on that specific combination only. For readability, we refer to such combinations as surgery types. In this chapter we provide examples of statistics for seven different surgery types, referred to using numbers 1 to 7. Table 5.1 provides a mapping of what procedure codes these seven surgery types represent. The third column is there only to emphasise that the surgery types are related to a specific surgeon, even though they are anonymised in this case.

Table 5.1: Surgery types referred to in this chapter

| **Surgery type number** | **Procedure code** | **Surgeon** |
|---|---|---|
| 1 | QDG20 | Surgeon 1 |
| 2 | NDM19 | Surgeon 2 |
| 3 | ACC51 | Surgeon 2 |
| 4 | NDM39 | Surgeon 2 |
| 5 | QDB10 | Surgeon 3 |
| 6 | ACC51 | Surgeon 4 |
| 7 | NGD11 | Surgeon 4 |

For every surgery it is specified whether the patient was elective or emergent. Figure 5.1 shows an example of how the empirical durations differ based on this characteristic, for

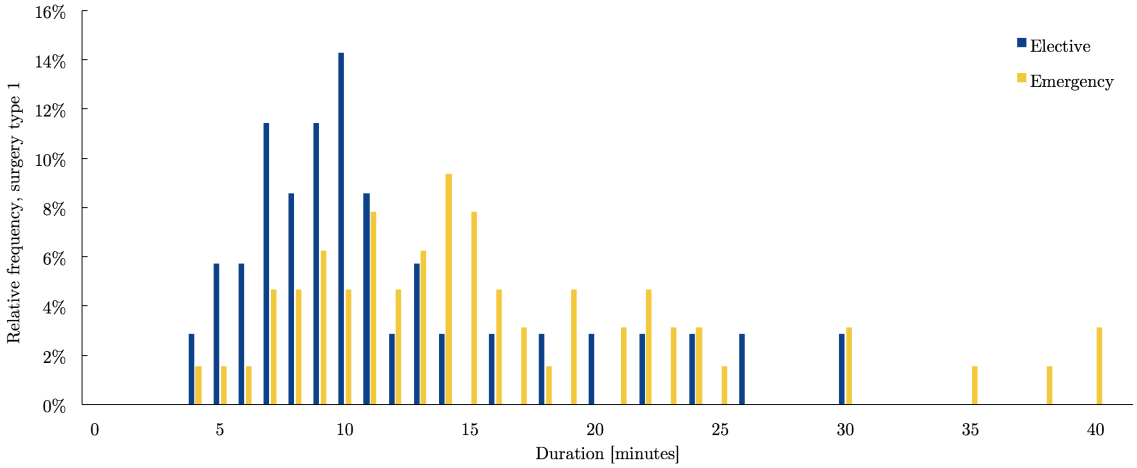



Figure 5.1: Durations for surgery type 1, separated by urgency

surgery type 1. The average duration for the elective surgeries of this specific surgery type is 12.6 minutes, while the average for the emergent surgeries is 16.1 minutes. Such differences exist for a lot of the surgeries and, for this reason, we distinguish between elective and emergent surgeries both in the analysis and when sampling in order to generate scenarios for use in the mathematical models.

### 5.2.2 Adjusting for trends

When analysing surgery durations for a given surgery over a period of ten years, one might expect the durations to be subject to a learning effect, especially when considering one single surgeon at a time. Technological improvements, as well as increasing tacit knowledge and experience, would be expected to cause the durations to decrease over time and, if so, this should be adjusted for when making predictions about surgeries in the present.

The learning effect in terms of production of a given good is often assumed to be driven by the cumulative amount produced of the good. Krajewski [74] uses a model on the form

$$d_{k+1} = d_1 k^b \tag{3}$$

where the direct labour hours for the $(k+1)$th unit, $d_{k+1}$, depend on the direct labour hours for the first unit, $d_1$, the cumulative amount produced, $k$, and a constant $b = \frac{log(r)}{log(2)}$, with $r$ being the learning rate. The reduction in time thus follows an exponential curve, with the learning effect being very high in the beginning and gradually diminishing with time as the total amount of units produced increases. The equivalent of this for our case would be to assume the learning effect to be driven by the number of times each surgeon has performed the given procedure in the past. However, as described above, we assume the effect to be not only due to the increased experience of that specific surgeon, but also due to technological advances and shared experience with other surgeons. Therefore, we introduce a time axis and let the learning effect depend on the point of time, assuming this

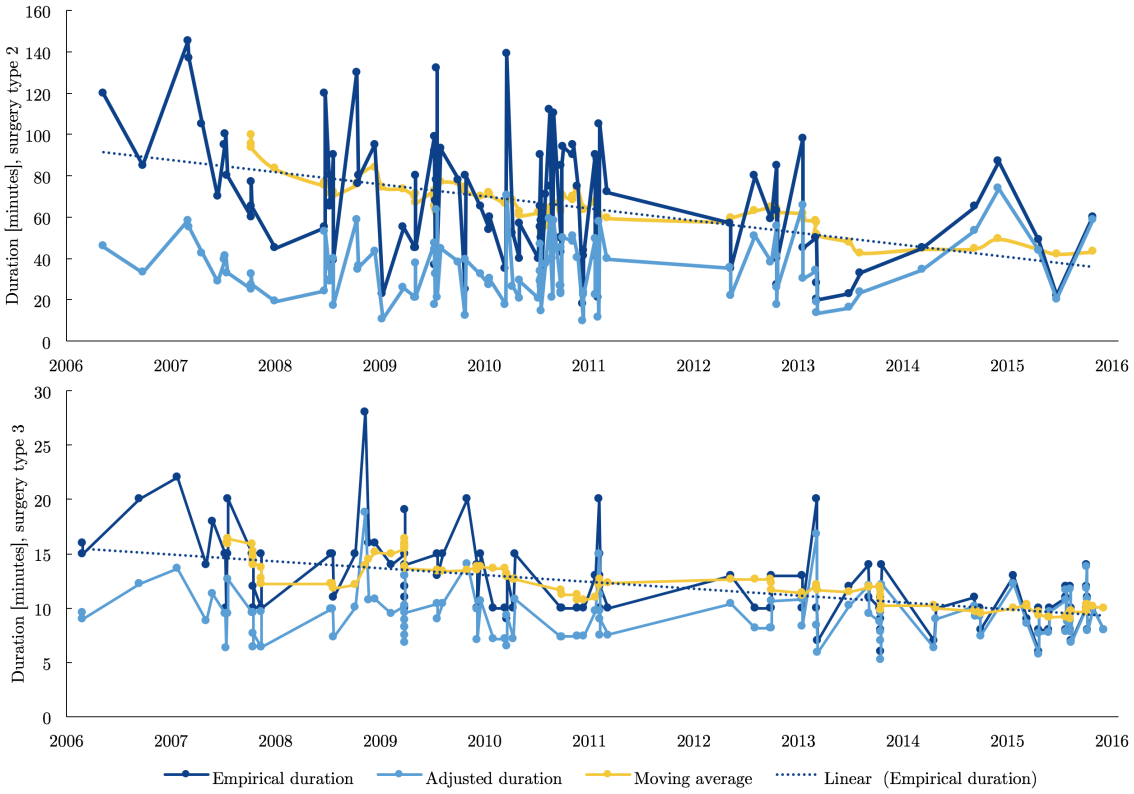



Figure 5.2: Adjusting for trend due to learning effect for surgery types 2 and 3

will account for both the surgeon-specific learning effect and the learning effect experienced by the hospital (or even the health sector) as a whole. For example, the dark blue line in the plots shown in Figure 5.2 connects the scatter plots of the durations of surgery types 2 and 3, respectively, with time on the horizontal axis. By observation, the durations tend to become shorter with time. We investigate this by calculating a simple moving average of the 10 most recent durations up to surgery number $k$ of a surgery type $s$, given by

$$SMA_{sk} = \sum_{i=k-10}^{k-1} \frac{d_{s,i+1}}{10} \tag{4}$$

with $k = [10, N_s]$, and $N_s$ being the total number of empirical durations for surgery type $s$. The yellow plots in Figure 5.2 show this moving average for surgery types 2 and 3. For surgery type 2, the moving average equals 99.7 minutes at $k = 10$, compared to 43.2 minutes at $k = N_2$. Evidently, the trend is quite considerable, and with this being the case for a lot of the surgeries analysed, the trend should be accounted for when making predictions about the duration of a surgery at a given time.

As mentioned above, learning curves tend to follow a logarithmic model with diminishing marginal reduction in time. Since we do not know where the ten year period for which we have data is placed in the course of learning (some procedures may have been performed for decades while other may be new), we make a simplification by assuming the effect

during the relevant ten year segment to be linear. The trend is thus described using a linear regression on the form

$$d_s(t) = a_s t_s + b_s \tag{5}$$

for every surgery type $s$. Regression coefficients $a_s$ and $b_s$ are calculated using a least squares method, meaning that the deviations from the trend are on average zero. Equation (5) provides an expectation for a surgery's duration at any day $t$ in the ten year interval, with $t = 1$ representing 1 January 2006. For surgery type 2 displayed in the upper plot in Figure 5.2, the regression coefficients are $a_2 = -0.01601$ and $b_2 = 93.2882$, with the slope $a_2$ indicating that on expectation the duration should decrease by 0.01601 minutes per day. This trendline, and the corresponding trendline for the lower plot, is shown as dotted lines in Figure 5.2. Evaluating equation (5) at $t = T + 1$, where $T$ is the last day of the ten year period such that $t = T + 1$ represents the present[1], we get the expected duration of surgery $s$ if it is to be performed in the present. We define this as

$$D_s = a_s(T + 1) + b_s \tag{6}$$

If we based the expectation on a flat average of all empirical durations, we would expect surgery type 2 to have a duration of 67.4 minutes. As has been commented on, the moving average changes considerably, making a flat average a poor prediction. The expectation from equation (6) equals 34.8 minutes, which to a larger extent concurs with the most recent empirical durations.

In Section 5.4 we will be using the moments of the set of empirical durations to generate scenarios for surgery durations as inputs in a stochastic model. Before calculating the moments, the data set should be adjusted such that it is representative of the present level of experience, knowledge and technology (which are all assumed to be components of what we describe as the learning effect). The adjustment uses the assumption of the process being what the literature refers to as a trend-stationary process [75, 76], typically given by

$$z_t = \mu + \beta_t + \epsilon_t \tag{7}$$

where $\mu + \beta_t$ is a deterministic mean based on a linear regression, and $\epsilon_t$ is a stationary stochastic process with zero mean. This is transferable to our case, with the deviations from the linear regression function (5) representing the stochastic process. We assume that the empirical durations, when adjusted for the learning effect, are stationary.

If we detrend the data by simply adjusting all durations down by the amount that, based

[1]Note that since we only have data until the end of 2015, we assume the present is 1 January 2016.

on the linear regression, is meant to come from the learning effect, the error terms, $\epsilon_t$, become unrealistically high in relative terms. For instance, a duration 40 minutes shorter than the expectation might have been feasible when the expectation was 90 minutes, but impossible if the expectation is 35 minutes. By visual inspection of historical durations, the variance appears to decrease with a decreasing mean. This was also verified with St. Olavs hospital staff, who argued that large absolute deviations from the expected duration is more common when the expected duration is high. Hence, we want to adjust the data points in a way such that the error terms relative to the mean are preserved. This is similar to what Tsay [76] suggests for adjusting for linear trends in historical data. For a given surgery $s$, the expected duration at time $t$ is given by the linear regression in equation (5). Letting $d^e_{ks}$ represent empirical duration number $k$ of surgery type $s$, performed at time $t_{ks}$, the deviation of $d^e_{ks}$ from the expectation is given by

$$\epsilon_{ks} = d^e_{ks} - a_s t_{ks} + b_s \tag{8}$$

which, in terms relative to the expected duration is

$$\epsilon^{\text{rel}}_{ks} = \frac{d^e_{ks} - a_s t_{ks} + b_s}{a_s t_{ks} + b_s} \tag{9}$$

We adjust for the learning effect while preserving the relative error by taking each empirical duration $d^e_{ks}$, finding its relative deviation $\epsilon^{\text{rel}}_{ks}$ from its expected duration, and multiplying $1 + \epsilon^{\text{rel}}_{ks}$ by the expected duration for a surgery performed at the present, given by equation (6). The formula for the trend-adjusted empirical duration of surgery number $k$ of surgery type $s$ is thus given by

$$d^{\text{adj}}_{ks} = D_s(1 + \epsilon^{\text{rel}}_{ks}) \tag{10}$$

The data used in the analyses[2] of this thesis is adjusted according to equation (10). For surgery 2 and 3, the adjusted data is shown by the light blue connected scatter plot in Figure 5.2, where we can observe that the errors from the mean are scaled based on the expectation at any time $t$.

### 5.2.3 Testing the hypothesis

According to the planners, some types of surgeries are more likely to be scheduled early than late. If, for example, major surgeries tend to be scheduled late in the afternoon, surgeries that start late will, on average, have a longer duration. Ignorance of this selection

[2]Note that only surgery types for which there is a significant number of empirical durations will be used, so we avoid the problem of adjusting for trends that have insufficient statistical basis.

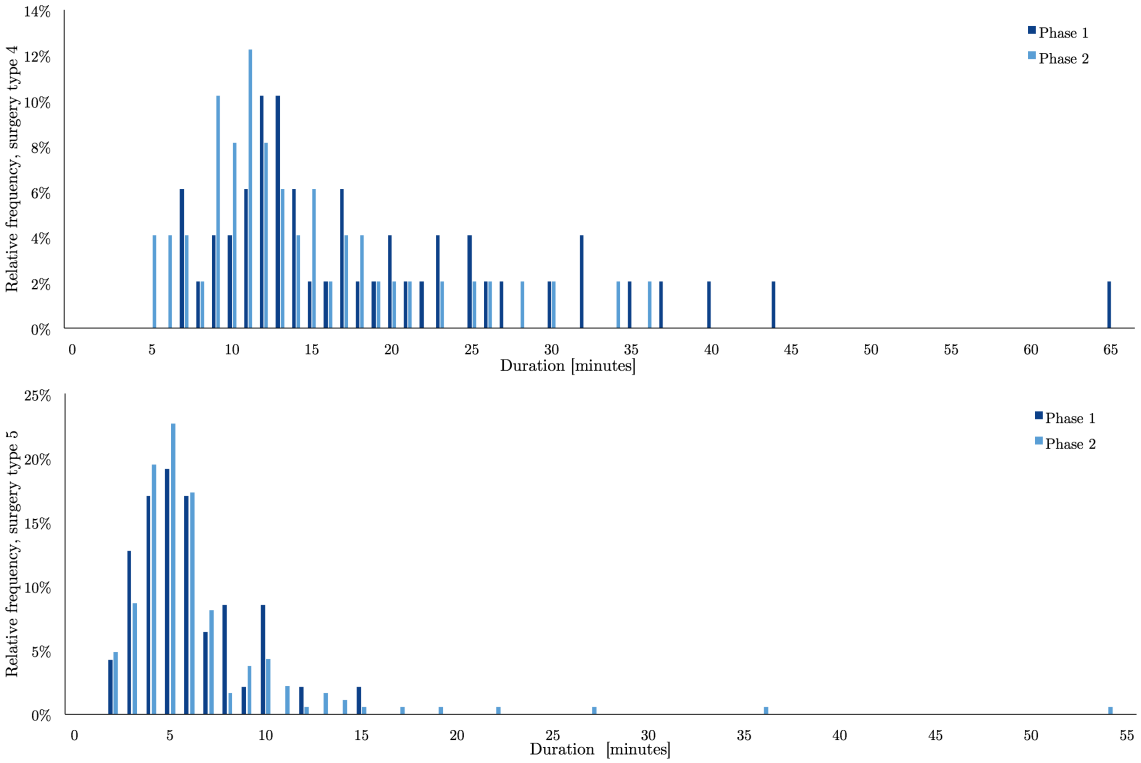



Figure 5.3: Frequency of durations for surgeries 4 and 5, per phase

bias when analysing the data could lead to the unjustified conclusion that a given surgery is likely to last longer if it is scheduled late in the afternoon. Therefore, in order to ensure comparability, we analyse analyse surgery durations per surgery type and not only per surgeon.

When testing the hypothesis we take, for a given combination of surgeon and procedure type, all empirical durations and divide them into two separate sets based on the start time of the surgery. The first set contains the duration of all surgeries that started before 11:30, while the other sets contain the rest. This point of time is set to coincide with the start of the lunch break, and out of the 55 514 elective surgeries in the data set, this gives two data sets containing 47.9% and 52.1% of the total samples, respectively. The time up until 11:30 is referred to as phase 1, and a surgery that is started before 11:30 is said to be performed in phase 1, even though it is not finished before the start of phase 2.

Figure 5.3 shows the distribution of durations in either phase for surgery type 4 and 5, respectively. For type 4, we can observe that surgeries in phase 2 tend to be shorter. Among the five joint highest durations, four were performed in phase 1, and the average differs substantially from one set to another. For type 5, in contrast, there is no clear tendency of durations differing between the two phases.

We want a test that can disprove, to a certain level of significance, that the durations in each phase for a given surgery are drawn from the same distribution, thus indicating that they may be from two different distributions. Two of the most commonly used tests

for independence of two samples are the two-sample Kolmogorov-Smirnov (KS) test and the Chi-square two-sample test. These have the advantage of being non-parametric and distribution free; they require no assumption on the distribution of the data from which the two samples have been drawn, and thus provide a non-biased test of independence.

Whereas the Chi-square test is used for categorical data, the Kolmogorov-Smirnov applies to continuous data. Technically, surgery durations are continuous, but in the process of being entered into the protocol by the nursing staff they are rounded to the nearest integer. This leads to the potential occurrence of ties in the data; multiple durations may be equal. If the number of ties is high, such that the samples are highly discrete, the data can be considered categorical, and the chi-square test is preferred. In our case, though, we avoid the problem of ties since the data becomes continuous when adjusted for trends according to equation (10). The trend adjustment is irrespective of phases, so a difference in duration between the two phases will be present also after the adjustment.

The two-sample Kolmogorov-Smirnov test [77] tests whether the underlying one-dimensional probability distributions of two samples A and B differ. The test is based on each sample's empirical distribution functions, $F_s$, defined as

$$F_s(i) = \frac{n_{is}}{N_s} \quad s \in \{A, B\}, \quad i \in [\min(A, B), \max(A, B)] \tag{11}$$

where $n_{is}$ is the number of observations in sample $s$ that are below or equal to $i$, and $N_s$ is the total number of observations in sample $s$. $F_s(i)$ therefore measures the fraction of the total number of observation that are below or equal to $i$, with $i$ ranging from the joint lowest observation to the joint highest observation in the two samples A and B.

The test statistic in the two-sample Kolmogorov-Smirnov test is defined as

$$D_{ss'} = sup|F_s(i) - F_{s'}(i)| \quad s, s' \in \{A, B\}, \quad s \neq s', \quad i \in [\min(A, B), \max(A, B)] \tag{12}$$

i.e. the highest absolute vertical distance between the two samples' cumulative empirical distribution functions. Figures 5.4 and 5.5 show the cumulative empirical distribution functions for the two surgeries displayed in Figure 5.3, and mark the d-statistic in yellow for each case. For surgery 4, when comparing the two distribution functions we can see that phase 1 tend to have longer durations than phase 2. The cumulative distribution function of phase 2 rises earlier, meaning a larger portion of the durations are short. This difference leads to a relatively large d-statistic. On the other hand, the two distributions for surgery 5 are quite similar, giving a lower d-statistic. When conducting the test, the null hypothesis is that the data is from the same distribution, and it is discarded if the p-value is below a specified alpha value.

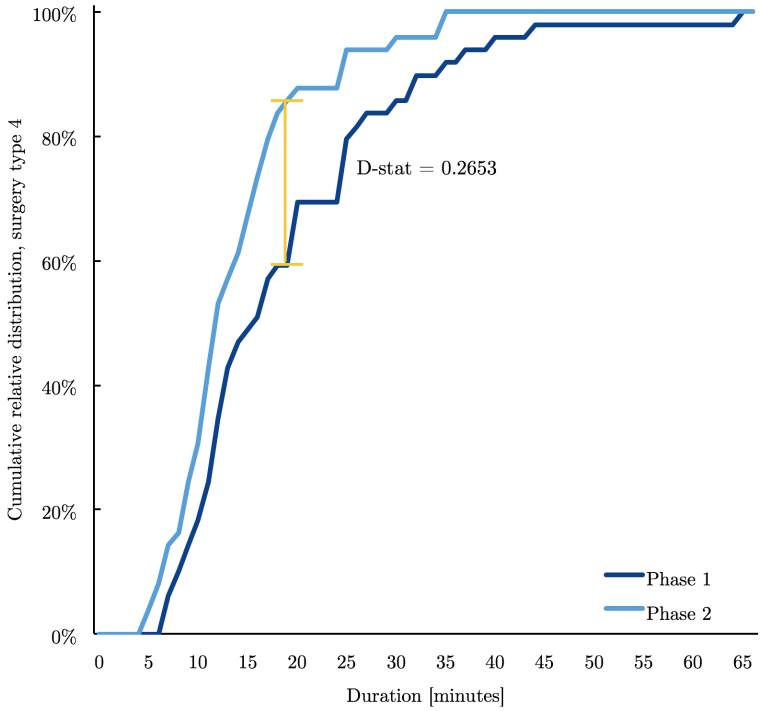



Figure 5.4: Cumulative distribution plots for the durations of surgery type 4

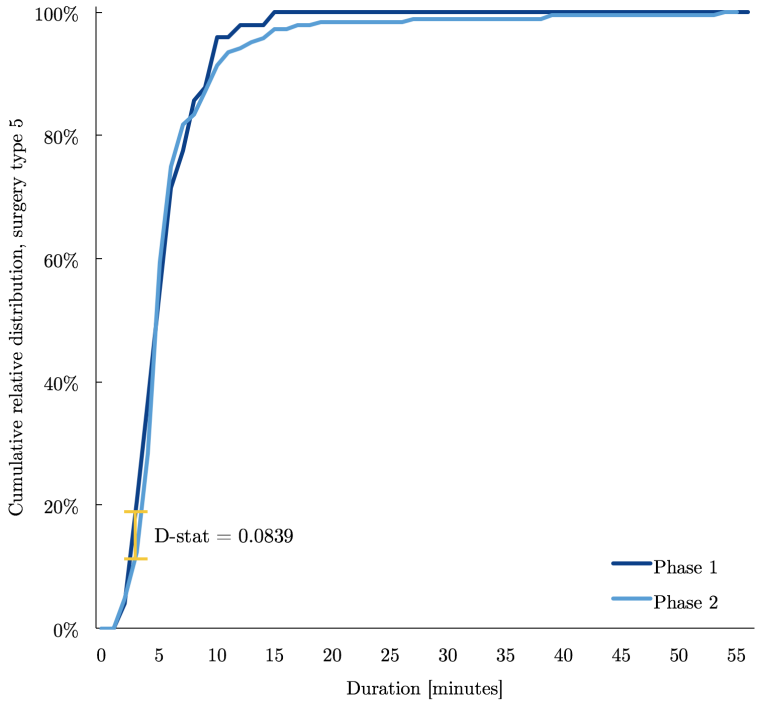



Figure 5.5: Cumulative distribution plots for the durations of surgery type 5

P-values have been calculated using a Monte Carlo simulation, where each scenario randomises the division of the duration into two separate sets, between which the d-statistic is calculated. The p-value then equals the fraction of the total d-statistics that are equal to or higher than the d-statistic we computed from the initial data set. It can therefore be interpreted as the chance of the d-statistic being as extreme or more extreme than the initial one, and a low p-value indicates that there is reason to believe that the underlying distributions behind the two samples are not the same. The p-value in the case of surgery 4 in Figure 5.3 is 0.041, indicating that it is unlikely that the two sets of durations come from the same underlying distribution. For surgery 5 the p-value is 0.631, meaning the test is inconclusive and the null hypothesis of the data coming from the same distribution remains.

Since the distinction between phases is used only in the Phase Model, which excludes all emergency patients, we only apply the test to elective surgeries. Out of the 55 514 elective surgeries in the data set, we choose to test the 200 combinations of surgeon(s) and procedure type that have the most occurences, i.e. those combinations for which we have the best statistical basis. These have on average 54 empirical durations (in comparison, Denton [19] has on average 21 samples per surgery type, without segmenting per surgeon), constituting roughly 20% of the total number of elective surgeries performed in the ten-year period.

Out of the 200 surgery types tested, the null hypothesis was discarded for 31, using a significance level of 0.10. These 31 surgery types are likely to have significantly different uncertainty in duration when they are performed in the two different phases, and are likely to be those surgeries for which the Phase Model provide the most value.

## 5.3 Selection of instances

As has been explained, the two problems presented in this thesis consider a single operating room on a single day. An instance is therefore a set of surgeries to be performed at a specific operating room on a specific day, with information on which surgical team is assigned to perform each surgery. In order to evaluate the resulting schedules, it is interesting to see how the model performs on instances that have in fact been experienced at St. Olavs Hospital. The data set includes surgeries from a total of 3 652 days (from 2006 to 2015) across a set of different operating rooms, with a total of 30 804 potential instances.

The evaluation of the two models explained in Chapter 6 will be made using different types of instances. The Phase Model will be evaluated on actual historical instances, whereas the Emergency Model will use three realistic instances designed to display the effects of the model. The following part explains the analysis performed in order to determine the instances to use for both models, both for the stability analyses and for the comparisons in the practical computational study. Note that when using real instances, the data has

been adjusted for trends in accordance with Section 5.2.2 to reflect the experience level on the relevant surgery date.

### 5.3.1 Instances for the Phase Model

Since the Phase Model excludes emergency patients, we need instances with elective patients only. Filtering out all instances that have one or more emergency patients during the day narrows down the number of instances from 30 804 to 18 009. In addition, we need instances where all surgeries during the day have been, historically, performed a significant number of times by the given surgeon, in order to be able to sample from a representative empirical sample of past surgery durations. If we require the number of past occurrences to be equal to or higher than ten in both phases, we are left with a sample of 1 917 instances, representing a total of 1 303 days across 18 operating rooms. 30% of these have only one surgery, 28% have two surgeries, and the remaining 42% have more than two surgeries. The problem of scheduling surgeries is trivial when there is only one surgery to be scheduled, and the effects of the two phases are believed to be more significant when the number of surgeries is higher, so we filter out all instances with fewer than three surgeries per day. The remaining 807 instances range from having one to having six *distinct* surgeries during the day. Again, the effects we want to analyse, and thus the value of the Phase Model, are expected to be higher when the surgeries have different uncertainty and statistical properties. If all surgeries have the same statistical properties, the order of the schedule will be of no relevance, and so we want to evaluate the model on instances exhibiting multiple different surgeries. Only considering instances with at least three distinct surgeries during the day leaves the data set with 413 instances. Moreover, we want to use those instances that include surgeries likely to have a significant difference in statistical properties between the two phases. Therefore, we make a mapping of the 31 surgeries for which we found the highest significance in Section 5.2.3, and use only those instances that have at least two unique surgeries among these 31, i.e. instances where at least two different surgeries are expected to have a significant difference between the two phases. This leaves us with 26 instances. Finally, the practical analysis that will follow in Chapter 7 evaluates the performance of St. Olavs' plan for each of these historical instances, on the scenario trees we generate. The necessary data on St. Olavs' plans was accessed through "OpPlan's" planning interface, where the relevant dates on the relevant operating rooms where looked up manually in the system. Six of the 26 instances had no data on how they were planned, resulting in a final sample of 20 instances to be used in the Phase Model and analysed further in Chapter 7. To summarise, we use those instances that contain

- only elective surgeries,
- only surgeries that has been performed at least ten times in each phase,

- at least three distinct surgery types,
- at least two surgeries among the 31 types found in Section 5.2.3,

and for which there exists data on how the day was planned. The 20 resulting instances include 37 distinct surgery types, 12 of which are among the 31 surgeries with the most significant phase-dependent difference in duration.

### 5.3.2 Instances for the Emergency Model

As for the Emergency Model, the available historical data is not sufficient in order to evaluate the model on real past instances. In particular, the data set does not contain information on what the plan looked like before emergency patients emerged, meaning we cannot compare our solution with what happened at the hospital on a given day. In the case of an emergency surgery, the hospital often reschedule one of the elective surgeries to another day to avoid overtime. Our model, on the other side, allows overtime and requires all scheduled surgeries to be performed on that day. The hospital might have adjusted their plan differently if cancelling were not an option, so the difference in cancel policy makes the comparison less valid. For the Emergency Model, we have therefore created three realistic instances of a day with four and five distinct elective surgeries, where a maximum of two additional emergency surgeries are to be scheduled.

## 5.4 Scenario generation

This section will first motivate for the choice of scenario generation technique by reviewing a few of the most relevant principles found in literature. Further, it will explain the essence of the chosen moment-matching algorithm, and evaluate the results of the resulting scenario generation.

Note that because knife-time is the only part of the total surgery duration that is considered stochastic, only this is used in the scenario generation. Deterministic durations of the pre- and post-phase are then added to the knife-time from the scenario generation, before the complete surgery durations are given as inputs to the optimisation models. This is in accordance with the explanation in Section 2.2.2.

### 5.4.1 Choice of scenario generation technique

One of the central elements of any stochastic model are the scenarios used to model the uncertainty, as described in Section 3.1. In our problems, the stochastic elements are the surgery durations and the number of emergency patients that arrive on a given day. The validity and practical value of the model is highly related to how realistically the generated scenarios represent the true stochastic processes. A good scenario generation

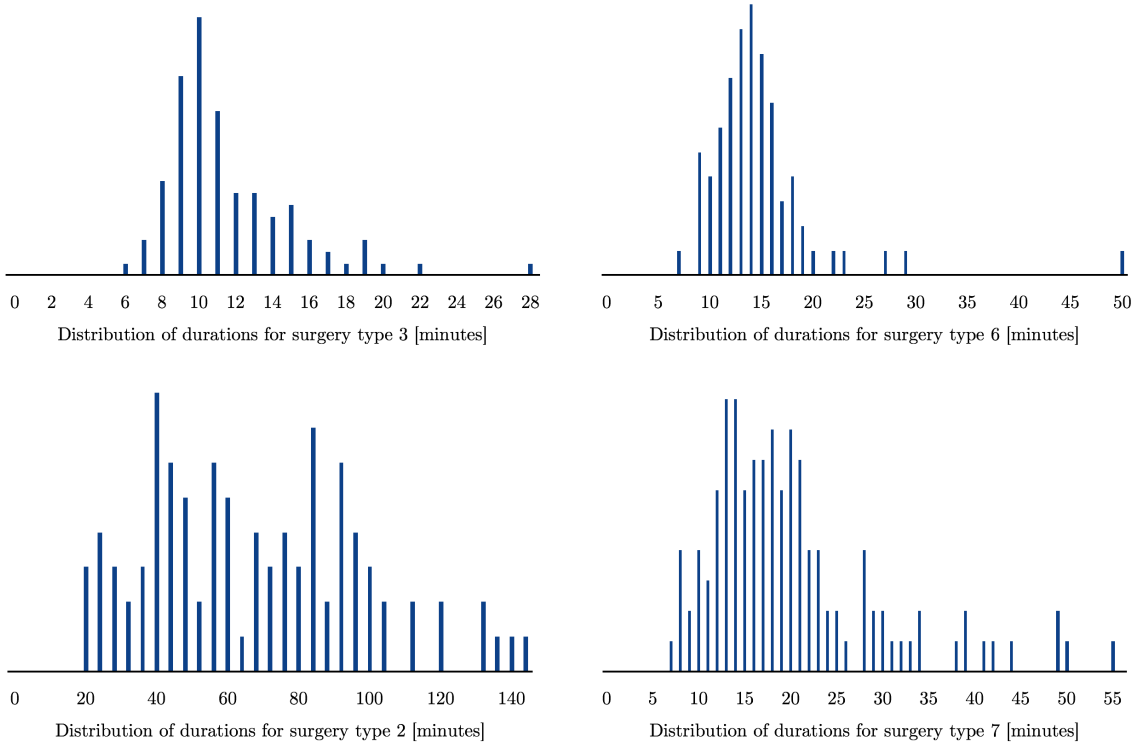


Figure 5.6: Distribution of empirical durations for surgery types 3, 6, 2 and 7. The vertical axes represents the relative frequency of each duration and is left out for readability.

method is therefore key to a good stochastic model, and the literature exhibit various different scenario generation techniques. Kaut and Wallace [78] provide an overview of common scenario generation methods, and discuss the evaluation of these for a given stochastic problem. In terms of sampling, one common decision is whether to sample from empirical data or from a fitted distribution. When sampling from empirical data, you avoid making any potentially wrongful assumptions on the nature of the distribution from which the data originates, but you need a high number of observations to be sure that the samples are in fact representative. Sampling from empirical data also requires the argument that the empirical data is a good representation of the future. On the other hand, sampling from a distribution completely relies on whether or not your assumption about the distribution is correct. Since most real-life processes are impossible to perfectly replicate using a known distribution, you are most likely to introduce a bias of some kind. Bratley [79] argues that quite often scenario generation is too concerned with fitting data to a distribution when, in fact, it would be more appropriate not to.

Figure 5.6 shows the distribution of durations for four different surgery types at St Olavs. Three of them exhibit similarities to a lognormal plot, which is a quite common assumption in literature. As discussed in Section 3.5, May [48] argues that a three-parameter lognormal fit is appropriate for surgery durations, and is cited by Mancilla [80], where lognormal surgery durations are used in a model for one surgeon on parallel operating rooms. Addis [81] also makes the assumption of lognormal surgery durations, in the environment of an advanced scheduling problem. However, for a lot of the surgeries in our data set, such as

surgery type 2 in Figure 5.6, this assumption is inaccurate, indicating that sampling from empirical data might be preferable to using a fitted distribution. In general, both sampling from empirical data and sampling from a specified fitted distribution is complicated when the variables are meant to be correlated. Firstly, you need to have enough data for the marginals distributions for all variables to be computed accurately with significance. It is also more complicated to maintain the desired marginal distributions of the resulting scenarios. In our case, however, we have assumed zero correlation. If we sample randomly, the correlation approaches zero as the number of scenarios increases, but we need to generate a very high amount of scenarios to achieve a natural correlation close to zero. Therefore, we want to be able to specify zero correlation explicitly.

Høyland et al. [82] propose a heuristic algorithm for scenario generation which ensures that the generated data matches the moments of the empirical data, and lets the user specify target correlations between the variables generated. While being based on empirical data, it has the advantage of letting the user modify the desired first four input moments of the generated variables if necessary. It uses the assumption that the distribution to be modelled can be described by its first four moments, requiring no assumptions about its nature. The algorithm is based on iterations between a cubic transformation solving a set of four non-linear equations in order to match the four first moments of the data, and a Cholesky transformation ensuring that the correlation is correct. Since the second transformation affects the higher moments, the algorithm iterates in such a way that the moments of the scenarios generated will match the original data after the Cholesky transformation is completed. It is an non-exact algorithm, but for a sufficiently low number of scenarios it is able to match both correlation and moments sufficiently. In terms of how many scenarios are needed to both replicate the first four moments and have zero correlation, it clearly outperforms a random scenario generation based on sampling from empirical data using Excel. Even with 10 000 scenarios, the random scenario generation in Excel results in correlations up to 0.1. The moment-matching algorithm, on the other hand, is able to almost exactly match the first four moments with close to zero correlation for a number of scenarios that is more than sufficiently low for the model to solve. This will be discussed in further detail in Section 5.4.3.

### 5.4.2 About the moment-matching algorithm

We use similar notation to the one used in the original paper by Høyland et al. [82] to briefly explain how the algorithm works. Let $n$ represent the number of random variables and $s$ the number of scenarios. $\tilde{X}$ is then a general $n$-dimensional random variable, and $\mathbb{X}$ is the $n \times s$-dimensional matrix of scenario outcomes, whose rows (vectors of outcomes for the $i$th variable) are referred to as $\mathbb{X}_i$. $\mathbb{P}$ is the row vector of scenario probabilities, specified by the user, and $\tilde{\mathcal{X}}$ is the discrete $n$-dimensional random variable given by $\mathbb{X}$ and $\mathbb{P}$. $\mathbb{E}[\tilde{X}]$ and $\mathbb{E}[\tilde{\mathcal{X}}]$ represent the vector of means of a random variable that is general or

discrete, respectively. The four target moments are denoted $\mu^*, \sigma^{2*}, \gamma^*$, and $\kappa^*$, with $R^*$ being the target correlation matrix. The goal is to generate scenarios with outcomes $\mathbb{Z}$, which together with probability vector $\mathbb{P}$ define the the discrete random variable $\tilde{\mathcal{Z}}$ (which is a discretisation of the theoretical random variable $\tilde{Z}$), with the specified target moments and target correlation. $\tilde{Y}$ denotes the intermediate variable, i.e. the intermediary results of all operations until the final transformation that results in $\tilde{Z}$. Note the difference in notation between random variables as abstract objects (such as $\tilde{X}$ and $\tilde{\mathcal{X}}$) as opposed to matrices of outcomes (such as $\mathbb{X}$).

The first step of the algorithm is to generate a standard normal random variable $\tilde{X}$ (having zero mean and unit variance). This variable is transformed using a cubic transformation on the form

$$\tilde{Y} = a + b\tilde{X} + c\tilde{X}^2 + d\tilde{X}^3 \tag{13}$$

to generate a variable $\tilde{Y}$, having moments $\mu \approx 0, \sigma^2 \approx 1, \gamma \approx \gamma^*$, and $\kappa \approx \kappa^*$ [3]. The cubic transformation in equation (13) is based on a method introduced by Fleishman [83] for generating a univariate non-normal variable with given first four moments, by solving a system of non-linear equations. In general, moment $j$ can be estimated by $E[(\tilde{Y} - E[\tilde{Y}])^j]$. Because of the standardisation, we have $E[\tilde{Y}] = 0$, so the expression for moment $j$ simplifies to

$$mom_j = E[\tilde{Y}^j] \tag{14}$$

By taking the expected value of either side of equation (13) we can express the four target moments as a system of equations dependent on the moments of the generated random variable $\tilde{X}$. Combining (13) with (14) gives the following system for the first four target moments

$$E[\tilde{Y}] = a + bE[\tilde{X}] + cE[\tilde{X}^2] + dE[\tilde{X}^3] \tag{15}$$

$$E[\tilde{Y}^2] = (a + bE[\tilde{X}] + cE[\tilde{X}^2] + dE[\tilde{X}^3])^2 \tag{16}$$

$$E[\tilde{Y}^3] = (a + bE[\tilde{X}] + cE[\tilde{X}^2] + dE[\tilde{X}^3])^3 \tag{17}$$

$$E[\tilde{Y}^4] = (a + bE[\tilde{X}] + cE[\tilde{X}^2] + dE[\tilde{X}^3])^4 \tag{18}$$

The algebraic extension of the latter equation includes an expression with $E[\tilde{X}^{12}]$, which due to relation (14) is equal to the twelfth moment of variable $\tilde{X}$. This, and the other

[3] The reason for the standardisation is that it makes the transformations far easier. It can be shown that neither the higher moments (skewness and kurtosis) nor the correlation are distorted when the algorithm finally transforms this back to having $\mu = \mu^*$ and $\sigma^2 = \sigma^{2*}$.

eleven moments of lower order, need to be computed for $\tilde{X}$. Note that the left hand sides, $E[\tilde{Y^k}], k = 1, ..., 4$, are taken as an input to the algorithm, but the first target moments are set to 0 and 1, respectively, as explained above. This leaves us with a system of four non-linear equations we can use to determine the four transformation parameters $a$, $b$, $c$ and $d$, which is done approximately using a least-squares method. When the transformation parameters are obtained, we use (13) to transform our $\tilde{X}$ to $\tilde{Y}$, with $\tilde{Y}$ having $\mu \approx 0, \sigma^2 \approx 1, \gamma \approx \gamma^*$, and $\kappa \approx \kappa^*$.

The other main transformation in the algorithm is a Cholesky transformation to ensure right correlation. Since changing the correlation distorts the moments of higher than second order, the algorithm iterates between the cubic transformation and the Cholesky transformation in order to reach both the target moments and the target correlation. The algorithm first transforms the variable set to having approximately zero correlation[4] and afterwards it transforms it again to have approximately target correlation. In order to impose zero correlation, $\mathbb{Y}$'s correlation matrix $R$ is decomposed using a Cholesky decomposition on the form $R = LL^T$, where $L$ is a lower triangular matrix. Then, a backward transformation on the form $\mathbb{Y} = L^{-1}\mathbb{Y}$ updates $\mathbb{Y}$ and we end up with close to zero correlation. This, however, changes the moments, so the algorithm performs another cubic transformation and iterates between these steps until the correlation is sufficiently close to zero. The next transformation also uses a Cholesky decomposition, but this time using the target correlation matrix. With $R^* = LL^T$, a forward transformation $\mathbb{Y} = L\mathbb{Y}$ is performed. In our case, $R^* = I$, so the Cholesky decomposition gives $L = I$, which means the transformation has no effect. The variables now have approximately right correlation, and moments $\mu \approx 0, \sigma^2 \approx 1, \gamma \approx \gamma^*$, and $\kappa \approx \kappa^*$, so the final step is transforming the variables back to the desired two first moments. A linear transformation

$$\mathbb{Z} = \alpha\mathbb{Y} + \beta \tag{19}$$

transforms the standardised variable $\tilde{\mathcal{Y}}$ to a variable $\tilde{\mathcal{Z}}$ with mean equal to $\beta$ and standard deviation equal to $\alpha$. This comes from the formula for standardising a random variable, and it can be shown that this distorts neither the correlation nor the higher moments. The $\beta$ term simply shifts all variables by a constant, whereas the $\alpha$ term scales the deviation by a constant factor, neither of which affect the skewness or the kurtosis. Setting $\beta = \mu^*$ and $\alpha = \sigma^{2*}$ we obtain a variable $\tilde{\mathcal{Z}}$ having correlation close to $R^*$ (= $I$ in our case) and first four moments close to $\mu^*, \sigma^{2*}, \gamma^*$, and $\kappa^*$, respectively.

[4]Note that although the variable is randomly generated it is likely to exhibit some correlation when the number of scenarios is limited. For an infinite number of scenarios the correlation of the randomly drawn variables $\tilde{X}$ would be zero.

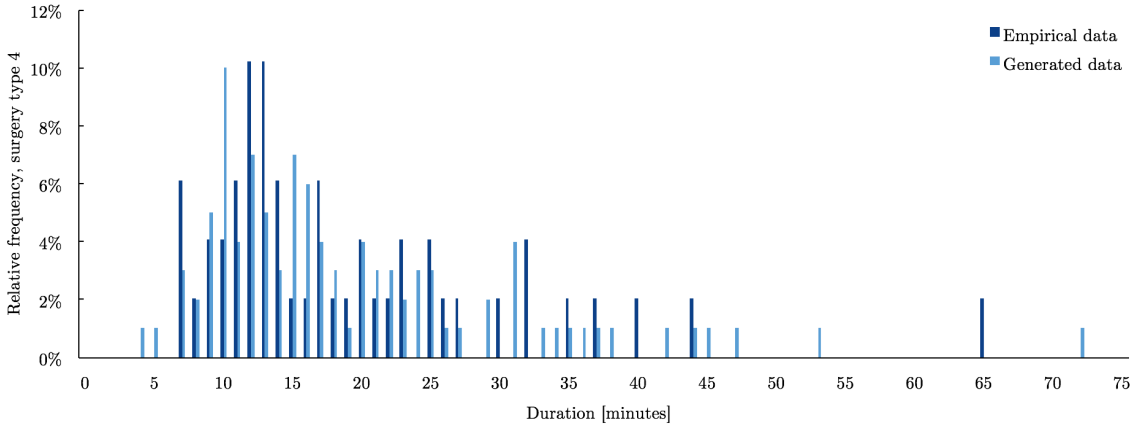



Figure 5.7: Comparison of empirical data and data generated by the moment-matching algorithm, for surgery type 4

### 5.4.3 About the results of the scenario generation

For the Phase Model, the distribution of the durations of a given surgery depends on the phase in which each surgery is started. When generating scenarios, we have split the durations for each surgery into two separate sets, and we thus generate two different independent variables for each surgery type. This means that for an instance with four surgeries, we generate eight independent variables, each representing a discretisation of the distribution of durations for each surgery in both phases. In order to evaluate the scenario generation we want to evaluate how well the data we have generated matches our empirical data. For instance, for surgery type 4, used in Figures 5.3-5.5, Figure 5.7 shows the empirical durations in phase 1 together with the corresponding durations generated by the algorithm. By observation, the data generated using the moment-matching algorithm seems to be a good representation of the empirical data, which is true for all surgeries in the instances we use. For each of the 37 surgeries we have compared the moments of the generated data to the moments of the empirical data. In addition, we have checked, for each of the 20 instances, the correlations between the variables generated, which were meant to be zero. Figures 5.8 and 5.9 show the results of both of these analyses. In Figure 5.9, we have computed the correlation between the variables within each instance and displayed them as a frequency plot. The magnitude of the relative error for all four moments are in the range $10^{-10}$ to $10^{-4}$, and the correlation is $10^{-3}$ at most. This is considerably better than what we could achieve by generating the same amount of scenarios manually. By using the algorithm we are able to maintain the correct statistical properties as well as approximately zero correlation while keeping the number of scenarios low enough for the model to solve in acceptable time.

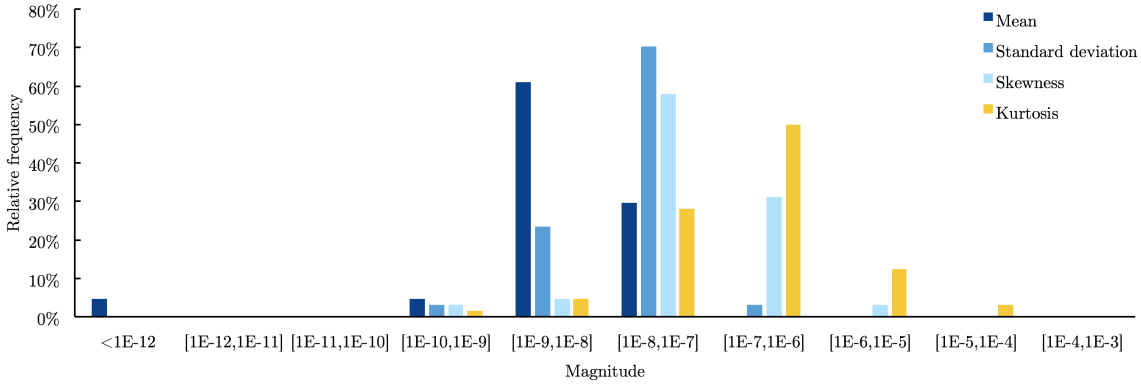



Figure 5.8: Relative errors of the moments of all variables generated

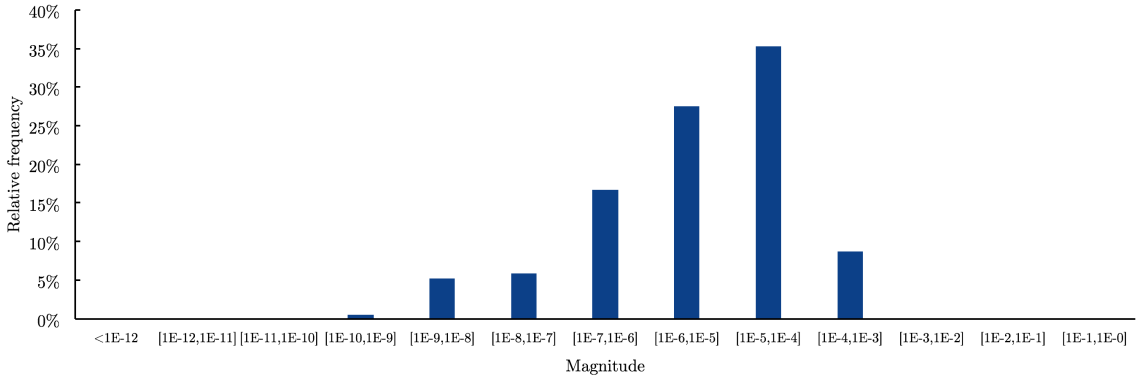



Figure 5.9: Correlation between all pairs of variables generated

# Chapter 6

# Model formulations

This chapter will present two mathematical models designed to solve the two surgery scheduling problems explained in Chapter 4. For each model, respectively, we will go through design considerations and notation before presenting the formulation in its entirety. Moreover, we will discuss challenges posed by the models, and propose different ways of strengthening the formulations and guiding the solution process. The basic structure of the models is inspired by the one presented by Mancilla and Storer [17].

## 6.1 Phase Model

The first mathematical model we present is a two-stage stochastic model with both integer and continuous decision variables. It incorporates stochastic surgery durations, where the probability distribution of the duration is dependent on the choice of actual start time of a surgery. This model only concerns elective patients.

### 6.1.1 Model design considerations

One of the most important parts of building a stochastic model is a description of the information structure, which requires reflections on when decisions have to be made and the information available at every decision. The uncertain elements in this model are the surgery durations, which are not known before each surgery finishes. This uncertainty affects the start of the next surgery, which needs to wait for the operating room to be ready. Assumption 1, described in Chapter 4, states that patients always arrive in time for their scheduled surgery start time. Since patients need to know their surgery appointment time in advance, scheduled start times must be decided before knowing the actual durations of the preceding surgeries on the surgery day. The first decision is therefore to decide scheduled start times for the given set of surgeries, with information about the probability

distribution of the duration of each surgery. Since all patients arrive in time, the first surgery of the day will always start as scheduled. Once the first surgery is finished, with its duration being known, the actual start time of the next surgery is decided. Note that this decision is a trivial one: If the scheduled start time of the next surgery has been reached, meaning that the patient has arrived, it starts immediately as there is no reason to make him wait. If not, it starts at the planned start time, i.e. as soon as the patient arrives. It would be less trivial if it was an option to change the sequence of the scheduled surgeries, because it would require an evaluation of which surgery to start next. However, changing the sequence of two surgeries would not be a relevant option since no other patients are ready for surgery[5]. Hence, as each surgery finishes, the next surgery is started as soon as possible, and once all surgeries are finished, potential overtime is calculated. The essence of the problem, then, is how to set the scheduled start time for all surgeries in order to minimise the expected weighted sum of waiting time, idle time and overtime. Setting scheduled start times for all surgeries also implies setting the sequence, and this will be used in the modelling.

The information structure explained in the above paragraph, including the decisions to be made, is illustrated by Figure 6.1 for an example of four surgeries, with a branching factor of two, meaning each surgery duration has two possible realisations. The internal nodes in the tree represent the decisions, described by the black text on the right hand side of the figure, while each new stage represents the revelation of a stochastic element, described by the blue text. With this structure, the number of stages equals the number of surgeries plus one. The dotted lines emphasise the nonanticipativity constraints, i.e. the decision levels with the same amount of information. The information structure can use any branching factor, where a higher factor gives a high increase in the number of scenarios. Given a branching factor $b$ that is equal for $n$ surgeries with independent durations, the number of possible scenarios is given by $b^{n+1}$. The path from the root node to a given leaf node, along a solid line, represents our scenarios, referred to by $\omega$.

The information structure in Figure 6.1 is the one we consider to most closely resemble the actual decision process. As noted above, the essence of the problem is setting the scheduled start times. The remaining decisions of how to set actual start times is determined trivially by the principle of starting each surgery as soon as possible, given the durations of preceding surgeries. Since the only relevant information when determining the actual start of a surgery are the durations of the preceding surgeries, it does not make a difference whether you know the durations of the following surgeries or not. For this reason, we can picture a simplified two-stage structure, as shown in Figure 6.2, where the durations of all surgeries are revealed at the same time, after scheduled start times are set. In both structures, the only information used for determining the start time of a surgery are the durations of the preceding surgeries, meaning that the simplification of

[5] Actually, in the case of being so far behind schedule that more than one patient have already arrived, it would be an option to change the sequence of these two, but this is assumed to be a special case.

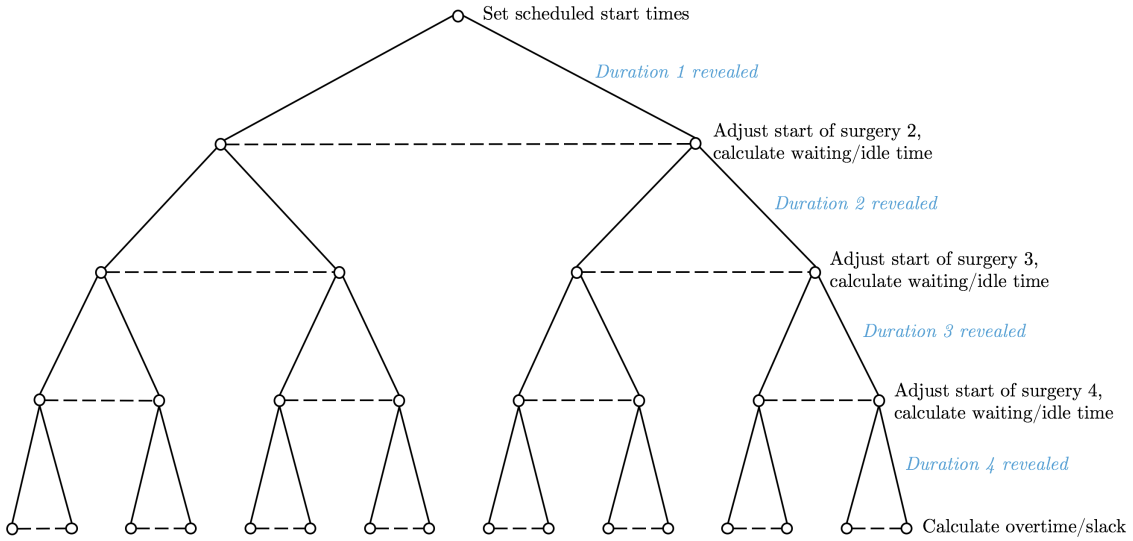



Figure 6.1: Example of complete information structure, with $n+1$ stages

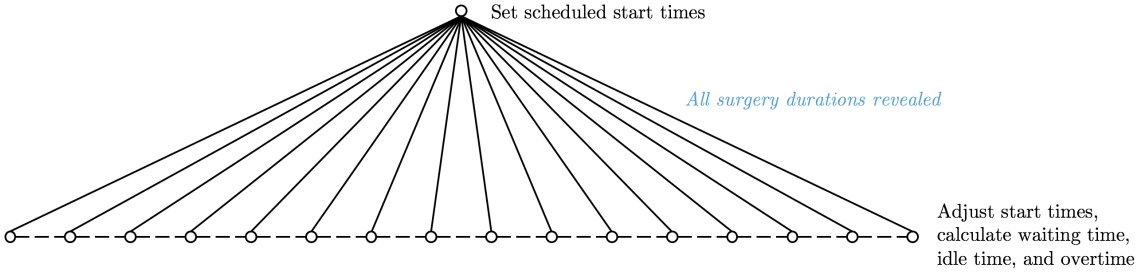



Figure 6.2: Simplified information structure with 2 stages

assuming the durations of the following surgeries to be known, as well, has no practical implications. The two-stage structure can be set up to have the same scenarios as the multi-stage structure and, from this point, when referring to the information structure of the Phase Model, we refer to the structure shown in Figure 6.2.

The Phase Model is designed to let the distributions from which the durations are generated depend on the start time of the surgeries. Section 5.2.3 divided the day into two phases to check if there was reason to believe the expected duration for some surgeries would be different based on whether it is scheduled before or after lunch. The same division of the day into two phases is used in the Phase Model. Thus, a given scenario does in fact give two surgery duration realisations; one for the duration in phase 1 and another for phase 2. Therefore, the scenario alone does not completely specify the surgery duration, but must be used in conjunction with the decision about in which phase the surgery is started. The decision-dependent uncertainty resulting from this is one of the complicating factors of the model.

### 6.1.2 Sets and indices

Let $N$ be the set of all surgeries that should be performed, and the set of positions in the sequence of surgeries. Surgeries are indexed by $j$, and positions by $i$. $N$ is defined as $N = \{1, 2, ..., n\}$, where $n$ is the number of surgeries to schedule. Further, let $H$ be the set of all phases throughout the day. Our formulations are, however, not completely general in the sense that this set cannot consist of more than two phases for the formulations to be correct. Lastly, let $\Omega$ be the set of all surgery duration scenarios. The described sets and corresponding indices are given in Table 6.1.

Table 6.1: Sets and indices in the Phase Model

| **Set** | **Description** |
|---|---|
| $N$ | Positions and surgeries, indexed by $i$ and $j$ |
| $H$ | Phases of the day, indexed by $h$ |
| $\Omega$ | Scenarios of surgery durations, indexed by $\omega$ |

### 6.1.3 Parameters

Let $D^{\omega}_{jh}$ be the duration of surgery $j$ if the surgery is in phase $h$ in scenario $\omega$. As discussed, the scenario partly indicates where to look up the duration, which is also dependent on the phase. $n$ denotes the last element of the set $N$. $d$ is the point in time which defines the end of the regular working day, meaning that any surgery time after this represents overtime.

Let $c^w_j$, $c^s_j$ and $c^l_j$ be the unit cost of waiting time, idle time and overtime, respectively, for surgery $j$. These can represent either monetary values or relative weights between the three measures. In the model formulation, these are found in the objective function.

$M^{\delta}_i$, $M^s_i$, $M^w_i$, $M^l_i$, and $M^{\text{late}}$ are used in big-M formulations and apart from the last M, the superscripts correspond to the variable they affect. Mathematically, any sufficiently large number will make the formulations valid. Lastly, let $P^{\text{split}}$ define the point in time which separates the first phase from the second phase.

An overview of the parameters found in the model formulation is given in Table 6.2.

### 6.1.4 Variables

The only variables that do not vary with the scenarios, are the $x_{ij}$ and $t_i$ variables. These constitute the first stage variables and define the planned surgery schedule. $x_{ij}$ are binary variables that define the sequence of surgeries. They take the value 1 if surgery $j$ is in position $i$, and 0 otherwise. The $t_i$ variables are continuous variables that state the

Table 6.2: Parameters in the Phase Model

| Parameter | Description |
|---|---|
| $D^{\omega}_{jh}$ | Duration of surgery $j$ in phase $h$ in scenario $\omega$ |
| $d$ | Point of time beyond which overtime is incurred |
| $n$ | Last element of the set $N$ |
| $c^w_j$ | Waiting time penalty for surgery $j$ |
| $c^s_j$ | Idle time penalty for surgery $j$ |
| $c^l_j$ | Overtime penalty for surgery $j$ |
| $M^{\delta}_i, M^s_i, M^w_i, M^l_i, M^{\text{late}}$ | Sufficiently large numbers |
| $P^{\text{split}}$ | Point in time defining the split between phase 1 and 2 |

scheduled start time of the surgery that is assigned to position $i$. The start time for the first surgery, $t_1$, is always set to 0.

There are a total of seven other variable groups that are part of the second stage. The first variables are $\tau^{\omega}_{ih}$, which set the phase of a surgery. More formally, $\tau^{\omega}_{ih}$ take the value 1 if the surgery is in position $i$ in phase $h$, and 0 otherwise. Since the decision of start time gives the phase, the decision-dependent uncertainty is modelled through the $\tau^{\omega}_{ih}$ variables, which in combination with the $D^{\omega}_{jh}$ matrix, select the appropriate surgery duration distribution to use.

To calculate the duration of the surgery in a given position, one need to to multiply the phase variable, the position variable and the duration matrix. To be able to keep our model linear, we introduce the variables $y^{\omega}_{ijh} = x_{ij}\tau^{\omega}_{ih}$, which can be linearised with some additional constraints. In this definition, $y^{\omega}_{ijh}$ vary with the scenario through the phase variables. However, the first stage decisions are conserved within the $x_{ij}$ variables, which are transferred to $y^{\omega}_{ijh}$.

The three variable groups $w^{\omega}_{ij}$, $s^{\omega}_{ij}$, and $l^{\omega}_{ij}$ are the waiting time, idle time and overtime, respectively. These are continuous variables, decided by the adjustments to the schedule made according to delays or early finishes in the preceding surgeries. If a surgery $j$ is scheduled to position $i$ in some scenario $\omega$, then $w^{\omega}_{ij}$ can take a positive value for the amount of time that surgery must wait between the scheduled start time and the actual surgery start time. Similarly, $s^{\omega}_{ij}$ are the amount of time after surgery $j$ in position $i$ that is not used for surgery, given that surgery $j$ is in position $i$. Lastly, $l^{\omega}_{ij}$ can be positive when surgery $j$ is in position $i$ and is the amount of time of that surgery that is after time $d$.

Variables $g^{\omega}$ are continuous slack variables that measure the earliness with respect to $d$ in a given scenario $\omega$. That is, they are the time between the end of the last surgery and the end of the regular working day. These variables are not penalised, and are used in the model formulation to balance time constraints. If there is any overtime in a given scenario, these are equal to 0.

The last variables are $\delta_i^\omega$. These are binary indicator variables used to correctly identify the overtime of each surgery, and are mathematically defined as

$$\delta_i^\omega = \begin{cases} 1 & \text{if } \sum_{j \in N} w_{ij}^\omega + t_i \geq d \\ 0 & \text{elsewhere} \end{cases} \tag{20}$$

$\delta_i^\omega$ take the value 1 when the entire surgery in position $i$ is in overtime, i.e., when the *actual* starting point of the surgery is greater than $d$. This relationship will be expressed as two linear constraints in the mathematical model. A brief summary of the variables is given in Table 6.3.

Table 6.3: Variables in the Phase Model

| **Variable** | **Description** |
|---|---|
| $x_{ij}$ | 1 if surgery $j$ is in position $i$ |
| $t_i$ | Scheduled start time of the surgery in position $i$ |
| $\tau_{ih}^\omega$ | 1 if the surgery in position $i$ is in phases $h$ in scenario $\omega$ |
| $y_{ijh}^\omega$ | 1 if surgery $j$ is in position $i$ and in phase $h$ in scenario $\omega$ |
| $w_{ij}^\omega$ | Waiting time when surgery $j$ is in position $i$ in scenario $\omega$ |
| $s_{ij}^\omega$ | Idle time after surgery $j$ when it is in position $i$ in scenario $\omega$ |
| $l_{ij}^\omega$ | Over time when surgery $j$ is in position $i$ in scenario $\omega$ |
| $\delta_i^\omega$ | 1 if entire surgery in position $i$ is in overtime in scenario $\omega$ |
| $g^\omega$ | Slack variable that measures the earliness with respect to $d$ in scenario $\omega$ |

### 6.1.5 Phase Model formulation

This section presents the mathematical model formulation of the scheduling problem with phase-dependent surgery durations. After the full model formulation, the objective and all constraints will be explained in detail. The reader is referred to the preceding sections for all notation definitions.

$$\min \sum_{\omega \in \Omega} \frac{1}{|\Omega|} \left[ \sum_{i \in N} \sum_{j \in N} (c_j^w w_{ij}^\omega + c_j^s s_{ij}^\omega + c_j^l l_{ij}^\omega) \right] \tag{21}$$

$$\text{s.t} \quad t_i - t_{i+1} - \sum_{j \in N} w_{i+1,j}^\omega + \sum_{j \in N} s_{ij}^\omega + \sum_{j \in N} w_{ij}^\omega = - \sum_{j \in N} \sum_{h \in H} D_{jh}^\omega y_{ijh}^\omega \quad i \in N \setminus \{|N|\}, \omega \in \Omega \tag{22}$$

$$t_n + \sum_{j \in N} w_{nj}^\omega - \sum_{i \in N} \sum_{j \in N} l_{ij}^\omega + g^\omega = - \sum_{j \in N} \sum_{h \in H} D_{jh}^\omega y_{njh}^\omega + d \quad \omega \in \Omega \tag{23}$$

$$\sum_{j \in N} l_{ij}^\omega \geq \sum_{j \in N} \sum_{h \in H} D_{jh}^\omega y_{ijh}^\omega - M_i^\delta (1 - \delta_i^\omega) \quad i \in N, \omega \in \Omega \tag{24}$$

$$\sum_{j \in N} l_{ij}^\omega \geq t_i + \sum_{j \in N} w_{ij}^\omega + \sum_{j \in N} \sum_{h \in H} D_{jh}^\omega y_{ijh}^\omega - d - M_i^\delta \delta_i^\omega \quad i \in N, \omega \in \Omega \tag{25}$$

$$M_i^{\delta}\delta_i^{\omega} \geq \sum_{j\in N} w_{ij}^{\omega} + t_i - d \quad i \in N, \omega \in \Omega \tag{26}$$

$$M_i^{\delta}(1-\delta_i^{\omega}) \geq d - (t_i + \sum_{j\in N} w_{ij}^{\omega}) \quad i \in N, \omega \in \Omega \tag{27}$$

$$\sum_{i\in N} x_{ij} = 1 \quad j \in N \tag{28}$$

$$\sum_{j\in N} x_{ij} = 1 \quad i \in N \tag{29}$$

$$s_{ij}^{\omega} \leq M^s x_{ij} \quad i \in N, j \in N, \omega \in \Omega \tag{30}$$

$$w_{ij}^{\omega} \leq M_i^w x_{ij} \quad i \in N, j \in N, \omega \in \Omega \tag{31}$$

$$l_{ij}^{\omega} \leq M_j^l x_{ij} \quad i \in N, j \in N, \omega \in \Omega \tag{32}$$

$$y_{ijh}^{\omega} \leq x_{ij} \quad i \in N, j \in N, h \in H, \omega \in \Omega \tag{33}$$

$$y_{ijh}^{\omega} \leq \tau_{ih}^{\omega} \quad i \in N, j \in N, h \in H, \omega \in \Omega \tag{34}$$

$$y_{ijh}^{\omega} \geq x_{ij} + \tau_{ih}^{\omega} - 1 \quad i \in N, j \in N, h \in H, \omega \in \Omega \tag{35}$$

$$t_i + \sum_{j\in N} w_{ij}^{\omega} \geq P^{\text{split}}(1-\tau_{ih}^{\omega}) \quad i \in N, \omega \in \Omega, h \in \{1\} \tag{36}$$

$$t_i + \sum_{j\in N} w_{ij}^{\omega} - P^{\text{split}} \leq M^{\text{late}}\tau_{ih}^{\omega} \quad i \in N, \omega \in \Omega, h \in \{2\} \tag{37}$$

$$\sum_{h\in H} \tau_{ih}^{\omega} = 1 \quad i \in N, \omega \in \Omega \tag{38}$$

$$x_{ij} \in \{0,1\} \quad i \in N, j \in N \tag{39}$$

$$\tau_{ih}^{\omega} \in \{0,1\} \quad i \in N, h \in H, \omega \in \Omega \tag{40}$$

$$t_1 = 0, \quad t_i \geq 0 \quad i \in N \tag{41}$$

$$\delta_i^{\omega} \geq 0 \quad i \in N, \omega \in \Omega \tag{42}$$

$$w_{ij}^{\omega} \geq 0, \quad s_{ij}^{\omega} \geq 0, \quad l_{ij}^{\omega} \geq 0 \quad i \in N, j \in N, \omega \in \Omega \tag{43}$$

$$y_{ijh}^{\omega} \geq 0 \quad i \in N, j \in N, h \in H, \omega \in \Omega \tag{44}$$

$$g^{\omega} \geq 0 \quad \omega \in \Omega, \quad \omega \in \Omega \tag{45}$$

The model minimises the objective function (21), which is a weighted sum of all waiting time, idle time, and overtime for elective patients. In a regular stochastic model, the objective contains the expected second-stage cost which, for discrete stochastic scenarios, can be calculated as a sum over all second-stage costs multiplied by the probability of that scenario. However, when we introduce decision-dependent uncertainty, the probability of a given scenario is not given by the regular expression $p_{\omega}$, but rather by a function $p(\omega, \tau_{ih}^{\omega})$. This probability multiplied with the weighted sum of waiting time, idle time, and overtime is clearly non-linear. To be able to utilise linear optimisation theory, we choose to set

$p(\omega, \tau_{ih}^{\omega}) = \frac{1}{|\Omega|}$. The probability of a given scenario is therefore constant and equal for all scenarios.

The objective function given by (21), and equations (22), (23), (28) and (29), are similar to those proposed by Mancilla and Storer [17]. Equation (22) defines the waiting time and idle time for every surgery and scenario, by balancing these times with the starting times and duration of all subsequent surgeries. This may be more easily interpreted by noting that $t_i + \sum_{j \in N} w_{ij}^{\omega}$ is the actual starting time of the surgery in position $i$. The equation then states that the difference between the actual starting time of two subsequent surgeries, plus any idle time between them, must equal the duration of the first of these two surgeries. Equation (23) gives a similar balance for the actual start of the last surgery, total overtime, any potential slack at the end of the day, surgery duration and the defined end of regular working hours. Equation (22) is adapted to handle the different phases. Similarly, equation (23) is adapted to handle overtime given for each patient. Compared to the formulation by Mancilla and Storer [17], the surgery specific overtime variables actually make equation (23) redundant, because additional constraints are needed to allocate the overtime appropriately among the surgeries. However, the model performs significantly better with these constraints, making the constraints good valid inequalities. For completeness, the constraints are kept in the model formulation instead of placed with other useful valid inequalities in Chapter 6.1.7.

Constraints (24) and (25) define the overtime for a given position and surgery. Each of these constraints contains a big-M part, depending on whether the entire surgery is in overtime or not. This means that at most one of the two inequalities can have a positive right hand side (RHS). If the entire surgery is in overtime, the $\delta_i^{\omega}$ variables are equal to 1 and constraints (24) state that the overtime must be at least equal to the surgery duration. On the other hand, if $\delta_i^{\omega}$ are zero, constraints (25) force the overtime to be greater than or equal to the part of the surgery that is in overtime (or zero if it ends before $d$). Because the objective is minimised, the overtime will be equal to the highest of the RHS of (24), the RHS of (25), and 0. Equation (26) utilises a big-M notation to force $\delta_i^{\omega}$ to 1 when the entire surgery in position $i$ is in overtime, while equation (27) makes $\delta_i^{\omega}$ take the value 0 in the opposite case. Equation (28) ensures that each surgery is assigned to a position and (29) makes sure every position has a surgery. Equations (30), (31) and (32) force the idle time, waiting time and overtime, respectively, to zero, if a surgery $j$ is not assigned to position $i$. Constraints (33), (34) and (35) linearise the relationship between $y_{ijh}^{\omega}$, $x_{ij}$ and $\tau_{ih}^{\omega}$. Equation (33) forces $y_{ijh}^{\omega}$ to 0 when the corresponding $x_{ij}$ is 0, while (34) forces $y_{ijh}^{\omega}$ to 0 when the corresponding $\tau_{ih}^{\omega}$ is 0. Each of these constraints is redundant when $x_{ij}$ or $\tau_{ih}^{\omega}$ are 1, respectively. When both $x_{ij}$ and $\tau_{ih}^{\omega}$ are 1, the constraints (35) force $y_{ijh}^{\omega}$ to take the value 1. When either $x_{ij}$ or $\tau_{ih}^{\omega}$ are 0, these constraints are redundant. Constraints (36) and (37) set the phase of the surgery in a given position based on the actual start time. More precisely, if the actual start time is before the split between the phases, the constraints (36) force $\tau_{ih}^{\omega}$ to 1 for phase $h = 1$. On the other hand, if the actual start time

is after the split between the phases, equation (37) sets $\tau_{ih}^{\omega}$ to 1 for phase $h = 2$. To make sure exactly one phase is selected for each position, constraints (38) are needed as well. The remaining constraints, in equations (39) - (45), are the variable domains.

The reader should note that the overtime formulations are not completely precise in all cases. In some rare occasions, the sum of overtimes for all surgeries is not equal to the total overtime. This anomaly appears because of the equality sign in balance constraints (23) in combination with the possibility of idle time between two surgeries in overtime. To balance these constraints, this idle time is also counted as overtime, but does not really belong to any of the surgeries in overtime. However, because constraints equations (24) and (25) are $\geq$ -constraints, the idle time can still be allocated to one of the surgeries. If there are several surgeries with overtime, this idle time is thus assigned to the surgery with the lowest cost of overtime. Because this problem rarely occurs and the amount of idle time is small when it happens, we do not introduce extra notation to handle these cases. In practice, this means that this special case of idle time is penalised both as any other idle time and as overtime.

### 6.1.6 Strengthening the big-M formulations

The big-M used in the MIP formulation may negatively affect the performance of the model if not appropriately set. The formulations should therefore be as tight as possible.

$M^s$, from equation (30) is given by

$$M^s = \max_{j \in N} \left\{ \max_{h \in H, \omega \in \Omega} \left\{ D_{jh}^{\omega} \right\} - \min_{h \in H, \omega \in \Omega} \left\{ D_{jh}^{\omega} \right\} \right\} \tag{46}$$

which is the largest value $s_{ij}^{\omega}$ can take. This will be the case if the surgery with the largest difference in possible duration, is scheduled to require its maximum duration, but actually ends up lasting its minimum duration.

From equation (31), $M_i^w$ is set to

$$M_i^w = \sum_{j=1}^{i-1} a_j \tag{47}$$

where $a_j$ is the $j$th largest value in

$$\max_{h \in H, \omega \in \Omega} \left\{ D_{jh}^{\omega} \right\} - \min_{h \in H, \omega \in \Omega} \left\{ D_{jh}^{\omega} \right\} \tag{48}$$

$M_i^w$ is therefore the largest possible value of $w_{ij}^{\omega}$. This happens when all previous surgeries have been scheduled to last their minimum duration, while they actually take their

maximum. The validity of these two big-M formulations are proved by Mancilla and Storer [80].

$M_j^l$ from equation (32) can be set to the maximum amount of overtime a surgery can suffer, which is equal to the longest surgery duration for that surgery. Mathematically, this can be expressed as

$$M_j^l = \max_{h=2,\omega\in\Omega} \left\{ D_{jh}^{\omega} \right\} \tag{49}$$

with $h = 2$, because overtime can only occur in phase two.

$M_i^{\delta}$ in constraints (26) and (27) can be strengthened to

$$M_i^{\delta} = \max_{\omega\in\Omega} \left\{ \sum_{j\in N} w_{ij}^{\omega} + t_i - d, d - \sum_{j\in N} w_{ij}^{\omega} + t_i \right\} \tag{50}$$

that is, the maximum of the difference between the latest possible scheduled start and $d$, and the difference between the earliest possible start time and $d$. This is equivalent to

$$M_i^{\delta} = \max \left\{ \sum_{j=1}^{i-1} b_j^{\max} - d, d - \sum_{j=1}^{i-1} b_j^{\min} \right\} \tag{51}$$

where $b_j^{\max}$ is the $j$th largest possible realisation of surgery durations for all surgeries, i.e. $j$th largest of $\max_{h\in H,\omega\in\Omega} D_{jh}^{\omega}$ and $b_j^{\min}$ is the $j$th smallest of $\min_{h\in H,\omega\in\Omega} D_{jh}^{\omega}$.

Similarly, $M^{late}$ in constraints (37) can be tightened to the latest possible start time less the parameter $P^{\text{split}}$. Using the same definition of $b_j^{\max}$ as above, this is given by

$$M^{\text{late}} = \sum_{j=1}^{n-1} b_j^{\max} - P^{\text{split}} \tag{52}$$

### 6.1.7 Valid inequalities

The model formulation in the previous section is sufficient to find optimal solution, but the LP-relaxation includes many infeasible solutions making the formulation inefficient. In order to to strengthen the formulation and LP relaxation of the formulation further, we introduce several valid inequalities.

Several papers concerning valid inequalities were discussed in Chapter 3. Applegate and Cook [32] look at several cuts for the job shop problems, which we adapt to the surgery scheduling problem. Different from the problem we face, the jobs in the job shop problem

need to be scheduled subsequently on several machines. It is in that case meaningful to discuss measures, such as the earliest possible starting time on a given machine, when generating cuts. On our single machine (operating room), this will simply be 0, which makes the cuts proposed by Applegate and Cook [32] much weaker in our formulation.

The first cut they consider is given by equation (53). In this equation, $E_{N\alpha}$ denote the earliest possible starting times of all jobs, $j \in N$, on machine $\alpha$, and the equation simplifies to equation (54), because $E_{N\alpha} = 0$ in our problem.

$$\sum_{i\in N}\sum_{j\in N}\sum_{h\in H} D_{jh}^{\omega} t_{ij} \geq E_{N\alpha}\sum_{j\in J}\sum_{h\in H} D_{jh}^{\omega} + \sum_{i\in N}\sum_{j\in N|j<i}\sum_{h\in H} D_{jh}^{\omega} D_{ih}^{\omega} \quad \omega \in \Omega \tag{53}$$

$$\sum_{i\in N}\sum_{j\in N}\sum_{h\in H} D_{jh}^{\omega} t_{ij} \geq \sum_{i\in N}\sum_{j\in N|j<i}\sum_{h\in H} D_{jh}^{\omega} D_{ih}^{\omega} \quad \omega \in \Omega \tag{54}$$

Specific for our problem, we have found that the following proposed inequalities improve the formulation. Most proofs follow the same structure, and we have therefore omitted them in the text. They can, however, be found in Appendix A.

**Proposition 6.1.** *The precedence inequality*

$$\tau_{i1}^{\omega} \geq \tau_{i+1,1}^{\omega} \tag{55}$$

*is valid for all* $i \in N \setminus \{|N|\}$ *and* $\omega \in \Omega$.

*Proof.* These constraints are easily confirmed using the definition of $\tau_{ih}^{\omega}$. Assume that for a given position $i$, the corresponding $\tau_{i1}^{\omega} = 0$. Then

$$\tau_{i1}^{\omega} = 0 \implies P^{\text{split}} \leq t_i + \sum_{j\in N} w_{ij}^{\omega}$$

From balance equation (22) we know that

$$t_i + \sum_{j\in N} w_{ij}^{\omega} \leq t_{i+1} + \sum_{j\in N} w_{i+1,j}^{\omega}$$

and thus

$$P^{\text{split}} \leq t_{i+1} + \sum_{j\in N} w_{i+1,j}^{\omega} \implies \tau_{i+1,1}^{\omega} = 0 \leq \tau_{i1}^{\omega}.$$

If $\tau_{i1}^{\omega} = 1$, then this cut does not restrict $\tau_{i+1,1}^{\omega}$. This cut is equivalent to $\tau_{i2}^{\omega} \leq \tau_{i+1,2}^{\omega}$. □

Proposition 6.1 state that if the surgery in position $i$ is in phase 2, then the surgery in position $i+1$ must also be in phase 2.

**Proposition 6.2.** *The inequalities*

$$t_{i+1} \geq t_i + \min_{h \in H, \omega \in \Omega} \sum_{j \in N} D_{jh}^{\omega} x_{ij} \tag{56}$$

*are valid for all* $i \in N \setminus \{|N|\}$ *and must be satisfied in the optimal solution.*

The inequalities in Proposition 6.2 cut several LP solutions by setting a minimum and maximum spread between surgeries in subsequent positions. The inequalities state that the surgery in position $i+1$ must start at least as long after the surgery in position $i$ as the shortest possible duration of the surgery in that position. If the starting intervals are closer, the surgery in position $i+1$ will introduce waiting time in every scenario, which is never optimal. In addition, they state that the surgery in position $i+1$ must start at least as long after the surgery in position $i$ as the longest duration realisation of the surgery in that position. If the starting intervals are further apart, this will introduce idle time in every scenario.

**Proposition 6.3.** *The following precedence constraints*

$$\delta_i^{\omega} \leq \delta_{i+1}^{\omega} \tag{57}$$

*are valid for all* $i \in N \setminus \{|N|\}$ *and* $\omega \in \Omega$.

Proposition 6.3 states that if the entire surgery in position $i$ is in overtime, then the entire surgery in position $i+1$ must be in overtime. This is a trivial result, and cuts away some possible configurations of $\delta_i^{\omega}$. However, because there are, generally, few surgeries with a positive $\delta_i^{\omega}$, this is not a very strong cut.

**Proposition 6.4.** *The equality constraints*

$$\sum_{h \in H} \sum_{i \in N} y_{ijh}^{\omega} = 1 \tag{58}$$

*is an expansion of equation* (28) *and must hold for all* $j \in N$ *and* $\omega \in \Omega$. *Similarly, expansion of equation* (29) *from the model formulation*

$$\sum_{h \in H} \sum_{j \in N} y_{ijh}^{\omega} = 1 \tag{59}$$

*must hold for all* $i \in N$ *and* $\omega \in \Omega$.

The following proof will show the validity of the first equation, while the second equation has an almost identical logic. The second part is included in Appendix A for completeness. These relationships are often too complex to be detected by general optimisation software within reasonable time, so we expect the explicit inclusion of these cuts to significantly strengthen the formulation.

*Proof.* For binary values of $y_{ijh}^{\omega}$, $x_{ij}$ and $\tau_{ih}^{\omega}$, Table 6.4 shows that equations (33)-(35) are

enough to linearise the relationship $y^{\omega}_{ijh} = x_{ij}\tau^{\omega}_{ih}$. The binding constraints are marked as active for each combination of values for $x_{ij}$ and $\tau^{\omega}_{ih}$.

Table 6.4: Linearisation of $y^{\omega}_{ijh}$

| $y^{\omega}_{ijh}$ | $x_{ij}$ | $\tau^{\omega}_{ih}$ | (33) | (34) | (35) |
|---|---|---|---|---|---|
| 0 | 0 | 0 | active | active | inactive |
| 0 | 0 | 1 | active | inactive | inactive |
| 0 | 1 | 0 | inactive | active | inactive |
| 1 | 1 | 1 | inactive | inactive | active |

From equation (29) in the model we know that

$$\sum_{j \in N} x_{ij} = 1 \quad i \in N$$

and from equation (38) we have

$$\sum_{h \in H} \tau^{\omega}_{ih} = 1 \quad i \in N, \omega \in \Omega$$

Combining this with the linearisation, we get

$$y^{\omega}_{ijh} = x_{ij}\tau^{\omega}_{ih} \quad i \in N, j \in N, h \in H, \omega \in \Omega \implies$$

$$\sum_{h \in H}\sum_{j \in N} y^{\omega}_{ijh} = \sum_{h \in H}\sum_{j \in N} x_{ij}\tau^{\omega}_{ih} \quad i \in N, \omega \in \Omega \implies$$

$$\sum_{h \in H}\sum_{j \in N} y^{\omega}_{ijh} = \sum_{h \in H} \tau^{\omega}_{ih} \sum_{j \in N} x_{ij} \implies y^{\omega}_{ijh} = 1 \cdot 1 = 1 \quad i \in N, \omega \in \Omega.$$

□

## 6.2 Emergency Model

The second mathematical model we present is a three-stage stochastic MIP. This model incorporates stochastic arrival of emergency patients, in addition to stochastic surgery durations.

### 6.2.1 Model design considerations

Similar arguments for a simplified information structure as those raised for the Phase Model, can be made for the Emergency Model. This means we can maintain the two-stage structure for the part of the information structure that is related to surgery durations. In addition, we model the number of emergency patients as an individual stage, giving a structure with three stages. In practice, elective patients are given an appointment time days or weeks ahead of the surgery, while the number of emergency patients that are

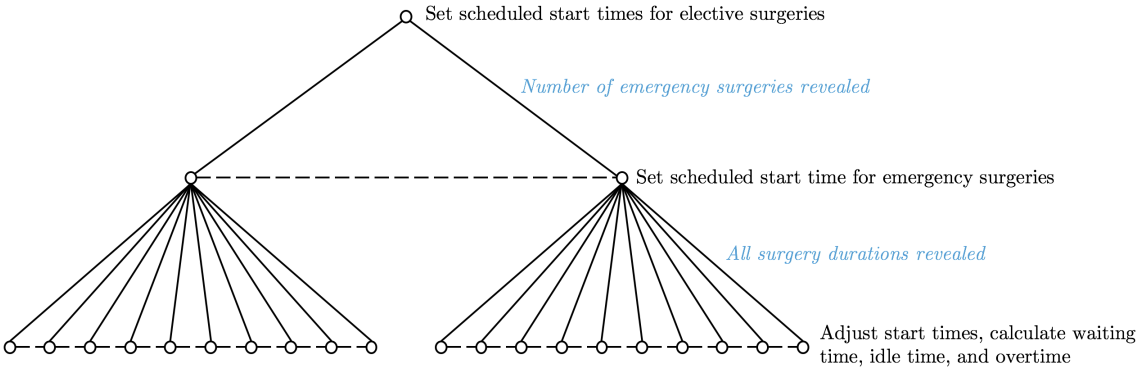



Figure 6.3: Stochastic information structure with three stages

transferred to the orthopaedic department is known by the start of a given day. Therefore, before the surgery durations are revealed, the planners are made aware of which, if any, emergency surgeries need to be performed that day. With this in mind, the planners set the scheduled start time for the newly revealed emergency patients. In the last stage, knowing all surgery durations, the actual start times are determined according to the same trivial rule of performing them as soon as possible, which determines all waiting time, idle time, and overtime.

The information structure described is shown in Figure 6.3. It has the same notation and colour scheme as in Section 6.1.1. The branching to the second stage is denoted by the emergency scenario index $\xi$, while the branching to the last stage is given by $\omega$. Thus, a given scenario is indexed by the combination of $\omega$ and $\xi$.

Due to Assumption 3, which states that all surgeries must be performed on the day they are planned, we require that the last surgery of the day cannot be an emergency surgery. In reality, some coordinators have a policy of deferring the last surgeries of the day if they are likely to lead to overtime, and if the last surgery is an emergency patient this would not be an option because of urgency considerations. Therefore, requiring the last surgery of the day to be an elective surgery, makes sure the schedules we create can be applied to the current practice.

### 6.2.2 Sets and indices

Let $N$ be the set of all elective surgeries and the possible positions of these, as before. Surgeries are still indexed by $j$, and positions by $i$. In this model, $i$ will also indicate the position of an elective surgery before which an emergency surgery can be placed. Let $Q_\xi$ be the set of all emergency surgeries $q$ and subpositions $u$ an emergency surgery can take before an elective surgery, in emergency scenario $\xi$. That is, for a given elective surgery position, the index $u \in Q_\xi$ impose an order on the emergency surgeries placed before that position. Together, the indices $i$ and $u$ define the possible positions for emergency surgeries. For example, one emergency surgery can be inserted as the first surgery ($u = 1$)

before the first elective surgery ($i = 1$), while another emergency surgery is scheduled as the second emergency surgery ($u = 2$) before the same elective surgery (still $i = 1$). The set $\Omega$ is the set of all scenarios of surgery durations, for both elective and emergency surgeries, and is indexed by $\omega$. Further, let the set $\Xi$ be the set of emergency scenarios, indexed by $\xi$.

The sets and indices described are summarised in Table 6.5.

Table 6.5: Sets and indices in the Emergency Model

| **Set** | **Description** |
|---|---|
| $N$ | Positions and surgeries, indexed by $i$ and $j$ |
| $\Omega$ | Scenarios of surgery durations, indexed by $\omega$ |
| $\Xi$ | Scenarios of the number of emergency surgeries to perform, indexed by $\xi$ |
| $Q_\xi$ | Emergency surgeries and subpositions in emergency scenario $\xi$, indexed by $q$ and $u$ |

### 6.2.3 Parameters

The parameters in the Emergency Model follow the same structure as in Section 6.1, but without the phase index. The additional notation is related to the emergency surgeries, indicated by superscript $E$.

Let $D_j^\omega$ be the duration of elective surgery $j$ in scenario $\omega$, while $D_q^{E\omega}$ is the duration of emergency surgery $q$ in scenario $\omega$. $d$ is still the point in time defining the end of the regular working day, so that any surgery time after this represents overtime. The probability of emergency scenario $\xi$ is given by $p_\xi$.

Let $c_j^w$, $c_j^s$ and $c_j^l$ be the unit cost of waiting time, idle time and overtime, respectively, for surgery $j$. Similarly, $c_q^{Ew}$, $c_q^{Es}$ and $c_q^{El}$ are the unit cost of waiting time, idle time and overtime for emergency surgery $q$. The last parameters are used in big-M formulations. All parameters used in the Emergency Model can be found in Table 6.6.

### 6.2.4 Variables

The Emergency Model has, in most problem instances, fewer binary variables than the Phase Model because the phases are not included. It has, however, more continuous variables, but is still a MIP model. The first-stage variables are exactly the same as in the Phase Model, and define the scheduled sequence and start times of all elective surgeries. In the second stage, the number of emergency patients has been revealed and need to be scheduled. To keep track of the scheduled emergency surgeries, let $z_{iqu}^\xi$ take the value 1 if emergency surgery $q$ is inserted as surgery number $u$ before the elective surgery in position $i$, in emergency scenario $\xi$. These variables are dependent on the emergency scenario to

Table 6.6: Parameters in the Emergency Model

| Parameter | Description |
|---|---|
| $D_j^{\omega}$ | Duration of surgery $j$ in scenario $\omega$ |
| $D_q^{E\omega}$ | Duration of emergency surgery $q$ in scenario $\omega$ |
| $d$ | Point of time beyond which overtime is incurred |
| $p_{\xi}$ | Probability of emergency scenario $\xi$ |
| $c_j^w$ | Waiting time penalty for surgery $j$ |
| $c_j^s$ | Idle time penalty for surgery $j$ |
| $c_j^l$ | Overtime penalty for surgery $j$ |
| $c_q^{Ew}$ | Waiting time penalty for emergency surgery $q$ |
| $c_q^{Es}$ | Idle time penalty for emergency surgery $q$ |
| $c_q^{El}$ | Overtime penalty for emergency surgery $q$ |
| $M^{\delta}, M^{E\delta}, M_i^s, M_i^w, M_i^l, M_i^{Es}, M_i^{Ew}, M_i^{El}$ | Sufficiently large numbers |

determine the number of patients. If emergency surgery $q$ neither is inserted as number $u$ before the elective surgery in position $i$ nor is defined for emergency scenario $\xi$, then $z_{iqu}^{\xi}$ take the value 0. $\phi_{iu}^{\xi}$ are defined as the scheduled start time for the emergency surgery inserted as number $u$ before the elective surgery in position $i$, in emergency scenario $\xi$. An important difference from the variables for elective patients, is that the number of possible positions for emergency surgeries exceeds the number of emergency surgeries. That is, given $v$ emergency surgeries and $n$ elective surgeries, then there are $v \cdot n$ possible positions for each emergency surgery. The consequence of this is that $\phi_{iu}^{\xi}$ will be defined for some positions with no emergency surgery assigned to them.

The remaining variables belong to the third stage and are, as before, the adjustments to the schedule based on the actual surgery durations. As in Section 6.1, the three variable groups $w_{ij}^{\omega\xi}$, $s_{ij}^{\omega\xi}$ and $l_{ij}^{\omega\xi}$ are the waiting time, idle time and overtime, respectively, for the elective surgery $j$ in position $i$ in scenario $\omega$ and $\xi$. Similarly defined, $w_{iqu}^{E\omega\xi}$, $s_{iqu}^{E\omega\xi}$ and $l_{iqu}^{E\omega\xi}$ are the waiting time, idle time and overtime for emergency surgery $q$ inserted as number $u$ before the elective surgery in position $i$ in scenario $\omega$ and $\xi$. The last continuous variables are the balance slack $g^{\omega\xi}$, which measure the earliness with respect to $d$ in scenario $\omega$ and $\xi$.

The only binary variables in the third stage are the indicator variables $\delta_i^{\omega\xi}$ and $\delta_{ru}^{E\omega\xi}$ for elective and emergency surgeries, respectively, indicating whether the entire corresponding surgery is in overtime.

All variables used in the Emergency Model are shown in Table 6.7.

Table 6.7: Variables in the Emergency Model

| Variable | Description |
|---|---|
| $x_{ij}$ | 1 if surgery $j$ is in position $i$ |
| $z^{\xi}_{iqu}$ | 1 if emergency surgery $q$ is inserted as number $u$ before elective surgery in position $i$ in emergency scenario $xi$ |
| $t_i$ | Scheduled start time of the surgery in position $i$ |
| $\phi^{\xi}_{iu}$ | Scheduled start time of the emergency surgery inserted as number $u$ before elective surgery in position $i$ in emergency scenario $\xi$ |
| $w^{\omega\xi}_{ij}$ | Waiting time when surgery $j$ is in position $i$ in scenario $\omega$ and $\xi$ |
| $s^{\omega\xi}_{ij}$ | Idle time after surgery $j$ when it is in position $i$ in scenario $\omega$ and $\xi$ |
| $l^{\omega\xi}_{ij}$ | Overtime when surgery $j$ is in position $i$ in scenario $\omega$ and $\xi$ |
| $w^{E\omega\xi}_{iqu}$ | Waiting time when emergency surgery $q$ is inserted as number $u$ before elective surgery in position $i$ in scenario $\omega$ and $\xi$ |
| $s^{E\omega\xi}_{iqu}$ | Idle time after emergency surgery $q$ when it is inserted as number $u$ before elective surgery in position $i$ in scenario $\omega$ and $\xi$ |
| $l^{E\omega\xi}_{iqu}$ | Overtime when emergency surgery $q$ is inserted as number $u$ before elective surgery in position $i$ in scenario $\omega$ and $\xi$ |
| $g^{\omega\xi}$ | Slack variable measuring the earliness with respect to $d$ in scenario $\omega$ and $\xi$ |
| $\delta^{\omega\xi}_{i}$ | 1 if the entire surgery in position $i$ is in overtime in scenario $\omega$ and $\xi$ |
| $\delta^{E\omega\xi}_{iu}$ | 1 if the entire emergency surgery inserted as number $u$ before elective position $i$ is in overtime in scenario $\omega$ and $\xi$ |

### 6.2.5 Emergency Model formulation

In this section, the mathematical formulation of the Emergency Model is presented. All the notation used is defined in the previous sections, with accompanying tables for quick reference. We first state the mathematical model and then explain the logical structure of the objective, constraints and discuss some mathematical caveats of the model.

$$\min \sum_{\xi\in\Xi}\sum_{\omega\in\Omega}\frac{p_\xi}{|\Omega|}\Bigg[\sum_{i\in N}\sum_{j\in N}(c^w_j w^{\omega\xi}_{ij} + c^s_j s^{\omega\xi}_{ij} + c^l_j l^{\omega\xi}_{ij}) + \sum_{i\in N}\sum_{q\in Q_\xi}\sum_{u\in Q_\xi}(c^{Ew}_q w^{E\omega\xi}_{iqu} + c^{Es}_q s^{E\omega\xi}_{iqu} + c^{El}_q l^{E\omega\xi}_{iqu})\Bigg] \quad (60)$$

$$\text{s.t} \quad t_i - t_{i+1} - \sum_{j\in N} w^{\omega\xi}_{i+1,j} + \sum_{j\in N} w^{\omega\xi}_{ij} + \sum_{j\in N} s^{\omega\xi}_{ij} + \sum_{q\in Q_\xi}\sum_{u\in Q_\xi} s^{E\omega\xi}_{i+1,q,u} = -\sum_{j\in N} D^{\omega}_j x_{ij} - \sum_{q\in Q_\xi}\sum_{u\in Q_\xi} D^{E\omega}_q z^{\xi}_{i+1,q,u} \quad i \in N\setminus\{|N|\}, \omega\in\Omega, \xi\in\Xi \quad (61)$$

$$t_n + \sum_{j\in N} w^{\omega\xi}_{nj} - \sum_{i\in N}\sum_{j\in N} l^{\omega\xi}_{ij} - \sum_{i\in N}\sum_{q\in Q_\xi}\sum_{u\in Q_\xi} l^{E\omega\xi}_{iqu} + g^{\omega\xi} = -\sum_{j\in N} D^{\omega}_j x_{nj} + d \quad \omega\in\Omega, \xi\in\Xi \quad (62)$$

$$\phi_{iu}^{\xi} + \sum_{q \in Q_\xi} w_{iqu}^{E\omega\xi} + \sum_{q \in Q_\xi} s_{iqu}^{E\omega\xi} + \sum_{q \in Q_\xi} D_q^{E\omega} z_{iqu}^{E\xi} \le t_i + \sum_{j \in N} w_{ij} \quad i \in N, \omega \in \Omega, \xi \in \Xi, u \in Q_\xi \quad (63)$$

$$\phi_{iu}^{\xi} + \sum_{q \in Q_\xi} w_{iqu}^{E\omega\xi} + \sum_{q \in Q_\xi} s_{iqu}^{E\omega\xi} + \sum_{q \in Q_\xi} D_q^{E\omega} z_{iqu}^{E\xi} \le \phi_{i,u+1}^{\xi} + \sum_{q \in Q_\xi} w_{iq,u+1}^{E\omega\xi} + M_i^{Ez}(1 - \sum_{q \in Q_\xi} z_{iq,u+1})$$
$$i \in N, \omega \in \Omega, \xi \in \Xi, u \in Q_\xi \quad (64)$$

$$t_i + \sum_{j \in N} w_{ij}^{\xi\omega} + \sum_{j \in N} s_{ij}^{\xi\omega} + \sum_{j \in N} D_j^{\omega} x_{ij} \le \phi_{i+1,u}^{\xi} + \sum_{q \in Q_\xi} w_{i+1,qu}^{E\omega\xi} + M_i^{Ez}(1 - \sum_{q \in Q_\xi} z_{iq,u+1})$$
$$i \in N \setminus |N|, \omega \in \Omega, \xi \in \Xi, u \in Q_\xi \quad (65)$$

$$\sum_{j \in N} l_{ij}^{\omega\xi} \ge \sum_{j \in N} D_j^{\omega} x_{ij} - M_i^{\delta}(1 - \delta_i^{\omega\xi}) \quad i \in N, \omega \in \Omega, \xi \in \Xi \quad (66)$$

$$\sum_{j \in N} l_{ij}^{\omega\xi} \ge t_i + \sum_{j \in N} w_{ij}^{\omega\xi} + \sum_{j \in N} D_j^{\omega} x_{ij} - d - M_i^{\delta} \delta_i^{\omega\xi} \quad i \in N, \omega \in \Omega, \xi \in \Xi \quad (67)$$

$$M_i^{\delta} \delta_i^{\omega\xi} \ge \sum_{j \in N} w_{ij}^{\omega\xi} + t_i - d \quad i \in N, \omega \in \Omega, \xi \in \Xi \quad (68)$$

$$M_i^{\delta}(1 - \delta_i^{\omega\xi}) \ge d - (t_i + \sum_{j \in N} w_{ij}^{\omega\xi}) \quad i \in N, \omega \in \Omega, \xi \in \Xi \quad (69)$$

$$\sum_{q \in Q_\xi} l_{iqu}^{E\omega\xi} \ge \sum_{q \in Q_\xi} D_q^{E\omega} z_{iqu}^{\xi} - M_i^{E\delta}(1 - \delta_{iu}^{E\omega\xi}) \quad i \in N, u \in Q_\xi, \omega \in \Omega, \xi \in \Xi \quad (70)$$

$$\sum_{q \in Q_\xi} l_{iqu}^{E\omega\xi} \ge \phi_{iu}^{\xi} + \sum_{q \in Q_\xi} w_{iqu}^{E\omega\xi} + \sum_{q \in Q_\xi} D_q^{E\omega} z_{iqu}^{\xi} - d - M_i^{E\delta} \delta_{iu}^{E\omega\xi} \quad i \in N, u \in Q_\xi, \omega \in \Omega, \xi \in \Xi \quad (71)$$

$$M_i^{E\delta} \delta_{iu}^{E\omega\xi} \ge \sum_{q \in Q_\xi} w_{iqu}^{E\omega\xi} + \phi_{iu}^{\xi} - d \quad i \in N, u \in Q_\xi, \omega \in \Omega, \xi \in \Xi \quad (72)$$

$$M_i^{E\delta}(1 - \delta_{iu}^{E\omega\xi}) \ge d - (\phi_{iu}^{\xi} + \sum_{q \in Q_\xi} w_{iqu}^{E\omega\xi}) \quad i \in N, u \in Q_\xi, \omega \in \Omega, \xi \in \Xi \quad (73)$$

$$\sum_{i \in N} x_{ij} = 1 \quad j \in N \quad (74)$$

$$\sum_{j \in N} x_{ij} = 1 \quad i \in N \quad (75)$$

$$\sum_{i \in N} \sum_{u \in Q_\xi} z_{iqu}^{\xi} = 1 \quad \xi \in \Xi, q \in Q_\xi \quad (76)$$

$$\sum_{q \in Q_\xi} z_{iqu}^{\xi} \le 1 \quad i \in N, u \in Q_\xi, \xi \in \Xi \quad (77)$$

$$s_{ij}^{\omega\xi} \le M^s x_{ij} \quad i \in N, j \in N, \omega \in \Omega, \xi \in \Xi \quad (78)$$

$$w_{ij}^{\omega\xi} \le M_i^w x_{ij} \quad i \in N, j \in N, \omega \in \Omega, \xi \in \Xi \quad (79)$$

$$l_{ij}^{\omega\xi} \le M_i^l x_{ij} \quad i \in N, j \in N, \omega \in \Omega, \xi \in \Xi \quad (80)$$

$$s_{iqu}^{E\omega\xi} \leq M^{Es} z_{iqu}^{\xi} \quad \xi \in \Xi, i \in N, q \in Q_\xi, u \in Q_\xi, \omega \in \Omega \tag{81}$$

$$w_{iqu}^{E\omega\xi} \leq M_i^{Ew} z_{iqu}^{\xi} \quad \xi \in \Xi, i \in N, q \in Q_\xi, u \in Q_\xi, \omega \in \Omega \tag{82}$$

$$l_{iqu}^{E\omega\xi} \leq M_i^{El} z_{iqu}^{\xi} \quad \xi \in \Xi, i \in N, q \in Q_\xi, u \in Q_\xi, \omega \in \Omega \tag{83}$$

$$x_{ij} \in \{0,1\} \quad i \in N, j \in N \tag{84}$$

$$t_i \geq 0, \quad t_1 \leq M_1^{Ez} \quad i \in N \tag{85}$$

$$\delta_i^{\omega\xi} \in \{0,1\} \quad i \in N, \omega \in \Omega, \xi \in \Xi \tag{86}$$

$$w_{ij}^{\omega\xi} \geq 0, \quad s_{ij}^{\omega\xi} \geq 0, \quad l_{ij}^{\omega\xi} \geq 0 \quad i \in N, j \in N, \omega \in \Omega, \xi \in \Xi \tag{87}$$

$$g^{\omega\xi} \geq 0, \quad \omega \in \Omega, \xi \in \Xi \tag{88}$$

$$z_{iqu}^{\xi} \in \{0,1\} \quad i \in N, \xi \in \Xi, q \in Q_\xi, u \in Q_\xi \tag{89}$$

$$\delta_{iu}^{E\omega\xi} \in \{0,1\} \quad i \in N, u \in Q_\xi, \omega \in \Omega, \xi \in \Xi \tag{90}$$

$$w_{iqu}^{E\omega\xi} \geq 0, \quad s_{iqu}^{E\omega\xi} \geq 0, \quad l_{iqu}^{E\omega\xi} \geq 0 \quad i \in N, \xi \in \Xi, q \in Q_\xi, u \in Q_\xi, \omega \in \Omega \tag{91}$$

$$\phi_{iu}^{\xi} \geq 0, \quad \phi_{iu}^{\xi} \leq M_i^{Ez} \quad i \in N, \xi \in \Xi, u \in Q_\xi \tag{92}$$

Except for extra notation, and some more complicating constraints, the basic structure of the Emergency Model is similar to the Phase Model.

In this model the objective function is given by equation (60). The objective is, as before, to find the optimal balance of waiting time, idle time and overtime. However, we also have to include these measures for the emergency surgeries. The objective is summed over all scenarios $\omega$ and $\xi$, and the probability function is assumed to be constant in $\omega$. However, in contrast to the Phase Model, the objective function would be linear even if the probability was dependent on the full scenario $\omega$ and $\xi$. There is no reason to give the surgery duration scenarios, $\omega$, different probabilities from the way we generate the scenarios, but it makes sense to allow different emergency scenarios to have different probability as the differences between the emergency scenario are much more significant. For example, zero emergency surgeries are more likely than one or two. Note that the durations of all surgeries are unaffected by the number of emergency patients.

Equations (61) - (65) constitute the balance constraints for both elective and emergency surgeries. Firstly, equation (61) defines the waiting time and idle time for every subsequent elective surgery. However, there may be scheduled emergency surgeries between elective surgeries, which must be included in the sum. This gives two extra terms compared to equation (22) in the elective model. It is worth noticing that only idle time and duration of potential emergency surgeries are included. This might not seem correct at first glance, but the waiting time of surgeries in the middle of a sequence of surgeries are already accounted for by the duration of the preceding surgery. By including the duration and idle time, the equation will sum from the actual start ($t_i + \sum_{j \in N} w_{ij}^{\omega\xi}$) of an elective surgery

to the actual start of the next elective surgery. Constraints (62) balance the actual start of the last surgery with overtime, duration of the last surgery, slack and the end of regular working hours. Because the emergency position variable, $z^{\xi}_{iqu}$ is 1 if emergency surgery $q$ is inserted *before* elective surgery $i$, we know that the last surgery on a given day must be elective. Therefore, equation (62) is correct in this model, but must include potential overtime in emergency surgeries that occurs before the last elective surgery. None of the two preceding constraints set the scheduled start time of the emergency surgeries because they start from an elective surgery and sum over the emergency surgeries.

Constraints (63) force the finish time of an emergency surgery to be equal or less to the actual start time of the next elective patient. Equation (64) forces the finish time of the emergency patient in one subposition to be less then actual start of the potential emergency patient in the proceeding position. Constraints (65) ensure that the actual start time of the first emergency is after the previous elective has finished. The two last equations require big-M formulations. This results from the fact that some possible positions for emergency patients are unoccupied. The big-M formulations make sure that the constraints only are binding if there is an emergency patient in the current position. Together, these three constraints balance the start time, waiting time, idle time and overtime of all the emergency patients.

Constraints (66) and (67) set the overtime of elective surgeries, equivalently to the Phase Model formulation, except that there is an additional scenario index. Equations (68) and (69) force the indicator variables $\delta_i^{\omega\xi}$ to 1 or 0 depending on whether the entire surgery in position $i$ is in overtime or not. Equations (70) - (73) are equivalent to Equations (66)-(69), but for the emergency surgeries. Equations (74) and (75) are also found in the Phase Model, and make sure each elective surgery is assigned to a position and that each position is assigned a surgery. Constraints (76) force every emergency surgery of an emergency scenario to be assigned to a position, while constraints (77) ensure each emergency position gets at most one emergency surgery. Because the number of possible positions for emergency surgeries exceeds the number of actual emergency surgeries, these cannot be equality constraints. The succeeding six constraints (78) - (83) force the idle time, waiting time and overtime for both elective and emergency surgeries, respectively, to 0 if there are no surgery scheduled for the corresponding position. If there is a surgery scheduled to that position, the big-M notation will make the constraints redundant. The remaining constraints (84) - (92) are the definitions of the variable domains.

### 6.2.6 Strengthening the big-M formulations

The big-M formulations in the constraints in the Emergency Model are similar to those of the Phase Model, but are stated here for completeness. To ease the equations, the following are defined

$$d_j = \max_{\omega \in \Omega} \left\{ D_j^{\omega} \right\} - \min_{\omega \in \Omega} \left\{ D_j^{\omega} \right\} \tag{93}$$

$$d_q^E = \max_{\omega \in \Omega} \left\{ D_q^{E\omega} \right\} - \min_{\omega \in \Omega} \left\{ D_q^{E\omega} \right\} \tag{94}$$

In addition, we define $Q^{\max}$ as the set of emergency patients in the scenario $\xi$ that has the largest number of emergency arrivals.

$M^s$ is given by

$$M^s = \max_{j \in N} \{ d_j \} + \sum_{q \in Q^{\max}} \max_{\omega \in \Omega} \left\{ D_q^{\omega} \right\} \tag{95}$$

This follows from the corresponding explanation in the Phase Model, and takes into account that the number and positions of emergency patients are uncertain. Thus, it may be optimal to allocate more idle time in order to be able to handle a maximum number and duration of emergency surgeries. This information is revealed at the same time for all emergency surgeries. The maximum possible idle time for emergency patients is, equivalently as for the Phase Model, given by

$$M^{Es\xi} = \max \left\{ \max_{j \in N} \{ d_j \}, \max_{q \in Q_\xi} \left\{ d_q^E \right\} \right\} \tag{96}$$

Conversely, the worst schedule for the elective patients in this scenario is when no idle time is allocated, which would cause waiting time equal to

$$M_i^w = \sum_{j=1}^{i-1} a_j + \sum_{q \in Q^{\max}} \max_{\omega \in \Omega} \left\{ D_q^{E\omega} \right\} \tag{97}$$

where $a_j$ is the $j$th largest value of $d_j$. The same holds for $M_i^{Ew}$, but in addition, information of the emergency arrival is revealed.

$$M_i^{Ew\xi} = \sum_{j=1}^{i-1} a_j + \sum_{q=1}^{|Q_\xi|-1} b_q \tag{98}$$

where $b_q$ is the $q$th largest value in $d_q^E$.

$M_j^l$ will remain the same the big-M for overtime in the Phase Model, and $M_q^{El}$ will, equivalently, be given by

$$M_q^{El} = \max_{\omega \in \Omega} \left\{ D_q^{E\omega} \right\} \tag{99}$$

$M_i^\delta$ is set to

$$M_i^\delta = \max\left\{\sum_{j=1}^{i-1} b_j^{\max} + \sum_{q\in Q^{\max}} \max_{\omega\in\Omega}\left\{D_q^{E\omega}\right\} - d, d - \sum_{j=1}^{i-1} b_j^{\min}\right\} \tag{100}$$

where $b_j^{\max}$ is the $j$th longest possible surgery durations for all surgeries, i.e. $j$th largest of $\max_{\omega\in\Omega} D_j^\omega$ and $b_j^{\min}$ is the $j$th smallest of $\min_{\omega\in\Omega} D_j^\omega$. $M_i^{E\delta}$ is, equivalently, given by

$$M_i^{E\delta\xi} = \max\left\{\sum_{j=1}^{i-1} b_j + \sum_{q=1}^{|Q_\xi|-1} \max_{\omega\in\Omega}\left\{D_q^{E\omega}\right\} - d, d - \sum_{j=1}^{i-1} b_j^{\min}\right\} \tag{101}$$

that is, the maximum difference between the latest or earliest possible start time and the parameter $d$.

$M_i^{Ez}$ is the latest start of any surgery, given by

$$M_i^{Ez} = \sum_{j=1}^{i-1} \max_{\omega\in\Omega}\left\{D_j^\omega\right\} + \sum_{q\in Q^{\max}} \max_{\omega\in\Omega}\left\{D_q^{E\omega}\right\} \tag{102}$$

### 6.2.7 Valid inequalities

Propositions 6.2 and 6.3 from the Phase Model hold in this formulation and may be further extended to include emergency patients. The structure of the propositions remain the same, and the full formulations are listed in Appendix A.

**Proposition 6.5.** *The inequalities*

$$\sum_{q\in Q_\xi} z_{iqu}^\xi \geq \sum_{q\in Q_\xi} z_{iq,u+1}^\xi \tag{103}$$

*are valid for all $i \in N$, $\xi \in \Xi$ and $u \in Q_\xi \setminus \left\{|Q_\xi|\right\}$ and must be satisfied in the optimal solution.*

The inequalities in Proposition 6.5 state that there must be an emergency surgery inserted as number $u$ before elective surgery $i$, before there can be a surgery inserted as number $u+1$ before the same elective surgery $i$. This is a logical cut, because it does not make sense to define an emergency surgery as inserted as number two ($u = 2$) before a given elective surgery $i$, if there is none inserted as number one ($u = 1$). Thus, this is added to remove symmetric solutions.

## 6.3 Challenges

The Phase Model and Emergency Model pose new challenges to the scheduling problem solved by Denton et al. [19] and Mancilla and Storer [80]. More variables and constraints complicate the structure and result in computationally harder problems. The main challenge is that there are continuous and integer variables in all stages. In the Phase Model, this is a result of our way of handling decision-dependent probability. In order to be able to use a general optimisation solver, the linearisation with binary variables ($\tau_{ih}^{\omega}$) is necessary. The Emergency Model introduces similar difficulties through the $z_{iqu}^{E\xi}$ variables in the second stage, resulting from how we define the timing of the revelation of information about emergency arrivals. Further, both models have binary overtime indicator variables that determine if the surgery starts after regular working hours. This could have been omitted and simplified in the same way as was done by Mancilla and Storer [80]. But, to make it possible to assign different costs to different surgeries, we think our formulation is a more appropriate reflection of the practice at St. Olavs Hospital.

There are few general properties for stochastic MIPs. Birge and Louveaux [84] state that the expected recourse function of an integer program is in general lower semi-continous, non-convex and discontinous. This implies that the usual form of duality is lost and the wide variety of decomposition methods that have been developed for stochastic linear problem in literature, break down when integer variables are introduced [85]. Therefore, we cannot use the same solution methods as Denton et al. [19] or Mancilla and Storer [80]. The development of stochastic MIP solution methods has attracted little attention in research and there are consequently few general efficient solution procedures [84].

Birge and Louveaux [84] state that solution methods for stochastic MIPs usually start with the use of the L-shaped method where the integrality constraints in the second-stage variables are relaxed. This is another challenge as the LP-relaxation in our formulations prove to be weak for the complete model. Weak LP-relaxations lead to excessive branching and long computation times [86]. Even if the solver quickly finds a good solution and thus a good upper bound, the lower bound will be far off. Consequently, it takes time to prove that a good solution has been found.

Further, a third challenge is the balance constraints that include many variables. These constraints couple many of the variables and may provide the solver with logical relationships between them. Decomposing our problem will deteriorate the structure and connections in our model. Birge and Louveaux [84] suggest methods which decompose the problem by stages or into one part with continuous variables and a second part with only integer variables. These were tested, but did not result in improvements in execution time.

Decomposition methods used on similar, but simpler, problems have been unsuccessful in achieving a significant speedup. The maximum number of surgeries to schedule at

St. Olavs Hospital is quite low and a formulation with some improving cuts is therefore enough to find the optimal solution within a reasonable time. The scalability is, at this point, quite low. To increase the solution time further on the cost of losing an optimality guarantee, we will evaluate and suggest several heuristics. These also mitigate some of the scalability problems. For bigger problem instances, more complex solution methods can be evaluated. Laporte and Louveax [31] introduced the integer L-shaped method for problems with binary variables in the first stage. These could be evaluated in future study.

## 6.4 Heuristics

This section describes the heuristics that will be evaluated in the computational study in Chapter 7. First we give the heuristic types a theoretical introduction, then we describe our use and implementation. The heuristics we look at are the Bailey-Welch rule, sorting surgeries in order of increasing mean and variance, local search, and simulated annealing.

In optimisation theory, heuristics are often categorised into groups, such as construction vs. local search heuristics, diversifying vs. intensifying, randomised vs. systematic, etc. [87]. We choose divide the heuristics we consider into two categories, based on the amount of optimisation performed by the heuristic; decision rules and heuristic searches. This does, to some extent, reflect the complexity of the heuristic and what decisions are optimised. The decision rules that will be discussed typically uniquely define a sequence of surgeries, but allow the start times to be optimised, or they fix the start times, allowing the sequence to be optimised. Some set both the start time and sequence based on surgery characteristics. The more complex heuristics we consider include local searches that explore the sequence of surgeries and then optimise the start times.

### 6.4.1 Bailey-Welch rule

The Bailey-Welch rule was briefly discussed in Chapter 3 and is often referred to in literature. Despite its simplicity, much literature reports good results using the rule. On the other hand Bosch and Dietz [27] claim sequencing rules based on patient characteristics do not give good performance. Therefore, we want to see how this decision rule performs on our problem.

The rule schedules $k$ surgeries to start at time 0, while the rest of the surgeries start at fixed intervals equal to the average surgery duration. The value of $k$ reflects the best trade-off between patient waiting time and idle time, and is thus dependent on the weights $c_j^w$, $c_j^s$ and $c_j^l$. Welch and Bailey [22] found that $k = 2$ gave the best trade-off. We will try with both $k = 1$ and $k = 2$. Because start times are defined for each position, this rule

is easily implemented by imposing restrictions on $t_i$. That is, for $i \in \{1, ..., k\}$, set $t_i = 0$. For all $i > k$, set $t_i - t_{i-1} = \mu$, where $\mu$ is the average surgery duration. The optimisation implementation is then run to find the optimal sequence of surgeries.

Because the start times are fixed, the phase variables are trivial to find as well, except for the cases where $t_i$ are close enough to $P^{\text{split}}$, such that the actual start time may be adjusted past $P^{\text{split}}$. The only remaining decisions are the sequence and the calculations of waiting time, idle time and overtime. We would expect this to simplify the optimisation, which should be reflected in execution time. For problems with more surgeries, the number of possible sequences increases fast, which for larger problems can be time consuming to optimise. However, because the scheduled start times are fixed, there are a lot fewer combination so evaluate.

### 6.4.2 Statistically based sequencing rules

Many statistical characteristics of patients are used in literature to guide decision rules. As pointed out in the literature review, Weiss [24] proved in 1990 that the optimal sequence of two surgeries is in order of increasing variance. According to Wang [16], the optimal sequence of surgeries are in order of increasing mean surgery durations, under a set of assumption. Denton et al. [19] also evaluated to sequence both in order of increasing mean and increasing variance. Both of these sequencing rules will be tested for our model.

When the sequence is fixed, the optimisation software is run to find the optimal start times, phases, waiting times, idle times and overtimes. However, there is very little flexibility for the optimiser when the sequence is fixed, because many of the binary variables are closely linked to the $x_{ij}$ variables from the sequence. This makes it solve quickly.

### 6.4.3 Local search

In a local search heuristic, the solutions are iteratively improved. From an initial solution, we make local modifications to find improved solutions. If no improving solution is found in a neighbourhood, we stop in a local optimum [87]. The neighbourhood can either be traversed systematically or randomly. There are also several ways to stop the search through a neighbourhood. One could for example implement a first improvement or best improvement criterion. The first approach searches the neighbourhood until it finds an improving solution, in contrast to the latter approach that performs an exhaustive search and selects the best solution if it is better than the current solution. Lundgren et al. [87] provide a general description of the local search heuristic, which is restated in Algorithm 6.1. A minimisation problem is assumed in this explanation. Note that the $x$ in this description represent a solution to the optimisation problem at hand and do not correspond with our $x_{ij}$ variables.

**Algorithm 6.1** General local search

1: Start from a feasible solution $x^{(0)}$ with cost $c(x^{(0)})$. Set $k = 0$.
2: Determine all points in the neighbourhood $N(x^{(k)})$.
3: If $c(x^{(k)}) \leq c(x)$ for all $x \in N(x^{(k)}) \implies$ stop.
4: Choose $x^{(k+1)} \in N(x^{(k)})$ such that $c(x^{(k+1)}) < c(x^{(k)})$.
5: Set $k := k + 1$ and go to Step 2.

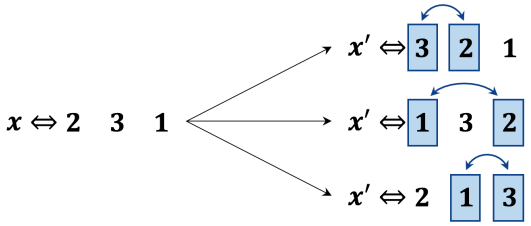



Figure 6.4: Two-swap neighbourhood of a sequence x, for three surgeries

Some definitions of neighbourhoods result in a large number of neighbours, which takes too much time to search through. It may also be difficult to compute the objective value of a neighbour if it, for example, requires solving a MIP.

We choose to test the performance of two different neighbourhoods. The neighbourhoods are also tested in an implementation of simulated annealing (SA), explained in the next section. The solutions are selected based on best improvement, because the neighbourhood sizes are manageable. However, neighbourhood sizes scale with the number of surgeries, and a first improvement approach might be more suitable for large problem instances.

The first neighbourhood we consider, is briefly mentioned by Bosch [28] and Denton and Gupta [18], and is a two-swap neighbourhood. Bosch reports results close to optimal for his problem instances. A mathematical description of this neighbourhood is given by equation (104). The size of the two-swap neighbourhood is $\binom{n}{2}$, which equals $\frac{n(n-1)}{2} = \mathcal{O}(n^2)$. An example of the neighbourhood is given in Figure 6.4 to visualise the neighbourhood. This shows the neighbours of a given sequence $x$, where the swaps are marked in blue boxes.

$$N(x) = \{x' | x'_{ib} = x'_{aj} = 1 \wedge x_{ij} = x_{ab} = 1 \wedge i \neq a \wedge j \neq b\} \tag{104}$$

A single swap neighbourhood would not work for our solutions, because if an $x_{ij} = 1$ is flipped to 0, the same $x_{ij}$ would have to be turned back to 1 to make the solution feasible (due to constraits $\sum_j x_{ij} = \sum_i x_{ij} = 1$).

The second neighbourhood we consider is a surgery-pair swap. The surgery-pair swap neighbourhood function chooses two pairs of subsequent surgeries in the sequence and swaps these pairs. That is, the first surgery of the first pair, takes the position of the first

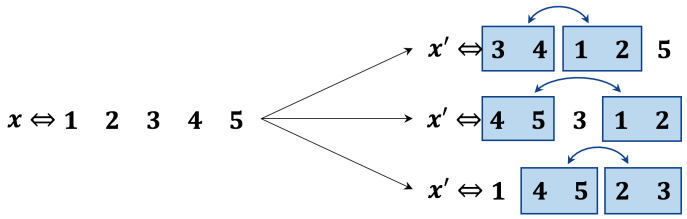


Figure 6.5: Surgery-pair swap neighbourhood of a sequence x, for five surgeries

surgery of the second pair and vice versa. Equivalently, the second surgery of the first pair switches position with the second surgery of the second pair. This neighbourhood is only useful for data instances of five or more surgeries, because there is only one way to select two surgery-pairs from four surgeries. An example of the neighbourhood is illustrated in Figure 6.5 and shows the neighbourhood of a given solution $x$. Again, the blue boxes indicate how the surgery pairs have been swapped. A mathematical description of the surgery-pair swap neighbourhood function is found in equation (105).

$$N(x) = \{x' | x'_{ib} = x'_{i+1,c} = x'_{aj} = x'_{a+1,h} = 1 \wedge x_{ij} = x_{i+1,h} = x_{ab} = x_{a+1,c} = 1 \\ \wedge\, i \neq a \neq i+1 \wedge j \neq b \neq j+1\} \quad (105)$$

The number of unique neighbours in a given neighbourhood can be found as follows. For a sequence of $n$ surgeries, there are $n-1$ ways of selecting two subsequent surgeries. If the two first surgeries are selected as the first pair, then there are $n-2$ surgeries left, which can be combined into $n-3$ pairs. Similarly, if the second and third surgery are chosen as the first surgery pair, there are only $n-3$ surgeries left to consider in the sequence and $n-4$ possible pairs. This gives a sum of $(n-3)+(n-4)+\cdots+1$. Because we are looking for unique neighbours, the pairs $(a,b)$ and $(c,d)$ are considered the same as $(c,d)$ and $(a,b)$. Thus, to calculate the number of neighbours, one can sum over the number of possible pairs *after* the first pair of surgeries in the sequence. Formalised, this gives the number of neighbours equal to $\sum_{i=3}^{n}(n-i)$, which can be written as $\frac{(n-2)(n-3)}{2} = \mathcal{O}(n^2)$. This has the same complexity as the first neighbourhood, but is clearly a lot smaller for small $n$. A comparison of the sizes for different $n$ is shown in Table 6.8. However, there is a trade-off between the neighbourhood size and the number of passes the algorithm makes through the neighbourhood calculation.

### 6.4.4 Simulated annealing

Lundgren et al. [87] describe metaheuristics as a mean to control and manage the search more systematically and efficiently, for example by directing the local search method to new parts of the feasible region and scan a larger area. Simulated annealing (SA) is a randomised metaheuristic, which can be categorised as both local and global search. It

Table 6.8: Comparison of the size of two-swap and surgery-pair swap neighbourhoods

| N | Two-swap | Surgery-pair swap |
|---|---|---|
| 3 | 3 | 0 |
| 4 | 6 | 1 |
| 5 | 10 | 3 |
| 6 | 15 | 6 |
| 8 | 28 | 15 |
| 10 | 45 | 28 |
| 20 | 190 | 153 |

was first proposed by Kirkpatrick et al. [88]. Different from other heuristics is that the method only examines one neighbour, and not all of them, before making a decision, and the neighbour selection is also random. The method accepts worse solutions with a probability proportional to the difference in objective value to the current solution. To make it converge, we gradually lower the probability of accepting worse solutions. SA is often used for problems where an acceptable local optimum is more important than the global optimum.

The heuristic can be described by Algorithm 6.2, as found in Lundgren et al. [87]. When selecting the parameters, it is important that the initial temperature, $T_0$, has a large value in the beginning, and that the reduction factor, $r$, does not reduce the search too quickly. For small values of the temperature, SA approximates a greedy algorithm.

**Algorithm 6.2** Simulated annealing

1: Start from a feasible solution $x^{(0)}$. Set $k = 0$.
2: Choose an initial temperature $T$, a reduction factor $r(0 < r < 1)$ and a maximum number of iterations $L$.
3: Randomly pick a neighbour $\hat{x} \in N(x^{(k)})$.
4: Let $\Delta = c(\hat{x}) - c(x^{(k)})$.
5: If $\Delta \leq 0$, set $x^{(k)} = \hat{x}$.
6: if $\Delta > 0$, set $x^{(k)} = \hat{x}$ with probability $e^{\frac{-\Delta}{T}}$.
7: Save the best solution found. Stop if $k = L$, otherwise set $T := rT$, $k := k + 1$ and go to Step 3.

# Chapter 7

# Computational study

This chapter describes the tests performed, to evaluate the model formulations and the resulting schedules. After a brief introduction to the hardware and software used, we will describe the problem test instances. The analyses that follow will in turn move from technical evaluations of the performance of the implementation, to a more practical perspective, where the potential of the stochastic considerations is evaluated.

We will begin the analyses with a discussion of the cost structure we have chosen in the objective functions, and how other combinations affect the models. Then, we will test for out-of-sample and in-sample stability. This gives an indication of how representative a given scenario tree size is for the true stochastic processes, as well as the robustness of the objective value with respect to the scenario generation. The subsequent experiments will evaluate the strength of the cuts added to the models and consider tuning the configuration of the optimiser. The next section will evaluate the proposed heuristics, before we finally assess the value of using stochastic programming in our case and the practical benefits of our model. Note that all values of waiting time, idle time and overtime are given in minutes, while run-time is given in seconds.

## 7.1 Hardware and software

All numerical experiments are performed on an Intel® Core™ i7-3770 computer with a quad-core CPU at 3,40Ghz and 16,0 GB RAM. The software used is FICO® Xpress Optimization Suite, with Xpress-IVE version 1.24.06, Xpress Mosel version 3.8.0 and Xpress Optimizer version 27.01.02.

Most experiments are run through the console from scripts written in Python 2.7.11 © Python Software Foundation.

Table 7.1: Overview of data instances for the Phase Model

| Instance | Number of surgeries | Relative surgery cost | Date | Operating room |
|---|---|---|---|---|
| Instance 1 | 4 | Variable | 17.06.2013 | 4 |
| Instance 2 | 4 | Variable | 24.01.2014 | BRN1 |
| Instance 3 | 4 | Constant | 30.01.2007 | DK6 |
| Instance 4 | 4 | Constant | 18.10.2010 | DK6 |
| Instance 5 | 4 | Variable | 20.12.2010 | DK6 |
| Instance 6 | 4 | Variable | 16.04.2012 | DK6 |
| Instance 7 | 5 | Constant | 13.09.2013 | DK6 |
| Instance 8 | 4 | Variable | 23.01.2014 | DK6 |
| Instance 9 | 5 | Variable | 27.10.2014 | DK6 |
| Instance 10 | 5 | Constant | 12.09.2013 | Stue 1 |
| Instance 11 | 5 | Variable | 16.09.2013 | Stue 1 |
| Instance 12 | 6 | Constant | 16.09.2014 | Stue 1 |
| Instance 13 | 4 | Variable | 13.10.2015 | Stue 1 |
| Instance 14 | 7 | Variable | 15.10.2015 | Stue 1 |
| Instance 15 | 6 | Constant | 17.03.2014 | Stue 2 |
| Instance 16 | 4 | Constant | 21.03.2014 | Stue 2 |
| Instance 17 | 5 | Constant | 06.10.2014 | Stue 2 |
| Instance 18 | 4 | Variable | 12.12.2014 | Stue 2 |
| Instance 19 | 5 | Variable | 28.04.2015 | Stue 2 |
| Instance 20 | 5 | Variable | 21.08.2015 | Stue 2 |

Table 7.2: Overview of data instances for the Emergency Model

| Instance | Number of elective surgeries | Number of emergency surgeries |
|---|---|---|
| Instance 1E | 5 | 2 |
| Instance 2E | 4 | 2 |
| Instance 3E | 4 | 2 |

## 7.2 Test instances

The process of deciding for which instances to run the models was described in Section 5.3. Table 7.1 shows the resulting 20 data instances for the Phase Model. It indicates how many surgeries there are in each of the instances, whether the relative surgery costs are constant or variable, along with the actual date and operating room of the instance. The Emergency Model uses three fictive instances, where up to two additional emergency surgeries may arrive, as was also described in Section 5.3. These are shown in Table 7.2. The probability for zero, one or two emergency patients are 68%, 23% and 9%, respectively.

## 7.3 Performance measures

This section presents the interpretation and values of the costs of waiting time, idle time and overtime used in the computational study. In addition, an analysis of different combinations of costs is conducted in the end of the section to get a deeper understanding of how changing the values impact the performance of the models and the resulting schedules.

### 7.3.1 Choice of cost combinations

Estimating monetary values for waiting time, overtime and idle time is an ongoing project at St. Olavs Hospital. This is a complex and comprehensive analysis and the calculations of these values are thus considered to be out of scope of this paper. Instead, relative costs are based on relevant literature, data analysis, and qualitative analysis in cooperation with hospital staff. These costs may not be an exact reflection of the actual situation at the hospital, but will give insights into on the trade-offs between waiting time, idle time and overtime.

The costs in the two model objectives in equations (21) and (60) are defined as follows: The waiting cost is interpreted as the disutility the patient experiences. The idle cost is understood as the opportunity cost of the surgical staff performing surgery. The overtime cost is the direct cost incurring when surgical staff have to work after regular opening hours. Waiting costs do not directly affect the hospital. Even though decisions are made from the hospital's point of view, patient waiting time is assigned a cost in order to reflect the considerations they need to take related to patient care.

When analysing the Phase Model, we have set the ratio between overtime and idle time ($c_j^l/c_j^s$) to 1.5, implying that the cost of the surgical staff is higher after regular working hours. This seems to be the most used ratio in literature, as described in Section 3.4 and is supported by the required minimum overtime payment in Norway [89]. The ratio between waiting time and idle time ($c_j^w/c_j^s$) determines how costly patients' waiting time is compared to surgical staff's idle time. In consultation with St. Olavs Hospital we have estimated patients' waiting time to be less important for the hospital than the surgical staff's idle time. The waiting cost for patients has been set to 0.5 while the idle cost is normalised to 1.

As noted in the literature review in Chapter 3, many papers assume the same cost for all surgeries. In agreement with the hospital, we consider it beneficial to weight surgeries differently depending on their resource consumption. The relative costs between surgeries may be calculated based on DRG values which are used by St. Olavs Hospital to classify surgeries into groups with similar complexity and resource consumption [90]. However, we do not have access to the values nor the attributes from which the DRG values are

calculated. Instead, the number of staff members participating in the surgery, which can be found in our data, is used as a measure of the resource usage. For most surgeries in the instances used, two or three surgical staff members are participating. This will be multiplied with the relative costs between waiting, idling and overtime. In other words, a surgery with a resource consumption of two have costs equal to $c_j^w = 1$, $c_j^s = 2$, and $c_j^l = 3$. For the tests with constant costs, the relative costs between surgeries are normalised to 1 resulting in $c_j^w = 0.5$, $c_j^s = 1$, and $c_j^l = 1.5$.

The costs in the Emergency Model are similar to those in the Phase Model. The only difference is the costs of overtime for the emergency patients, which are set to three times that of the idle time. This is meant to reflect the inclination to avoid scheduling emergency patients late in the day, knowing that the planners might want to defer the last surgeries to avoid overtime. Deferring emergency surgeries is not an option due to their urgency, so the high overtime cost of overtime reflects this consideration.

### 7.3.2 Cost analysis

The evaluation of waiting time, idling time and overtime varies greatly both in literature and between different hospitals. In addition, as the costs are set based on qualitative judgements, it may be preferable to change these with new or better information. As was pointed out in the preceding section, we have chosen the cost combination referred to as Cost combination 1 in Table 7.3. To get a deeper understanding of how different cost combinations affect our models and the resulting schedules, all cost combinations listed in the table are tested in this section, both with variable and constant costs between surgeries. These combinations are chosen in order to provide a representative selection of what is used in literature.

Tests on the Emergency Model show similar results as for the Phase Model, and are therefore not repeated.

Table 7.3: Cost combinations of waiting, idling and overtime in the Phase Model

| **Cost combination** | **Cost of waiting** | **Cost of idling** | **Cost of overtime** |
|---|---|---|---|
| 1 | 0.5 | 1 | 1.5 |
| 2 | 1 | 1 | 1 |
| 3 | 1 | 0 | 1.5 |
| 4 | 0.1 | 1 | 1.5 |
| 5 | 0.01 | 1 | 1.5 |

The results are listed in Table 7.4. This show the average waiting time, idle time and overtime denoted in minutes, across all instances. The average execution time is included to accentuate how the cost structures affect the computational complexity. If an optimal solution is not found within 1 800 seconds, the execution is stopped.

Table 7.4: Average waiting time, idle time, and overtime resulting from solving the Phase Model with five different cost structures, with the average execution time

| | Variable cost | | | | Constant cost | | | |
|---|---|---|---|---|---|---|---|---|
| **Cost** | **Wait** | **Idle** | **Overtime** | **Exe** | **Wait** | **Idle** | **Overtime** | **Exe** |
| 1 | 23.3 | 5.6 | 6.5 | 955.0 | 22.6 | 5.9 | 6.5 | 909.6 |
| 2 | 15.5 | 10.5 | 7.2 | 842.5 | 15.0 | 10.5 | 7.1 | 799.3 |
| 3 | 4.3 | 198.9 | 7.8 | 276.8 | 4.6 | 200.4 | 7.5 | 330.3 |
| 4 | 47.0 | 1.1 | 5.7 | 747.9 | 49.4 | 1.1 | 5.4 | 536.9 |
| 5 | 71.5 | 0.1 | 5.6 | 451.7 | 85.3 | 0.1 | 5.3 | 453.7 |

The results demonstrate that, when the ratio $c_j^s/c_j^w$ increases, the amount of waiting time for patients increases. For example, for cost combination 5, 100 minutes of waiting is equally expensive as 1 minute of idling. This creates a schedule where it is preferable to schedule the start times closer together to reduce the probability of a surgery finishing before the next is ready to start, i.e. reduce the probability of idle time. Oppositely, for cost combination 3, where idling costs are 0, a schedule where the start times are more widely distributed is preferable, reducing the probability of delaying the next surgery and thus the probability of waiting time.

The amount of overtime is approximately constant on average. This is because overtime will only be scheduled when it is difficult to avoid, i.e. when the total duration of the surgeries is long. This is an effect of overtime being the most expensive cost in almost all combinations. In cases with a high probability of overtime, scheduled start times are often set closer together to reduce the expected makespan of the day and the probability of overtime. This creates a schedule with less idle time and more waiting time. Comparing the performance of instances where overtime occurs with the performance across all instances, show that the average waiting time increases by 20% and idle time decreases by 20% on average for cost combination 1, which supports this reasoning.

The results of these tests validate the intuition that when more trade-offs are taken into consideration, the problem is more complex. That is, cost structures 3 and 5 effectively only have to balance two measures, which gives shorter execution times than the other combinations. Further, this also holds for the comparison between variable and constant costs between surgeries. The execution time is on average lower for the latter.

Lastly, cost combination 1, which will be used for the rest of the analyses, is clearly the most demanding to solve, giving the longest execution times. Thus, if any of the other cost structures would be chosen instead, the model would probably run faster, and the following discussions of execution times can therefore be considered as informal upper bounds.

## 7.4 Stability testing

This section presents theory and analyses of the stability of the Phase Model and of the Emergency Model. All theory about stability testing is gathered from King et al. [9].

Stability testing is used to determine the likelihood that tests evaluate the optimisation model, rather than the scenario generation procedure. In this way, stability testing tries to answer if the discretisation of scenarios is good. The objective is to rule out the possibility that the results of the optimisation model are just random or systematic side effects of a poor scenario generation procedure. The main assumption is that there is an underlying true problem we approximate, but cannot solve. Commonly used tests are in-sample and out-of-sample stability tests, in addition to bias testing. Bias testing is usually impossible to execute, unless you can solve the true problem, and the theory about this is therefore excluded.

To describe the stability tests, we define $\mathcal{T}$ to be a two-stage scenario tree, where the optimisation problem can be written as

$$\min_x f(x; \mathcal{T})$$

The $x$ refers to the first stage variables and it is implicit that we calculate the expectation over the second stage variables, $y(\mathcal{T})$. The true optimisation problem, which we want to approximate, is given by

$$\min_x f(x; \xi)$$

King et al. [9] argue that stochastic programs tend to have flat objective functions, which means that very different solutions can have approximately the same objective value. To avoid the problem of comparing two different solutions with almost equal objective value, stability is measured by the objective value. Thus, for both of the following stability tests, it is the objective values that are compared for different solutions.

### 7.4.1 Out-of-sample stability

Out-of-sample stability means that the true objective value corresponding to solutions from scenario trees of different size are approximately the same. Thus, we test that the scenario generation procedure has not created an incorrect stability that is not really there.

If the scenario generation procedure is run several times on the same data, it will produce many different scenario trees, which we denote by $\mathcal{T}_i$. For each of these, the optimisation model is solved, producing equally many optimal solutions, denoted by $\hat{x}_i$. Mathematically, this means that

$$f(\hat{x}_i; \xi) \approx f(\hat{x}_j; \xi)$$

This test is easier to solve than the true optimisation problem, because we test with a fixed $\hat{x}_i$ and then add expectations of the second stage with respect to the true distribution. Solving $f(\hat{x}_i;\xi)$ then turns into solving a large number of independent second stage problems.

For out-of-sample stability, the situation is more complicated with multiperiod trees. A solution based on one tree cannot simply be evaluated in another one, as the nodes beyond the root do not coincide. One solution is to only implement the root node decisions and re-run the model with an updated tree and fixed root node decisions. That is, for multistage scenario trees $\mathcal{T}_1$ and $\mathcal{T}_2$, find the corresponding solutions $\hat{x}_1$ and $\hat{x}_2$. Then fix the root node solutions from $\mathcal{T}_2$ and run them on $\mathcal{T}_1$ and vice versa. If the two objective values are approximately the same, the method is out-of-sample stable.

We will use the results from the stability tests to decide the appropriate number of scenarios to use for later analyses. The data instances we use for the Phase Model vary in size from four to seven surgeries on a given day. Among these, we have chosen to run the stability tests on a set of instances with different number of surgeries, $n$. Thus, the stability tests for the Phase Model are run on instances 1, 2, 7, 9, 12, 15 and 14, in order of increasing number of surgeries. For the Emergency Model, we test stability related to the number of surgery duration scenarios, not the emergency patient scenarios. This is because the emergency scenarios not are a result of a scenario generation procedure, but consist of probabilities of emergency patient arrivals, which are found based on statistical analyses.

The out-of-sample stability tests for the Phase Model are run with 40-100 (in steps of 10), 125, 150, and 200 scenarios. For smaller scenario sizes, the scenario generation procedure does not converge, that is, the number of scenarios is too small to match the first four statistical moments of the data appropriately. Each combination of data instance and scenario size are run five times with different scenario trees. For both models, the first stage solutions, i.e. the surgery sequence and start times of elective surgeries, are recorded together with the objective value, $f(\hat{x}_i;\mathcal{T}_i)$. Then we generate a large scenario tree of 10 000 duration scenarios as an approximation of the true scenario tree in the Phase model. For each run $i$, we fix the first-stage solutions and run these on the large scenario tree to find $f(\hat{x}_i;\xi)$. Further, we calculate the average distance between the objective value from a run on the smaller scenario trees, $\mathcal{T}_i$, and the objective value from a run on $\xi$.

The average over five runs of each combination of instance and scenario size for the Phase Model is plotted in Figure 7.1. From the plots, one can see a significant improvement of the absolute distance as the number of scenarios approaches 100. Between 100, 125, 150, and 200 scenarios, the improvements are mostly small, and certainly diminishing. For most of the data instances, the relative distance is on average 4.0% for 100 scenarios, which we consider acceptable. This is not shown in the figure, but can be found in Appendix B. We

will from this point onward run experiments with 100 scenarios on the Phase Model. An example of the data from the out-of-sample stability tests of the Phase Model is given in Table 7.5. This shows the average over the five runs for data instance 12, for each tested scenario size.

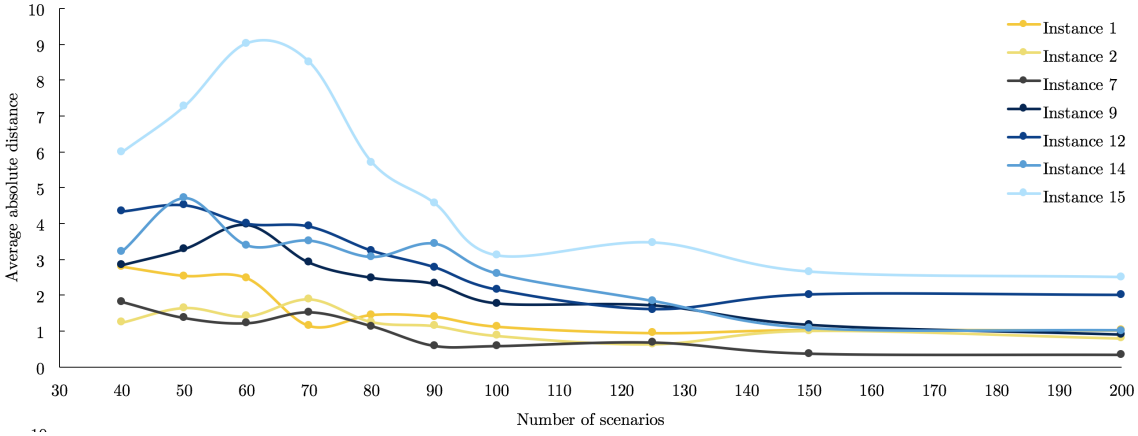



Figure 7.1: Average absolute distance to 10 000 scenario run, over five runs, for each data instance in the Phase Model

Table 7.5: Out-of-sample results for instance 12. Distance to the objective value when run on the true scenario tree, $\xi$, is given as an average over five run.

| **Number of scenarios** | **Average distance** | **Average relative distance %** |
|---|---|---|
| 40 | 4.3 | 8.6 |
| 50 | 4.5 | 10.3 |
| 60 | 4.0 | 9.0 |
| 70 | 3.9 | 8.8 |
| 80 | 3.2 | 7.2 |
| 90 | 2.8 | 6.2 |
| 100 | 2.1 | 4.7 |
| 125 | 1.6 | 3.1 |
| 150 | 2.0 | 4.4 |
| 200 | 2.0 | 4.3 |

Figure 7.2 shows the out-of-sample results for the Emergency Model. Again, the plot displays the average absolute distance between the objective value from an instance run and the objective value in an approximation of the true scenario tree where the first-stage solutions are fixed. The values are average values over five runs for each data point. In contrast to the Phase Model, we had to use a smaller scenario tree in the out-of-sample test for this model, due to long execution times. For a 10 000 scenario tree, even after an hour of execution time, we experienced large optimality gaps. However, with a limit of 1 800 seconds on the execution time, we could run the 2 000 scenario tree with resulting gaps in a 1-2% range. The out-of-sample stability tests on the Emergency Model are run with 40 to 200 (in steps of 10) scenarios. The axes in the plot are scaled equally as in the plot for the Phase Model, to compare the stability. The relative distance to the objective

value ranges from 1.5% for 40 scenarios to 0.1% for 200 scenarios. Based on the relative distances, and the number of scenarios required for the plot to seemingly converge, we conclude that 60 scenarios is stable enough for our further tests, with an average relative distance of 0.7% across the three emergency instances.

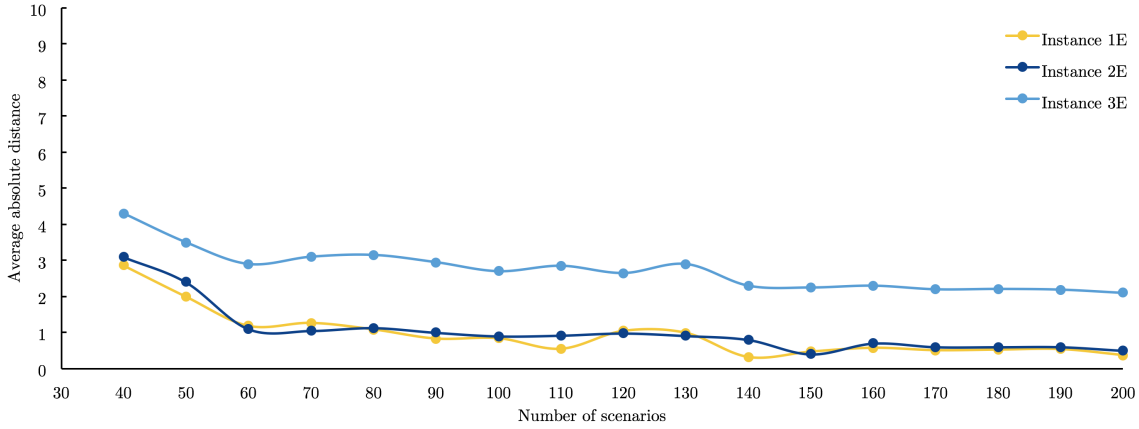



Figure 7.2: Average absolute distance to 2 000 scenario run, over five runs, for each data instance in the Emergency Model

### 7.4.2 In-sample stability

In-sample stability represents a test of a model's internal consistency. That is, if we have in-sample stability, it does not matter which scenario tree, $\mathcal{T}_i$, we use. Mathematically, this means that we have an in-sample stable model if

$$f(\hat{x}_i; \mathcal{T}_i) \approx f(\hat{x}_j; \mathcal{T}_j)$$

This ensures that running the scenario generation procedure and then the optimisation procedure yields the same objective value if we repeat it with the same data and a newly generated scenario tree. An in-sample unstable model cannot be properly tested, as the results it produces are random.

To test for in-sample stability, we have run each data instance five times with different scenario trees. The objective value from each run, $f(\hat{x}_i)$ is recorded and compared across the five runs, for each data instance. Specifically, the CV is calculated over the five runs as a measure of the dispersion of the relative variance of the five runs, to evaluate if $f(\hat{x}_i; \mathcal{T}_i) \approx f(\hat{x}_j; \mathcal{T}_j)$. The results of the tests are shown in Table 7.6 for the Phase Model. The first column shows the data instance and the second column shows the coefficient of variance, given in percentage. The last column shows the average of all the reported data instances.

The largest CV for the Phase Model is 2.97% for data instance 14. If this would have come from a normal distribution, it would mean there is approximately 68%[6] chance of

[6]From a table of the Standard Normal Cumulative Distribution Function, $\Phi(1) - \Phi(-1)$.

Table 7.6: In-sample results for the Phase Model. CV is calculated over five runs with different scenario trees.

| **Problem instance** | **Coefficient of variance, %** |
|---|---|
| Instance 1 | 2.83 |
| Instance 2 | 1.48 |
| Instance 7 | 1.50 |
| Instance 9 | 0.93 |
| Instance 12 | 1.01 |
| Instance 14 | 2.97 |
| Instance 15 | 1.80 |
| Average | 1.79 |

getting an objective value within the interval $\mu \pm 2.97\%$. We deem this as a sufficiently small CV and stable for our analysis purposes.

Table 7.7: In-sample results for the Emergency Model. CV is calculated over five runs with different scenario trees.

| **Problem instance** | **Coefficient of variance, %** |
|---|---|
| Instance 1E | 0.21 |
| Instance 2E | 0.35 |
| Instance 3E | 0.58 |

The CV for the Emergency Model is only 0.21%, 0.35%, and 0.58% for emergency instances 1E, 2E and 3E, respectively, for 60 scenarios. Because the CV is the standard deviation divided by the mean, a higher mean in the Emergency Model implies a lower CV, given a constant standard deviation. Nevertheless, the difference between the max and min objective value across the five runs is very low, which indicates that the the Emergency Model is in-sample stable.

## 7.5 Improved implementation

In this section we will perform an evaluation, in terms of improved execution times and lower gaps, of the cuts we proposed in the model formulation Chapter 6. In addition, we will discuss how adjustments of the optimiser branching strategies will affect the solution time.

### 7.5.1 Strength of valid inequalities

All valid inequalities introduced for the Phase Model in Section 6.1.7 and the Emergency Model in Section 6.2.7 are tested here. To evaluate the strength of these additional constraints, we will test each valid inequality in turn and report the optimality gap and

execution time. The instances are also tested with all five valid inequalities, to evaluate their collective improvement of the model.

The tests of the valid inequalities in the Phase Model are run on all data instances. We report the maximum and average relative optimality gap improvement when the model reaches maximum execution time of 1 800 seconds. If the model is solved to optimality within the time limit, the maximum and average relative execution time improvement are reported. The improvements are calculated relative to runs with no valid inequalities. We have chosen to report the maximum relative improvements instead of an exhaustive list of improvements for all combinations of instance and cut, because we are interested in valid inequalities that give a speedup for some instances and not necessarily for all. However, it is important that the cuts do not severely degrade the performance of other instances. The average improvements have thus been included to quantify the general improvement.

Table 7.8: Strength of valid inequalities in the Phase Model. The results get the maximum and average relative improvements across all instances, given in percent.

| **Valid inequality** | **Equation** | **Max time** | **Avg time** | **Max gap** | **Avg gap** |
|---|---|---|---|---|---|
| Overtime balance | (23) | 70.2 | 8.3 | 21.6 | 5.6 |
| Expanded $x$ sum | (58),(59) | 92.0 | 48.0 | 79.1 | 16.7 |
| Phase ($\tau$) precedence | (55) | 67.1 | 18.5 | 74.1 | 12.6 |
| Minimum start spread | (56) | 84.8 | 30.0 | 88.5 | 30.2 |
| Overtime ($\delta$) precedence | (57) | 11.6 | 1.3 | 2.3 | 0.7 |
| All | N/A | 89.5 | 39.4 | 90.0 | 41.7 |

The test results for the Phase Model are given in Table 7.8. The results show significant speedups for several of the cuts. For the smaller data sets, the model solves to optimality within the time limit. The execution time improvements are therefore only based on the results from these tests.

The expanded $x$ sum provides the best improvements, and gives up to 92% shorter execution times when the solution is found optimal in less than 1 800 seconds. It is easy to see that this cut is correct from equations (28) and (29) in the Phase Model. However, as stated in Chapter 6.3 about formulation challenges, these relationships between $x_{ij}$ and $y^{\omega}_{ijh}$ can be hard for a general MIP optimiser to detect. By including the constraints explicitly, the model gets much tighter relationships between these variables, which results in the observed speedup. The overtime precedence inequalities give only minor execution time improvements for the tested data sets, on average. With all inequalities included, we experience 39% shorter execution times on average. This is lower than with only the expanded $x$ sum inequality. A possible explanation for this may be that the combination of inequalities changes how the optimisation software searches for solutions, which may affect search times.

The improvements of the optimality gaps are calculated over all runs that reach maximum execution time. This is the case for all the larger problem instances with six and seven surgeries. First of all, one can see that the overtime precedence inequality improves the optimality gap a bit. Second, as mentioned in the problem formulation chapter, the overtime balance constraints provide a tighter formulation and improves the optimality gap with 6% on average. Similar, the expanded $x$ sum inequalities lower the optimality gap by 17%. These improvements are observed because the constraints bind several variables together, revealing important logical relationships between the variables to the optimiser. The best gap improvements, both in maximum and average terms, come from the minimum start time spread inequalities, with 89% and 30% improvements, respectively.

In contrast to the execution time improvements, the combination of all the inequalities gives a significantly tighter formulation than any single inequality does, when considering the optimality gap. With all inequalities included, we achieve 42% relative lower optimality gaps when the model has run for 1 800 second without finding the optimal solution. All cuts are therefore included in the preceding analysis of the Phase Model.

The tests of the valid inequalities in the Emergency Model are run on all emergency data instances. The tested valid inequalities include the one discussed in Chapter 6.2.7 and those that are transferable from the Phase Model and adapted to the emergency setting in Appendix A. The results shown are the relative improvements for each emergency data instance, relative to runs with none of the cuts included.

Table 7.9: Strength of valid inequalities in the Emergency Model. The results are the relative improvements for instances 1E and 2E, given in percent.

| **Valid inequality** | **Equation** | **Run time 1E** | **Run time 2E** |
|---|---|---|---|
| Start time spread | (112) - (113) | -34.7 | 26.1 |
| Subposition precedence | (103) | -18.9 | -31.7 |
| Overtime ($\delta$) precedence | (114) - (116) | -123.0 | 4.4 |
| All | N/A | 3.8 | -14.5 |

Table 7.9 provides a report of the valid inequalities for the Emergency Model. These tests are allowed only 900 seconds of execution time, because the Emergency Model, in general, runs faster than the Phase Model. Data instance 3E, however, reached maximum computation time and achieved very low gaps for all cuts. That is, for long run times, the cuts did not provide any significant improvement of the optimality gap and this data instance is therefore excluded in the table. For data instances 1E and 2E, execution time improvements are stated for each cut. Both data instances are solved to optimality within 2 seconds of each other and the relative improvements are therefore comparable.

Interestingly, the cuts behave differently for the two data instances. For instance 1E every cut makes the solution time increase rather than decrease, but all together they still improve solution time. The least damaging cut is the subposition precedence inequality. On the other hand, for instance 2E, the total effect of the cuts is negative, even though

two of three cuts improve the solution time. Because of the inconsistency in the results, we would have to perform more extensive testing of the cuts, on a lot more data instances to determine their average effect. Therefore, none of the three valid inequalities are included in the continuation of the computational study.

### 7.5.2 Branching

To speed up the solution process we look at some implementation finesses and settings in the optimiser. In a branch-and-bound search, the optimiser uses certain rules to determine both the next variable and the value to branch on. The choice of branching order and strategy can sometimes be derived more readily with intuition and comprehension of the model, than with general mathematics and rules. However, the exact effect of different configurations may not be as expected, and a trial-and-error approach is often used.

In the Phase Model, we believe the branching order should start with the $x_{ij}$ variables. This is because these are the most determinant variables and restrict the rest of the problem. Thus, by branching on these, we make the most significant decisions first. Because we have $i \cdot j$ number of $x_{ij}$ variables and exactly $i$ of them are equal to 1 in the optimal solution, a branching strategy that tries to branch up to 1 first will be the most restrictive. This is the first branching strategy we test. The second strategy consists of branching on all the first-stage variables, i.e. $x_{ij}$ and $t_i$, where $x_{ij}$ is given a higher priority than $t_i$.

To test the effect of the branching rules, we run all data sets that can be solved to optimality within 1 800 seconds. By letting the optimiser use the default branching strategy, the average execution time of the data instances is 204 seconds. By using the first branching strategy, we are able to decrease the average over the same data instances to 103 seconds. The second strategy performed almost identical. Several other branching rules are tested, with varying degrees of speedup. However, none of these were able to match the performance improvements from the first mentioned branching rules, and the exact numbers are therefore omitted. The strategy of branching on $x_{ij}$ is used in the succeeding analysis of the Phase Model.

Similar branching combinations are tested for the Emergency Model. We first try branching on the $x_{ij}$ and then on both $x_{ij}$ and $t_i$. The decision of where to place the emergency surgeries may heavily affect the performance, which can make it beneficial to branch on this early in the search. Thus, we also try to branch on the second-stage variables $z^{\xi}_{iuq}$ and $\phi^{\xi}_{iu}$. As for the Phase Model, the sequencing variables, i.e. $x_{ij}$ and $z^{\xi}_{iuq}$, are given a higher priority than $t_i$ and $\phi^{\xi}_{iu}$. The binary variables are branched up to 1, as this is more restrictive.

Different combinations of branching strategies are tested on all emergency instances, all of which solve to optimality. Solving the model with default branching settings uses 120

seconds on average, while branching on both $x_{ij}$ and $z^{\xi}_{iuq}$ reduces the execution time to 78 seconds on average. None of the other combinations of branching show any promising results. For example, branching on $x_{ij}$ and $t_i$ results in an increase in execution time to 219 seconds on average. The strategy of branching on both $x_{ij}$ and $z^{\xi}_{iuq}$ is used in the succeeding analysis.

## 7.6 Heuristics

In this section, we will compare the optimal solutions of the model with the solutions we get by using the decision rules and heuristics described in Chapter 6.4. The motivation for testing heuristics is twofold. Firstly, we want to evaluate the effect of heuristics and decision rules found in literature and at St. Olavs Hospital. We want to investigate whether simple decision rules are enough to provide decent schedules, or if more complex optimisation-based heuristics are needed. Secondly, we want to evaluate the trade-offs between optimal solutions and shorter execution times.

The heuristics will only be tested for the Phase Model. The decision rules are developed based on the characteristics of uncertain duration and not of uncertain arrival of patients, so these are not considered appropriate for the Emergency Model. The Emergency model already has low execution times for all the data instances, and preliminary tests reveal that both local search and simulated annealing give higher objective values and higher execution times than solving the stochastic program to optimality for this model.

There are several parameters in the heuristics that can be adjusted. Before we compare the heuristic solutions with the optimal solution, we will run some tests to tune these parameters. For the Bailey-Welch rule, this means determining the parameter $k$, which is the number of surgeries with scheduled start time 0. For the two search heuristics, we will experiment with the initial sequence and the neighbourhood function. Specifically for simulated annealing, we must decide the parameters $T$ and $r$. An initial configuration of these is suggested and will be tested against some variations in the same value range.

### 7.6.1 Optimal parameter for the Bailey-Welch decision rule

The average objective values found when using the Bailey-Welch decision rule, for all data instances, with $k = 1$ and $k = 2$, are 78 and 239, respectively. The optimal value depends on the ratio between the costs of waiting and idling. A relatively higher waiting than idle cost, will give results in favour of $k = 1$, as $k = 2$ implies more waiting. For the chosen cost combinations, the results show a clear benefit of using $k = 1$. The Bailey-Welch rule with $k = 2$ shows an interesting trend; when there are no overtime, the surgeries are scheduled in order of increasing mean. This may be explained by the core definition of this decision rule, scheduling surgeries at even intervals with length equal to the overall mean surgery

duration. Thus, by scheduling the surgeries with lowest mean first, the added flexibility of having a patient waiting can be utilised, and the total waiting time decreases. On the other hand, if surgeries with duration higher than the mean are scheduled first, there would be even more waiting time imposed for all succeeding surgeries. These trade-offs are, however, not as straight forward when overtime is present. For the Bailey-Welch rule with $k = 1$, the sequence of surgeries do not follow this pattern, because there is not the same amount of waiting accumulated from the start of the day. Equivalently, this is also the main driver for the objective value for $k = 1$.

When we report the performance of the Bailey-Welch decision rule in the following sections, we will be using $k = 1$, based on this test.

### 7.6.2 Performance of neighbourhood functions

The performance of the two neighbourhood functions, two-swap and surgery-pair swap, are compared in Table 7.10 for all instances in the Phase Model for both the local search and simulated annealing. Two-swap performs better in terms of objective value for 96% of the instances in both heuristics. The execution time, however, is better for the surgery-pair swap across all instances and heuristics. This follows from the size of the neighbourhoods, which was discussed in Chapter 6.4. Two-swap has a larger neighbourhood and thus evaluates a higher number of sequences than surgery-pair swap. Consequently, there is a trade-off between the chance of finding better solutions and the increase in execution time. As will be discussed later, the two-swap gives a significant speedup compared to solving the model to optimality, and is thus useful. However, if a further decrease of the execution time is crucial, the surgery-pair swap neighbourhood function should be applied.

Table 7.10: Comparison of the neighbourhood performance for the search heuristics, tested for all data instances. $N1$ denotes the two-swap neighbourhood, while $N2$ denotes the surgery-pair swap.

| | **Objective value** | | | | **Execution time** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Local search** | | **SA** | | **Local search** | | **SA** | |
| **Size** | **N1** | **N2** | **N1** | **N2** | **N1** | **N2** | **N1** | **N2** |
| 4 | 38.9 | 38.9 | 38.9 | 38.9 | 38.9 | 38.9 | 38.9 | 38.9 |
| 5 | 36.7 | 36.7 | 36.7 | 36.7 | 36.7 | 36.7 | 36.7 | 36.7 |
| 6 | 241.2 | 241.2 | 241.2 | 241.2 | 241.2 | 241.2 | 241.2 | 241.2 |
| 7 | 47.2 | 47.2 | 47.2 | 47.2 | 47.2 | 47.2 | 47.2 | 47.2 |
| Average | 58.8 | 69.3 | 58.8 | 65.5 | 164.3 | 32.2 | 177.4 | 54.7 |

Because the searches can converge to a local optimum, we also test the sensitivity to the initial sequence. For the two-swap neighbourhood, every possible sequence can be reached in $n - 1$ iterations, when $n$ is the number of surgeries. Every data instance, with five random neighbourhoods that are $n - 2$ iterations away from each other, is tested. That is, each initial sequence we test, is the maximum distance away from the other initial

sequences. We test each of these initial sequences for both the local search method and simulated annealing, and compare the objective values found.

For both algorithms, there are only a few instances that do not give the same optimal solution, and in these cases, there are only small standard deviations across the objective values. This might be explained by the flat objective function often found in stochastic programs, and does not necessarily imply insensitivity. However, the low number of runs that differs from each other for a given data instance, is a strong implication of insensitivity towards the initial sequence. Because the heuristics are not sensitive to the choice of start point, we simply start the heuristics with a random start sequence to reduce the solution time compared to using a construction heuristic to find a good starting sequence.

### 7.6.3 Tuning of simulated annealing

In the simulated annealing algorithm, one needs to decide several parameters: the initial temperature, the reduction factor and the number of iterations. A too high number of iterations, will simply make the algorithm greedy because the temperature becomes too low to accept worse solutions. If one in addition hashes all previous solutions, repeatedly solving the same problem in a local optimum will go extremely fast. Thus, we only need a number of iterations that is high enough. A few test runs show that 100 iterations comply with these criteria.

To test different combinations of initial temperature and reduction factor, we approximate a range of appropriate values. The initial temperature should be in a range such that the algorithm has a diversified behaviour in the early iterations. That is, the expression $e^{-\frac{\Delta}{T}}$ should give a high probability of accepting solutions with objective values that are slightly worse than the incumbent solution. As a starting point this probability was set to 90%. Experiments on our problem instances show that solutions found early in the optimisation process usually have values in the range [100, 120]. Thus, given that the incumbent solution has an objective value of 100, a candidate solution of 120 has 90% chance of being accepted in the first iterations. This gives $e^{\frac{-(120-100)}{T_0}} = 0.90$, and further the initial temperature, $T_0 \approx 190$. The idea of the algorithm, is that it should turn into an intensifying search after some iterations. To find a suitable value for $r$, we therefore test with an almost greedy search after 50 iterations. This means the probability of accepting a worse solution, say with a value 1 higher than the current best solution, should be low, for example 1%. By using the expression for acceptance probability, this gives $e^{\frac{-1}{T}} = 0.01 \implies T_{50} \approx 0.2$. Thus, an initial $r$ could be chosen such that $T_0 * r^{50} = 0.2 \implies r = \frac{0.2}{190}^{\frac{1}{50}} \approx 0.87$. This is close to $r = 0.85$ found by, for example, Park and Kim [91].

To find the best parameters for our problem, we test several combinations of parameters. Specifically, we test all initial temperatures, $T_0$ in the set $\{100, 150, 190, 200, 250\}$ together with all reduction factors, $r$, in the set $\{0.8, 0.85, 0.87, 0.9, 0.95\}$. Table 7.11 shows the

average and standard deviation of the objective values found for each data instance across all 25 different combinations of $T_0$ and $r$. To compress the size of the table, all instances with standard deviation equal to zero are omitted. All instances with zero standard deviation of the objectives are insensitive to the choice of $T_0$ and $r$. Thus, to evaluate the parameters, we must look at how they affect the remaining instances. Many of these instances have very small standard deviations. The deviations from the best objective value found come from a few runs, that have found slightly worse solutions. If we, for all of

Table 7.11: Tuning of simulated annealing parameters

| **Data instance** | **Average objective value** | **Standard deviation** |
|---:|---:|---:|
| 11 | 39.17 | 0.12 |
| 15 | 41.74 | 0.25 |
| 16 | 53.95 | 0.17 |
| 17 | 102.77 | 4.51 |
| 19 | 48.74 | 1.79 |
| 21 | 382.61 | 0.85 |
| 23 | 32.40 | 0.14 |
| 25 | 27.11 | 0.32 |
| 26 | 28.97 | 1.86 |

these parameter combinations, look at the average relative deviation of the objective value found to the best objective for that data instance, the combinations $(T_0 = 150, r = 0.8)$, $(T_0 = 100, r = 0.85)$ and $(T_0 = 190, r = 0.85)$ have 0%, 0% and 0.8% average relative deviation, respectively. The rest of the combinations have a deviation larger than 3%. Among the three combinations with the lowest relative deviations, we will use the parameters that provide the most diversified search. This is achieved by the last combination. From this point tests of simulated annealing will be using initial temperature of 190 and a reduction factor of 0.85.

### 7.6.4 Analysis of heuristics

The results for the decision rules and heuristics are shown in Tables 7.12 and 7.13, which list the objective value and the execution time for each instance. Bailey-Welch, sort by variance, sort by mean, local search and simulated annealing are presented as BW, variance, mean, local and SA, respectively, in these tables. Every heuristic and decision rule, with the chosen parameters, are run for all instances. These results are compared to the objective values of the optimal solution, solved using branch-and-bound. All solution methods are run with the same scenarios trees for each instance, and the maximum execution time is set to 1 800 seconds.

From Table 7.12, it is evident that the sort by variance decision rule performs best on average in terms of objective value with values close to the optimal solution. This method

Table 7.12: Objective values for the branch-and-bound method, heuristic and decision rules

| Size | Instance | Optimal | BW | Variance | Mean | Local | SA |
|---|---|---|---|---|---|---|---|
| 4 | Instance 1 | 25.6 | 31.9 | 25.6 | 25.6 | 25.6 | 25.6 |
| | Instance 2 | 28.5 | 109.0 | 28.9 | 31.6 | 28.5 | 28.5 |
| | Instance 3 | 38.9 | 50.8 | 39.3 | 38.9 | 141.9 | 141.9 |
| | Instance 4 | 22.1 | 24.9 | 22.1 | 45.1 | 22.1 | 22.1 |
| | Instance 5 | 37.5 | 46.4 | 37.8 | 61.7 | 37.5 | 37.5 |
| | Instance 6 | 39.2 | 45.5 | 41.5 | 56.2 | 39.2 | 39.2 |
| | Instance 8 | 24.5 | 40.3 | 24.5 | 52.0 | 24.5 | 24.5 |
| | Instance 13 | 17.4 | 46.3 | 17.4 | 26.7 | 17.4 | 17.4 |
| | Instance 16 | 26.7 | 28.3 | 27.6 | 36.7 | 26.7 | 26.7 |
| | Instance 18 | 26.2 | 30.9 | 26.2 | 30.7 | 26.2 | 26.2 |
| 5 | Instance 7 | 31.8 | 69.5 | 38.3 | 59.5 | 31.8 | 31.8 |
| | Instance 9 | 45.7 | 46.7 | 47.2 | 47.6 | 45.7 | 45.7 |
| | Instance 10 | 41.7 | 107.6 | 41.7 | 47.9 | 41.7 | 41.7 |
| | Instance 11 | 53.9 | 84.1 | 58.0 | 63.9 | 53.9 | 53.9 |
| | Instance 17 | 31.8 | 52.9 | 32.5 | 39.7 | 31.8 | 31.8 |
| | Instance 19 | 26.5 | 91.4 | 27.7 | 39.9 | 26.5 | 26.5 |
| | Instance 20 | 25.3 | 40.4 | 25.3 | 32.7 | 25.3 | 25.3 |
| 6 | Instance 12 | 69.9 | 148.9 | 81.5 | 104.3 | 100.4 | 100.2 |
| | Instance 15 | 381.9 | 480.0 | 418.1 | 431.5 | 381.9 | 382.6 |
| 7 | Instance 14 | 48.4 | 84.8 | 55.2 | 66.2 | 48.4 | 48.4 |
| Average | | 52.2 | 83.0 | 55.8 | 66.9 | 58.8 | 58.9 |

found the best solution of all heuristics and decision rules in 40% of the instances. It is the only solution method that for no instances has the largest deviation from the optimal objective value, and thus proves to be quite robust for our data. The performance supports the intuition that if a surgery duration is hard to predict, meaning it has a large variation, setting it later in day reduces the risk of impacting a large number of subsequent surgeries. However, this may not be desirable when there is a risk of overtime, as this has a high cost. This tendency is shown in our results. As the number of surgeries increase, i.e. as the expected duration of the day increases, the relative deviation from the optimal solution increases from 2% to 14% as the number of surgeries increases from four to seven. In other words, sorting by variance performs well for a small number of surgeries, but the performance degrades with an increasing number of surgeries. The sort by variance solves to optimality almost instantly for all instances, shown in Table 7.13. This is a result of the sequencing variables already being fixed, which reduces the size and complexity of the optimisation problem. The combination of short execution time and low objective values may explain the popularity of this scheduling strategy.

Sort by mean provides similar results as the sort by variance decision rule. However, it only gives the best solution in 4% of the instances and has on average 31% deviation from the optimal solution. These results indicate that the variance does not necessarily increase

Table 7.13: Execution time for the branch-and-bound method, heuristics and decision rules

| Size | Instance | Optimal | BW | Variance | Mean | Local | SA |
|---|---|---|---|---|---|---|---|
| 4 | Instance 1 | 34.5 | 18.1 | 0.2 | 0.1 | 6.8 | 8.1 |
| | Instance 2 | 19.5 | 16.6 | 0.4 | 0.4 | 6.9 | 10.0 |
| | Instance 3 | 427.5 | 23.1 | 1.2 | 0.9 | 159.5 | 134.4 |
| | Instance 4 | 36.1 | 14.2 | 0.4 | 0.3 | 4.6 | 12.5 |
| | Instance 5 | 142.7 | 25.0 | 0.5 | 2.3 | 18.4 | 22.5 |
| | Instance 6 | 1800.5 | 21.5 | 18.6 | 0.4 | 144.3 | 161.2 |
| | Instance 8 | 39.1 | 12.3 | 0.5 | 0.5 | 5.9 | 7.8 |
| | Instance 13 | 21.9 | 15.6 | 0.1 | 0.2 | 950.1 | 1097.7 |
| | Instance 16 | 17.9 | 27.5 | 0.2 | 0.2 | 104.5 | 222.1 |
| | Instance 18 | 16.7 | 22.7 | 0.3 | 0.4 | 130.5 | 129.8 |
| 5 | Instance 7 | 94.6 | 71.5 | 1.8 | 2.0 | 12.0 | 18.1 |
| | Instance 9 | 583.3 | 35.6 | 2.2 | 2.7 | 8.0 | 7.5 |
| | Instance 10 | 1799.6 | 45.1 | 5.5 | 3.2 | 6.6 | 7.9 |
| | Instance 11 | 1800.1 | 72.5 | 5.3 | 15.5 | 76.6 | 166.8 |
| | Instance 17 | 339.7 | 54.1 | 2.4 | 2.5 | 60.9 | 196.0 |
| | Instance 19 | 1800.1 | 39.9 | 3.8 | 1.3 | 739.4 | 478.2 |
| | Instance 20 | 689.8 | 56.4 | 2.6 | 3.0 | 351.3 | 257.2 |
| 6 | Instance 12 | 1799.4 | 147.6 | 5.2 | 5.2 | 99.7 | 111.1 |
| | Instance 15 | 1799.3 | 247.1 | 4.5 | 3.8 | 51.9 | 223.6 |
| 7 | Instance 14 | 1799.6 | 1799.3 | 9.8 | 8.1 | 347.7 | 274.8 |
| Average | | 753.1 | 138.3 | 3.3 | 2.6 | 164.3 | 177.4 |

with the duration of the surgery. If a long surgery has a low variance, setting this in the beginning of the day may be preferable as its duration is easier to predict. In terms of execution time, the decision rule performs well. This follows from the same explanation as given for the sort by variance. The results of these two decision rules are in accordance with the results of Denton et al. [19] mentioned in the literature review in Chapter 3.

Simulated annealing and local search perform similarly for all instances. They also provide the best solution in 90% of the runs. However, for certain instances, their performance degrades severely compared to sort by variance and thus do not prove to be equally stable for our instances. As opposed to sort by variance, simulated annealing and local search experience a diminishing deviation from the optimal value as the number of surgeries increases. It deviates by 36% on average for instances with four surgeries, but gives the optimal solution for the instance with seven surgeries. As we only have one instance with seven surgeries, it may be a coincidence that it finds the optimal solution, but the trend holds for instances with both five and six surgeries as well. This may indicate that a good sequence cannot easily be calculated based on statistical measures for more complex instances. Both heuristics have higher execution time than the decision rules. It is worth noting that this is still an improvement from the branch-and-bound method, with an average reduction of 77% of the execution time.

Local search and simulated annealing perform similarly for all instances. This may be because they use the same neighbourhood. Even though simulated annealing incorporates randomisation, this proves to have little effect for these instances. As they also have similar execution times, this may be an indication that they investigate a similar amount of possible sequences.

Obvious from the results of the objective values, Bailey-Welch has the worst performance of all heuristics and decision rules tested. It does not show any consistency in the results as the number of surgeries increase. It performs slightly better than local search and simulated annealing in terms of execution time, which may be explained by the fact that the start times are fixed, reducing the level of optimisation compared to the other two. Both high computational times and objective values invalidates this as an appropriate solution method for the cost structure we consider.

The most considerable speedup is expected for the instances that reach the maximum execution time of 1 800 second, and would work longer if allowed. By eliminating these from the calculations, the solution methods still prove to outperform the branch-and-bound method. While the optimal solution is found after 230 seconds on average, the Baily-Welch, sort by variance, sort by mean, local search and simulated annealing finish after 39, 1, 1, 142 and 164 seconds, respectively.

## 7.7 Value of information

A commonly used criterion for determining the importance of uncertainties in mathematical programs is the expected value of perfect information ($EVPI$) [92]. The $EVPI$ reveals the potential worth of more accurate duration forecasts. However, in situations exhibiting external uncertainty, it may not be possible to gather more information about the future, and it may be more relevant for decision makers to know the expected value of planning with uncertainty compared to the expected value case [92]. The measure called the value of the stochastic solution ($VSS$) quantifies this. Further, we want to quantify both the effect of planning with phases and emergency patients. We call the first measure the expected value of planning with phases ($EVPP$) and the second the expected value of planning with emergency patients ($EVPE_i$). This section will provide definitions of the defined measures. Further, an analyses of these are conducted for both the Phase Model and the Emergency Model.

### 7.7.1 The expected value of perfect information and the value of stochastic solution

We let $SP$ denote the objective value of the stochastic program and $WS$ the objective value of the wait-and-see problem, i.e. the expected value when all scenarios are solved

individually as deterministic problems. Then, the expected value of perfect information for a minimisation problem is given by

$$EVPI = SP - WS \tag{106}$$

as stated by Birge and Louveaux [92]. This determines how much one is willing to pay to obtain perfect information of the future [93]. Further, we let the $EEV$ denote the expectation of the expected value solution, i.e. the objective value of the problem where all stochastic parameters are replaced with their expected values and the first-stage solution from this is evaluated in the stochastic model. The $VSS$ denotes the value of solving a stochastic model compared to a deterministic one [92] and is given by

$$VSS = EEV - SP \tag{107}$$

Also for multi-stage programs with $t = 1, ..., T$ stages, the $VSS$ can be obtained with Equation (107), by solving the expected value case and fixing variables in all stages except the last. A more appropriate interpretation of the information structure may be obtained by the $VSS^D$ by fixing variables at each stage dynamically [93]. With this approach, values are updated when more information is revealed. We define $G_t$ as the set of scenario groups at stage $t$; two scenarios belong to the same group in a given stage provided that they have the same realisations of the uncertain parameters up to that stage. In other words, all child nodes for a certain node at stage $t$ belongs to the same scenario group. The $EV_g$ is the problem for scenario group $g$ where all random parameters in subsequent stages are estimated by their expected values, while all variables in previous stages are fixed to their optimal values obtained in the chain of $EV_g$ for $g \in G_\tau$, $\tau = 1, ..., t-1$. Let $Z^g_{EV}$ be the optimal value for $EV_g$ and $p^g$ be the probability of scenario group $g$. Then the expected result in $t$ of using the dynamic solution of the average scenario, the $EDEV_t$, is given by

$$EDEV_t = \sum_{g \in G_t} p^g Z^g_{EV} \quad t = 1, ..., T \tag{108}$$

The value of the dynamic stochastic solution is given by

$$VSS^D = EDEV_T - SP \tag{109}$$

### 7.7.2 The expected value of planning with phases and expected value of planning with emergency patients

The $EVPI$ and the $VSS$ compare the stochastic model to the extremities of solving the problem with perfect information and solving the problem simply without considering the uncertainty. To evaluate how the Phase Model and the Emergency Model perform, we propose one additional measure for each model. We will use these to quantify the value that the additional information of phases and emergency patients provides.

We let the $SPNP$ denote the expected objective value of a stochastic program without phases. The $SP^P$ is the objective value of the stochastic program with phases[7]. Similar to the expression for the $VSS$, we define

$$EVPP = SPNP - SP^P \tag{110}$$

This is the expected objective value when planning without phases less the expected objective value from planning with phases.

Similarly, we want to analyse how the Emergency Model performs compared to scheduling either exactly zero, one or two emergency patients. We let the $SPEP_i$ denote the expected objective value of a stochastic program scheduling $i$ emergency patients, while $SP^E$ is the objective value of the stochastic program with emergency patients. The expected value of the stochastic solution compared to planning with exactly $i$ emergency patients, is then

$$EVPE_i = SPEP_i - SP^E \tag{111}$$

### 7.7.3 Value of the Phase Model

The $EVPI$, the $VSS$, and the $EVPP$ are calculated for all problem instances for the Phase Model. All tests are run to optimality and the results are stated in Table 7.14. This shows the average values for each size of problem instances. To use the appropriate amount of information when calculating the $SPNP$, we find the average statistical moments for a full day, without differentiating between phases, for the same instances as listed in Table 7.1. This means that the scenario generation procedure must be run separately to generate the input data to find the $SPNP$ and $SP$. Therefore, in-sample instabilities can affect the results slightly and variations of up to 2.9%[8] between the two cannot be considered significant.

[7]The $SP^P$ and the $SP^E$ are only noted here to accentuate the difference between the stochastic solution from the Phase Model and the Emergency Model, respectively. When there is no risk of confusion, $SP$ will be used.

[8]From in-sample stability tests for the Phase Model

Table 7.14: Average $EVPI$, $VSS$ and $EVPP$ for all problem instance sizes for the Phase Model

| **No. of surgeries** | **SP** | **WS** | **EEV** | **SPNP** | **EVPI** | **VSS** | **EVPP** |
|---|---|---|---|---|---|---|---|
| 4 | 28.2 | 0.0 | 135.8 | 37.8 | 28.2 | 107.6 | 9.6 |
| 5 | 37.9 | 0.0 | 168.4 | 44.0 | 37.9 | 130.6 | 6.1 |
| 6 | 107.0 | 74.5 | 361.2 | 259.2 | 32.5 | 254.2 | 152.2 |
| 7 | 36.7 | 0.0 | 243.0 | 73.1 | 36.7 | 206.3 | 36.4 |

The $WS$ is close to zero for most instances, except for a few of the larger sets. This happens because in a deterministic environment with perfect information, all surgeries will be scheduled back-to-back, starting from time $t_1 = 0$, with no risk of suffering from waiting time or idle time. In these cases, there are no costs and the objective value is zero. However, in some scenarios, it is impossible to avoid overtime because the sum of realised surgery durations exceeds the duration of the regular working hours. This is shown in the average of the $WS$ for instances with six surgeries.

In the cases where the $WS$ is larger than zero, the $SP$ experience a proportional increase. This means that the $EVPI$ is similar for all instances and the value of perfect information compared to the stochastic model is approximately constant when the number of surgeries increases. With an average cost of 1 per minute of waiting and 2 per minute of idling, perfect information will on average reduce the waiting time by 34 minutes or idle time by 17 minutes.

The $VSS$ varies a lot more than the $EVPI$, ranging from 106 to 254. Using a stochastic instead of a deterministic model may result in a reduction of 254 minutes waiting time or 127 minutes idle time. On average, our stochastic model reduces the cost by 77% compared to the expected value model. Because we only have a few data instances with six and seven surgeries, it is hard to tell whether the average values in Table 7.14 are representative for all data instances of this size, but the table still shows a trend. A higher number of surgeries lead to a higher $VSS$. This result is strengthened with the following reasoning. On a day with two surgeries, allocating a smaller amount of time than the realised duration of the first surgery may cause a delay for maximum one succeeding surgery. When the number of surgeries increases, a delay in the first surgery may delay several succeeding surgeries. This indicates that for our instances, the stochastic solution is more valuable for a higher number of surgeries.

The $EVPP$ is displayed in the last column of Table 7.14. As for the value of the stochastic solution, we expect the value of planning with phases, which gives a more detailed description of the underlying uncertainty, to be more valuable for a higher number of surgeries. This is, to some extent, supported by the results shown in the table. The highest average $EVPP$ is observed for data instances with six surgeries. Considering all data instances, the average $EVPP$ is 17, which relative to the average $SPNP$ is an improvement of 30%. However, the range between the instances is broad. For the instances with six surgeries,

planning with phases may reduce the total waiting time by 152 minutes on average, or the total idle time by 76 minutes, while for four surgeries, the total waiting time will only be reduced by 10 minutes. We conclude that the inclusion of phases will, for some instances, provide valuable information that can improve the schedule. For other instances on the other hand, it provides little additional value.

### 7.7.4 Value of the Emergency Model

Average values of the $EVPI$ and the $VSS$ for each size of problem instances in the Emergency Model are stated in Table 7.15. The model is run to optimality for all test instances. Since the surgery durations do not depend on emergency arrivals, the same test instances are used for the calculation of the $EVPE_i$, where the uncertainty in emergency arrivals has been removed. The first-stage solution from these tests are then used on the complete scenario tree to obtain the objective value of the $SPEP_i$.

In contrast to the tests of the Phase Model, all instances will suffer from overtime as the sum of durations in the majority of the scenarios are higher than $d$. For the instances with five elective surgeries, all scenarios suffer from overtime, and this is the case for some of the instances with four elective patients. The increased cost as a result of overtime is obvious from the values of the $WS$, which are larger than zero.

Table 7.15: Average $EVPI$ and $VSS$ for the all problem instances in the Emergency Model

| **Instance** | **SP** | **WS** | $\mathbf{EDEV_T}$ | **EVPI** | $\mathbf{VSS^D}$ |
|---|---|---|---|---|---|
| Instance 1E | 558.1 | 329.8 | 870.8 | 228.3 | 312.7 |
| Instance 2E | 211.9 | 64.3 | 258.1 | 147.5 | 46.3 |
| Instance 3E | 241.9 | 124.3 | 398.2 | 117.5 | 156.4 |

Both the $EVPI$ and the $VSS$ increase when the amount of expected overtime increases in our test instances. As we only test instances with four and five elective surgeries, it is difficult to determine whether this is valid in general. It does, however, comply with the reasoning for the Phase Model, that the value of information increases with the number of surgeries. The Emergency Model reduces the cost of the expected value solution by 38% on average. The average values of the $EVPI$ and the $VSS$ are 164 and 153, respectively. Stated in terms of operational time, solving the problem stochastically may reduce the average total waiting time by 153 minutes or the idle time by 76 minutes from a deterministic schedule.

Table 7.16 shows the results of the $SPEP_i$ and the $EVPE_i$. The $SPEP_i$ can be interpreted as the value of a deterministic strategy which schedules $i$ emergency patients and $n$ elective patients, as if there were $n + i$ elective patients. Deterministic in this context refers to a deterministic number of patients. The deviation between the $SP$ and the $EVPE_0$ is 0.6% on average. This shows that the current strategy at St. Olavs Hospital, that assumes a deterministic arrival of zero emergency patients, will give approximately the

same results as planning for the uncertainty in patient arrival. The other two strategies, which deterministically schedule $n+1$ and $n+2$ patients, perform significantly worse than the first strategy, in this stochastic environment. This shows that with the probabilities of patient arrival used in our model, the current scheduling strategy used at St. Olavs Hospital, which does not consider potential emergency patients, may be appropriate.

Table 7.16: $SPEP_i$ and $EVPE_i$ for all emergency instances

| **Instance** | **SP** | $\mathbf{SPEP_0}$ | $\mathbf{SPEP_1}$ | $\mathbf{SPEP_2}$ | $\mathbf{EVPE_0}$ | $\mathbf{EVPE_1}$ | $\mathbf{EVPE_2}$ |
|---|---|---|---|---|---|---|---|
| Instance 1E | 558.1 | 565.4 | 788.3 | 995.8 | 7.3 | 230.2 | 437.7 |
| Instance 2E | 211.9 | 212.5 | 284.0 | 505.0 | 0.6 | 72.2 | 293.1 |
| Instance 3E | 241.9 | 244.0 | 493.2 | 799.0 | 2.1 | 251.3 | 557.1 |

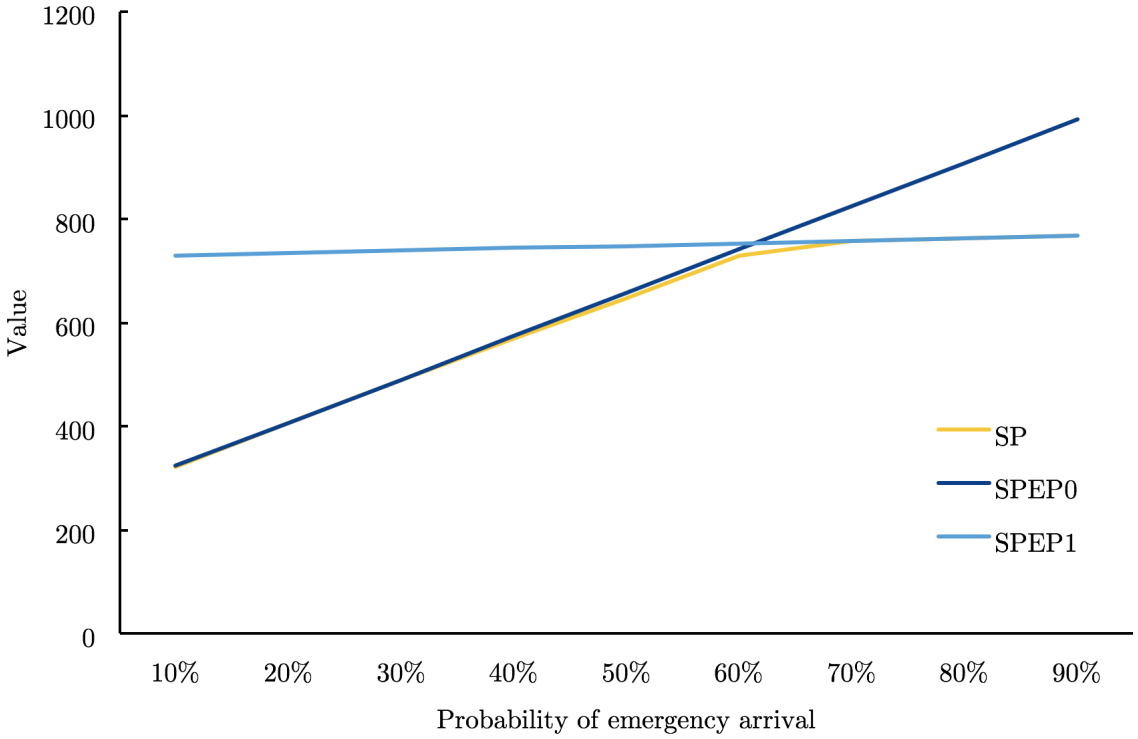
The probability of emergency patient arrivals is estimated over a long period of time. This may change during certain seasons and holidays, as mentioned in Chapter2. To investigate the impact on the scheduling strategy when the probability of emergency arrival changes, we run the same tests with varying probabilities. These are run on instances with the same elective patients, but with uncertain arrival of one emergency patient with probability ranging from 10% to 90%. The results are displayed in Figure 7.3. The figure shows that when the probability is below 62%, our stochastic model performs similarly to the strategy that schedules zero emergency patients. However, when the probability of a emergency patient arrival exceeds 62%, the optimal strategy changes. For these cases, our model performs similarly to a strategy that schedules $n+1$ elective patients. In other words, the model makes schedules close to the extreme cases of exactly zero or one emergency patients. With knowledge of this percentage, it may not be necessary to incorporate the uncertain emergency arrival in a stochastic model.

The conclusion is that the choice of optimal scheduling strategy, i.e. the number of patients to schedule, is not trivial. This depends both on the probabilities of arrival and on the characteristics of the surgery durations. This is complicated further as the number of possible emergency patients increases. The stochastic model and information about emergency arrivals are still necessary to determine which of the possible deterministic strategies is the best. Even if the probabilities of emergency patients are unchanged, we have shown that a slightly better solution can be found by using the Emergency Model.

## 7.8 Practical analysis

In this section, we will evaluate the potential effects of using optimisation techniques for St. Olavs Hospital. We will discuss the trade-offs made by the optimisation program and how it can improve the scheduling strategy at St. Olavs Hospital. For the Phase Model we will compare our solutions with actual schedules, and we will try to extract a decision rule that can be implemented on the basis of statistical data analysis. Comparisons with actual schedules are difficult to make for the Emergency Model, and will be based on realistic,

Figure 7.3: $SP$, $SPEP_0$ and $SPEP_1$ for different probabilities of the arrival one emergency arrival




but constructed schedules.

### 7.8.1 Phase Model

To evaluate the potential gain of new scheduling strategies at St. Olavs Hospital, we compare the performance of the schedules made by St. Olavs Hospital for our data instances, with the performance of the schedules made by our optimisation program. This means that the surgery sequence and start time of the schedule from St. Olavs Hospital has been evaluated in our scenario tree, which is assumed to be a good representation of reality. Because we do not have data to justify a start time different from zero for the first surgery, all surgery start times are adjusted back so that the first surgery starts at time zero for the scheduled made by St. Olavs Hospital. The planners at St. Olavs Hospital also take many practical aspects into consideration when determining the order of surgeries. To incorporate some of these, we also evaluate the performance of an optimised schedule that use the same order of surgeries as planned by St. Olavs Hospital, but can change the start times as long as the plan comply with this order. The results of these tests are shown in Table 7.17. The table shows the decomposition of the objective value into waiting time, idle time and overtime for every data instance. The three columns for each performance measure indicates the performance of the plan made by St. Olavs Hospital, by fixing only the sequence and a fully optimised schedule, respectively.

Table 7.17: Performance of schedule planned by St. Olavs Hospital compared to the performance of the optimised start times, and fully optimised schedules

| Instance number | Total wait time | | | Total idle time | | | Total overtime | | |
|---|---|---|---|---|---|---|---|---|---|
| | StOlav | Half | Full | StOlav | Half | Full | StOlav | Half | Full |
| 1 | 110 | 19 | 17 | 49 | 6 | 4 | 1 | 0 | 0 |
| 2 | 74 | 13 | 12 | 19 | 3 | 4 | 0 | 0 | 0 |
| 3 | 212 | 15 | 18 | 0 | 10 | 6 | 0 | 0 | 0 |
| 4 | 124 | 20 | 20 | 7 | 3 | 3 | 0 | 0 | 0 |
| 5 | 170 | 34 | 25 | 0 | 9 | 6 | 0 | 0 | 0 |
| 6 | 153 | 21 | 21 | 0 | 10 | 9 | 1 | 0 | 0 |
| 7 | 323 | 22 | 17 | 0 | 10 | 7 | 0 | 0 | 0 |
| 8 | 62 | 42 | 18 | 20 | 10 | 3 | 0 | 0 | 0 |
| 9 | 38 | 37 | 30 | 21 | 7 | 8 | 0 | 0 | 0 |
| 10 | 465 | 50 | 23 | 0 | 12 | 6 | 0 | 0 | 0 |
| 11 | 499 | 29 | 24 | 0 | 22 | 15 | 0 | 0 | 0 |
| 12 | 55 | 48 | 32 | 130 | 6 | 7 | 181 | 18 | 7 |
| 13 | 80 | 14 | 12 | 28 | 3 | 3 | 0 | 0 | 0 |
| 14 | 21 | 39 | 25 | 113 | 8 | 6 | 116 | 4 | 4 |
| 15 | 800 | 108 | 73 | 0 | 15 | 4 | 107 | 121 | 98 |
| 16 | 195 | 20 | 19 | 0 | 5 | 4 | 0 | 0 | 0 |
| 17 | 98 | 23 | 18 | 67 | 5 | 5 | 29 | 0 | 0 |
| 18 | 57 | 35 | 16 | 58 | 5 | 5 | 0 | 0 | 0 |
| 19 | 95 | 34 | 15 | 6 | 11 | 6 | 2 | 0 | 0 |
| 20 | 23 | 23 | 16 | 35 | 7 | 5 | 0 | 0 | 0 |
| Avg. | 182.7 | 32.3 | 22.6 | 27.7 | 8.4 | 5.8 | 21.9 | 7.2 | 5.5 |

From the results, the first observation that can be made, is the major improvement of the average amount of waiting time, idle time and overtime. This is true for all the three measures of schedule quality. The improvement can be decomposed into two steps: improvements from optimising the start times given a fixed sequence of surgeries, and improvements from optimising the start times and sequence. The first of these two is, by far, the most significant. That is, St. Olavs Hospital has a huge improvement potential even by only optimising the intervals of surgeries. This means that considerations regarding the position of a given surgery (see Other considerations in Section 2.2.3), which have so far been excluded, can be adhered to because the positions remain as they were in the actual schedules. If there are no such considerations, meaning there are no strong arguments for keeping the sequence, the gains from optimising the sequence can be realised as well.

The weight combination we have used for the cost of waiting time, idle time and overtime does not necessarily reflect St. Olavs Hospital's considerations completely. However, in 11 of 20 data instances, the solution from the optimisation program is strictly better than the schedule made by St. Olavs Hospital, in the sense that all three performance measures are equal or better. This means that, regardless of the preferred weight combination, the optimised schedule will perform better. The improvements of the remaining 9 data instances, where there is one measure that is worse off in the optimised solution, are

Table 7.18: Difference in performance between the schedule made by St. Olavs Hospital and the optimised schedule

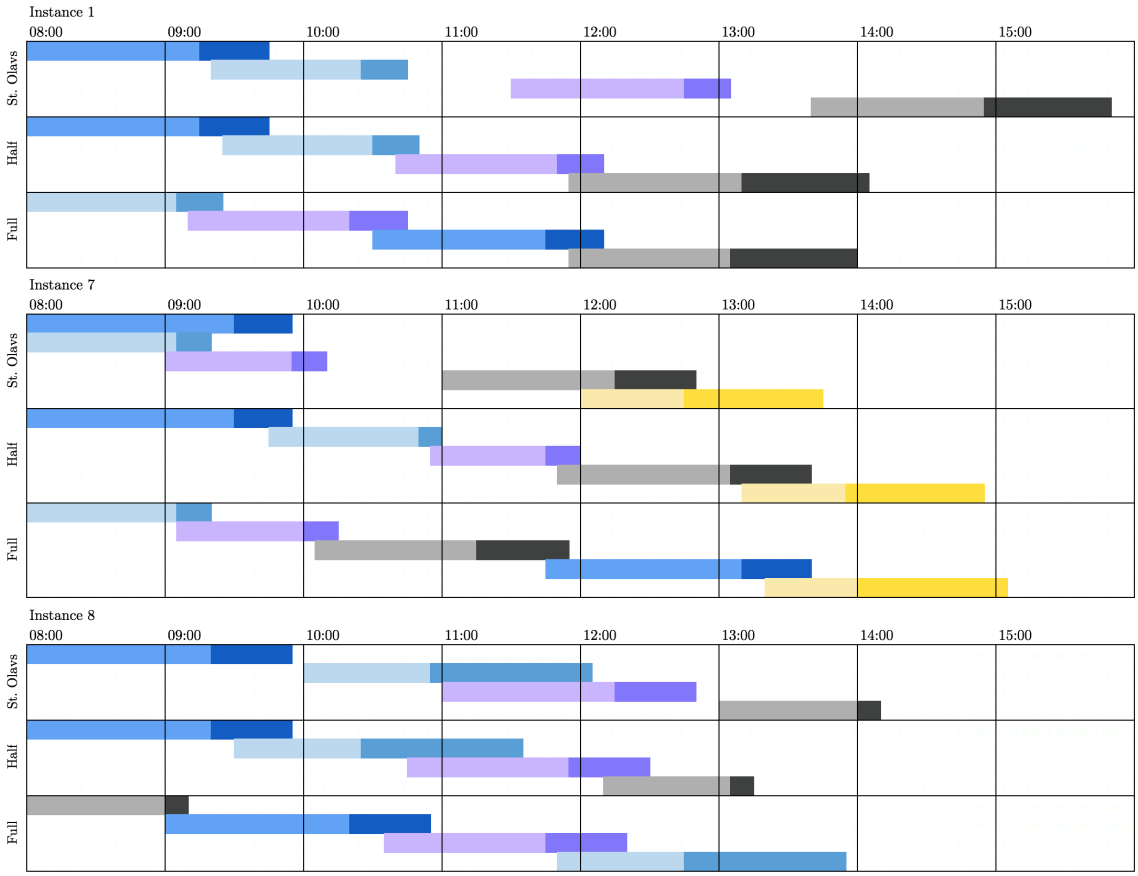
| **Instance** | **Waiting time** | **Idle time** | **Overtime** | **Trade off** |
|---|---|---|---|---|
| Instance 3 | 194 | -6 | 0 | 32 wait/idle |
| Instance 5 | 145 | -6 | 0 | 24 wait/idle |
| Instance 6 | 132 | -9 | 1 | 15 wait/idle |
| Instance 7 | 306 | -7 | 0 | 44 wait/idle |
| Instance 10 | 442 | -6 | 0 | 74 wait/idle |
| Instance 11 | 475 | -15 | 0 | 32 wait/idle |
| Instance 14 | -4 | 107 | 112 | 27 idle/wait, 28 over/wait |
| Instance 15 | 727 | -4 | 9 | 182 wait/idle |
| Instance 16 | 176 | -4 | 0 | 44 wait/idle |

more debatable. The difference in each measure is shown in Table 7.18 for each of these instances. The last column shows the ratio between the improving measure and the worsening measure. Even though the trade-offs depend on the weights of each measure, one can argue that reducing the amount of waiting time by 132 minutes at the expense of 9 minutes of idle time is a reasonable trade-off. This is even the trade-off with the lowest ratio. The best trade-off reduces total waiting time by 727 minutes at a cost of 4 minutes of idle time, meaning a reduction of 182 minutes of waiting time per extra minute of idle time. Most of the improvements come from reductions in waiting time. This may indicate that St. Olavs Hospital associates a lower cost with waiting time than our cost combination. However, in instance 14, there is clearly scheduled too much idle time, which results in excessive overtime. This can be almost completely avoided by careful scheduling, as shown in Table 7.18 for the row of instance 14. Based on the large improvements relative to the the measures that are worsened, we argue that the optimised schedule performs significantly better than the plan made at St. Olavs Hosptial for all problem instances evaluated.

For instances 1, 7, and 8, Figure 7.4 visualises the results of the three different scheduling strategies discussed. The first is the schedule made by St. Olavs Hospital, the second is the result from only optimising start times, given the sequence from the first, and the third one is the result from the full optimisation. Each colour indicates a different surgery, to show how they are rearranged. The darker part of each bar indicates the 90% confidence interval of the duration of the surgery. The start of the dark part shows the lowest possible duration in the confidence interval, while its end shows the highest possible. The average surgery duration will therefore be located somewhere in the dark shade.

An interesting observation that can be made is that, for several of the schedules made by St. Olavs Hospital, two surgeries are scheduled to start at time zero. This is to ensure that either of the surgeries start at the beginning of the day to not waste time. After a couple of surgeries, there is scheduled a large slack to absorb some of the delays this leads

Figure 7.4: Comparison of the schedule made by St. Olavs Hospital, the schedule with optimised start times, and the schedule from the full optimisation



to. Obviously, this kind of scheduling strategy will impose a lot of waiting in the schedule, as shown in Table 7.18. This scheduling strategy is similar to the Bailey-Welch decision rule we evaluated in Section 7.6, which in general gave rather poor performance. Thus, the main improvement in these cases comes from reducing waiting time.

By inspecting the mean surgery durations (not shown in the figure), we also observe that many of the intervals from one surgery start time to the next coincide with the mean surgery duration of the surgery. This is interesting when trying extract a decision rule, and a closer analysis supports this finding. For all the data instances, we calculated the average absolute difference between the scheduled start time intervals and the mean surgery duration in each interval. These distances range from a low of 0.5 to 5.7, with an average value of 2.6 minutes. Further, when the full problem is optimised, the visualisations of the schedules show a tendency of setting the surgeries with the higher variance (larger confidence intervals) later in the day. This also complies with the good performance we saw with the decision rule that sorts the surgeries in order of increasing variance in Section 7.6. The order does not follow this rule strictly, but a linear regression of the variances of the surgery durations reveals a clear positive slope for all the regression lines, further supporting the observation. Based on these discoveries, we suggest a final decision rule, where the surgeries are scheduled in order of increasing variance, with start time intervals equal to the mean duration of each surgery.

The final results we present for the Phase Model, are the performance of the proposed variance sort-mean interval decision rule, referred to as the var-mean rule. An advantage of this decision rule is that it determines both the sequence and start times. Thus, it can be used without any interaction with a complete optimisation program. Because the decision rule base the surgery order on the variance, we have included the sort by variance decision rule for comparison in addition to the optimal stochastic solution. These are shown in Table 7.19. Note that the values of the optimal stochastic program objective and the sort by variance objective are repeated from Table 7.12.

The results show that the sort by variance decision rule is better the var-mean decision rule for every instance. This is obvious and will always be the case, because both fix the sequence the same way, bu only the former has the flexibility of optimising the start times. Compared to the optimal solution, the performance of the var-mean rule varies from a 1% to 30% optimality gap.

Similar to the sort by variance decision rule, the var-mean rule gives a good performance in most cases. Therefore, it might function well as an easy rule of thumb. However, as with the optimisation program, its performance relies on an extensive data analysis. In addition, this decision rule was derived by studying the resulting schedules from the optimisation program. It is therefore sensitive to the cost combination we use. For example in the case where the cost of waiting time is evaluated to close to zero, the optimal solution would be to schedule all patients to arrive at time zero, such that there is always a patient ready

Table 7.19: The objective values of the var-mean decision rule, the stochastic program, and the sort by variance decision rule

| **Instance** | **Stochastic program** | **Sort by variance** | **Var-mean** |
|---|---|---|---|
| Instance 1 | 25.58 | 25.58 | 25.95 |
| Instance 2 | 28.52 | 28.89 | 30.63 |
| Instance 3 | 38.88 | 39.32 | 44.68 |
| Instance 4 | 22.12 | 22.12 | 24.53 |
| Instance 5 | 37.48 | 37.80 | 39.53 |
| Instance 6 | 39.15 | 41.45 | 43.71 |
| Instance 7 | 31.76 | 38.29 | 41.28 |
| Instance 8 | 24.54 | 24.54 | 27.99 |
| Instance 9 | 45.69 | 47.21 | 49.79 |
| Instance 10 | 41.69 | 41.72 | 42.10 |
| Instance 11 | 53.91 | 58.01 | 59.71 |
| Instance 12 | 69.89 | 81.54 | 87.06 |
| Instance 13 | 17.42 | 17.42 | 17.81 |
| Instance 14 | 48.37 | 55.20 | 55.91 |
| Instance 15 | 381.92 | 418.08 | 443.07 |
| Instance 16 | 26.68 | 27.61 | 28.23 |
| Instance 17 | 31.85 | 32.49 | 33.44 |
| Instance 18 | 26.17 | 26.17 | 26.40 |
| Instance 19 | 26.48 | 27.71 | 30.06 |
| Instance 20 | 25.26 | 25.26 | 25.51 |
| Average | 52.17 | 55.82 | 58.87 |

when a surgery finishes.

A first step to taking advantage of the potential scheduling improvements can be to implement the suggested var-mean decision rule. However, this will neither capture the benefit of planning with phases nor from fully taking the uncertainty into account. To realise the full potential, one can use the optimisation techniques discussed in this paper.

### 7.8.2 Emergency Model

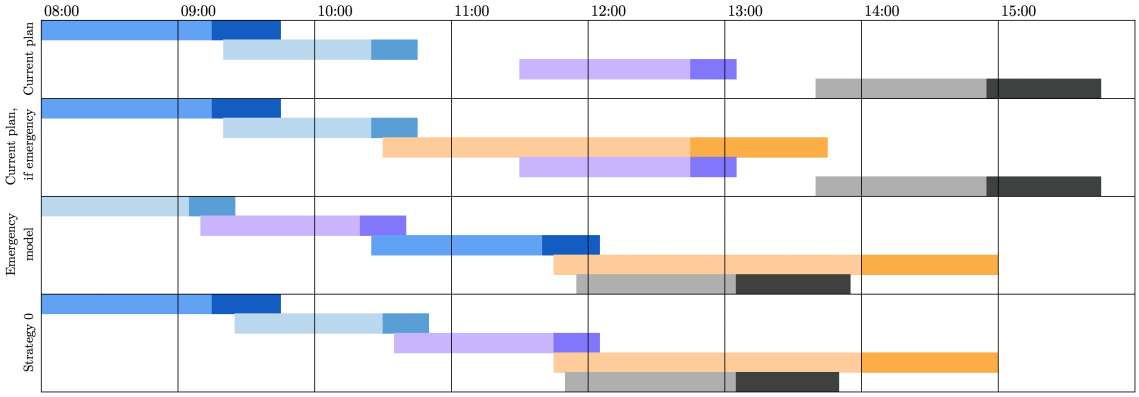
Similar analyses are conducted for the Emergency Model. Since the planners often defer elective surgeries when emergency patients arrive, information about the planned schedule before patient deferrals proved difficult to retrieve. We are therefore not able to compare these results with actual schedules at the hospital. As the patient coordinators claim to use the same scheduling strategy, independently of the potential emergency arrival, we have used the same elective patients as for the Phase Model to be to perform a comparison. These instances are regenerated with 60 scenarios, no phases and the appropriate moments. The emergency patients, their respective arrival probabilities and the uncertain surgery durations remain the same as for the emergency instances used in previous analysis.

To conduct this analysis, the elective start times from the planned schedules have been fixed in the Emergency Model, to find the position for the emergency patients. This has the advantage of optimising the positions of the potential emergency patients and may thus be a better solution than what would have been scheduled at St. Olavs Hospital. This schedule is compared to the optimal schedule from the Emergency Model. Lastly, we have included the described strategy that schedules $n+0$ elective patients (Strategy 0) and then handles the emergency cases on demand.

Figure 7.5 shows the four described schedules for one of the constructed test instances. The visual representation follows the same structure as for the Phase Model. Specific to this figure, is the emergency surgery, indicated by the orange colour. The first schedule in the figure shows the original plan made by St. Olavs Hospital, while the second includes the optimal positioning of the emergency patient in this schedule. The third schedule is the solution from the Emergency Model. The last is, as described, the resulting schedule for using the two-stage Strategy 0. Because all instances provided similar results, only the results from this instance is displayed.

The trade-offs that are made between waiting, idling and overtime in the Emergency Model are similar to those of the Phase Model, and will not be restated. However, we will discuss how the emergency arrival affect the schedules. The Emergency Model improves the planned schedule by reducing waiting time and idle time by 75 and 40 minutes, respectively, increasing overtime by 6 minutes for this instance on average. This is the difference between schedules 2 and 3 in Figure 7.5. The most important insight, however, comes from the comparison between schedules 3 and 4. The objective values for these two are

Figure 7.5: Comparison of the schedule planned by St. Olavs Hospital, the schedule with optimised start times, the schedule from full optimisation, and the schedule from using Strategy 0




about the same, as discussed in Section 7.7.4 regarding the value of the Emergency Model. The exact numbers are not shown in the figure, but one can see that the spread and the potential overtime are very similar. This means most of the potential gain can be realised by using Strategy 0, which is close to the current scheduling strategy at St. Olavs Hospital. Note, however, that this is only regarding the number of patients, and there is still much to gain by using stochastic planning for the surgery durations. The small differences in the schedules, such as the minor displacement of the last elective surgery and the reordering of the sequence will still provide further improvements, which can be harvested by using the complete Emergency Model.

The arrival of additional patients seems to be a larger disruption than can be adjusted for by rescheduling the elective surgeries. The consequence is that it is sufficient to plan for an exact number of arrivals, instead of allocating flexibility to absorb it. We claim this indicates that the emergency surgeries should be handled at an earlier decision level, for example in the advance scheduling problem. However, given that this is not taken into account, the Emergency Model can give the exact number of arrivals one should anticipate.

## 7.9 Further discussion and future research

In this section we will discuss the impact of the main assumptions explained in Chapter 4, and possible sources of error. Many of these are interesting to consider in future work, which will be discussed in the end of this section.

The assumption of all patients arriving exactly in time for their scheduled surgery start can be expected to have two different effects on our plan. First of all, disregarding the aspect of no-shows and late arrivals implies that, in reality, our planning strategy is likely to lead to more idle time than what our results suggest. However, since the assumption also removes the possibility of patients arriving early and the option of calling patients to

make them do so, some idle time could most likely be eliminated by the hospital in a real situation. The overall effect of these two arguments is therefore ambiguous.

The assumption of all other resources being available, on the other hand, removes potential bottlenecks that, in reality, could have led to waiting time, idle time and overtime. This simplification will therefore pull in the direction of our results giving an optimistic estimate.

Further, the assumption requiring all surgeries to be performed on the day they are scheduled, is a restriction because it deprives the planners of the option of postponing surgeries to another day when this is considered desirable. The effect of this assumption on the objective value will therefore be conservative.

Lastly, there is an implicit assumption that the scenarios we generate are a good representation of the reality. This is a central assumption in all stochastic models, and a potential source of error. If this is not true, the comparisons between the schedules from the optimisation program and the schedules made by St. Olavs Hospital are not meaningful. This is because the optimised schedules are both created and evaluated based on scenario trees of the same structure (though not the same scenario trees), while the schedules from St. Olavs Hospital are only evaluated on these scenarios. If the scenarios are completely unrealistic, the optimised schedules can seem to perform well, and wrongfully indicate large improvements from the schedules made by St. Olavs Hospital, even though this will not be the case in reality.

A potential source of errors is the low number of test instances we use. This arise mainly due to the vast number of possible combinations of surgeries and the strict statistical criteria we demand of the data. The implication of this is that the average improvements we report may not be scalable. Also, there is no guarantee that the results for the Emergency Model is completely transferable to reality. Some seemingly inconsistent results might have shown more consistency if we had more instances to test. The data instances selected for testing the Phase Model include surgeries that are expected to have a significant difference in the probability distribution for the two phases. Therefore, the potential benefit of planning with phases presented, should be considered optimistic. The conclusion is, nevertheless, that for certain surgeries and surgeon combinations, there are clear advantages to differentiating between the two phases.

The two models we propose manage to solve most of the data instances to optimality in reasonable time. However, a few of the largest data sets for the Phase Model experience some optimality gap when allowed to run for 3 600 seconds. The number of surgeries on a given day has a practical limitation at St. Olavs Hospital and the solution time is not an issue. If the model formulations should be generalised to be suitable for larger problem instances, or problems with similar structure in other contexts, exact solution methods are important challenges to explore for future research. If more extensions of the models need larger scenario trees to ensure stability, decomposition approaches for these integer

stochastic programs may also be relevant for future work. Lastly, we make assumptions regarding the decision dependent uncertainty that appears in the Phase Model. To the best of our knowledge, no literature exists that solves similarly structured problems with a continuous transition of probability distributions, i.e. using uncertainty with a continuous relationship with some of the decision variables. This will remove some of the mathematical oddities that occur because there is a discrete change of probability distribution at the point that separates the phases, but will at the same time require even more complex data analyses to model the uncertainty. This is the last major research area we want to point future work towards.

# Chapter 8

# Conclusion

Surgery scheduling is a complex and often manual process which, for the case hospital, is largely determined by rules of thumb and tacit knowledge held by individual coordinators. The uncertain environment of hospitals makes surgery schedules prone to disruptions, and with surgeries representing an important part of a hospital's activities, disruptions are costly and can lead to major dissatisfaction among patients and staff.

The objective of this thesis has thus been to provide insights to how uncertainty should be accounted for when scheduling surgeries, to minimise the consequence of waiting time, idle time and overtime. The main problem considered has been how to schedule a set of surgeries with uncertain duration at a given operating room on a given day. We have applied operations research and formulated two separate stochastic mathematical models that solve two variations of this problem, where the uncertainty is represented by scenarios generated through moment-matching. The problem addressed by the first model was motivated by a hypothesis that the uncertainty of a surgery's duration may depend on its start time, which complicates the mathematical formulation by introducing the aspect of decision-dependent uncertainty. The second model solves a multi-stage stochastic mathematical problem, incorporating uncertain arrival of emergency patients in addition to the uncertain surgery durations, a combination that is yet to receive much attention in literature.

We have argued that the duration of a surgical procedure may depend on the surgeon performing it, and generated scenarios using the moments of the empirical durations for specific combinations of procedure and surgeon. In addition, the data has been adjusted for the trend of surgeon efficiency to increase with experience. This has enabled us to make more accurate estimates of surgery durations, meaning that the value of the solutions we have presented is not only due to solving the problem stochastically but may be, partly, due to improved estimates.

Various attempted decompositions proved unsuccessful due to the aspect of decision-dependent uncertainty and because our formulations require integer variables in all stages. We have, however, strengthened the formulations by adding several valid inequalities, and guided the solution process through branching and different heuristics, significantly enhancing the computational performance. The valid inequalities were very effective for the first model we proposed. For the largest problem instances, that reached a maximum execution time of 1 800 seconds, the average gap was reduced by 42% after adding all five valid inequalities. For the instances that reached optimality, we showed that the most effective cut proved to be one strengthening the link between two of the sequencing variables, reducing the average run-time by 39% across all instances.

Through a practical computational study, we have combined common approaches from the literature with conclusions drawn from the results of our optimisation programs. This has enabled us to extract a simple but effective decision rule that can be implemented in elective surgery scheduling without use of optimisation. The rule implies sequencing the surgeries by ascending variance, and setting start times with intervals equal to each surgery's mean duration. Attaining the full potential we have identified requires the application of stochastic programming, but most of it can be captured using the suggested decision rule. For the Emergency Model, results indicate that the optimal scheduling strategy, when considering potential emergency patients, depends both on the probabilities of arrival and on the characteristics of the surgery durations. The model makes schedules that resemble schedules resulting from deterministically planning for an integer number of emergency patients. Since this, essentially, equals planning for additional elective patients, the suggested decision rule applies to this problem as well.

Moreover, we have shown that, for some surgery types, the aspect of surgery durations depending on their start time can in fact be worth accounting for. This was shown for the surgery types whose statistical properties were, according to a two-sample Kolmogorov-Smirnov test of independence, most likely to differ based on the time of day. Hence, the effects will in general not be prevalent for all surgeries. As for all decision support, the potential gain from using the proposed Phase Model should therefore be traded off versus the cost of implementing the consideration.

The overall results from our stochastic models suggest that for the test instances, St. Olavs Hospital could have reduced the average total waiting time, idle time and overtime by 160, 22, and 16 minutes per day, respectively. Note that since all results are based on a relatively limited sample of instances, extrapolating the potential results to a larger scale should be done with care. Even so, the practical conclusion of this thesis is that there are significant potential gains from applying operations research to handle the uncertainty in surgery scheduling.

# References


[1] Atle Riise and Edmund K Burke. Local search for the surgery admission planning problem. *Journal of Heuristics*, 17(4):389–414, 2011.

[2] Helsedirektoratet. Innsatsstyrt finansiering 2015. *Innsatsstyrt finansiering*, 2015.

[3] Daiki Min and Yuehwern Yih. An elective surgery scheduling problem considering patient priority. *Computers & Operations Research*, 37(6):1091–1099, 2010.

[4] St. Olavs Hospital, Trondheim University Hospital. Retrieved from: *https://stolav.no/st-olavs-hospital-trondheim-university-hospital,*.

[5] Store Norske Leksikon. Retrieved from: *https://snl.no/ortopedisk_kirurgi*.

[6] St. Olavs Hospital, Trondheim University Hospital. Retrieved from: *https://stolav.no/avdelinger/klinikk-for-ortopedi-revmatologi-og-hudsykdommer/ortopedisk-avdeling-oya*.

[7] Lawrence W Robinson and Rachel R Chen. Scheduling doctors' appointments: optimal and empirically-based heuristic policies. *Iie Transactions*, 35(3):295–307, 2003.

[8] Julia L Higle and Stein W Wallace. Sensitivity analysis and uncertainty in linear programming. *Interfaces*, 33(4):53–60, 2003.

[9] Alan J King and Stein W Wallace. *Modeling with stochastic programming*. Springer Science & Business Media, 2012.

[10] Julia L Higle. Stochastic programming: optimization when uncertainty matters. *Cole Smith J (ed) Tutorials in operations research*, pages 30–53, 2005.

[11] Lars Hellemo, Paul I Barton, and Asgeir Tomasgard. Stochastic programming with decision-ependent probabilities. *Unpublished*, 2015.

[12] Dinh-Nguyen Pham and Andreas Klinkert. Surgical case scheduling as a generalized job shop scheduling problem. *European Journal of Operational Research*, 185(3):1011–1025, 2008.

[13] John R Charnetski. Scheduling operating room surgical procedures with early and late completion penalty costs. *Journal of Operations Management*, 5(1):91–102, 1984.

[14] Sakine Batun, Brian T Denton, Todd R Huschka, and Andrew J Schaefer. Operating room pooling and parallel surgery processing under uncertainty. *INFORMS journal on Computing*, 23(2):220–237, 2011.

[15] Patrick P Wang. Static and dynamic scheduling of customer arrivals to a single-server system. *Naval Research Logistics (NRL)*, 40(3):345–360, 1993.

[16] Patrick P Wang. Optimally scheduling N customer arrival times for a single-server system. *Computers & Operations Research*, 24(8):703–716, 1997.

[17] Camilo Mancilla and Robert Storer. A sample average approximation approach to stochastic appointment sequencing and scheduling. *IIE Transactions*, 44(8):655–670, 2012.

[18] Brian Denton and Diwakar Gupta. A sequential bounding approach for optimal appointment scheduling. *Iie Transactions*, 35(11):1003–1016, 2003.

[19] Brian Denton, James Viapiano, and Andrea Vogl. Optimization of surgery sequencing and scheduling decisions under uncertainty. *Health care management science*, 10(1):13–24, 2007.

[20] Ho-Yin Mak, Ying Rong, and Jiawei Zhang. Appointment scheduling with limited distributional information. *Management Science*, 61(2):316–334, 2014.

[21] Norman TJ Bailey. A study of queues and appointment systems in hospital outpatient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 185–199, 1952.

[22] JD Welch and Norman TJ Bailey. Appointment systems in hospital outpatient departments. *The Lancet*, 259(6718):1105–1108, 1952.

[23] Chrwan-Jyh Ho and Hon-Shiang Lau. Minimizing total cost in scheduling outpatient appointments. *Management science*, 38(12):1750–1764, 1992.

[24] Elliott N Weiss. Models for determining estimated start times and case orderings in hospital operating rooms. *IIE transactions*, 22(2):143–150, 1990.

[25] Eric Marcon and Franklin Dexter. Impact of surgical sequencing on post anesthesia care unit staffing. *Health Care Management Science*, 9(1):87–98, 2006.

[26] Sabine Sickinger and Rainer Kolisch. The performance of a generalized bailey–welch rule for outpatient appointment scheduling under inpatient and emergency demand. *Health care management science*, 12(4):408–419, 2009.

[27] Peter M Vanden Bosch and Dennis C Dietz. Minimizing expected waiting in a medical appointment system. *Iie Transactions*, 32(9):841–848, 2000.

[28] Peter M Bosch. Scheduling and sequencing arrivals to a stochastic service system. Technical report, DTIC Document, 1997.

[29] Guido C Kaandorp and Ger Koole. Optimal outpatient appointment scheduling. *Health Care Management Science*, 10(3):217–229, 2007.

[30] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–193, 1906.

[31] Gilbert Laporte, François V Louveaux, and Luc van Hamme. Exact solution to a location problem with stochastic demands. *Transportation Science*, 28(2):95–103, 1994.

[32] David Applegate and William Cook. A computational study of the job-shop scheduling problem. *ORSA Journal on computing*, 3(2):149–156, 1991.

[33] Egon Balas. *On the facial structure of scheduling polyhedra*. Springer, 1985.

[34] Martin E Dyer and Laurence A Wolsey. Formulating the single machine sequencing problem with release dates as a mixed integer program. *Discrete Applied Mathematics*, 26(2):255–270, 1990.

[35] Robert Johannes Maria Vaessens, Emile HL Aarts, and Jan Karel Lenstra. Job shop scheduling by local search. *INFORMS Journal on Computing*, 8(3):302–317, 1996.

[36] Brecht Cardoen, Erik Demeulemeester, and Jeroen Beliën. Sequencing surgical cases in a day-care environment: an exact branch-and-price approach. *Computers & Operations Research*, 36(9):2660–2669, 2009.

[37] Irem Ozkarahan. Allocation of surgical procedures to operating rooms. *Journal of Medical Systems*, 19(4):333–352, 1995.

[38] Tugba Cayirli and Emre Veral. Outpatient scheduling in health care: a review of literature. *Production and Operations Management*, 12(4):519–549, 2003.

[39] Brant E Fries and Vijay P Marathe. Determination of optimal variable-sized multiple-block appointment systems. *Operations Research*, 29(2):324–345, 1981.

[40] Qingxia Kong, Chung-Yee Lee, Chung-Piaw Teo, and Zhichao Zheng. Scheduling arrivals to a stochastic service delivery system using copositive cones. *Operations research*, 61(3):711–726, 2013.

[41] Tugba Cayirli, Emre Veral, and Harry Rosen. Designing appointment scheduling systems for ambulatory care services. *Health care management science*, 9(1):47–58, 2006.

[42] Tugba Cayirli, Emre Veral, and Harry Rosen. Assessment of patient classification in appointment system design. *Production and Operations Management*, 17(3):338–353, 2008.

[43] Christos Zacharias and Michael Pinedo. Appointment scheduling with no-shows and overbooking. *Production and Operations Management*, 23(5):788–801, 2014.

[44] TF Keller and DJ Laughhunn. An application of queuing theory to a congestion problem in an outpatient clinic. *Decision Sciences*, 4(3):379–394, 1973.

[45] Kenneth J Klassen and Thomas R Rohleder. Scheduling outpatient appointments in a dynamic environment. *Journal of operations Management*, 14(2):83–101, 1996.

[46] James R Swisher, Sheldon H Jacobson, J Brian Jun, and Osman Balci. Modeling and analyzing a physician clinic environment using discrete-event (visual) simulation. *Computers & operations research*, 28(2):105–125, 2001.

[47] Diwakar Gupta and Brian Denton. Appointment scheduling in health care: Challenges and opportunities. *IIE transactions*, 40(9):800–819, 2008.

[48] Jerrold H May, David P Strum, and Luis G Vargas. Fitting the lognormal distribution to surgical procedure times. *Decision Sciences*, 31(1):129–148, 2000.

[49] Alfonso Soriano. Comparison of two scheduling systems. *Operations Research*, 14(3):388–397, 1966.

[50] Kum Khiong Yang, Mun Ling Lau, and Ser Aik Quek. A new appointment rule for a single-server, multiple-customer service system. *Naval Research Logistics (NRL)*, 45(3):313–326, 1998.

[51] Chrwan-Jyh Ho and Hon-Shiang Lau. Evaluating the impact of operating conditions on the performance of appointment scheduling rules in service systems. *European Journal of Operational Research*, 112(3):542–553, 1999.

[52] Robert M O'Keefe. Investigating outpatient departments: implementable policies and qualitative approaches. *Journal of the Operational Research Society*, pages 705–712, 1985.

[53] Walton M Hancock, Paul F Walter, Roy A More, and Noah D Glick. Operating room scheduling data base analysis for scheduling. *Journal of medical systems*, 12(6):397–409, 1988.

[54] David J Robb and Edward A Silver. Scheduling in a management context: Uncertain processing times and non-regular performance measures*. *Decision Sciences*, 24(6):1085–1108, 1993.

[55] Birger Jansson. Choosing a good appointment system—a study of queues of the type (d, m, 1). *Operations Research*, 14(2):292–312, 1966.

[56] Chtng-Jong Liao, C Dennis Pegden, and Matthew Rosenshine. Planning timely arrivals to a stochastic production or service system. *IIE transactions*, 25(5):63–73, 1993.

[57] John R Simeoni. An efficient approach to solving the optimal control of arrivals problem. Technical report, DTIC Document, 1994.

[58] Liming Liu and Xiaoming Liu. Dynamic and static job allocation for multi-server systems. *IIE transactions*, 30(9):845–854, 1998.

[59] Tore W Jonsbråten, Roger JB Wets, and David L Woodruff. A class of stochastic programs with decision dependent random elements. *Annals of Operations Research*, 82:83–106, 1998.

[60] Jitka Dupacová. Optimization under exogenous and endogenous uncertainty. *University of West Bohemia in Pilsen*, 2006.

[61] Shabbir Ahmed. *Strategic planning under uncertainty: Stochastic integer programming approaches*. PhD thesis, University of Illinois at Urbana-Champaign, 2000.

[62] Kannan Viswanath, Srinivas Peeta, Sibel F Salman, et al. Investing in the links of a stochastic network to minimize expected shortest path. length. Technical report, Purdue University, Department of Economics, 2004.

[63] Vikas Goel and Ignacio E Grossmann. A class of stochastic programs with decision dependent uncertainty. *Mathematical programming*, 108(2-3):355–394, 2006.

[64] Vikas Goel and Ignacio E Grossmann. A stochastic programming approach to planning of offshore gas field developments under uncertainty in reserves. *Computers & chemical engineering*, 28(8):1409–1429, 2004.

[65] Vikas Goel, Ignacio E Grossmann, Amr S El-Bakry, and Eric L Mulkay. A novel branch and bound algorithm for optimal development of gas fields under uncertainty in reserves. *Computers & chemical engineering*, 30(6):1076–1092, 2006.

[66] Bora Tarhan, Ignacio E Grossmann, and Vikas Goel. Stochastic programming approach for the planning of offshore oil or gas field infrastructure under decision-dependent uncertainty. *Industrial & Engineering Chemistry Research*, 48(6):3078–3097, 2009.

[67] Vijay Gupta and Ignacio E Grossmann. Solution strategies for multistage stochastic programming with endogenous uncertainties. *Computers & Chemical Engineering*, 35(11):2235–2247, 2011.

[68] Yigal Gerchak, Diwakar Gupta, and Mordechai Henig. Reservation planning for elective surgery under uncertain demand for emergency surgery. *Management Science*, 42(3):321–334, 1996.

[69] Mehdi Lamiri, Xiaolan Xie, and Shuguang Zhang. Column generation approach to operating theater planning with elective and emergency patients. *IIE Transactions*, 40(9):838–852, 2008.

[70] Mehdi Lamiri, Xiaolan Xie, Alexandre Dolgui, and Frédéric Grimaud. A stochastic model for operating room planning with elective and emergency demand for surgery. *European Journal of Operational Research*, 185(3):1026–1037, 2008.

[71] Mehdi Lamiri, Johann Dreo, and Xiaolan Xie. Operating room planning with random surgery times. In *Automation Science and Engineering, 2007. CASE 2007. IEEE International Conference on*, pages 521–526. IEEE, 2007.

[72] Gerhard Wullink, Mark Van Houdenhoven, Erwin W Hans, Jeroen M van Oostrum, Marieke van der Lans, and Geert Kazemier. Closing emergency operating rooms improves efficiency. *Journal of Medical Systems*, 31(6):543–546, 2007.

[73] Nordic Medico-Statistical Committee. Nomesco classification of surgical procedures, 1996.

[74] Lee J. Krajewski, Manoj K. Malhotra, and Larry P Ritzman. *Operations Management: Processes and Supply Chains, Global Edition*. Pearson, 2016.

[75] William H. Greene. *Econometric Analysis*. Pearson, 2012.

[76] Ruey S Tsay. *Analysis of Financial Time Series*. Wiley, 2010.

[77] Van-Nam Huynh, Yoshiteru Nakamori, Jonathan Lawry, and Masahiro Inuiguchi. *Integrated Uncertainty Management and Applications*. Springer-Verlag Berlin Heidelberg, 2010.

[78] Michal Kaut and Stein W Wallace. Evaluation of scenario-generation methods for stochastic programming. *Pacific Journal of Optimization*, 2003.

[79] Paul Bratley, Bennet L Fox, and Linus E Schrage. *A guide to simulation*. Springer Science & Business Media, 2011.

[80] Camilo Mancilla and Robert H Storer. Stochastic sequencing and scheduling of an operating room. *Theses and Dissertations, Lehigh University, Department of Industrial and Systems Engineering (November 14, 2009)*, 2009.

[81] Bernardetta Addis, Giuliana Carello, and Elena Tànfani. A robust optimization approach for the Advanced Scheduling Problem with uncertain surgery duration in Operating Room Planning - an extended analysis. working paper or preprint, 2014.

[82] Kjetil Høyland, Michal Kaut, and Stein W. Wallace. A heuristic for moment-matching scenario generation. *Computational Optimization and Applications*, 24(2):169–185, 2003.

[83] Allen I. Fleishman. A method for simulating non-normal distributions. *Psychometrika*, 43(4):521–532, 1978.

[84] John R Birge and Francois Louveaux. *Introduction to stochastic programming*. Springer Science & Business Media, 2011.

[85] Gyana R Parija, Shabbir Ahmed, and Alan J Kingf. On bridging the gap between stochastic integer programming and mip solver technologies. *INFORMS Journal on Computing*, 16(1):73–83, 2004.

[86] Kelly Easton, George Nemhauser, and Michael Trick. Cp based branch-and-price. In *Constraint and Integer Programming*, pages 207–231. Springer, 2004.

[87] Jan Lundgren, Mikael Rönnqvist, and Peter Värbrand. *Optimization*. Studentlitteratur, 2010.

[88] Scott Kirkpatrick, Mario P Vecchi, et al. Optimization by simmulated annealing. *science*, 220(4598):671–680, 1983.

[89] Arbeidstilsynet. Retrieved from: *http://www.arbeidstilsynet.no/fakta.html?tid=78157*.

[90] Helsedirektoratet. Retrieved from: *https://helsedirektoratet.no/finansieringsordninger/innsatsstyrt-finansiering-isf-og-drg-systemet/drg-systemet*.

[91] Moon-Won Park and Yeong-Dae Kim. A systematic procedure for setting parameters in simulated annealing algorithms. *Computers & Operations Research*, 25(3):207–217, 1998.

[92] John R Birge. The value of the stochastic solution in stochastic linear programs with fixed recourse. *Mathematical programming*, 24(1):314–325, 1982.

[93] Laureano F Escudero, Araceli Garín, María Merino, and Gloria Pérez. The value of the stochastic solution in multistage problems. *Top*, 15(1):48–64, 2007.

# Appendix

# Appendix A

# Propositions and proofs

The following are the proofs of the validity of the valid inequalities stated in the Phase Model.

*Proof of Proposition 6.2.* From the model formulation, we have balance equation (22) stating

$$t_i - t_{i+1} - \sum_{j \in N} w^{\omega}_{i+1,j} + \sum_{j \in N} s^{\omega}_{ij} + \sum_{j \in N} w^{\omega}_{ij} = - \sum_{j \in N} \sum_{h \in H} D^{\omega}_{jh} y^{\omega}_{ijh} \quad i \in N \setminus \{|N|\}, \omega \in \Omega$$

if we solve for $t_{i+1}$, we get that

$$t_{i+1} + \sum_{j \in N} w^{\omega}_{i+1,j} = t_i + \sum_{j \in N} s^{\omega}_{ij} + \sum_{j \in N} w^{\omega}_{ij} + \sum_{j \in N} \sum_{h \in H} D^{\omega}_{jh} y^{\omega}_{ijh} \quad i \in N \setminus \{|N|\}, \omega \in \Omega$$

From the variable domain definitions we have $s^{\omega}_{ij}, w^{\omega}_{ij} \geq 0$. Thus, removing these terms must leave the RHS less than or equal to the actual start time of surgery in position $i+1$:

$$t_{i+1} + \sum_{j \in N} w^{\omega}_{i+1,j} \geq t_i + \sum_{j \in N} \sum_{h \in H} D^{\omega}_{jh} y^{\omega}_{ijh} \quad i \in N \setminus \{|N|\}, \omega \in \Omega$$

In the shortest possible duration realisation of the surgery in position $i$, there will not be any waiting for the surgery in position $i+1$ in the optimal solution because one always can achieve less waiting in this case by setting a higher $t_{i+1}$ without risking any more idle time. The shortest possible duration is found across all scenarios and phases, which makes it possible to substitute the $y^{\omega}_{ijh}$ with $x_{ij}$. That is, when

$$\sum_{j \in N} \sum_{h \in H} D^{\omega}_{jh} y^{\omega}_{ijh} = \min_{h \in H, \omega \in \Omega} \sum_{j} D^{\omega}_{jh} x_{ij}$$

we get

$$\sum_{j \in N} w^{\omega}_{i+1,j} = 0$$

Thus,

$$t_{i+1} \geq t_i + \min_{h \in H, \omega \in \Omega} \sum_{j} D^{\omega}_{jh} x_{ij} \quad i \in N \setminus \{|N|\}$$

□

*Proof of Proposition 6.3.* Assume that for a given position $i$, the corresponding $\delta_i^\omega = 1$. Then from equation (27) we get

$$\delta_i^\omega = 1 \implies d \leq t_i + \sum_{j \in N} w_{ij}^\omega$$

From balance equation (22) we know that

$$t_i + \sum_{j \in N} w_{ij}^\omega \leq t_{i+1} + \sum_{j \in N} w_{i+1,j}^\omega$$

and therefore

$$d \leq t_{i+1} + \sum_{j \in N} w_{i+1,j}^\omega \implies \delta_{i+1}^\omega = 1.$$

If, on the other hand, $\delta_i^\omega = 0$, then the following is always true

$$0 \leq \delta_{i+1}^\omega$$

Thus,

$$\delta_i^\omega \leq \delta_{i+1}^\omega$$

is a valid inequality for all $i \in N \setminus \{|N|\}, \omega \in \Omega$. □

*Proof of Proposition 6.4, part 2.* The linearisation of

$$y_{ijh}^\omega = x_{ij} \tau_{ih}^\omega \quad i \in N, j \in N, h \in H, \omega \in \Omega$$

was shown in Chapter 6.1.7. From equation (28) we know that

$$\sum_{i \in N} x_{ij} = 1 \quad j \in N$$

and from equation (38) we have

$$\sum_{h \in H} \tau_{ih}^\omega = 1 \quad i \in N, \omega \in \Omega$$

Combining this with the linearisation, we get

$$y_{ijh}^\omega = x_{ij} \tau_{ih}^\omega \quad i \in N, j \in N, h \in H, \omega \in \Omega \implies$$

$$\sum_{h \in H} \sum_{i \in N} y_{ijh}^\omega = \sum_{h \in H} \sum_{i \in N} x_{ij} \tau_{ih}^\omega \quad j \in N, \omega \in \Omega \implies$$

$$\sum_{h \in H} \sum_{i \in N} y_{ijh}^\omega = \sum_{h \in H} \tau_{ih}^\omega \sum_{i \in N} x_{ij} \implies y_{ijh}^\omega = 1 \cdot 1 = 1 \quad j \in N, \omega \in \Omega$$

□

The following are the valid inequalities used in the Emergency Model. The structure of these are similar to the Phase Model, but is adapted to the Emergency Model notation and definitions.

**Proposition A.1.** *The following inequalities restrict the intervals between surgery start times*

$$\phi^{\xi}_{iu} + \min_{\omega \in \Omega} \sum_{q \in Q^{\xi}} D^{E\omega}_{q} z^{\xi}_{iqu} \leq \phi^{\xi}_{i,u+1} + M^{Ez}_{i}(1 - \sum_{q \in Q^{\xi}} z^{\xi\omega}_{iq,u+1}) \tag{112}$$

$$t_i + \min_{\omega \in \Omega} \sum_{j \in N} D^{\omega}_{j} x_{ij} \leq \phi^{\xi}_{i+1,u} + M^{Ez}_{i}(1 - \sum_{q \in Q^{\xi}} z^{\xi\omega}_{i+1,qu}) \tag{113}$$

*are valid for all* $i \in N, u \in Q^{max}, \xi \in \Xi$ *and must be satisfied in the optimal solution.*

**Proposition A.2.**

$$\delta^{\xi\omega}_{iu} \leq \delta^{E\xi\omega}_{i,u+1} + (1 - \sum_{q \in Q^{\xi}} z^{\xi\omega}_{iq,u+1}) \tag{114}$$

$$\delta^{\xi\omega}_{i} \leq \delta^{E\xi\omega}_{i+1,u} + (1 - \sum_{q \in Q^{\xi}} z^{\xi\omega}_{i+1,qu}) \tag{115}$$

$$\delta^{E\xi\omega}_{iu} \leq \delta^{\xi\omega}_{i} \tag{116}$$

*are valid for all* $i \in N, u \in Q^{max}, \omega \in \Omega, \xi \in \Xi, q \in Q^{\xi}$ *and must be satisfied in the optimal solution.*

# Appendix B

# Out-of-sample stability results

The following tables show the out-of-sample results for both models. The average values are across five runs on different scenario trees of different size. The Phase Model results are shown first, while the three last tables are from the Emergency Model.

Out-of-sample results for instance 1

| **# of scenarios** | **Avg. distance** | **Avg. relative distance %** |
|---|---|---|
| 40 | 2.785 | 11.5% |
| 50 | 2.532 | 10.6% |
| 60 | 2.467 | 10.0% |
| 70 | 1.144 | 4.4% |
| 80 | 1.448 | 5.7% |
| 90 | 1.401 | 5.4% |
| 100 | 1.124 | 4.4% |
| 125 | 0.946 | 3.6% |
| 150 | 1.032 | 4.0% |
| 200 | 1.020 | 4.0% |

Out-of-sample results for instance 2

| **# of scenarios** | **Avg. distance** | **Avg. relative distance %** |
|---|---|---|
| 40 | 1.238 | 4.4% |
| 50 | 1.645 | 5.8% |
| 60 | 1.409 | 5.0% |
| 70 | 1.885 | 6.9% |
| 80 | 1.254 | 4.4% |
| 90 | 1.144 | 4.1% |
| 100 | 0.870 | 3.0% |
| 125 | 0.640 | 2.2% |
| 150 | 1.011 | 3.5% |
| 200 | 0.800 | 2.8% |

Out-of-sample results for instance 7

| # of scenarios | Avg. distance | Avg. relative distance % |
|---|---|---|
| 40 | 1.811 | 5.0% |
| 50 | 1.359 | 3.7% |
| 60 | 1.213 | 3.3% |
| 70 | 1.520 | 4.2% |
| 80 | 1.128 | 3.1% |
| 90 | 0.576 | 1.5% |
| 100 | 0.569 | 1.5% |
| 125 | 0.670 | 1.8% |
| 150 | 0.357 | 1.0% |
| 200 | 0.325 | 0.9% |

Out-of-sample results for instance 9

| # of scenarios | Avg. distance | Avg. relative distance % |
|---|---|---|
| 40 | 2.838 | 5.8% |
| 50 | 3.291 | 6.9% |
| 60 | 3.968 | 8.5% |
| 70 | 2.910 | 6.0% |
| 80 | 2.485 | 5.1% |
| 90 | 2.317 | 4.8% |
| 100 | 1.770 | 3.6% |
| 125 | 1.713 | 3.5% |
| 150 | 1.168 | 2.4% |
| 200 | 0.895 | 1.8% |

Out-of-sample results for instance 12

| # of scenarios | Avg. distance | Avg. relative distance % |
|---|---|---|
| 40 | 4.343 | 8.6% |
| 50 | 4.518 | 10.3% |
| 60 | 4.000 | 9.0% |
| 70 | 3.920 | 8.8% |
| 80 | 3.242 | 7.2% |
| 90 | 2.780 | 6.2% |
| 100 | 2.149 | 4.7% |
| 125 | 1.601 | 3.1% |
| 150 | 2.014 | 4.4% |
| 200 | 2.001 | 4.3% |

Out-of-sample results for instance 14

| # of scenarios | Avg. distance | Avg. relative distance % |
|---|---|---|
| 40 | 3.210 | 9.2% |
| 50 | 4.707 | 15.4% |
| 60 | 3.386 | 10.6% |
| 70 | 3.519 | 11.3% |
| 80 | 3.061 | 9.6% |
| 90 | 3.435 | 10.6% |
| 100 | 2.598 | 8.1% |
| 125 | 1.829 | 4.8% |
| 150 | 1.075 | 3.2% |
| 200 | 1.010 | 3.0% |

Out-of-sample results for instance 15

| # of scenarios | Avg. distance | Avg. relative distance % |
|---|---|---|
| 40 | 5.980 | 4.8% |
| 50 | 7.264 | 5.8% |
| 60 | 9.025 | 7.3% |
| 70 | 8.500 | 6.9% |
| 80 | 5.705 | 4.6% |
| 90 | 4.567 | 3.5% |
| 100 | 3.107 | 2.5% |
| 125 | 3.465 | 2.5% |
| 150 | 2.653 | 2.1% |
| 200 | 2.500 | 1.9% |

Out-of-sample results for emergency instance 1E

| # of scenarios | Avg. distance | Avg. relative distance % |
|---|---|---|
| 40 | 2.869 | 1.0 |
| 50 | 1.999 | 0.7 |
| 60 | 1.186 | 0.4 |
| 70 | 1.268 | 0.5 |
| 80 | 1.088 | 0.4 |
| 90 | 0.832 | 0.3 |
| 100 | 0.841 | 0.3 |
| 110 | 0.546 | 0.2 |
| 120 | 1.046 | 0.4 |
| 130 | 0.992 | 0.4 |
| 140 | 0.317 | 0.1 |
| 150 | 0.468 | 0.2 |
| 160 | 0.574 | 0.2 |
| 170 | 0.507 | 0.2 |
| 180 | 0.523 | 0.2 |
| 190 | 0.537 | 0.2 |
| 200 | 0.371 | 0.1 |

Out-of-sample results for emergency instance 2E

| # of scenarios | Avg. distance | Avg. relative distance % |
|---|---|---|
| 40 | 3.187 | 1.5 |
| 50 | 2.445 | 1.2 |
| 60 | 1.123 | 0.6 |
| 70 | 1.140 | 0.5 |
| 80 | 1.172 | 0.6 |
| 90 | 1.009 | 0.5 |
| 100 | 0.909 | 0.4 |
| 110 | 0.973 | 0.5 |
| 120 | 0.989 | 0.5 |
| 130 | 0.929 | 0.5 |
| 140 | 0.898 | 0.4 |
| 150 | 0.439 | 0.2 |
| 160 | 0.790 | 0.4 |
| 170 | 0.666 | 0.3 |
| 180 | 0.617 | 0.3 |
| 190 | 0.620 | 0.3 |
| 200 | 0.502 | 0.3 |

Out-of-sample results for emergency instance 3E

| **# of scenarios** | **Avg. distance** | **Avg. relative distance %** |
|---|---|---|
| 40 | 4.353 | 1.5 |
| 50 | 3.520 | 1.2 |
| 60 | 2.950 | 1.0 |
| 70 | 3.192 | 1.1 |
| 80 | 3.169 | 1.1 |
| 90 | 2.965 | 1.1 |
| 100 | 2.766 | 1.0 |
| 110 | 2.934 | 1.0 |
| 120 | 2.714 | 0.9 |
| 130 | 2.946 | 1.0 |
| 140 | 2.364 | 0.8 |
| 150 | 2.304 | 0.8 |
| 160 | 2.378 | 0.8 |
| 170 | 2.218 | 0.8 |
| 180 | 2.211 | 0.8 |
| 190 | 2.272 | 0.8 |
| 200 | 2.158 | 0.7 |