BLOODY FAST BLOOD COLLECTION

Sem van Brummelen

Promotie commissie

Voorzitter/secretaris:	Prof. dr. P.M.G. Apers *University of Twente, Enschede, the Netherlands*
Promotors:	Prof. dr. R.J. Boucherie *University of Twente, Enschede, the Netherlands*
	Prof. dr. N.M. van Dijk *University of Twente, Enschede, the Netherlands*
	Prof. dr. W.L.A.M de Kort *University of Amsterdam, Amsterdam, the Netherlands*
Leden:	Prof. dr. H. van den Berg *University of Twente, Enschede, the Netherlands*
	Prof. dr. J.T. Blake *Dalhousie University, Halifax, Canada*
	Prof. dr. ir. E.W. Hans *University of Twente, Enschede, the Netherlands*
	Dr. K. van den Hurk *Sanquin, Amsterdam, the Netherlands*
	Prof. dr. I. van Nieuwenhuyse *KU Leuven, Leuven, Belgium*
	Prof. dr. R. Nunez-Queija *University of Amsterdam, Amsterdam, the Netherlands*

Ph.D. thesis, University of Twente, Enschede, the Netherlands
Center for Telematics and Information Technology (No. 17-446, ISSN 1381-3617)
Center for Healthcare Operations Improvement and Research

This research was in part conducted at and financially supported by the Sanquin Blood Supply Foundation by means of project No. PPOC-14-DS-04.

The distribution of this thesis is financially supported by Sanquin Research, Amsterdam, the Netherlands

Printed by Ipskamp printing, Enschede, the Netherlands
Cover design: Lisa Klinkenberg, Hoorn, the Netherlands

ISBN 978-90-365-4428
DOI 10.3990/1.9789036544283

BLOODY FAST BLOOD COLLECTION

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
Prof. dr. T.T.M. Palstra,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op donderdag 14 december 2017 om 14.45 uur

door

Samuel Pieter Josephus van Brummelen

geboren op 23 november 1990
te Uithoorn, Nederland

Dit proefschrift is goedgekeurd door de promotors:

Prof. dr. R.J. Boucherie
Prof. dr. N.M. van Dijk
Prof. dr. W.L.A.M. de Kort

Voorwoord

Nu iets meer dan acht jaar geleden begon ik aan mijn bachelor econometrie in Amsterdam. Waar sommige studenten al een duidelijk beeld hadden waar ze naar toe werkten vanaf het eerste college, behoorde ik tot de groep die het over zich heen liet komen, en niet verder vooruitkeek dan het einde van de studie. Halverwege mijn master Operations Research vroeg mijn, naar later bleek, promotor Nico me na een college nog even te blijven zitten. Hij vroeg of ik weleens had nagedacht over een promotie, en wist mogelijk nog wel een project voor me, een samenwerking tussen de Universiteit Twente en Sanquin. Ongeveer 5 jaar later schrijf ik dit dankwoord, en heb ik een erg leuk, interessant en leerzaam promotietraject achter de rug.

Ik wil graag beginnen al mijn begeleiders, Nico, Wim, Richard en Katja, te bedanken voor de vrijheid die ik tijdens de afgelopen jaren heb gehad. Ik heb mijn eigen onderzoek vorm mogen geven, nieuwe ideeën mogen uitwerken en fouten mogen maken. Dat alles heeft me erg veel geleerd en mede gemaakt tot wie ik nu ben.

Nico, tijdens het schrijven van mijn bachelor thesis heb ik je een beetje leren kennen. Toen Martijn en ik tijdens onze afsluitende presentatie vertelden dat wat wij hadden gedaan nieuw was, vroeg je direct wanneer het paper zou worden geschreven. Hoewel dit paper er nooit is gekomen, ben je nu wel mede-auteur van al mijn artikelen. Ik wil je erg graag bedanken voor je eindeloze enthousiasme en het in mijn gestelde vertrouwen tijdens het begeleiden van zowel mijn master thesis als mijn promotie.

Wim, allereest wil ik je heel erg bedanken voor de warme ontvangst bij Donorstudies. Ik voelde me direct welkom in Nijmegen en later in Amsterdam. Ik heb het altijd erg gewaardeerd dat je aan bijna alle wiskunde in dit proefschrift tijd hebt besteed, om de methode te doorgronden. Als je niet stiekem een studie wiskunde hebt gedaan, moet ook jij wat hebben geleerd van mijn promotie, want van alle papers begreep je op zijn minst in hoofdlijnen hoe de methode werkt. Daarnaast wil ik je ook bedanken voor de betrokkenheid en de feedback op alles wat ik je heb toegezonden.

Richard, ook jou wil ik graag bedanken voor het in mij gestelde vertrouwen. Ik heb onze samenwerking aan Hoofdstuk 2 van dit proefschrift gewaardeerd, en wil je graag bedanken voor de bijdrage aan dit proefschrift en hoofdstuk 2 in het bijzonder.

Katja, ik heb je betrokkenheid bij dit project zeer gewaardeerd. Onze (ongeveer) tweewekelijkse gesprekken werkten erg goed om me los te trekken van de computer en samen na te denken over de praktische toepasbaarheid van het onderzoek. Ook heb ik veel geleerd van de verschillen tussen onderzoek in de wiskunde en onderzoek in de epidemiologie, en ik denk dat mijn proefschrift sterker is geworden door jouw

epidemiologische blik.

I would also like to thank all my committee members, Hans van den Berg, John Blake, Erwin Hans, Katja van den Hurk, Inneke van Nieuwenhuyse and Sindo Núñez, for the time invested in my thesis and defense. I would also like to thank all of you for the valuable comments.

Rosa en Maurits, ik ben blij dat jullie 14 december naast me staan. De borrels, pannenkoekenavonden en oio-uitjes zijn hoogtepunten uit mijn PhD tijd. Ook wil ik jullie bedanken voor de gezelligheid tijdens congressen en aansluitende vakanties. Ongetwijfeld tot de volgende borrel of het volgende etentje!

Alle collega's van CHOIR, SOR en DMMP wil ik graag bedanken voor de koffiepauzes, lunch(wandelingen), en uitjes. Daarnaast wil ik nog een paar mensen persoonlijk noemen.

Ingeborg en Nardo, overbuurvrouw en -man, bedankt voor alle gesprekken, zowel over werk als wat minder werkgerelateerde onderwerpen. Daarnaast wil ik je, Ingeborg, graag bedanken voor het nalezen van vele stukken, waaronder, maar zeker niet beperkt tot, delen van dit proefschrift!

Gréanne, gefeliciteerd met je proefschrift! Bedankt voor de samenwerking bij de organisatie van ons symposium.

Ingeborg, Gréanne, Maartje, Aleida, Nardo, Eline en Shiya, bedankt voor alle gesprekken, discussies en gezelligheid op de CHOIR kamer.

Anne en Corine, bedankt voor de spelletjes, wandelingen en algemene gezelligheid in Lunteren. Hoewel grapjes over de onbereikbaarheid van De Werelt meer dan eens zijn langsgekomen, vond ik het toch altijd gezellig.

Joost, samen hebben we bij Sanquin de Operations Research binnengebracht. Bedankt voor de samenwerking, met een mooi hoofdstuk (en toekomstig paper) als resultaat.

Nu is het een kleine stap om ook de rest van Sanquin Donorstudies te bedanken. Hoewel ik niet heel vaak aanwezig was, hebben jullie me altijd het gevoel gegeven een vol onderdeel van de groep te zijn. Ik heb het geluk gehad om van twee groepen onderdeel uit te maken. Ook jullie wil ik graag bedanken voor de koffiepauzes, lunches en uitjes. Ook hier wil ik graag nog een paar (oud) Donorstudies collega's persoonlijk noemen:

Karlijn, bedankt voor alle potjes tennis en aansluitende gezelligheid.

Esther (ik hoop dat je het me niet kwalijk neemt dat ik je tussen Donorstudies noem), bedankt voor alle gezamelijke fietstochtjes!

Lisa, bedankt voor het ontwerpen van de kaft van dit proefschrift.

Ook wil ik graag iedereen van Sanquin, inclusief Sanquin als geheel, bedanken voor het steunen van dit onderzoek.

Tot slot wil ik ook graag alle familie en vrienden bedanken. Pappa en Mamma, wie had gedacht dat ik toch nog op het snijvlak van jullie werkvelden terecht zou komen? Eigenlijk hebben jullie allebei wel een punt, zowel de zorg als het programmeren heeft bijzonder leuke kanten. Bedankt voor de steun tijdens mijn gehele opleiding.

Daarnaast wil ik jullie bedanken voor de nieuwsgierigheid die jullie me al van jongs af aan hebben bijgebracht, en waarzonder ik lang niet zo ver was gekomen.

Job, ik geloof niet dat er iemand is waarmee ik meer (soms zinloze) discussies heb gevoerd. Het heeft me geleerd altijd kritisch te zijn, ook op mezelf, wat me vaak heeft geholpen.

Opa's en oma's, Willy, Annie, Piet en Jo, ik prijs mezelf gelukkig dat jullie er alle vier nog zijn. Mede via de opvoeding van mijn ouders, ben ik ook aan jullie dank verschuldigd voor dit proefschrift. Omdat ik nu dreig in een oneindige iteratie terecht te komen, zal ik het bij twee generaties laten.

Lieve Wendy, bedankt voor alle kleine dingetjes (denk bijvoorbeeld aan het meelezen en corrigeren van een aantal stukken van dit proefschrift). Maar vooral: bedankt voor alle afleiding op momenten dat ik dit proefschrift (of andere dingen) niet kon loslaten. Ik heb onzettend veel zin om samen door Vietnam te gaan reizen!

Sem
Enschede, november 2017.

Contents

Part I

Introduction

Chapter 1

S.P.J. van Brummelen. Introduction to Bloody Fast Blood Collection.

Chapter 2

N.M. van Dijk, S.P.J. van Brummelen, and R.J. Boucherie. Uniformization: Basics, extensions and applications. *Performance Evaluation, accepted*.

CHAPTER 1

Introduction

1.1 A short history of blood donation and transfusion

The first blood transfusion - receiving blood - and the first blood donation - giving blood, were performed on dogs by dr. Richard Lower in 1665. Blood was directly transferred from one dog to another. In the following year, similar experiments were performed with different animals, including transfusions between different species of animals. Although most of these experiments were successful, i.e. the receiving animal remained or became healthy, people at the time still largely thought the qualities of humans were determined by their blood, so transfusions between humans were still out of the question.

However, this did not rule out transfusions with human recipients. The first transfusions with human recipients of blood were even founded in the same belief that blood determines one's qualities. These transfusions were aimed at curing mental illnesses, and not, as might seem obvious, as a cure for excessive bleeding. The first transfusion with a human recipient was carried out in 1667, by Jean-Baptiste Denis in Paris, transfusing blood of lambs and calves. Later the same year, dr. Lower transfused a 22-year old student in Cambridge with the blood of a sheep. Although both these patients reportedly survived their transfusions, multiple other patients died, and the practice of transfusions soon fell out of favor for approximately 150 years.

In 1818, the first human to human blood transfusion was reported. James Blundell transfused blood to women suffering from "postpartum hemorrhage", i.e. bleeding after childbirth. He also suggested to only use human blood, as his experiments with transfusion between different animal species all ended in death for the transfused animal. Although it was known that blood was not compatible between species, all of these initial transfusions happened without the knowledge of blood types. Blood clotting when blood of different species is mixed was described in 1875 by Landois. Karl Landsteiner first described the same effects when mixing blood of humans in 1901. He discovered the ABO-system (see Table 1.1), for which he was awarded the Nobel prize. Later, the Rhesus D (often indicated with a + or - after the ABO indication) and other blood groups were discovered.

Blood transfusions still had to deal with severe limitations. Blood platelets are activated as soon as blood leaves the human body, and start inducing blood clotting.

Table 1.1 Compatibility of donors and recipients. '✓' indicates that a transfusion is possible, '×' indicates that a transfusion is very likely to cause clotting of blood, usually resulting in death.

		Blood type recipient			
		O	A	B	AB
Blood type donor	O	✓	✓	✓	✓
	A	×	✓	×	✓
	B	×	×	✓	✓
	AB	×	×	×	✓

This causes blood to quickly develop fibrinogen clots (turn into some sort of unusable gel). As long as there was nothing available to stop this process, the amount of blood that could be transfused was very limited, and blood could not be stored. Alexis Carrel developed a surgical technique to be able to transfuse more blood, first used in 1908, for which he too received the Nobel prize. Richard Lewinsohn introduced sodium citrate as a first anti-clotting solution. Very high, toxic levels of the solution were already used in laboratories for the same purpose, but he proposed experimenting with much lower levels to store blood. This blood was only stored for hours, but the addition of dextrose to the solution made storage for weeks possible. Similar solutions are still used for the long term storage of blood, up to 42 days.

The introduction of anti-clotting solutions made the introduction of blood banks possible. Although blood banks are now often tasked with collection, testing, storing and distribution of blood and derivative products, the first blood banks directed blood donors to hospitals in need of blood. The first blood bank of this type was established in London in 1921 by Percy Oliver. (Section based on [87, 129, 186])

1.1.1 The Netherlands

Following the example of Percy Oliver in 1921, Dr. H.C.S.M van Dijk established the first blood bank in the Netherlands in Rotterdam. Blood banks in The Hague and Utrecht soon followed. Even in this earliest stage, the conscious decision was taken that donors in the Netherlands should be voluntary and non-remunerated, a principle that still stands. During the run-up to the Second World War, the demand for blood rose, and facilities were opened in Rotterdam and Amsterdam. The facility in Amsterdam survived the war and developed into the Central Laboratory for Blood Transfusion Services, or CLB for short.

In 1947 the CLB started producing pharmaceuticals from blood plasma. In the following years the CLB expanded quickly. A laboratory for blood typing was built, and diagnostic and scientific research into blood transfusion was started. In 1962, the amount of scientific research within the CLB had reached a level that a separate foundation was created, the Karl Landsteiner Foundation. By this time, the CLB was doing 700,000 tests per year, making it the largest diagnostic facility in the Netherlands.

The number of blood banks in the Netherlands, responsible for directing donors to hospitals, had grown to 110 by 1973. With the introduction of centralized storage

of blood, this number was reduced to 22 in 1982. These new blood banks were all independent, and were tasked with collection, storage and distribution of blood. In 1998 a new law on blood supply, "Wet inzake bloedvoorziening", was implemented. This led to the foundation of Sanquin — from the Latin word for blood: sanguis — by merging the CLB with these blood banks. At this point, the number of blood banks was reduced to nine. In 2001, the number of blood banks was again reduced, leaving four blood banks, still in operation today: North-East, North-West, South-East and South-West. (Section based on [135])

1.1.2 Sanquin

Sanquin has been established by law as the organization responsible for collection, production, storage and distribution of all blood and related products in the Netherlands. In 2016, Sanquin had 2821 employees, working in over 130 locations. In total, over 720,000 donations were collected by Sanquin. Currently, 343,112 people are registered as a blood donor in the Netherlands. In addition to the blood bank, the largest division, Sanquin has five other divisions. The first is a large facility fractionating blood products from blood plasma for both national and international usage. Second, a diagnostics division testing donations and other blood samples. A third division produces reagents used in blood typing. Fourth, Sanquin operates a tissue bank that stores bone and other human tissue. Finally, Sanquin also has a large Research division. This research division mainly does scientific research into transfusion medicine and immunology. The research presented in this thesis has also been supported by and has taken place in close collaboration with this division.

The main focus of this thesis, however, is on the blood collection activities of the Sanquin blood bank. Different types of donations are collected, but two types form the overwhelming majority of donations. The first is the whole blood donation. This is the simplest possible donation. A needle is injected, and 500 ml of blood is transferred into a bag, which contains some anti-clotting solution. Although a healthy human can easily lose 500 ml of blood, it can take a human body months to replenish the cells in the whole blood donation. Therefore, at least 56 days have to pass between two whole blood donations, and maximum number of donations per year has been set: 3 for females and 5 for males. In 2016, a total of 420,163 whole blood donations were collected by Sanquin.

The second major type of donation is the plasma donation. Plasma is the fluid that contains the blood cells. Plasma also contains proteins and other substances that can be used to produce pharmaceuticals. With a plasma donation, blood is collected in a centrifuge. The plasma is filtered out, and the remaining cells are passed back into the donor body. Although this procedure takes longer, it is less invasive in the long run, as plasma is replenished much quicker than blood cells. Plasma can therefore be donated every two weeks. In 2016, a total of 306,402 plasma donations were collected by Sanquin.

Donations at Sanquin blood collection sites still take place on a voluntary, non-remunerated basis, as recommended by the World Health Organization. This has multiple reasons, not the least of which that paying for donations might attract

unwanted donors, and might cause donors to lie about their eligibility to donate. However, voluntary donors want and deserve the best possible service. An important aspect of offering the best possible service is the minimization of waiting times that a donor may experience at a blood collection site.

1.2 Motivation

Every year, approximately four million trauma, oncology, hematology, and obstetric patients, of all ages, in Europe require blood transfusions. Moreover, several millions of immune compromised, clotting factor deficient, and other patients are treated with plasma-derived pharmaceuticals, for which approximately four million kilogram of plasma needs to be collected every year. The supply of these blood products depends on blood banks having access to sufficient healthy and motivated donors.

As required for the safety of both the donor and the recipient of the blood donation, donors receive a limited health check, which could be seen as a small compensation. Donors also occasionally receive small gifts after some number of donations. However, this is not at all proportionate to both the advantages gained by the recipient of the donation and the time and effort put in by donors. Sanquin and other blood banks therefore rely on altruism of donors. However, if donors have negative associations with the blood bank, donors might not be willing to make further donations.

Ferguson [81] has done a literature survey on the return behavior of donors, and finds that among organizational factors, waiting time at the collection site is the most consistent negative influence on the return behavior of donors. More recently, McKeever et al. [132] confirmed the negative association of long waiting times with the probability that a donor returns for a subsequent donation.

Non-returning donors can cause substantial problems for Sanquin. Recruiting new donors requires far more effort and is more expensive than inviting an existing donor back. A potential new donor first has to be convinced to become a blood donor, a process that requires a time investment at the very least. Before the first actual donation, the donor visits a collection site and goes through a screening process to determine eligibility. Additionally, if too many donors have negative associations with the blood bank, this could cause general goodwill decrease, making both retaining and recruiting donors much more difficult.

Clearly, Sanquin must be concerned about waiting times experienced by donors at collection sites. The easiest way to improve waiting times has always been to expand the capacity. At Sanquin this could imply either more collection sessions, more collection sites or more capacity during current collection sessions. All of these solutions would require additional investments by Sanquin, which is not possible given budget constraints.

The only remaining option to decrease waiting times is to use the existing capacity of Sanquin's collection sites more effectively. For this purpose, this thesis presents a number of approaches to compute, predict and decrease long waiting times at collection sites, without the need for increased capacity.

1.3 A blood collection site

Most of the research in this thesis focuses on analyzing and improving the service and efficiency at blood collection sites in the Netherlands. Here, it is important to note that collection sites throughout the world have similar layouts, structures and processes, and the methods can be applied in other countries as well. Blood collection sites in the Netherlands come in two main varieties: fixed sites and mobile sites. Fixed collection sites are located in major cities in the Netherlands. These locations have at least a few sessions every week. Most fixed collection sites collect both whole blood and plasma donations. Mobile sites are located in towns and small cities, and are visited by trucks, on average, once a month. The number of visits can vary between every two weeks and a couple of times per year, depending on the population in the service area. Upon arrival, the trucks deploy a fully equipped collection site. Mobile sites only collect whole blood donations.

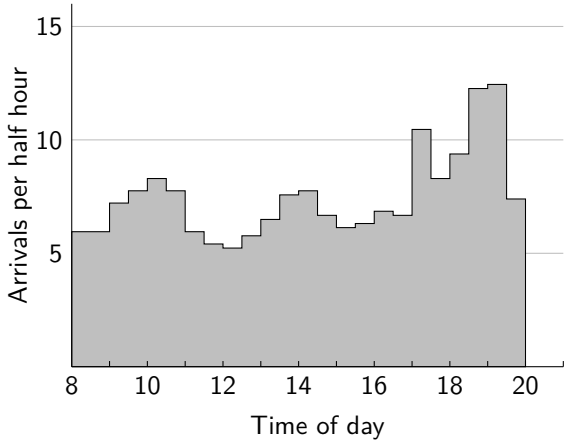
Opening times and days of Dutch blood collection sites vary between collection sites. There is consistency though, as opening times are always one or a combination of the seven different collection sessions shown in Table 1.2. For staff scheduling, an extra half hour before and a half hour or hour after the session is added to the shift. This extra time is necessary to set up equipment when starting a collection session and to clear the collection site and shut down the equipment at the end of the collection session. A session is usually divided into one to three shifts, with a shift covering a morning, afternoon or evening. Each shift is covered by 6 to 12 staff members, depending on the size, measured in the number of donations beds and interview rooms, of the collection site, and to a far lesser extent, the time of day. This means that the total number of staff members that is present at the collection site may change during the day. But even for long collection sessions the total number of staff members present changes only slightly.

Table 1.2 Session types at Sanquin and their opening hours (M=morning, A=afternoon, E=evening).

session name	opening hours
M1	8.00 - 11.00
M2	8.00 - 12.00
MA	8.00 - 15.30
AE	12.30 - 20.00
E1	16.00 - 20.00
E2	17.00 - 20.00
MAE	8.00 - 20.00

At collection sites, two main types of staff member are always present: general staff members and one physician. All tasks described in the description of the collection process below can be executed by general staff members. A physician always has to be present in case of complications during a donation (e.g. fainting). The physician also has to be present to answer questions of donors and general staff members in case the eligibility of a donor is non-trivial. In addition, the physician is

Figure 1.1 Typical arrival pattern of walk-in whole blood donors for a collection site that is opened the whole day (MAE session).



also responsible for the interview of new donors. It is important to note that, in the Netherlands, the first donation does not involve a donation, and therefore has been excluded from this thesis.

The number of arriving donors is managed differently for whole blood and plasma donors. For whole blood donors, Sanquin decides on how many donors to invite to come in for a donation every week. This is currently a manual decision and is done by setting a collection goal for every collection site. This goal is based on the current stock - and by extension the expected stock - of blood products. As the probability of no-show per collection site is known, the goal is then divided by the probability that a donors shows up to determine the number of invitations that will be send out to donors. In this invitation, a specific date and time are not specified, but a two week period for the donation is mentioned instead. After receiving an invitation, which does specify a collection site, a donor is free to decide when to donate and whether to donate at all. A donor is also free to donate at a different collection site than specified on the invitation. All of these uncertainties result in strongly time-varying arrivals. However, clear patterns do show up. The arrival patterns differ between session types. However, even though the absolute number of arrivals change from day to day, the ratios between hours is largely constant for a session type. An example of an arrival pattern is shown in Figure 1.1, which shows the average number of arriving donors for every half hour during an MAE session.

Plasma donors, in contrast, must make an appointment for their donation. This gives Sanquin much more control in the arrivals of plasma donors to collection site. Sanquin aims, and mostly succeeds, in spreading these arrivals uniformly throughout the day.

When a whole blood or plasma donor arrives at a Dutch blood collection site, the donor will first go to the Registration desk. Depending on the collection site in question and the time of day, there might be a short queue before the registration desk. After the possible queue, the arrival of the donor is recorded and the potential

donor is handed a questionnaire. The donor is asked to fill out this questionnaire, which mostly includes questions regarding the donor's health and eligibility to donate blood. After the questionnaire is filled out, the donor deposits the questionnaire at the registration desk, and takes a seat in a waiting room.

The donor now has to wait for a staff member to pick up the donor. In this queue, plasma donors, identified by a different color questionnaire, are serviced with priority. When the donor is picked up, the staff member takes the donor to an interview room and discusses the questionnaire with the donor. Subsequently, the staff member tests the pulse, blood pressure and Hemoglobin (Hb) level of the donor. If neither the interview, nor the tests, give an indication for ineligibility for a blood donation, the donor is directed to the donation room, and is again asked to wait to be picked up by a staff member. On average, the interview and tests take about six minutes. Note that the interview can be done by a general staff member, except for a first time visit. The interview stage usually is a bottleneck in the process, as there is only a limited number of interview rooms. These interview rooms are also used for the much longer lasting interviews at a first visit. Aside from the priority received by plasma donors, the interview is the same for whole blood and plasma donors.

When the donor is picked up from the waiting area of the donation room, the donor is guided to a donation chair. For plasma donations more equipment is required than for a whole blood donation. For this reason, a fixed collection site usually has a number of donation chairs that already have the plasma equipment set up, and will not be used for whole blood donations. Usually, staff members are assigned to either whole blood or plasma in the donation room, and sometimes the plasma donation chairs will even be in a different room than the whole blood chairs. This largely separates the donation stage of the process for whole blood and plasma donors.

Setting up the machine and connecting it to the donor takes more time for plasma donations than for whole blood. After starting the donation, while the actual donation is ongoing, no staff member is directly required, unless complications occur. The donation machine signals the staff members when 500ml of whole blood or 660ml of plasma has been collected and the donation is finished. The donor then waits for a staff member to uncouple the donation equipment. On average, the collection process takes approximately fifteen minutes for a whole blood donation, and 45 minutes for a plasma donation. After the donation, the donor is offered a refreshment, before leaving the collection site.

1.4 Literature

Specific literature on blood collection sites is sparse, as confirmed by the literature review on blood management by Baş et al. [20]. A first aspect of the analysis of blood collection sites is to determine the arrival pattern of walk-in donors. Bosnes et al. [32] and Testik et al. [173] both focus on determining and predicting the arrival pattern of blood donors. Testik et al. also determine the minimal number of required staff members based on these arrival patterns. Blake and Shimla [29] also determine minimal staffing requirements for blood collection sites by modeling blood collection

sites as a series of M/M/c queues.

Simulation models have frequently been used for the actual analysis of blood collection sites. Pratt and Grindon [154] were the first to use a simulation model for the analysis of a blood collection site. They tested a few scenarios with respect to arrivals and scheduling strategies. Brennan et al. [38] developed a simulation model for the American Red Cross blood collection sessions to reduce waiting times. The Red Cross was concerned that long waiting times would reduce the willingness of donors to return for subsequent donations. Several strategies were tested and are presented in this paper. Michaels et al. [136] use a similar simulation model to improve donor scheduling at the American Red Cross.

Alfonso et al. [7, 8] also used a simulation model. They described a French blood collection site as a Petri net, and turned this Petri net formulation into a discrete event simulation model. The model is used to test several scenarios for the blood collection site, based on three different arrival patterns.

Bretthauer and Côté [39] developed a method to determine the required capacity of Health care systems based on a mathematical programming approach. One of the two test cases included in their paper is based on a blood collection site. De Angelis et al. [15] studied the allocation of servers at health care systems by combining simulation and optimization. They also used a blood collection site to demonstrate the practical application of their method.

Alfonso et al. [6] present a Mixed Integer Non Linear Program to schedule appointments at blood collection sites. Their method takes waiting at the blood collection site into account based on a Petri net formulation of the blood collection site. The arrivals of whole blood donors without appointments are combined with appointment based arrivals for plasma and platelet donors. Alfonso et al. [9] also study the problem of scheduling donors, this time combined with capacity planning. They formulate the problem as a mathematical programming model, and evaluate the results with a simulation model.

1.5 Thesis outline

Following this introduction we introduce the technical method of Uniformization in Chapter 2. A chapter is devoted to the method because it is one of the underlying methods for many of the approaches presented in this thesis. Chapters 3, 4, 6 and 7 are based on the method. Chapter 2 discusses the basics of uniformization, as well as several extensions and applications of the method. each extension is supported by numerical examples. A broader scope of uniformization is presented in Chapter 2 than is required for the subsequent chapters. However, it does provide the opportunity to give an overview of the other possibilities with the method, and insights in the intuition behind the method.

Part II: Evaluation, contains two chapters that discuss methods to compute and evaluate waiting times and queues at blood collection sites. Chapter 3 contains three main results. First, it provides a closed form expression for the queueing distributions at blood collection sites in steady state, under an exponential assumption. Second,

it proves that a standard expression for M/M/c queues can be used to determine the waiting time distribution at the individual stations of blood collection sites. Third, a numerical procedure is given to compute the total delay time distribution of all stations at the collection site combined. Chapter 4 then shows a potential approach to include time-dependencies into the transient computation of queue length distributions at blood collection sites. The research presented in this chapter strongly depends on uniformization.

The first three chapters of Part III: Optimization, provide structured approaches to decrease waiting time at blood collection sites. Chapter 5 proposes a staff scheduling approach that bases required staff levels on expected waiting time at a collection site throughout the day. Its simultaneous utilization of flexible shift lengths ensures that no extra staff are required, while fostering waiting time reductions in most cases. Chapter 6 introduces a Markov Decision Process to reallocate staff members during a collection session, based on the number of donors present at the collection site. Chapter 7 shows how appointments can be introduced for whole blood donors to distribute arrivals of donors more equally over the day, and shows the effects for the other donors at the collection site.

The final chapter of Part III, Chapter 8, proposes a method to improve the inventory management of red blood cells. In the proposed method, the red blood cell unit that is used to fulfill a requested unit, is based on both the age and rarity of the red blood cell units available.

Finally, the thesis will be concluded with a general conclusion in Chapter 10. This will summarize the results of all chapters combined. Both the opportunities to implement the results from this thesis, and the opportunities for future research will be discussed.

CHAPTER 2

Uniformization: Basics, extensions and applications

2.1 Introduction

In this chapter, we will present a computational method to transform continuous time systems, such as blood collection sites, to discrete time systems: Uniformization. The method will be used in several chapters in this thesis. This chapter presents the basics of the method, as well as several extensions and applications. The chapter is meant as an overview of the method, and not all of the extensions and applications are related to the remaining thesis.

Continuous-time Markov chains are widely applicable for modelling practical situations that evolve continuously in time with jumps or changes at specific epochs, with applications in, e.g., telecommunications, computer systems, manufacturing, material handling, inventory theory, maintenance and reliability. Over the last decades, uniformization introduced in [117] has been shown to be a powerful tool for performance analysis of systems modelled by continuous-time Markov chains, see, e.g., [96, 99, 134].

Uniformization, also referred to as randomization, or Jensen's method, transfers continuous-time Markov chains (CTMCs) into discrete-time Markov chains (DTMCs). As a result, for the uniformized chain steady state equations as well as iterative computation of the transient distribution (from discrete-time point to the next discrete-time point) can be applied directly in line with standard DTMCs. For the important case of transient analysis of the CTMC over a finite time horizon, the uniformization approach transfers the CTMC into a discrete Possionian matrix expansion. As this expansion allows for an infinite number of Poisson steps, some form of computational approximation, e.g., by tail or state space truncation, will necessarily be involved, even when the state space itself is finite. As a result, a large number of papers on uniformization in literature is devoted to effective computation of transient performance measures.

The uniformization technique seems to be perceived to be restricted to CTMCs with (i) time-homogeneous and (ii) uniformly bounded transition rates. The first condition is justified for steady state situations, but is less realistic for transient analysis. Transition rates usually remain bounded over finite time intervals, but the second condition can easily be violated in practical situations, for example, it

already fails for the infinite server queue. In addition, in literature, uniformization techniques seem, often, to be applied without a formal justification or an explicit argument for the results to be exact or approximate. This chapter provides an overview of exact and approximate uniformization results beyond time-homogeneous and bounded transition rates.

Uniformization is an appealing technique for performance evaluation of CTMCs as it uses a discrete-time Markov chain to obtain the continuous-time transition matrix. As will be shown in Section 2.3.2, for a conservative and irreducible CTMC X_t with countable state space S and generator $\mathbf{Q} = (q(i,j),\ i,j \in S)$ such that $\sum_{j \neq i} q(i,j) \leq B < \infty$, the transition matrix $\mathbf{P}_t$ with elements $\mathbf{P}_t(i,j) = \mathbb{P}(X_t = j | X_0 = i)$ can be written as:

$$\mathbf{P}_t(i,j) = \sum_{k=0}^{\infty} \frac{(tB)^k}{k!} e^{-tB} \mathbf{P}^k(i,j), \quad i,j \in S,\ t > 0, \tag{2.1}$$

where $\mathbf{P}^k$ is the k-th matrix power of the one-step transition probability matrix $\mathbf{P} = \mathbf{I} + \frac{1}{B}\mathbf{Q}$ of a DTMC, the so-called uniformized Markov chain. Analysis of a DTMC, in general, is much less involved than analysis of a CTMC. As a consequence, uniformization in its standard form (2.1) is often applied to obtain

1. average or stationary results,
2. transient results, and
3. cumulative rewards

for CTMCs.

The uniformized Markov chain has the same transition structure as the CTMC. Therefore, equivalence of *average or stationary results* for the uniformized Markov chain and the CTMC seem to be intuitively obvious. In Section 2.3.3 we will make this explicit showing that both Markov chains have the same generator. Note that also for average results uniformization is numerically appealing since we may obtain these results via iterative computation of $\mathbf{P}^k = \mathbf{P}\mathbf{P}^{k-1}$, $k = 1, 2, 3, \ldots$, whereas for the CTMC we have to solve a possibly large or unbounded system of equations using, e.g., a Gauss-Seidel method [170].

Obtaining *transient results*, such as the explicit distribution $\mathbb{P}(X_t = j | X_0 = i)$ at time t, is, perhaps, the best known application of uniformization. To this end, from (2.1), we may obtain $\mathbb{P}(X_t = j | X_0 = i)$ by iterative computation of $\mathbf{P}^k$ and observing that the CTMC makes k steps until time t according to a Poisson process with rate B of which some steps result in dummy transitions. Interpretation of this result as thinning of the Poisson process with rate B suggest the generalisation of standard uniformization to *exact uniformization for* a CTMC with *time-inhomogeneous transition rates* $\mathbf{Q}_t = (q_t(i,j)\ i,j \in S)$, reflecting, e.g., arrival patterns, or service speed fluctuations. This generalisation will be presented in Section 2.5.

The uniformized Markov chain may also be used to obtain *cumulative rewards*. In Section 2.6 we will first consider the CTMC that incurs reward at rate $r(i)$ while

residing in state i. Uniformization then allows evaluation of the total reward $\mathbf{W}_t$ at time t via the k-step reward $\mathbf{W}^k$ of the uniformized Markov chain by analogy with (2.1). Uniformization for rewards is also most appealing to obtain the average reward for a CTMC in the stationary regime. In particular, the DTMC directly enables use of computational bounds, such as the Odoni bounds that are well-known for Markov decision processes [148]. Uniformization for cumulative rewards may also be extended to the time-inhomogeneous case, as will be illustrated in Section 2.6.4.

To illustrate uniformization beyond CTMCs with bounded transition rates and countable state spaces, we also consider *approximate uniformization for unbounded transition rates* in Section 2.7 and *exact uniformization for continuous state variables* in Section 2.8. The transition rates for the infinite server queue are unbounded. We show that an approximate uniformization technique that introduces a DTMC by analogy for the uniformized Markov chain for states $0, \ldots, N$ and uses the discrete-time Markov jump chain for states $N+1, N+2, \ldots$ yields an approximation that is asymptotically exact for large N. Section 2.7 presents results indicating that this approximate uniformization approach is asymptotically exact for large N for general CTMC with unbounded rates.

Uniformization samples time at Poisson rate and uses a DTMC that makes transitions at the epochs of the Poisson process to evaluate performance measures for CTMCs. For a process with a continuous state space we may also consider uniformization with respect to the continuous state space. As an illustration, Section 2.8 considers a uniformization procedure for stochastic service networks with non-exponential service times. Note that such processes need the residual or spent service times to be included in the state description to have the Markov property. Via the hazard rates of the service times we consider a Poisson process that samples the service times and consider the transitions of the uniformized model that makes transitions at the epochs of this Poisson process. We show that the equilibrium distribution of the uniformized model coincides with that of original process. This result opens a route to new applications of uniformization to continuous state variables.

This chapter is meant as an expository chapter to provide a basis for uniformization and its generalizations, as well as to shed some light on computational issues. Some remarks on numerical evaluation and comparison between uniformization and time-discretization are included in Sections 2.3 – 2.7. First, a brief survey of the literature is included in Section 2.2. In line with literature, Section 2.2 mainly considers numerical approaches to exact and approximate uniformization. The Poissonian expression for the transient probabilities (2.1) includes an infinite Poisson summation, since the number of Poisson epochs in an interval of length t is unbounded. Unless an analytic form can be found for the k-step transition probabilities $\mathbf{P}^k$, for computational purposes a truncation for this Poisson summation is required to evaluate (2.1). There is a vast literature on its numerical consequences, see Section 2.2.1. The state space might be infinite, either through a continuous-state description or, as more common in performance evaluation, through a discrete but enumerable state space, see Section 2.2.5 that indicates that countable state spaces are mainly addressed in the setting of Markov decision processes. Other important cases addressed

in Section 2.2 include unbounded transition rates and steady state detection.

The remainder of this chapter is structured as follows. *Standard uniformization* is addressed in Section 2.3 in a self-contained manner, including embedding in the general setting of continuous-time Markov processes, a formal justification, some intuitive views and different interpretations that form the basis for the generalizations in subsequent sections. Section 2.4 presents a numerical illustration of standard uniformization for a web server application and a comparison with a time-discretization. Sections 2.5 to 2.8 then provide a number of extensions. These sections are set up identically and are in parallel, with theoretical results first, followed by numerical support (excluding Section 2.8). *Exact uniformization for time-inhomogeneous transition rates* is introduced in Section 2.5, and Section 2.6 considers *exact uniformization for reward models*. Section 2.7 presents *approximate uniformization for unbounded transition rates*, and Section 2.8 extends uniformization in time to *exact uniformization for continuous state variables for non-exponential networks*. Although most of the theoretical results in these sections are not new, the aim of the chapter is to Introduce the method of uniformization and possibly stimulate further research into the uniformization method. Finally, Section 2.9 completes the chapter with some remarks on possible further developments of the uniformization technique both in theory and applications.

2.2 Literature

This section provides a brief overview of literature on uniformization highlighting the special cases of uniformization that are addressed in this chapter.

2.2.1 Standard uniformization

Jensen [117] introduced the basic uniformization method, as explained in more detail in section 2.3, in 1953. Grassmann [95] compares uniformization to Runge-Kutta and Liou's method for computing transient distributions of Markovian queueing systems and finds uniformization superior to these methods. Some numerical experiments for (at the time considered large) queueing systems are shown. An implementation of uniformization for computing transient distributions is presented in [94] and extended to compute the waiting time distribution of an M/M/1 queue where the next job to receive service is randomly selected from the queue. Gross and Miller [99] present algorithms to compute unifomization results and some additional transient performance measures of a Markov process, such as expected sojourn time averages and the expected number of events. Motivated by the need to compute delay times and first passage times in queueing networks, Melamed and Jadin [134] discuss a method to bound the time spent in a specified set of states in a CTMC, before moving to another specified set of states. The method is then applied to a tandem queueing network, for which the bounds on the sojourn time are computed. Reibman and Trivedi [157] compare uniformization to both an implicit and explicit numerical solution to the underlying differential equations. Uniformization is shown to be more accurate

at lower computational cost, except for very stiff models, i.e., models where states have out-rates of greatly varying magnitude. For very stiff models, uniformization is outperformed by implicit differential equation solution algorithm.

For Markov reward processes, Reibman et al. [158] compare several approaches. Uniformization is considered an efficient algorithm to obtain transient state probabilities for non-stiff models, while for stiff models implicit solutions to the differential equations are preferred. For the distribution of cumulative rewards, uniformization is again the method of choice, if the model has a low number of distinct reward rates.

2.2.2 Time-inhomogeneous uniformization

Uniformization is well suited to be applied to time inhomogeneous systems. Schwarz et al. [164] recently surveyed the literature on performance analysis of time inhomogeneous queueing systems including time inhomogeneous uniformization. The survey mentions that the method has two major advantages: it can be applied to any Markovian queueing system, and it can be used to compute the entire distribution. In a comparison of uniformization with five other methods for the $M(t)/M/s(t)$ queue by Ingolfsson et al. [111], uniformization is shown to be almost as accurate as an exact differential equation solver, but uses less than half of the computational time. In contrast, some approximations such as the modified offered load approximation [116] may be much faster, but less accurate. Creemers et al. [54] use uniformization to analyze inhomogeneous multi-server queues with phase-type distributed inter arrival, service and abandonment times. Dormuth et al. [71] compare uniformization to the backwards Euler method for a time inhomogeneous single server queue with phase-type distributed service time and shows that both methods perform well. Andreychenko et al. [14] introduce a method for the computation of infinite-state time inhomogeneous CTMCs through uniformization. Their method, similar to an adaptive uniformization technique, for the next time-step only considers states where the majority of the probability mass is located.

A theoretically exact method to determine transient distributions for time inhomogeneous CTMCs is developed in Van Dijk [64], and discussed in more detail in Section 2.5. The work is continued and implemented numerically in [141] and [16]. Rindos et al. [162] suggest a method to convert a time inhomogeneous CTMC in a homogeneous CTMC that may then be analyzed via uniformization.

2.2.3 Steady state detection

Muppala and Trivedi [142] introduce a method to reduce the computational efforts of uniformization. They suggest the use of steady state distributions instead of computing all vector-matrix multiplications if the difference between iterations i and $i-m$ is small enough. The method is demonstrated by applying it to a closed queueing network based on a computer system. Malhotra et al. [128] compare uniformization with steady state detection to a third and second order implicit solution method to the differential equations. The methods are evaluated based on their accuracy and computational cost when solving stiff CTMCs. For mildly stiff models, uniformization

is the method of choice as it has lower computational cost. For very stiff models, the implicit solution method is preferred.

2.2.4 Adaptive uniformization

Adaptive uniformization is introduced by Van Moorsel and Sanders [139]. The uniformization rate under adaptive uniformization is based on the states the process may reach in a particular number of jumps, whereas this rate is based on the complete state space under standard uniformization. Under adaptive uniformization the uniformization rate may be lower, leading to a potential reduction of the computation time. The main computational savings of adaptive uniformization are in limiting the size of the space of states the process may reach, and therefore also in the computational load of the matrix-vector multiplications. Unfortunately, in most cases the distribution of the number of transitions in intervals is not Poisson. Adaptive uniformization is computationally more intricate than standard uniformization. Diener and Sanders [59] numerically compare different adaptive uniformization methods and find that so-called layered uniformization gives the lowest roundoff errors. Depending on the problem and its size, layered uniformization is much faster than standard uniformization. Didier et al. [58] present a faster, although slightly less accurate, version of adaptive uniformization that seems especially useful for biochemical reactions. Adaptive uniformization is most useful if the number of states the process can be in is small, usually a short time after the process started. As this number of states increases with time, the computation time of adaptive uniformization increases, and standard uniformization becomes the faster method. [140] suggests using adaptive uniformization up to some time threshold and then switching to standard uniformization to take advantage of both methods.

In addition, as the uniformization rates are based on the states the process may reach in a finite number of steps, adaptive uniformization may allow to invoke uniformization for systems with unbounded rates, as in demonstrated in Section 2.7.

2.2.5 Unbounded Markov decision processes

Guo et al. [101] survey recent developments for Markov decision processes (MDP). Here uniformization may be invoked to deduce optimal decisions for a CTMC from its DTMC counterpart. However, this is not directly possible for systems with unbounded transition rates. Blok et al. [31] discuss unbounded rates both for discrete-time and continuous-time MDP. Their advised course of action for an continuous-time MDP with unbounded rates is to apply some perturbation and then apply uniformization. Bhulai et al. [27] introduce the first general method, Smoothed Rate Truncation (SRT), for this perturbation that conserves the structural properties of the original model. SRT is based on linear smoothing of unbounded rates to obtain a finite set of recurring states. Section 2.6 considers uniformization to evaluate rewards.

2.3 Standard uniformization

2.3.1 Markov generators

This expository chapter deals with uniformization for continuous-time Markov processes. To put this well-known concept in somewhat wider perspective, let us first briefly present the notion of a generator. Based on the Markovian property of memorylessness, continuous-time Markov processes can be characterized by their infinitesimal generator $\mathbf{A}$, see, e.g., [75, 89]. With a discrete or continuous state represented by x, with bounded or unbounded state space S, and X_t denoting the state of the system at time t, the infinitesimal generator $\mathbf{A}$ is defined as an operator $\mathbf{A}f$ for arbitrary real valued functions $f : S \to \mathbb{R}$, by:

$$\mathbf{A}f(x) = \frac{d}{dt}\left[\mathbb{E}(f(X_t|X_0 = x)) - f(x)\right]. \tag{2.2}$$

As one well-known case, diffusion processes are characterized by:

$$\mathbf{A}f(x) = a(x)\frac{d}{dx}f(x) + \frac{1}{2}\sigma^2(x)\frac{d^2}{dx^2}f(x), \tag{2.3}$$

reflecting a state dependent drift as well as a continuously adjusted Brownian motion component. Such processes might typically be of interest in performance evaluation to model highly random varying arrivals (e.g. Levy input) streams or highly fluctuating service speeds. More common in perfomance modelling - essentially based on an underlying exponential structure - is the generator of a *pure Markov jump process* (see [89]). For arbitrary state space S, the generator of a Markov jump process is characterized by:

$$\mathbf{A}f(x) = \int_S q(x;dy)\left[f(y) - f(x)\right], \tag{2.4}$$

where $q(x, dy)$ represents a transition rate density function for state x:

$$q(x;C) = \lim_{\Delta t \to 0}\frac{1}{\Delta t}\mathbb{P}(X_{\Delta t} \in C \mid X_0 = x), \quad x \notin C. \tag{2.5}$$

Mixtures of (2.3) and (2.4) as Markov jump-diffusion processes are also conceivable. Within queueing theory and the wide application area of performance evaluation, the Markov jump process usually has a discrete state space and is generally referred to as a continuous-time Markov chain (CTMC). In this case, with discrete state space S and real valued functions $f : S \to \mathbb{R}$, the operator representation (2.4) reduces to:

$$\mathbf{A}f(i) = \sum_{j \in S} q(i,j)\left[f(j) - f(i)\right], \quad i \in S. \tag{2.6}$$

In this chapter, as it is meant to be of main interest for system performance evaluation, uniformization will primarily be tailored to the CTMC case (2.6). For the discrete state space CTMC case the operator $\mathbf{A}$ will be identified with the generator matrix $\mathbf{Q}$.

2.3.2 Standard uniformization

Consider a continuous-time, conservative and irreducible Markov chain (CTMC) X_t with countable state space S and transition rates

$$q(i,j), \quad i \neq j,\ i,j \in S,$$

for a transition from a state i into another state j and for $i \in S$

$$-q(i,i) = \sum_{l \neq i} q(i,l).$$

Let $\mathbf{Q}$ be the corresponding matrix of transition rates. We assume these rates to be uniformly bounded (in literature also referred to as uniformizable), i.e., for some finite constant $B < \infty$ and all $i \in S$

$$q(i) = \sum_{j \neq i} q(i,j) \leq B. \tag{2.7}$$

We define the transition probability matrix $\mathbf{P}$ by

$$\mathbf{P}(i,j) = \begin{cases} q(i,j)/B, & j \neq i, \\ 1 - \sum_{l \neq i} q(i,l)/B, & j = i, \end{cases} \tag{2.8}$$

or

$$\mathbf{P} = \mathbf{I} + \frac{1}{B}\mathbf{Q}, \tag{2.9}$$

where B is a uniformization rate that is not required to be equal to the maximum exit rate from any state i, but can be any number satisfying (2.7).

Let $\mathbf{P}_t$ denote the transition matrix of the CTMC with elements $\mathbf{P}_t(i,j) = \mathbb{P}(X_t = j | X_0 = i)$, π_c the steady-state distribution of the CTMC and π_d the steady-state distribution of the DTMC with one-step transition matrix $\mathbf{P}$. The following result was first shown by Jensen [117], and can be found in other references, see, e.g., [96, 99, 134]. It is generally referred to as uniformization or randomization.

Result 3.1 (Standard uniformization) *The steady-state distribution π_c of the CTMC and π_d of the DTMC with one-step transition matrix $\mathbf{P}$ coincide:*

$$\pi_c(i) = \pi_d(i), \quad i \in S.$$

In addition, for all $i,j \in S$ and $t > 0$:

$$\mathbf{P}_t(i,j) = \sum_{k=0}^{\infty} \frac{(tB)^k}{k!} e^{-tB} \mathbf{P}^k(i,j), \tag{2.10}$$

where $\mathbf{P}^k$ represents the k-th matrix power of the one-step transition probability matrix $\mathbf{P}$.

For selfcontainedness, but also as starting point for the generalizations presented in subsequent sections, below we include proofs for the equivalence of the CTMC and its uniformized DTMC. The proof for the steady state case uses basic balance equations. The proof for the transient case uses exponential expansion. We provide an alternative proof for the transient case that it is based on convergence results for processes.

Proof
The steady state equivalence of the steady-state distributions π_c of the original CTMC and π_d of the discrete-time Markov chain with one-step transition matrix $\mathbf{P}$ is straightforward, noting that the steady state distributions π_c and π_d are the unique solution (up to normalization) of the global balance equations for the continuous-time and discrete-time Markov chains:

$$\begin{aligned} 0 &= \pi_c \mathbf{Q}, \\ \pi_d &= \pi_d \mathbf{P}. \end{aligned} \tag{2.11}$$

Substituted in detail, for $j \in S$:

$$\pi_c(j) \sum_{i \neq j} q(j,i) = \sum_{i \neq j} \pi_c(i) q(i,j), \tag{2.12}$$

and

$$\begin{aligned} \pi_d(j) &= \sum_i \pi_d(i) p(i,j) \\ &= \sum_{i \neq j} \pi_d(i) q(i,j) \frac{1}{B} + \pi_d(j) - \pi_d(j) \sum_{i \neq j} q(j,i) \frac{1}{B}. \end{aligned}$$

As the solution of (2.12) is unique up to a multiplicative constant, it must be that $\pi_c = \pi_d$.

The result for the transient case (2.10) can be demonstrated via substitution of (2.9) into the general expression $\mathbf{P}_t = \mathrm{e}^{\mathbf{Q}t}$, also see Interpretation 3.3 below. To this end, observe that

$$\begin{aligned} \mathbf{P}_t &= \mathrm{e}^{\mathbf{Q}t} = \mathrm{e}^{B(\mathbf{P}-\mathbf{I})t} \\ &= \mathrm{e}^{B\mathbf{P}t} \mathrm{e}^{-B\mathbf{I}t} \quad \text{[which is allowed as } \mathrm{e}^{B\mathbf{P}t} \text{ and } \mathrm{e}^{-B\mathbf{I}t} \text{ commute]} \\ &= \sum_k \frac{(-B\mathbf{I}t)^k}{k!} \sum_k \frac{(B\mathbf{P}t)^k}{k!} \\ &= \sum_k \frac{(-Bt)^k}{k!} \mathbf{I} \sum_k \frac{(B\mathbf{P}t)^k}{k!} \\ &= \mathrm{e}^{-Bt} \sum_k \frac{(-Bt)^k}{k!} \mathbf{P}^k, \end{aligned}$$

which concludes the proof. □

Proof via the generator for the transient case
Result 3.1 can also be concluded for the transient case invoking general limit theorems, by showing that:

$$\frac{\mathbf{P}_{\Delta t}(i,j) - \mathbb{1}_{\{j=i\}}}{\Delta t} \to q(i,j) \quad \text{for} \quad \Delta t \to 0 \tag{2.13}$$

in strong convergent sense (that is, uniformly in all i, j) and by applying general results from literature (cf. [75, 88]) which state that (2.13) uniquely determines an underlying stochastic process (in the sense of a probability law on the space of right-continuous sample paths: $D([0,\infty])$. The convergence (2.13) is readily shown by writing:

$$\begin{aligned}\mathbf{P}_{\Delta t}(i,j) = \mathbb{1}_{\{j=i\}}\Big(1 - \sum_{l\neq i} q(i,l)\Delta t + o(\Delta t)\Big) + \\ \mathbb{1}_{\{j\neq i\}}\Big(q(i,j)\Delta t + o(\Delta t)\Big) + o(\Delta t),\end{aligned} \tag{2.14}$$

where a function $f(x) = o(x)$ if $\lim_{x\to 0} f(x)/x = 0$. □

Remark 3.1 (Continuous state case) A similar proof for the transient continuous-state case might be provided by more extended notation. For non-exponential queueing networks, a continuous-state description and corresponding uniformization is presented in Section 2.7. □

Remark 3.2 (Numerical computation) It is often necessary to restrict the range of outcomes of the Poisson distribution for which the probability $\mathbf{P}^k(i,j)$ is calculated. Some approaches are common in literature. One method, referred to as the Fox-Glynn method, introduced in [84], provides a stable algorithm to compute Poisson probabilities. Uniformizaton is the method of choice for the evaluation of the matrix exponential for transient probabilities in CTMCs for Stochastic Model Checking.

Several tools have been introduced with the aim of automating Stochastic Model Checking like Prism [124], Interactive Markov Chains [107] and PEPA [108]. See [21] for a general introduction to model checking. Generally, all these software tools are focused towards the calculation of probability vectors and use the Fox-Glynn method.

A widely used method is scaling and squaring, see, e.g., [147]. Here we observe that we can write

$$\mathbf{P}_{2t} = e^{\mathbf{Q}(2t)} = \left(e^{\mathbf{Q}t}\right)\left(e^{\mathbf{Q}t}\right) = \left(\mathbf{P}_t\right)^2$$

and thus calculate $\mathbf{P}_{t_0}$ for some reasonable value of t_0 and calculate $\mathbf{P}_t$ by successive squaring for large values of t with a limited number of matrix multiplications. In many cases the full matrix is not needed and substantial computational savings can then be obtained using only matrix-vector operations.

An obvious way to evaluate (2.10) is to truncate the sum:

$$\mathbf{P}_t^{(N)}(i,j) \approx \sum_{k=0}^{N} \frac{(tB)^k}{k!} e^{-tB} \mathbf{P}^k(i,j).$$

With $||\cdot||$ the supremum norm $||A|| = \sup_i \sum_j |a_{ij}|$ for any matrix $A = (a_{ij})$:

$$||\mathbf{P}_t - \mathbf{P}_t^{(N)}|| = \sum_{k=N+1}^{\infty} \frac{(tB)^k}{k!} e^{-tB}.$$

As a consequence, for any $t \geq 0$ this truncation converges to $\mathbf{P}_t$ as $N \to \infty$, but

$$\lim_{N\to\infty} \sup_{t\geq 0} ||\mathbf{P}_t - \mathbf{P}_t^{(N)}|| = 1,$$

so convergence is not uniform in t. The approximation performs badly for fixed N and large enough t. If the process has an equilibrium distribution we may use this distribution $\pi = \pi_d$ in the approximation, see [117],

$$\mathbf{P}_t(i,j) \approx \mathbf{P}_t^{(N)}(i,j) = \pi_d(j) + \sum_{k=0}^{N} \frac{(tB)^k}{k!} e^{-tB} (\mathbf{P}^k(i,j) - \pi_d(j)),$$

with truncation error satisfying

$$\lim_{N\to\infty} \sup_{t\geq 0} ||\mathbf{P}_t - \mathbf{P}_t^{(N)}|| = 0,$$

so that the truncation level N can be chosen such that the approximation error has a specified accuracy for all $t \geq 0$. □

2.3.3 Interpretations

The equivalence of the CTMC and the uniformized DTMC has several interpretations that we present below to provide an intuitive explanation of uniformization and as basis for some of the generalisations in subsequent sections.

Interpretation 3.1 (Overrelaxation by dummy jumps) One way to interpret (2.8) is that dummy transitions $i \to i$ are introduced as possible events, while the holding times up to a next event (which may include dummy events) have been uniformized to be the same for all states i to be exponential with uniformization rate B. Given that an event occurs a transition takes place proportional to the transition rates at that instant. □

Interpretation 3.2 (Poisson thinning) Another way to look at uniformization is based on the fact that events generated by a Poisson process can be seen as a series of times drawn from a continuous uniform distribution of the time horizon. Hence,

if k events have taken place, the epochs of these events are spread according to a k-fold uniform distribution. Once these epochs are sampled, the conditional jump probabilities are proportional to the corresponding rates at these epochs. This will be used in section 2.5. □

Interpretation 3.3 (Same generator – backward and forward equations) A more technical description considers uniformization as an equivalent approach to the original CTMC through the generator. As argued in Section 2.3.1, a generator determines a process. By analogy with the standard exponential function which is uniquely determined by its exponential coefficient μ through its derivative, $\frac{d}{dt}\left[e^{\mu t}\right] = \mu\left[e^{\mu t}\right]$, the transition probability matrix $\mathbf{P}_t$ with elements $\mathbf{P}_t(i,j)$ for transition from state i into state j, over a period of time t is determined by its generator through

$$\mathbf{P}_t = e^{\mathbf{Q}t} \tag{2.15}$$

or through the *backward Kolmogorov equations* (see, e.g., [126, p. 311])

$$\frac{d}{dt}\mathbf{P}_t = \mathbf{Q}\mathbf{P}_t, \quad t > 0, \tag{2.16}$$

or in detailed form, for $i, j \in S$,

$$\frac{d}{dt}\mathbf{P}_t(i,j) = \sum_l q(i,l)\mathbf{P}_t(l,j), \quad t > 0. \tag{2.17}$$

Introducing a DTMC with transition matrix

$$\mathbf{P} = \mathbf{I} + \Delta\mathbf{Q}, \quad \text{with } \Delta \leq 1/B \tag{2.18}$$

and regarding Δ as a time-increment, the discrete-time analog of (2.16) is

$$\left[\mathbf{P}^{k+1} - \mathbf{P}^k\right]/\Delta = \left[\mathbf{P} - \mathbf{I}\right]\mathbf{P}^k/\Delta \overset{(2.18)}{=} \mathbf{Q}\mathbf{P}^k, \tag{2.19}$$

which implies that the generator of the CTMC and DTMC, given in (2.16) and (2.19), are identical and given by $\mathbf{Q}$.

For a uniformizable CTMC the solution $\mathbf{P}_t$ of the backward Kolmogorov equations coincides with that of the *forward Kolmogorov equations* that may also be directly obtained from the generator (2.15) (see, e.g., [49, Theorem II.18.3], [126, p. 311]):

$$\frac{d}{dt}\mathbf{P}_t = \mathbf{P}_t\mathbf{Q}, \quad t > 0, \tag{2.20}$$

or in detailed form, for $i, j \in S$,

$$\frac{d}{dt}\mathbf{P}_t(i,j) = \sum_l \mathbf{P}_t(i,l)q(l,j), \quad t > 0. \tag{2.21}$$

For the discrete-time analogon of the forward Kolmogorov equations (2.20) consider a DTMC $X_{d,t}$ at times $t = k\Delta$, with one-step transition matrix $\mathbf{P}$ and transition matrix $\mathbf{P}_{d,t}$ over time t, then

$$\frac{[\mathbf{P}_{d,t+\Delta} - \mathbf{P}_{d,t}]}{\Delta} = \frac{[\mathbf{P}^{k+1} - \mathbf{P}^k]}{\Delta} = \frac{\mathbf{P}^k[\mathbf{P} - \mathbf{I}]}{\Delta} = \frac{\mathbf{P}_{d,t}\,[\mathbf{P} - \mathbf{I}]}{\Delta} \overset{(2.18)}{=} \mathbf{P}_{d,t}\mathbf{Q},$$

which implies that the discrete-time analogon also has generator $\mathbf{Q}$. □

Interpretation 3.4 (No order terms – time-discretization) For Δ sufficiently small, instead of using the continuous-time transition probabilities

$$\mathbf{P}_\Delta(i,j) = q(i,j)\Delta + o(\Delta), \tag{2.22}$$

we might simply ignore the order terms in Δ and only consider the transition terms directly proportional to the length of time Δ, i.e., the transition rates $q(i,j)$. Accordingly, as a transition matrix over a fixed time length $h \leq 1/B$, we might use:

$$\mathbf{P} = \mathbf{I} + h\mathbf{Q}. \tag{2.23}$$

This might be used in a time-discretization approach, with discrete-time analog X_{kh} of the continuous-time process X_t.

As the generators for the continuous-time and discrete-time processes are equal, average performance measures should also be equal per unit time. For finite time or transient measures, however, differences will still appear and accumulate due to time discrepancies. Results will thus be approximate, as will be illustrated in Section 2.4.3.2. □

Interpretation 3.5 (Global balance equations) For computational purposes the forward Kolmogorov equations (2.21) are more common as they allow for a straightforward interpretation as balance equations balancing probability flux into and out of the states. Recall that if $\mathbf{Q}$ is conservative, (2.21) then reads

$$\frac{d}{dt}\mathbf{P}_t(i,j) = \sum_{l \neq j} \left\{\mathbf{P}_t(i,l)q(l,j) - \mathbf{P}_t(i,j)q(j,l)\right\}, \quad t > 0.$$

In particular, the global balance equations are well-known as

$$\sum_{l \neq j} \pi_c(j)q(j,l) = \sum_{l \neq j} \pi_c(l)q(l,j)$$

for all j, independent of initial state i at time 0.

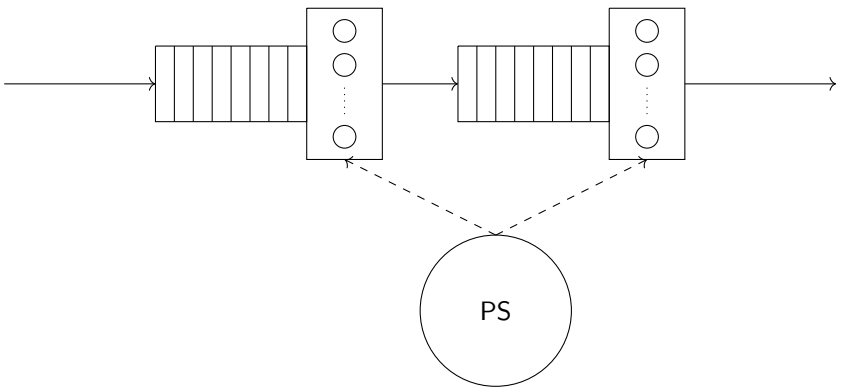
The backward Kolmogorov equations have a dynamic programming structure, and turn out to be most useful to describe rewards. In the following discussion, we will use the forward Kolmogorov equations to determine state probabilities (see Sections 2.5 and 2.7) and the backward Kolmogorov equations to determine rewards (see Section 2.6). □

2.4 Example and numerical results

2.4.1 Web server tandem model

Consider the web server tandem model described in [76, 179, 180] depicted in Figure 2.1. The model is based on requests arriving at a web server. The system is a two-station tandem queue, where all jobs in service share one common server. Station $i \in \{1, 2\}$ has limit c_i on the number of jobs that can be in service simultaneously - these numbers represent the maximum number of so-called threads.

Figure 2.1 The web server tandem model (picture is based on [180]). PS indicates a processor sharing server. It serves a maximum of c_1 and c_2 jobs at the first and second stations respectively.



Let $\Phi_i(n_1, n_2)$ be the proportion of the total service speed attributed to station $i \in \{1, 2\}$ in state (n_1, n_2). The transition rates are:

$$q((n_1, n_2), (n_1', n_2')) = \begin{cases} \lambda, & (n_1', n_2') = (n_1 + 1, n_2), \\ \mu_1 \Phi_1(n_1, n_2), & (n_1', n_2') = (n_1 - 1, n_2 + 1), \\ \mu_2 \Phi_2(n_1, n_2), & (n_1', n_2') = (n_1, n_2 - 1), \end{cases} \tag{2.24}$$

where λ is the arrival rate and μ_1 and μ_2 are the maximum service rates of station 1 and 2, respectively. For $c_1 = c_2 = \infty$ the system can be seen as a standard processor sharing system with

$$\Phi_i(n_1, n_2) = \frac{n_i}{n_1 + n_2}, \quad i = 1, 2.$$

The system can then be shown to have a product form steady state distribution, see, e.g., [24, 47, 180], with $\pi(0, 0)$ the normalizing constant:

$$\pi(n_1, n_2) = \pi(0, 0) \lambda^{n_1 + n_2} \binom{n_1 + n_2}{n_1} \prod_i \left[\frac{1}{\mu_i} \right]^{n_i}, \quad n_1 \geq 0, n_2 \geq 0. \tag{2.25}$$

From this expression we readily compute performance measures of the form

$$G = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \pi(n_1, n_2) g(n_1, n_2), \tag{2.26}$$

where G may represent mean queue lengths, queue length tails, excess probabilities or the effective service speed of either of the two stations. As an additional appealing feature, in this particular processor sharing case, this product form can be shown to be insensitive, i.e., not to depend on the service distributions other than via their means (e.g. [47, 172]).

For finite numbers of threads c_1 and c_2, however, an analytic result is much harder to obtain. For example, it is shown in [180] that for service sharing specified by:

$$\Phi_i(n_1, n_2) = \frac{\min\{n_i, c_i\}}{\min\{n_1, c_1\} + \min\{n_2, c_2\}} \tag{2.27}$$

a product form expression cannot be obtained. Numerical computation to evaluate performance measures is thus of interest, where uniformization will then be very useful.

2.4.2 Numerical evaluation

Equation (2.10) can be numerically evaluated by truncating the sum at level K. It can then be used to approximate the queue length distribution as:

$$\pi_t = \sum_{k=0}^{K} \pi_0 \frac{(tB)^k}{k!} e^{-tB} \mathbf{P}^k, \quad t \geq 0, \tag{2.28}$$

where π_0 is the intial distribution. This, in turn, can be used to compute performance measures.

The algorithm used in the remainder of this section does not use matrix powers. First define $\pi^{(k)}$ as the queue length distribution after exactly k transitions, and $\pi_t^{(k)}$ as the not normalized queue length distribution if at most k transitions take place during time t. Initialize these two as follows:

$$\begin{aligned} \pi_t^{(0)} &= \mathbf{0}, \\ \pi^{(0)} &= \pi_0. \end{aligned} \tag{2.29}$$

We may now iteratively compute, for $k = 1, 2, \ldots$,

$$\pi^{(k)} = \pi^{(k-1)} \mathbf{P}, \tag{2.30}$$

$$\pi_t^{(k)} = \pi_t^{(k-1)} + \frac{(tB)^k}{k!} e^{-tB} \pi^{(k)}, \tag{2.31}$$

converging to π_t for $k \to \infty$. As this infinite limit is impossible in numerical computations, the total number of iterations will be limited by K that we have defined as the smallest K that satisfies:

$$1 - \sum_{k=0}^{K} \frac{(tB)^k}{k!} e^{-tB} < 10^{-6}. \tag{2.32}$$

2.4.3 Results

This section contains numerical results illustrating the speed of convergence to steady state and comparison of uniformization and time-discretization. For the numerical computations, we will set a maximum of N_1 and N_2 to the number of jobs present in the first and the second station, respectively.

2.4.3.1 Convergence to steady state for varying service rates

For the first computational experiment, the web server tandem model was started empty. The maximum number of jobs in the first and second station are $N_1 = N_2 = 99$. The maximum number of available threads c_1 and c_2 are set to 5. The arrival rate $\lambda = 1$, while the service rate was varied. The transient queue length distribution was determined until the steady state was reached, where we consider the system to have reached steady state if the difference between two distributions that are one time-step apart is sufficiently small:

$$\|\pi_{t-1} - \pi_t\| < 10^{-6}. \tag{2.33}$$

This condition was checked after every time-unit, starting with $t = 0$. The average number of jobs in the system in steady state is given in Table 2.1, and the time at which the steady state was reached is shown in Table 2.2. Note that the steady state for $\mu_1 = \mu_2 = 0.5$ was not reached within 10000 time-units, the maximum number of time-units for this numerical experiment.

In the case that the queue is unstable, the steady state is, of course, determined by the maximum number of jobs allowed in the system. If queue 1 is unstable, the second queue might become stable due to the fact that the effective arrival rate to the second station is the service rate at the first station. This leads to around 100 jobs in the system, as the first station is full and the second station can directly serve the jobs arriving from the first station. If the second station is unstable, the second station will become full, and as a result block jobs arriving from the first station, causing this station to also become full, resulting in around 200 jobs in the system. The time until the stationary state is reached is the longest if the second queue has

Table 2.1 Average number of jobs present in the system. * indicates the queue is unstable, ** indicates the queue i is unstable if the other queue j has service speed $\mu_j \leq 2$ and *** indicates queue 2 is stable if the service speed is $\mu_1 < \mu_2$.

		μ_2					
		0.5***	1***	2** ***	3	5	10
μ_1	0.5*	NR	101.48	99.808	99.264	98.788	98.4
	1*	196.64	147.87	100.15	98.31	95.706	90.185
	2**	197.28	195.91	74.526	4.8025	2.3028	1.4934
	3	197.45	196.45	5.153	1.9995	1.1414	0.7643
	5	197.52	196.62	2.3473	1.1437	0.66666	0.42856
	10	197.56	196.44	1.5013	0.76481	0.42857	0.25

Table 2.2 Time units until the steady state is reached. * indicates the queue is unstable, ** indicates the queue i is unstable if the other queue j has service speed $\mu_j \leq 2$ and *** indicates queue 2 is stable if the service speed is $\mu_1 < \mu_2$.

		μ_2					
		0.5***	1***	2** ***	3	5	10
μ_1	0.5*	NR	249	286	308	330	350
	1*	725	7208	577	828	1352	2594
	2**	400	692	5579	327	113	65
	3	431	956	326	84	38	24
	5	461	1565	113	38	18	11
	10	490	3247	65	24	11	5

an effective in-rate equal to the out-rate, i.e., if $\mu_1 = \mu_2 \leq 0.5\lambda$. In all other cases, either one of the queues quickly becomes saturated, or the system moves to a stable steady state.

2.4.3.2 Uniformization and time-discretization

Following Interpretation 3.2, under uniformization the process can be interpreted to make a Poisson number of steps until time T and these steps are uniformly distributed over the interval $(0, T)$. Following Interpretation 3.3, the distribution can also be approximated using time-discretization, that splits the time until the time horizon T into time steps with length h defined as:

$$0 \leq h = \frac{1}{\beta B} \leq \frac{1}{B}. \tag{2.34}$$

For this section, assume $B = \max_{i \in S} q(i)$. At each time step of length h one transition takes place. However, this can be a dummy transition where the state does not change. The transition matrix $\mathbf{P}_d$ for time-discretization is defined as:

$$\mathbf{P}_d = \mathbf{I} + h\mathbf{Q} \tag{2.35}$$

and the queue length distribution computed with time-discretization $\pi_{d,t}$ at time t can be initialized and computed as

$$\begin{aligned} \pi_{d,0} &= \pi_0, \\ \pi_{d,t+h} &= \pi_{d,t}\mathbf{P}_d. \end{aligned} \tag{2.36}$$

In a formal sense, time-discretization can be interpreted in either of two ways: i) by splitting the time period T in fixed intervals of length h or ii) as the distribution after an Erlang distributed time period with T/h exponential phases each with rate $1/h$. This gives the Erlang distribution an expected length of T and a variance of Th. So, since $\lim_{h\to 0} Th = 0$, time-discretization is an exact method for $h \to 0$.

We have compared time-discretization and uniformization in Table 2.3 for the web server tandem model starting empty. Similar to the previous section, we have used arrival rate $\lambda = 1$ and $c_1 = c_2 = 5$. For this experiment, the service rates have been fixed to $\mu_1 = \mu_2 = 2.5$. The maximum number of jobs at both stations has been set to $N_1 = N_2 = 999$. To show the difference between uniformization and time-discretization, the approaches are compared at $T = 1$. Table 2.3 shows the Euclidian distance between the queue length distributions π and $\pi^{(disc)}$, and the average queue length for both methods, for different values of β, varying between 1 and 100. Table 2.3 also shows the time to compute the distribution at time $T = 1$.

Table 2.3 Uniformization compared to time-discretization.

β	1	2	5	10	20	50	100
$\|\pi^{(disc)} - \pi\|$	0.0433	0.0162	0.0066	0.0033	0.0016	0.0007	0.0003
Average queue time-discretization	0.7813	0.7681	0.7547	0.7505	0.7484	0.7471	0.7467
Average queue uniformization	0.7462	0.7462	0.7462	0.7462	0.7462	0.7462	0.7462
Computational time time-discretization	0.17	0.26	0.46	0.84	1.68	3.71	7.06
Computational time, uniformization	0.21	0.21	0.21	0.21	0.21	0.21	0.21

As can be seen in Table 2.3, time-discretization is faster for very low values of β, but the distance between the distributions is substantial. As β increases, the average queue length of time-discretization converges to the result of uniformization, and the distance between both queue length distributions converges to 0. However, the computational cost for high precision of time-discretization is substantial.

2.5 Exact uniformization for time-inhomogeneous transition rates

Performance measures are often of interest over a finite period of time during which the system parameters such as arrival rates, service speeds, service availabilities and

capacities might be varying. The period may represent some specific time interval, such as a day or a user session, during which arrivals follow some pattern, e.g. a bursty start or a peaky ending. Rather than using an average parameter estimate, it is more realistic to use time-inhomogeneous performance estimates. One pragmatic way to do so would be to simply distinguish different time segments at which parameters are assumed constant and apply uniformization iteratively over these segments, as will be illustrated in Section 2.5.2. This section first considers general time-inhomogeneous uniformization in Section 2.5.1.

2.5.1 General time-inhomogeneous uniformization

Consider a time-inhomogeneous CTMC with transition rates

$$q_t(i,j) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathbb{P}(X_{t+\Delta t} = j \mid X_t = i), \quad j \neq i, t \geq 0, \tag{2.37}$$

where these rates are assumed to be right-continuous in t for any fixed i, j. Furthermore, for given fixed Z assume that for some constant $B < \infty$, all $i \in S$ and $t \leq Z$:

$$q_t(i) = \sum_{j \neq i} q_t(i,j) \leq B. \tag{2.38}$$

Let $\mathbf{P}_{s,t}(i,j) = \mathbb{P}(X_t = j \mid X_s = i)$ denote the transition probability to observe the system in state j at time t given that it was in state i at time s. Furthermore, for all $t \leq Z$ define the uniformized transition matrix $\mathbf{M}_t$ similar to (2.8) by:

$$\mathbf{M}_t(i,j) = \begin{cases} q_t(i,j)/B, & j \neq i, \\ 1 - \sum_{l \neq i} q_t(i,l)/B, & j = i. \end{cases} \tag{2.39}$$

The following result can then be proven in line with Interpretation 3.2. The proof is based on the existence of a unique semigroup that satisfies strong regularity infinitesimal conditions in combination with a so-called minimal construction for Markov jump processes (cf. [88]).

Result 5.1 (Time-inhomogeneous uniformization) *For all $s \leq t \leq Z$ and all $i, j \in S$:*

$$\begin{aligned} \mathbf{P}_{s,t}(i,j) = \sum_{k=0}^{\infty} \frac{(t-s)^k B^k}{k!} e^{(t-s)B} \\ \int_{\{t_1 \leq t_2 \leq \cdots \leq t_k\}}^{t} \int_{s}^{t} \cdots \int_{s}^{t} \mathbf{M}_{t_1} \mathbf{M}_{t_2} \ldots \mathbf{M}_{t_k}(i,j)\, d\bar{\mathbf{H}}(t_1, t_2, \ldots, t_k), \end{aligned} \tag{2.40}$$

where $d\bar{\mathbf{H}}(t_1, t_2, \ldots, t_k)$ *is the density of a* k*-dimensional uniform distribution of the order statistics* $t_1 \leq t_2 \leq \cdots \leq t_k$ *at* $[s,t] \times [s,t] \times \cdots \times [s,t] \subset \mathbb{R}^k$ *and* $\mathbf{M}_{t_1}\mathbf{M}_{t_2}\ldots\mathbf{M}_{t_k}$ *is the standard matrix product of the transition matrices from expression* (2.39) *at times* $t_1, t_2, \ldots, t_k$*.*

Proof
Similar to the alternative proof via the generator for the transient case of Result 3.1 for the time-homogeneous case, the proof follows by showing, for both the original CTMC and the uniformized Markov chain, that

$$\begin{aligned} &\mathbf{P}_{t,t+\Delta t}(i,j)/\Delta t \to q_t(i,j), && j \neq i, \\ &\left(\mathbf{P}_{t,t+\Delta t}(i,i) - 1\right)/\Delta t \to -\sum_{j\neq i} q_t(i,j), && j = i. \end{aligned} \tag{2.41}$$

Original CTMC: By standard construction for CTMCs, the transition probabilities can be constructed iteratively (the so-called minimal process construction):

$$\begin{aligned} \mathbf{P}_{s,t}(i,j) &= \sum_{k=0}^{\infty} \mathbf{P}^k_{s,t}(i,j) \\ \mathbf{P}^0_{s,t}(i,j) &= \mathbb{1}_{\{j=i\}} \exp\left[-\int_s^t q_u(i)\, du\right] \\ \mathbf{P}^{k+1}_{s,t}(i,j) &= \int_s^t \exp\left[-\int_s^v q_u(i)\, du\right] \left[\sum_{m\in S} q_v(i,m)\mathbf{P}^k_{v,t}(m,j)\right] dv, \end{aligned} \tag{2.42}$$

from which the convergence (2.41) is readily verified.

Uniformized Markov chain: By analogy, we may define the 'uniformized' continuous-time Markov chain through transition matrices from time s to time t by the construction:

$$\begin{aligned} \mathbf{U}_{s,t}(i,j) &= \sum_{k=0}^{\infty} \mathbf{U}^k_{s,t}(i,j) \\ \mathbf{U}^0_{s,t}(i,j) &= \mathbb{1}_{\{j=i\}} \exp\left[-(t-s)B\right] \\ \mathbf{U}^{k+1}_{s,t}(i,j) &= \int_s^t \exp\left[-(v-s)B\right] B \left[\sum_m \mathbf{M}_v(i,m)\mathbf{U}^k_{v,t}(m,j)\right] dv, \end{aligned} \tag{2.43}$$

Then, on the one hand, by straightforward calculus we may show that

$$\left(\mathbf{U}^0_{s,s+\Delta s}(i,j) + \mathbf{U}^1_{s,s+\Delta s}(i,j)\right)/\Delta s \to \begin{cases} q_s(i,j), & j \neq i, \\ 1 - \sum_{l\neq i} q_s(i,j), & j = i, \end{cases} \tag{2.44}$$

while, on the other hand,

$$\begin{aligned}
\mathbf{U}_{s,t}^k &= \int_s^t e^{-B(v_1-s)} B\mathbf{M}_{v_1} \mathbf{U}_{v_1,t}^{k-1}\, dv_1 \\
&= \int_s^t e^{-B(v_1-s)} B\mathbf{M}_{v_1} \left[\int_{v_1}^t e^{B(v_2-v_1)} B\mathbf{M}_{v_2} \mathbf{U}_{v_2,t}^{k-2}\, dv_2 \right] dv_1 \\
&= \int_s^t \int_{v_1}^t \dots \int_{v_{k-1}}^t \big[B^k e^{-B(v_1-s)} e^{-B(v_2-v_1)} \dots e^{-B(v_k-v_{k-1})} e^{-B(t-v_k)} \\
&\quad \mathbf{M}_{v_1}\mathbf{M}_{v_2}\dots\mathbf{M}_{v_k} \big]\, dv_k \dots\, dv_2\, dv_1 \\
&= B^k e^{-B(t-s)} \int_s^t \int_{v_1}^t \dots \int_{v_{k-1}}^t \mathbf{M}_{v_1}\mathbf{M}_{v_2}\dots\mathbf{M}_{v_k}\, dv_k \dots\, dv_2\, dv_1 \\
&= B^k e^{-B(t-s)} \frac{(t-s)^k}{k!} \int_s^t \int_{t_1}^t \dots \int_{t_{k-1}}^t \mathbf{M}_{t_1}\mathbf{M}_{t_2}\dots\mathbf{M}_{t_k}\, d\bar{\mathbf{H}}(t_1,t_2,\dots,t_k).
\end{aligned} \tag{2.45}$$

By combining these results, (2.42) and (2.45), with sufficient uniqueness theorems (cf. [75, 88]), i.e., the uniqueness of a corresponding semigroup of transition probabilities concluded, see [88, p. 347-353, p. 364-366], the proof is completed. □

Remark 5.1 Expression (2.40) can be simplified by substituting

$$d\bar{\mathbf{H}}(t_1,t_2,\dots,t_k) = \frac{k!}{(t-s)^k}\, dt_1 \dots\, dt_k. \tag{2.46}$$

The form (2.40), however, directly corresponds to Interpretation 3.2 and the form (2.10) in the homogeneous case. Clearly, (2.40) reduces to (2.10) if $\mathbf{M}_t = \mathbf{P}$ for all t. □

Remark 5.2 (Truncation) By analogy with Remark 3.2, with $\mathbf{P}_{s,t}^{(N)}$ the truncated version of (2.40) at level $k = N$ and with $\|.\|$ the standard supremum norm:

$$\|\mathbf{P}_{s,t}^{(N)} - \mathbf{P}_{s,t}\| \le \sum_{k=N+1}^{\infty} \frac{(t-s)B^k}{k!} e^{-(t-s)B} \le \frac{(t-s)^N B^N}{N!}, \tag{2.47}$$

that may be used for $t-s$ sufficiently small. □

Remark 5.3 (Discrete approximation) Expression (2.40) can be impractical as it requires storage of a continuum of matrices and to perform integration. A discrete-grid approximation can be used where we assume that the transition rate matrix $\mathbf{Q}_t = (q_t(i,j))$, with diagonal elements $-q_t(i)$, satisfies a Lipschitz grid condition:

$$\|\mathbf{Q}_{nh+\Delta t} - \mathbf{Q}_{nh}\| \le \Delta t K \qquad \text{for all } \Delta t \le h \text{ and } nh \le Z, \tag{2.48}$$

where h is some fixed gridsize $h \leq 1/B$ and K some constant. Let $n_i = [t_i h^{-1}]$ where $[x]$ is the integer such that $[x] \leq x < [x]+1$, and denote by $\bar{\mathbf{H}}(n_1, n_2, \ldots, n_k)$ the probability mass function of the order statistics $n_1 \leq n_2 \leq \cdots \leq n_k$ of a k-dimensional uniform distribution at $\{n, n+1, \ldots, m-1\}^k$, where $n = [sh^{-1}]$ and $m = [th^{-1}]$. Now let $\mathbf{P}^h_{s,t}$ be the discrete-grid approximation matrix defined by

$$\begin{aligned}\mathbf{P}^h_{s,t} = &\sum_{k=0}^{\infty} \frac{(t-s)^k B^k}{k!} e^{-(t-s)B} \\ &\sum_{n_1=n}^{m-1} \sum_{n_2=n_1}^{m-1} \cdots \sum_{n_k=n_{k-1}}^{m-1} [\mathbf{M}_{n_1 h} \mathbf{M}_{n_2 h} \ldots \mathbf{M}_{n_k h}]\, \bar{\mathbf{H}}(n_1, n_2, \ldots, n_k).\end{aligned} \tag{2.49}$$

Then

$$\|\mathbf{P}^h_{s,t} - \mathbf{P}_{s,t}\| \leq h(t-s)K. \tag{2.50}$$

□

Remark 5.4 (Computation by simulation) Monte-Carlo simulation can be used to evaluate (2.40) as follows. First truncate (2.40) at level L. Next compute the Poisson probabilities for each $k \leq L$. For fixed k now do the following:

1. Generate k uniform random numbers $x_i \in [s, t]$.
2. Take their order statistics $t_1 = \bar{x}_1 \leq t_2 = \bar{x}_2 \leq \cdots \leq t_k = \bar{x}_k$.
3. Compute (approximate or simulate) the matrix product: $\mathbf{M}_{t_1} \mathbf{M}_{t_2} \ldots \mathbf{M}_{t_k}$
4. Repeat Step 1 – Step 3 for a prescribed number of times and compute sample averages. □

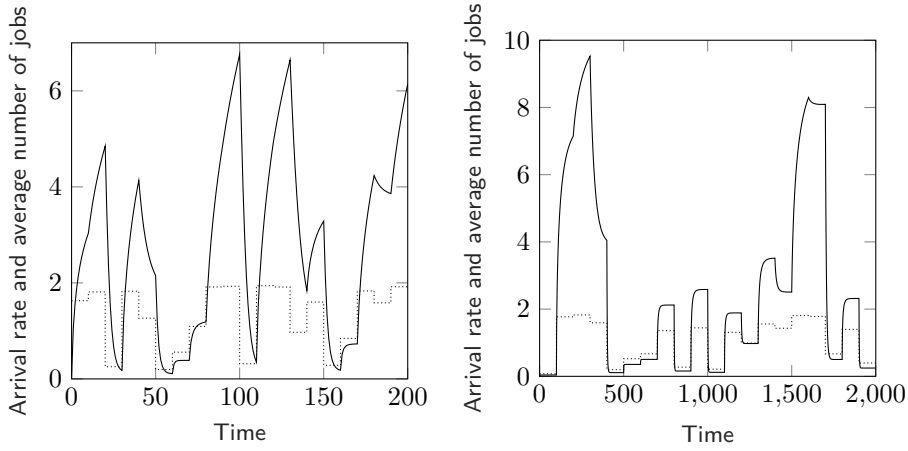
2.5.2 Piece-wise constant transition rates

A practical approach to time-inhomogeneous systems is the piece-wise time-independent approximation, i.e., with transition rates constant over time-intervals $(t_1, t_2), (t_2, t_3), \ldots, (t_{n-1}, t_n)$. In this case, standard uniformization can be applied to each of these time-intervals (t_l, t_{l+1}). This approach is the basis for Chapter 4, and a detailed description of the approach and its application to blood collection sites can be found in Chapter 4. Here, we will summarize the approach, and show the application to the web server tandem model.

For every interval (t_l, t_{l+1}), start by computing $\mathbf{P}_{t_1,t_2}$ by (2.8) using the time-dependent transition rates $q_{t_l,t_{l+1}}(i,j)$. Given some initial distribution π_0, $\pi_{t_{l+1}}$ can be iteratively computed by:

$$\pi_{t_{l+1}} = \sum_{k=0}^{K} \pi_{t_l} \mathbf{P}^k_{t_l,t_{l+1}} \frac{(B(t_{l+1}-t_l))^k}{k!} e^{-B(t_{l+1}-t_l)}, \tag{2.51}$$

Figure 2.2 Average number of jobs in the web server tandem model (solid line) for changing arrival rate (dotted line) for different intervals between arrival rate changes.



(a) The arrival rate changes every 10 time units.

(b) The arrival rate changes every 100 time units.

where K can be determined through (2.32). For the computation of (2.51), equations (2.29), (2.30) and (2.31) can again be applied to compute $\pi_{t_{l+1}}$ without the need for matrix products.

The results of this approach are depicted in Figures 2.2a and 2.2b for the web server tandem model. The maximum number of threads are set to $c_1 = c_2 = 5$, and the service rates are set to $\mu_1 = \mu_2 = 4$. The maximum number of jobs per station is set to $N_1 = N_2 = 99$. The arrival rate changes over time, and is drawn from a uniform distribution between 0 and 2. Both queues start empty. For Figure 2.2a, the arrival rate changes every 10 time units, and for Figure 2.2b the arrival rate changes every 100 time units. The figures show the average number of jobs in the system (solid line), and the arrival rate (dotted line).

The system is unstable for $\lambda_t = 2$, but as we are drawing from a continuous distribution, $\lambda_t < 2$ for all t.

2.6 Exact uniformization for reward models

For applied purposes one is generally interested in, or satisfied with, just one or a limited number of special performance measures, like a workload, a system or server utilization, a loss or congestion percentage, another threshold measure or some idleness or starvation probability, rather than the full probability distribution at the expense of high computational costs. This is where the concept of expected rewards emerges. It is to be kept in mind though that a different reward computation will be required for each different measure.

By using some appropriate reward rate function, the computation of the expected

cumulative reward can be straightforward and far more efficient by just using the uniformization matrix. In fact, as shown in Section 2.6.4, in its approximate discrete-time version we can directly incorporate the time-inhomogeneous case. In addition, for the average case even computational bounds can be kept track of, as will be shown in Section 2.6.5.1.

But first, as for the inhomogeneous case, let us first respond to the more theoretical question whether the principle of uniformization also remains exact for such measures. Again, as shown in Section 2.6.1, the answer is affirmative.

The preceding sections presented uniformizaton to evaluate the time-dependent transition matrix with elements $\mathbf{P}_t(i,j) = \mathbb{P}(X_t = j|X_0 = i)$ from which performance measures may be obtained. This section presents a direct uniformization approach to obtain performance measures from a reward structure. We first give a full description of the time-homogeneous case in Sections 2.6.1 and 2.6.2 followed by a numerical illustration in Section 2.6.3. Finally, in Sections 2.6.4 and 2.6.5 we cover the computationally more pragmatic time-discretized and more general time-inhomogeneous approximation, along with an approximate error bound statement.

2.6.1 Reward structure

Uniformization as presented in Section 2.3.2 may be used, for example, to keep track of state visits to obtain the average reward or absorption probabilities (see, e.g., [96, 99, 134]). For sojourn times or cumulative rewards a more direct uniformization approach can be followed. By analogy with uniformization for the state distribution presented in Section 2.3.2, below we present a direct uniformization for rewards.To this end, let

$$\begin{aligned} &r(i) && \text{:be a reward rate in state } i \text{ and} \\ &\mathbf{H}_s f(i) = \sum_j \mathbf{P}_s(i,j) r(j) && \text{:be the expected reward at time } s \text{ given} \\ & && \text{initial state } i. \end{aligned} \tag{2.52}$$

In this section we consider the expected cumulative reward function $\mathbf{W}_T$ over a finite time period $[0,T]$, where for all $t \geq 0$ the function $\mathbf{W}_t$ is defined by:

$$\begin{aligned} \mathbf{W}_t &= \int_0^t \mathbf{H}_s r \, ds \qquad \text{or more detailed:} \\ \mathbf{W}_t(i) &= \int_0^t \mathbf{H}_s r(i) \, ds = \mathbb{E}\left[\int_0^t r(X_s) \, ds | X_0 = i \right]. \end{aligned} \tag{2.53}$$

Different cumulative performance measures can be covered such as in the web server tandem example from Section 2.4, with $i = (n_1, n_2)$. A number of reward functions and their resulting performance measure are, for example:

$r(n_1, n_2)$	$= n_2$	the mean queue length at station 2
$r(n_1, n_2)$	$= I(n_2 > c_2)$	the total excess load for station 2
$r(n_1, n_2)$	$= \Phi_2(n_1, n_2)$	the effective service rate from station 2
$r(n_1, n_2)$	$= I(n_2 < c_2)$	the time until the level c_2 is reached in a modified process that absorbs at $n_2 = c_2$

From (2.53) we can represent the total reward function $\mathbf{W}_t$ by:

$$\frac{d}{dt}\mathbf{W}_t = r + \mathbf{Q}\mathbf{W}_t, \quad t \geq 0. \tag{2.54}$$

A way to look at this equation is to consider the extended operator $\bar{\mathbf{Q}}$:

$$\frac{d}{dt}\mathbf{W}_t = \bar{\mathbf{Q}}\mathbf{W}_t, \quad \text{with } \bar{\mathbf{Q}}f = r + \mathbf{Q}f \tag{2.55}$$

The relations (2.53) and (2.55) suggest two possible approaches for computation in line with Section 2.3. Again, a straightforward one by time-discretization, which will be approximative, or by an extended version of uniformization, which might be exact. These will be outlined in the next two subsections.

2.6.2 Uniformization for time-homogeneous reward processes

Let $\mathbf{P}$ the uniformization matrix (2.8) and define the discrete-time expectation operator $\mathbf{H}$ by, for $i \in S$,

$$\begin{aligned} \mathbf{H}^0 f(i) &= f(i), \\ \mathbf{H} f(i) &= \sum_j \mathbf{P}(i,j) f(j), \\ \mathbf{H}^k f(i) &= \sum_j \mathbf{P}^k(i,j) f(j), \end{aligned} \tag{2.56}$$

Let $\mathbf{W}^n(i)$ represent the expected cumulative reward for the uniformized Markov chain over n steps, each of time length $1/B$, with one-step rewards $r(j)/B$ if the system is in state j. Hence, $\mathbf{W}^0 = \mathbf{0}$ and $\mathbf{W}^n$ is given by:

$$\mathbf{W}^n(i) = \frac{1}{B}\sum_{k=0}^{n-1} \mathbf{H}^k r(i), \quad i \in S. \tag{2.57}$$

The following result, taken from [62], is the analogue of Result 3.1 and presents an exact uniformized expression for the cumulative reward. Result 6.1 is closely

related to relations for hitting probabilities and sojourn times in [93, 134, 158].

Result 6.1 (Reward uniformization) *For all $i \in s$ and $t \geq 0$:*

$$\mathbf{W}_t(i) = \sum_{k=1}^{\infty} e^{-tB} \frac{(tB)^k}{k!} \mathbf{W}^k(i) \tag{2.58}$$

Proof
For notational convenience, define $p(k,\nu) = e^{-\nu}\nu^k/k!$. Starting with (2.53) and then using (2.52), (2.10), (2.56), the Gamma-Possion relation $\int_0^t \lambda p(m-1,\lambda s)\, ds = \sum_{k=m}^{\infty} p(k,\lambda t)$ and (2.57), for all $i \in S$ and $t > 0$:

$$\begin{aligned}
\mathbf{W}_t(i) &= \int_0^t \mathbf{H}_s r(i) ds = \int_0^t \sum_j \mathbf{P}_s(i,j) r(j) ds \\
&= \int_0^t \sum_j \sum_{k=0}^{\infty} p(k,sB) \mathbf{P}^k(i,j) r(j) ds = \int_0^t \sum_{k=0}^{\infty} p(k,sB) \mathbf{H}^k r(i) ds \\
&= \sum_{k=0}^{\infty} \left[\sum_{m=k+1}^{\infty} \frac{1}{B} p(m,tB) \right] \mathbf{H}^k r(i) = \sum_{m=1}^{\infty} p(m,tB) \mathbf{W}^m(i),
\end{aligned}$$

which completes the proof. □

Remark 6.1 (Numerical computation) As in the standard uniformization case in Section 2.3, expression (2.58) is exact. However, similar to Remark 3.2, for actual computation the Poisson tail approximation and possibly a state space truncation may be used.

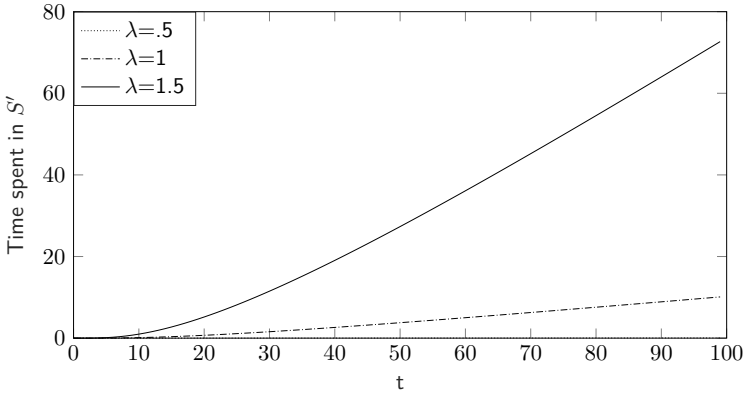
2.6.3 Numerical illustration: sojourn time and hitting probability for a set of states

Let us illustrate the reward uniformization approach by keeping track of the total time that the system will be in a set of states $S' \subset S$ within a time-interval of length t. To this end, first define vector r with elements $r(i)$ as follows:

$$r(i) = \begin{cases} 1, & i \in S', \\ 0, & i \notin S', \end{cases} \tag{2.59}$$

and introduce $\mathbf{W}_t^{(k)}$ and $\mathbf{W}^{(k)}$ analogous to (2.29) in Section 2.4.2 as the cumulative reward after at most k transitions and reward after exactly k transitions respectively,

Figure 2.3 Average time spent in S' when the system is empty.



and initialize these as follows:

$$\begin{aligned} \mathbf{W}_t^{(0)} &= \mathbf{0}, \\ \mathbf{W}^{(0)} &= r. \end{aligned} \tag{2.60}$$

By (2.58), we can then compute the time spent in S' before time t. This is shown in Figure 2.3 for the web server tandem model with $\mu_1 = \mu_2 = 2.5$, $c_1 = c_2 = 5$, $N_1 = N_2 = 100$ and varying λ, if the system starts empty for a maximum number of 5 threads at either station, i.e.,

$$S' = \{(n_1, n_2) \in \mathbb{N} | n_1 > c_1 \text{ or } n_2 > c_2\} \subset S. \tag{2.61}$$

As alternative measure of practical interest, we may also compute the probability to hit a state in S' at least once before time t. To this end, consider the following iterative procedure, for $k = 1, 2, \ldots$,

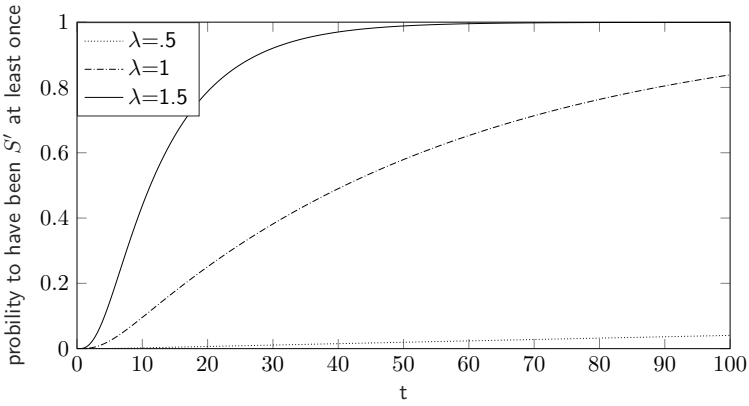
$$\mathbf{W}^{(k,*)} = r + \mathbf{P}\mathbf{W}^{(k-1)}, \tag{2.62}$$

$$\mathbf{W}^{(k)}(i) = \max\{\mathbf{W}^{(k,*)}(i), r(i)\}, \tag{2.63}$$

$$\mathbf{W}_t^{(k)} = \mathbf{W}_t^{(k-1)} + \frac{(tB)^k}{k!} e^{-tB} \mathbf{W}^{(k)}. \tag{2.64}$$

Here, $\mathbf{W}^{(k,*)}$ is introduced as a intermediate variable to avoid confusion. As in Section 2.4.2, $\mathbf{W}_t^{(k)}$ converges to $\mathbf{W}_t$ the probability to hit S' within time t for $k \to \infty$. For numerical computation we will use K as defined in (2.32) to truncate k.

Figure 2.4 Probability to hit subset S' at least once starting in an empty system



As numerical illustration, consider the web server tandem model with $\mu_1 = \mu_2 = 2.5$ and $c_1 = c_2 = 5$, and S' given in (2.61). Figure 2.4 shows the probability to reach a state in S' at least once before time t between 0 and 100 under three different values for the arrival rate λ.

2.6.4 Time-discretization approximation

Following the intuitive interpretation of uniformization as sketched in Interpretation 3.4, we can also rewrite the cumulative reward expression as time-discretization with time length $h \leq 1/B$. This can be done by regarding the iterative steps of uniformization as a time-increment of arbitrary length h. By expression (2.57) the expected cumulative reward functions $\mathbf{W}^n$ can then also be represented in an iterative manner as backward Kolmogorov equations:

$$\begin{aligned}
\mathbf{W}^{n+1}(i) &= \sum_{k=0}^{n} \mathbf{H}^k r(i) \\
&= \mathbf{H} \left[\sum_{k=1}^{n} \mathbf{H}^{k-1} r \right] (i) + \frac{1}{B} r(i) \\
&= \mathbf{H} \left[\sum_{k=0}^{n-1} r \right] (i) + \frac{1}{B} r(i) \\
&= \frac{1}{B} r(i) + \mathbf{H}\mathbf{W}^n(i)
\end{aligned} \tag{2.65}$$

written as a stochastic dynamic programming relation without actions:

$$\mathbf{W}^{n+1}(i) = \frac{1}{B} r(i) + \sum_{j \neq i} \mathbf{P}(i,j) \mathbf{W}^n(j). \tag{2.66}$$

By this principle and in line with Section 2.5, we can also directly present its more general time-inhomogeneous form. To this end, for given finite time horizon Z and all $t \leq Z$, let

$$\begin{aligned} \mathbf{W}_{t,Z}(i) &= \mathbb{E}\left[\int_t^Z r(X_s)\, ds | X_t = i\right] \\ &= \int_t^Z \left[\sum_l \mathbf{P}_{t,s}(i,l) r(l)\right] ds \end{aligned} \tag{2.67}$$

be the expected reward during the time interval for the time-inhomogeneous CTMC as given in Section 2.5, given that the system was in state i at time t. As a time-inhomogeneous analog of (2.65), choose a large N, such that $h = Z/N \leq 1/B$. By allowing a time-dependence and recalling matrices M_{nh} defined by (2.39) at times $t = nh$, we can now define the discrete-time functions $\bar{\mathbf{W}}^n$ for $n = 0, 1, 2, \ldots, N-1, N$ by

$$\begin{aligned} \bar{\mathbf{W}}^N &= 0 \\ \bar{\mathbf{W}}^n &= hr(i) + \sum_{l \neq i} \mathbf{P}(i,l) \bar{\mathbf{W}}^{n+1} \end{aligned} \tag{2.68}$$

The following approximation result then applies, as can be concluded from [60] in a more general controlled setting.

Result 6.2 *For arbitrary Z, with $h \leq 1/B$,*

$$\|\bar{\mathbf{W}}^n - \mathbf{W}_t\| \leq hC_Z \leq \frac{1}{B} C_Z, \quad nh \leq t < (n+1)h,\ t \leq Z. \tag{2.69}$$

As a special case, for a time-homogeneous CTMC, with $\mathbf{W}^n$ as by (2.65) *it leads to*

$$\|\mathbf{W}^n - \mathbf{W}_t\| \leq \frac{1}{B} C_Z, \quad nh \leq t < (n+1)h,\ t \leq Z. \tag{2.70}$$

Remark 6.2 (Time-dependent growth) The constant C_Z will depend on the length of the time horizon Z. A first rough estimate of C_Z, in line with the exponential or, in general, semigroup representation as in [75] gives, for some constant C independent of Z:

$$C_Z \sim e^{CZ} \tag{2.71}$$

A more careful investigation, under fairly general assumptions, among which a uniformly bounded reward rate r, could even lead to

$$C_Z \sim C\, Z^2 \tag{2.72}$$

which is in line with [181]. By also studying so-called bias or relative gain terms, depending on the system and the reward of interest, see, e.g., [67], this can even be brought down to

$$C_Z \sim C\, Z. \tag{2.73}$$

□

2.6.5 Two applications

2.6.5.1 Average computational bounds

We may also obtain bounds for computing the average reward or performance measures of a CTMC by application of a well-known result in Markov decision theory, often referred to as Odoni-bounds from [148], also see, e.g., [174, pp. 191-193], or [155], for the discrete-time case that just as well apply to uncontrolled DTMCs. These bounds are attractive to computationally bound convergence speeds. To do so, for W^n as in (2.66) for the homogeneous case let

$$\begin{aligned} m_n &= \min_i |\mathbf{W}^n(i) - \mathbf{W}^{n-1}(i)|, \\ M_n &= \max_i |\mathbf{W}^n(i) - \mathbf{W}^{n-1}(i)|. \end{aligned} \tag{2.74}$$

By combination of the uniformization Result 3.1 and these Odoni bounds, assuming existence of the average reward

$$G = \lim_{t\to\infty} \frac{1}{t}\mathbf{W}_t(i), \quad \text{for all } i \in S, \tag{2.75}$$

the following result can be concluded. It makes the Odoni bounds applicable to CTMCs.

Result 6.3 *For arbitrary $h \leq 1/B$ and bounded reward rate r: m_n is non-decreasing in n and M_n is non-increasing in n and*

$$h^{-1} m_n \leq G \leq M_n h^{-1}. \tag{2.76}$$

2.6.5.2 Error bounds

For performance measures of systems that do not have a closed form expression for the equilibrium distribution we may evaluate these performance measures for an approximate system and then develop bounds on the approximation error. Such approximate systems may be obtained through, e.g., state space truncation, relaxing state space restrictions or modification of the transition rates. We may also use error bounds when the data for, e.g., the service rates or interarrival times is imprecise.

Based on the uniformization results of this section, error bounds may be developed by bounding so-called bias-terms $[\mathbf{W}^k(j) - \mathbf{W}^k(i)]$ for the corresponding uniformized systems. The discrete-time (iterative) reward relation (2.66) may enable an analytical expression for the bounds on these bias-terms. These, in turn, yield analytical bounds on the approximation error, for more details see [34, Chapter 9].

2.7 Approximate uniformization for unbounded transition rates

The uniformization procedure is usually based on uniformly bounded exit rates, $q(i)$, from all states, $i \in S$. Uniformization as presented in the previous sections relies on over-relaxation of the exit rates from the states. This section shows that the uniform boundedness condition can be relaxed. To this end, a form of under-relaxation will be applied via a simple pragmatic adjustment of the one-step transition probabablity matrix of the uniformized Markov chain. As a price to pay, though, the uniformization approach will no longer be exact. Section 2.7.1 considers the infinite server queue to illustrate the approach. Section 2.7.2 presents the general result. The results are illustrated for the infinite server queue in Section 2.7.3.

2.7.1 Uniformization for the infinite server queue

Consider the infinite-server queue with Poisson arrival rate λ and exponential service rate μ per server, with transition rates:

$$q(i,j) = \begin{cases} \lambda, & j = i+1, \\ i\mu, & j = i-1. \end{cases} \tag{2.77}$$

As the service rates are unbounded, the infinite server queue violates the uniformizability condition (2.7) for any B so that we cannot define the uniformization matrix (2.8). We may, however, provide an approximate uniformization approach. To this

end, consider a fixed large N and define the matrix $\mathbf{P}^{[N]}$ as follows:

$$\text{for } i \leq N: \quad \mathbf{P}^{[N]}(i,j) = \begin{cases} \dfrac{\lambda}{\lambda + N\mu}, & j = i+1, \\ \dfrac{i\mu}{\lambda + N\mu}, & j = i-1, \\ 1 - \dfrac{\lambda + i\mu}{\lambda + N\mu}, & j = i, \end{cases} \tag{2.78}$$

$$\text{for } i > N: \quad \mathbf{P}^{[N]}(i,j) = \begin{cases} \dfrac{\lambda}{\lambda + i\mu}, & j = i+1, \\ \dfrac{i\mu}{\lambda + i\mu}, & j = i-1. \end{cases}$$

In this transition matrix $\mathbf{P}^{[N]}$, for $i \leq N$, the probability of jumping out of state i is $\frac{\lambda+i\mu}{\lambda+N\mu}$, whereas for $i > N$, this probability is 1.

2.7.2 Approximate uniformization result

The idea from the infinite server example can readily be generalized. Following the steps for standard uniformization, let $B < \infty$ be arbitrarily large, and let

$$J(i) = \min \left\{ 1, \sum_{j \neq i} q(i,j)/B \right\}$$

replace (2.7). Define the transition probability matrix $\mathbf{P}^{[B]}$ by

$$\mathbf{P}^{[B]}(i,j) = \begin{cases} 1 - J(i), & j = i, \\ J(i) \left[\dfrac{q(i,j)}{\sum\limits_{l \neq i} q(i,l)} \right], & j \neq i, \end{cases} \tag{2.79}$$

replace (2.8), i.e., with probability $J(i)$ a jump will take place in state i and given that a jump takes place, the transition probability is proportional to the corresponding transition rates in that state. The following result is proven in [63]. The proof is omitted as the details are rather technical and do not provide additional insight into the result.

For given initial distribution π_0, let $\pi_t = \pi_0 \mathbf{P}_t$ be the probability distribution of the original CTMC at time t and

$$\pi_t^{[B]} = \pi_0 \sum_{k=0}^{\infty} \frac{(tB)^k}{k!} \mathrm{e}^{-tB} \left(\mathbf{P}^{[B]} \right)^k, \quad t > 0. \tag{2.80}$$

the probability distribution of the approximate uniformized DTMC governed by the one-step transition matrix $\mathbf{P}^{[B]}$. An approximate result can then be concluded. To

this end, let $\mu : S \to \mathbb{R}$ with $\mu(i) \geq 1$ and μ non-decreasing. This function can be seen as a bounding function. Let a μ-norm be defined by, for $f : S \to \mathbb{R}$,

$$||f||_\mu = \sup_i |f(i)/\mu(i)|.$$

The following approximation result now applies for finite time-horizon Z.

Result 7.1 (Approximate uniformization with unbounded rates) *Let r be a reward rate such that for some constants $B_Z^{(1)}$ and $B_Z^{(2)}$:*

$$\begin{aligned} ||\pi_t^{[B]} r||_\mu &\leq B_Z^{(1)}, \quad t \leq Z, \\ ||\pi_t r||_\mu &\leq B_Z^{(2)}, \quad t \leq Z. \end{aligned}$$

Then for some constant C_Z and all $t \leq Z$:

$$||\pi_t^{[B]} r - \pi_t r|| \leq \frac{1}{B} C_Z. \tag{2.81}$$

Proof
The result may be readily obtained from [63] as follows. It essentially shows that the generator $\mathbf{Q}$ of the original CTMC and $\mathbf{Q}^{[B]}$ of the approximate Markov chain satisfy

$$||\mathbf{Q}^{[B]} f - \mathbf{Q} f||_\mu \sim \frac{1}{B} ||f||_\mu$$

for each f with $||f||_\mu < \infty$. □

Remark 7.1 The constant C_Z depends on the time-horizon Z. By analogy with Remark 6.2 we may find different bounds for C_Z. □

Remark 7.2 Note the resemblance between (2.70) and (2.81): the transformation in the transition rate resulting in (2.81) could be regarded to be of similar order as the transformation in time resulting in (2.70). □

2.7.3 Bounds for the infinite server queue

For the infinite server queue with reward rate $r(n) = n$, to evaluate the mean number of servers utilized, we may use $\mu(n) = 1 + n$. Result 7.1 then gives

$$||\pi_t^{[N]} r - \pi_t r||_\mu \sim \frac{1}{N} C_2.$$

Using bias-term results provided in [34, Chapter 9], for the infinite server queue we may also show that

$$||\pi^{[N]} r - \pi r||_\mu \sim \frac{1}{N},$$

yielding an asymptotic bound for the average reward, i.e., mean queue length, for $N \to \infty$.

2.7.4 Numerical illustration

For the infinite server queue with arrival rate λ and service rate μ and $\rho = \lambda/\mu$, in steady state, the probability $\pi(i)$ that i jobs are present in the system is given by:

$$\pi(i) = e^{-\rho}\frac{\rho^i}{i!}, \quad i \geq 0. \tag{2.82}$$

By comparing expression (2.82) and the numerical solution of the uniformized DTMC desribed by expression (2.78), the quality of the approximation (2.80) can be evaluated. In our numerical experiment, we consider an $M/M/\infty$ system with arrival rate $\lambda = 1$ and service rate $\mu = 0.1$. Table 2.4 shows the difference between the numerical approximation $\pi^{[N]}$ and the exact solution π for different values of N. For a numerical solution, the state space has to be truncated as well by limiting the maximum number of jobs in the system. Table 2.4 shows the quality of the approximation for different values of N and the state-space truncation.

Table 2.4 Difference between exact and approximate uniformized Markov Chain $\|\pi^{[B]} - \pi\|$.

State space	N						
Truncation	1	2	4	8	16	32	64
1	0.708	0	0	0	0	0	0
2	0.673	0.673	0	0	0	0	0
4	0.597	0.596	0.592	0	0	0	0
8	0.336	0.335	0.334	0.327	0	0	0
16	0.036	0.036	0.036	0.029	0.01348	0	0
32	0.034	0.034	0.034	0.029	0.00109	5.36E-09	0
64	0.034	0.034	0.034	0.029	0.00109	1.66E-10	5.80E-16
128	0.034	0.034	0.034	0.029	0.00109	1.66E-10	5.81E-16
256	0.034	0.034	0.034	0.029	0.00109	1.66E-10	5.82E-16
512	0.034	0.034	0.034	0.029	0.00109	1.66E-10	5.85E-16

The main observation from Table 2.4 is that the distance decreases strongly when N exceeds the point where the effective out-rate is larger than the effective in-rate, i.e., $N * \mu > \lambda$.

2.8 Exact uniformization with continuous state variables for non-exponential networks

In stochastic service networks, the assumption of exponentially distributed service times is introduced to allow for a tractable Markovian description of the network.

Generally distributed service times may be included in the Markovian description by keeping track of residual or spent service times, which, however, will generally lead to continuous-state differential equations that usually cannot be solved unless special detailed balance conditions are preserved, see, e.g., [33, 172]. A more pragmatic approach is to assume that the service times have a phase-type distribution, which allows a Markovian description of the network and in some cases closed form expressions for the equilibrium distribution, see, e.g., [125], [146]. These results enable approximate computational results for performance measures. Coxian distributions and mixtures of Erlang distributions are dense within the class of all distributions with non-negative support. Invoking additional weak convergence results for the corresponding processes, results under the assumption of phase-type distributions then carry over to those for arbitrary distributions [109].

This section takes an alternative approach for a special but general framework of stochastic service networks: it considers an extended uniformization procedure that applies to non-exponential stochastic service networks. As non-exponential times are involved, the Markovian property, essential to the uniformization approach, is lost unless the received or residual service times are contained in the state description resulting in a continuous state space. This section presents an equivalence result for the stationary distribution of the original model with state-dependent jump rates and a modified model with state-independent jump rates by analogy with the uniformization relation between the state-dependent transition or jump rates and the uniformized process with constant jump rates as presented in Section 2.3. This result is of practical interest as it may enable one to reduce the simulation or numerical computation of a non-exponential complex network to that of a Markov chain. The result can be interpreted as uniformization for the continuous service time. It is intuitively appealing and may already have been used by practitioners.

2.8.1 Model

Consider a stochastic network with a fixed number of M jobs (for convenience of presentation, we restrict the model to a fixed number of jobs as in a closed queueing network). A state $[L, T]$ with $L = (l_1, \ldots, l_M)$ and $T = (t_1, \ldots, t_M)$ denotes for each job i the current job mark l_i of job i with $l_i \in S$, where S is a countable space of possible jobmarks and t_i the amount of service that job i has received since its last service completion. For example, in queueing network applications a jobmark l can be of the form $l = (r, j, p)$ with r the type number of the job, j the station at which it is present and p its service position at this queue, while t is the amount of service that the job has already received at that station.
The law of motion is determined by the characteristics:

$$\begin{array}{ll} F_l(.) & \text{: distribution functions} \\ s_i([L,T]) & \text{: service rates (speeds)} \\ p_i(l|[L,T]) & \text{: transition probabilities} \end{array} \tag{2.83}$$

as follows. Whenever a job changes its jobmark to l it requires a random amount of service with distribution function F_l, independent of the other jobs and services received before. The rate at which jobs are being served, however, is state-dependent: when the system is in state $[L,T]$, the service rate, i.e., the amount of service per unit of time provided to job i, is $s_i([L,T])$, $i=1,\ldots,M$. When the system is in state $[L,T]$ and job i completes its service, its jobmark is changed to l' with probability $p_i(l'|[L,T])$. The jobmark of the other jobs thereby remain unchanged. Note that the transition probabilities and the service rates depend on the amount of service received by the jobs. The law of motion allows modeling of various service disciplines, including Processor Sharing and FCFS, as well as state-dependent transition probabilities including blocking of jobs due to capacity restrictions at the queues.

Under the following assumptions, the system can be uniformized with respect to the continuous parameter for the received amount of service of the jobs.

Assumptions

1. For all l, the function $F_l(t)$ is absolutely continuous for $t\in(0,\infty)$ with density function $f_l(t)$. Hence, its failure rate is well defined by $f_l(t)/[1-F_l(t)]$ for all $t\in(0,\infty)$. We introduce the notation

$$d_i([L,T]) = s_i([L,T])\frac{f_{l_i}(t_i)}{1-F_{l_i}(t_i)} \tag{2.84}$$

2. For some constant $D<\infty$ and all $[L,T]$:

$$d([L,T]) = \sum_i d_i([L,T]) \le D \tag{2.85}$$

2.8.2 Uniformized stochastic network

The law of motion is now defined as follows. Let B be some finite number with $B\ge D$ and assume that at exponential inter arrival times with parameter B a so-called "jump" occurs. When a jump occurs while the system is in state $[L,T]$, with probability

$$d_i([L,T])p_i(l'|[L,T])/B \tag{2.86}$$

this state will change to $[L',T'] = [L,T]-(l_i,t_i)+(l',0)$ with $l'_j=l_j$ and $t'_j=t_j$ for all $j\neq i$ but $l'_i=l'$ and $t'_i=0$, where i can be any of the jobs $i=1,\ldots,M$, representing the state equal to $[L,T]$, except for job i. With probability

$$1-d([L,T])/B \tag{2.87}$$

the state remains unchanged, i.e., is no real transition takes place, such that the state directly after the jump is still $[L,T]$. Note that (2.86), summed over all i and

(2.87) sum up to 1 and thus represent a probability.

Remark 8.1 The formulation (2.86) and (2.87) resemble the standard uniformization technique when all distributions $F_l(.)$ are exponential, but now uniformized with respect to the statespace variable t. □

Assume that both the original and uniformized model have a unique stationary density function at one and the same irreducible set S which we denote by $\pi_1(L,T)$ and $\pi_2(L,T)$ respectively. We have the following result. This result can be adapted from a wider setting in [61]. Let us give a short self-contained proof in line with section 2.3.1 on Markov generators.

Result 8.1

$$\pi_1(L,T) = \pi_2(L,T), \quad (L,T) \in S. \tag{2.88}$$

Proof
By carefully working out the infinitesimal characteristics (of the continuous-state expanded version of (2.13), see, e.g., [33], for a system related to the given stochastic network description), it can be shown that the infinitesimal generator for both the orignal description, i.e., (2.84), and for the uniformized description, i.e., (2.86) and (2.87), are indentical as expressed by, for $g : S \to \mathbb{R}$,

$$\begin{aligned}\mathbf{A}g([L,T]) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \left[\sum_{L'} \int_{T'} \mathbf{P}_{\Delta t}([L',dT'])g([L',T']) - g([L,T]) \right] \\ &= \sum_{i=1}^{M} \left[\frac{d}{dt_i} g([L,T]) + \sum_{l'} p_i(l' \,|[L,T])g([L,T] - (l_i,t_i) + (l',0)) \right].\end{aligned}$$

The uniqueness of a corresponding transition probability semigroup based on these infinitesimal characteristics or generator for arbitrary continuous-time and continuous-state Markov processes, see [88, 89], then completes the proof. □

Remark 8.2 (Simulation/computation) As continuous distributions are involved, the actual simulation or computation of the transition probabilities may still lead to technical complications. In simulation, the rejection method may come in handy. In computing, a discretization or approximation either by exponential phase-type distributions, or by using discrete-time grids (cf. [60]) for the service times, seems most natural. □

2.9 Concluding remarks

This chapter has provided a mathematical and intuitive review of uniformization technique and some of its extensions. The extensions to Markov chains with time-dependent transition rates and with unbounded transition rates open a wealth of

possible applications of the uniformization technique. Next to the application of uniformization to performance measures that can be evaluated via the state distribution, we have highlighted that uniformization may also be directly applied to Markov reward processes. Furthermore, we have shown that uniformization might be used not only for time-discretization, but also for state space discretization.

The basic theory supporting the evaluation of performance measures based on standard uniformization for CTMCs seems to be well-developed. In contrast, for the extensions highlighted in this chapter there is ample room for extensions of the basic theory. For example, uniformization for systems with time-dependent *and* unbounded rates, or uniformization *both* in time and in space may turn out to be of considerable theoretical and practical interest. As systems studied in practice become more and more involved, extension of the uniformization technique to facilitate their performance analysis seems of utmost importance.

Part II

Evaluation

Chapter 3

S.P.J. van Brummelen, W.L. de Kort and N.M. van Dijk. Waiting time computation for blood collection sites. *Operations Research for Health Care* 7:70-80, 2015.

Chapter 4

S.P.J. van Brummelen, W.L. de Kort and N.M. van Dijk. Queue length Computation of Time-Dependent Queueing Networks and its Application to Blood Collection. *Operations Research for Health Care*, accepted.

CHAPTER 3

Waiting time computation for blood collection sites

3.1 Introduction

To analyze the realistic logistical situation at blood collection sites, it is, first of all, important to be able to compute the waiting times in a reliable and replicable manner. Although a substantial number of queueing models exist, these methods first have to be adapted and further developed for application in blood collection sites. These methods should also be usable to determine the necessary staffing level before a session starts, and to show how interventions might affect waiting times. This chapter will provide methods to compute waiting times at blood collections sites.

The methods are inspired by blood collection sites, but can be extended to more generic queueing systems. More patricularly to Jackson networks - networks of queues with probabilistic routing. The model can be applied within a health care setting, such as within an emergency department or an organ donation system, or even more general production and service systems.

This chapter will be structured as follows. First, we will discuss the relevant literature in Section 3.1. We will then introduce the model in more technical detail (Section 3.3). Next, we will show that a product form expression applies (Sections 3.4.1 and 3.4.2). This will lead to a waiting time distribution for each of the separate stations of the intake process (Section 3.4.3). We will then illustrate that the total delay remains intriguing and non-trivial. A numerical method to compute the total delay time distribution will be presented (Section 3.5). Both the analytic results and numerical method will be applied to a test case, based on a real blood collection site, and different scenarios will be compared (Section 3.6). The chapter will end with a conclusion (Section 3.7).

3.2 Literature

For an overview of literature on blood collection sites, the reader is referred to Section 1.4. The remainder of this section will only discus literature on methods and models relevant for the methods developed and discussed in this chapter.

In general, analytic or closed form results for total waiting or delay times in tandem structures, which is the most realistic structure to model a blood collection site, appear to be rather limited. Even for the 'simple' case of just a tandem queue with two queues, such results only seem to be available for special situations or under special assumptions, such as identical servers at both stations, single server cases or overtaking-free assumptions and infinite capacities; e.g. see the book by Boxma and Daduna [35] and references therein.

Product form results for queueing networks without finite capacity constraints are well-known in literature since the pioneering work by Jackson [114] for Jacksonian networks. In particular, Gordon and Newell [91] have explicitly presented a product form solution for unlimited serial or tandem structures. Exact product form results for systems with finite capacity constraints, however, are limited. In Gordon and Newell [92] only closed tandem structures - with a fixed total number of jobs circulating - are studied along with finite constraints in specific cases, under the assumption that these constraints are small or large.

In Jackson [115], an extension of his classical 1957 paper [114], finite capacity constraints are incorporated by either a total number dependent arrival rate or by lower limit service rates for each station. As a special case, the product form preservation is also argued by either instantaneous triggering of new arrivals or by service deletion. However, an upper limit for just one station, as in Section 3.4.2, is not included.

In Kelly [118] and Pittel [153], the inclusion of finite constraints is only justified by a specific routing assumption: the product forms remain valid with finite truncations, provided the system has a so-called reversible routing. A tandem system, like the application in this chapter, is excluded as its routing structure is strictly not reversible.

Product forms have also been applied to practical situations. Although product forms have not been applied to or justified for blood collection sites, they have been applied in health care settings, e.g. see Xie et al.[103] and Yom-Tov [188].

All of these earlier references verify the product form result by the global balance or Kolmogorov equations. None of these references explicitly mention the more detailed verifications for each station separately as shown in a straightforward proof presented in Section 3.8.

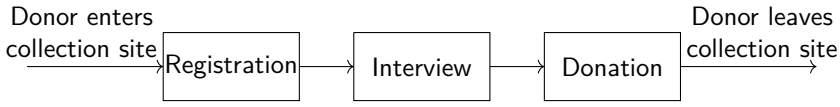
As such, the product form result that will be reported in this chapter, at least from an application point of view, can be regarded as new. To some extent, mainly focusing on finite limitations in a tandem queue, it is also new from a more technical point of view. The exact product form result that will be presented for the unlimited case also leads to a marginal waiting time distribution.

Given the practical and generic character of tandem structures and finite capacity constraints in assembly line and production systems, the literature has paid considerable attention to approximation methods for infinite and finite tandem systems. Work on this topic has been reported on in, among others, the book on queueing networks with blocking by Perros [152], the excellent survey on manufacturing flow line systems by Dallery and Gerschwin [55], the well-known QNA method by Whitt [182] and the early, elegant paper by Buzacott [42] to capture interaction based on

Pollazchek- Khintchine's formula. Without exception, these approximation procedures use some form of decomposition by which the service stations are regarded as separate with interaction between the stations incorporated by coefficients of variation. These coefficients link variability in inter-arrival and service times. Some of these procedures also include a number of iterations for adjustments. In Section 3.4.4 the well-known approximation method QNA (the Queueing Network Analyzer), developed by Whitt [182], will be briefly presented. All of these approaches, however, provide approximations for just the mean delay and mean waiting times and not for complete distributions. This chapter, in contrast, will also provide an algorithm to numerically compute a total delay distribution, which has not been reported in literature.

3.3 Model description

Figure 3.1 Schematic representation of a collection site



The queueing model of a blood collection site that will be used in this chapter is shown in Figure 3.1. In line with the description given in Section 1.3, we will model the blood collection site by a tandem queue with three stations: Registration (station 1), Interview and testing (station 2) and Donation (station 3). Donors may have to wait at each of these stations. We also include the possibility that the total number of donors present in the collection site may not exceed some limited number M. When this number has been reached, arriving donors are rejected, i.e. kindly asked to return in a next session. Clearly, the realistic unlimited case, $M = \infty$, remains included. We also refer to Section 3.4.2 for the possibility of finite constraints. The parameters and variables used are listed in Table 3.1. Note that we distinguish between outputs waiting time W, which does not include service time and Delay time T, which does include service times.

For analytic purposes, we assume all service times and inter-arrival times to be exponential. Since donors are free to choose when to donate blood the assumption of exponential inter-arrival times - an arrival is independent of the time since the last arrival - for whole blood donations seems justified. Although exponential service times are a less accurate assumption, the methods discussed in this chapter require this assumption, along with most exact methods from queueing theory. Most methods not requiring this assumption only approximate mean delay and waiting times. Even these non-exponential methods rely on decomposition results - which requires exponential assumptions - to justify the possible use of decomposition, as will be discussed in Section 3.4.4. This section introduces one of these approximate methods, that relaxes the exponential assumption.

Table 3.1 Parameters, variables and outputs

λ	Arrival rate at Registration station
μ_q	Service rate at station q
τ_q	Mean service time at station q $(= 1/\mu_q)$
s_q	number of staff members at station q
ρ_q	utilization at station q, $= \lambda/s_q * \mu_q$
M	A limitation on the total number of donors in the system ($M = \infty$ is included)
N_q	A limitation on the number of donors at station q ($N_q = \infty$ is included)
n_q	Number of donors at station q
n	Vector: (n_1, n_2, n_3)
W	Waiting time in the system (W_q for station $q = 1, 2, 3$)
T	Delay time in the system (T_q for station $q = 1, 2, 3$)

3.4 Exact Product Form

3.4.1 Product Form

This Section will present a product form expression - also referred to as a separable network in the situation of unlimited capacities. Roughly speaking, the term product form reflects that a joint workload or queue length distribution can be obtained for an entire queueing network by factorizing terms for each individual station, as if these stations can be seen as independent stations in isolation. A more detailed discussion of this term can be found in literature, e.g. see Kelly [118], Perros [152], van Dijk [65].

Theorem 1 shows that a blood collection site, as described in Section 3.3, indeed has a product form solution. This specific application of product forms for blood collection sites has not been reported explicitly in the literature. For explicit verification of the global balance equations by individual equations for each of the stations and for the possible inclusion of finite constraints, a direct proof is provided in Section 3.8.

For presentational simplicity, finite constraints N_q are excluded in the theorem. The implications of including these constraints will be discussed in Section 3.4.2.

Theorem 1. *Let $n = (n_1, n_2, n_3)$ with n_q the number of donors at station $q = 1, 2, 3$. Then the steady state distribution π_n - the probability that the system is in state n, with $\pi_{\mathbf{0}}$ a normalizing constant where $\mathbf{0} = (0, 0, 0)$, is given by:*

$$\pi_n = \pi_{\mathbf{0}} \prod_{q=1}^{3} \left[\left(\frac{\lambda}{\mu_q} \right)^{n_q} \frac{1}{min\{n_q, s_q\}!} \left(\frac{1}{s_q} \right)^{[n_q - s_q]^+} \right]. \tag{3.1}$$

Here $[n_q - s_q]^+ = max\{0, n_q - s_q\}$. The normalizing constant of expression (3.1)

is given by:

$$\pi_{\mathbf{0}} = \left[\sum_{n_1=0}^{M} \sum_{n_2=0}^{M-n_1} \sum_{n_3=0}^{M-n_1-n_2} \prod_{q=1}^{3} \left[\left(\frac{\lambda}{\mu_q} \right)^{n_q} \frac{1}{min\{n_q, s_q\}!} \left(\frac{1}{s_q} \right)^{[n_q - s_q]^+} \right] \right]^{-1} \tag{3.2}$$

In the infinite case ($M = \infty$), we can rewrite expression (3.1) to

$$\pi_n = \prod_{q=1}^{3} \pi_{q,n_q} = \prod_{q=1}^{3} \left[\pi_{q,0} \left(\frac{\lambda}{\mu_q} \right)^{n_q} \frac{1}{min\{n_q, s_q\}!} \left(\frac{1}{s_q} \right)^{[n_q - s_q]^+} \right] \tag{3.3}$$

with π_{q,n_q} the marginal distribution - with normalizing constant $\pi_{q,0}$ - of the number of donors present at station $q = 1, 2, 3$. This distribution and normalizing constant are equal to that for a standard M/M/s queue (e.g. see Cooper [53] and the proof in Section 3.8).

Theorem 1 shows that the solution can indeed be seen as a factorization of terms for individual stations. In Section 3.8, we present a straightforward proof of this theorem by showing balance for each station separately.

From a mathematical point of view and given the system description in Section 3.3 without capacity constraints, the product form in Theorem 1 is not new and it can already be concluded from the classic literature on Jackson networks (see Jackson [114, 115]). However, for more restricted cases, such as a finite common constraint or finite constraints for each station separately, this is less clear.

For a total capacity constraint, the product form has been shown to be applicable by Jackson [115]. However, this reference is not explicit about balance equations for each station separately (as we have shown in Section 3.8). A total capacity constraint can also be concluded from Kelly [118] by using the concept of quasi reversibility.

For the inclusion of capacity limitations on individual stations, references are even less clear. In Kelly [118] and Pittel [153] the inclusion of such finite constraints is only justified by a reversible routing assumption. The product remains valid provided the system has a so-called reversible routing. This implies that if the routing probability of moving from station q to station q', $p_{qq'} > 0$, then necessarily $p_{q'q} > 0$. Tandem systems clearly do not have a reversible routing.

Nevertheless, as argued in Section 3.4.2, the proof in Section 3.8 can be extended to included finite limitations on individual stations.

3.4.2 The product form and the extension with finite limitation

The total number of donors at the Donation station could be limited, reflecting a finite number of beds and a restricted waiting capacity, by including a maximum N_3. If the Donation station is congested, the Interview station will be stopped. To preserve analytical feasibility, an additional and somewhat artificial assumption is required, stopping the Registration station and arrivals when the Donation station

is congested. Under this modification, the product form expression (3.1) remains valid. This can be seen directly by adding an indicator function $\mathbb{1}_{(n_3<N_3)}$ in both the left and right hand side of (3.10.1), (3.10.2) and (3.10.3) in Section 3.8. The only change occurs in the normalization constant $\pi_{\mathbf{0}}$. If N_1, N_2, N_3 or $M < \infty$, this normalizing constant can be computed by restricting the summation to the set of admissible states:

$$S = \{n \quad | \quad n_q \leq N_q, n_1 + n_2 + n_3 \leq M\} \tag{3.4}$$

Despite its slightly unrealistic additional assumption, this analytic result might be quite useful to establish reasonable approximations for queue lengths and possibly a safe estimate (upper bound) for the congestion probability of the collection site, in line with the result in Van Dijk and Kortbeek [66].

It is even possible to truncate each station separately by finite numbers N_q for station $q \in \{1, 2, 3\}$. The product form (3.1) remains valid if the normalization constant is restricted to the truncated state space and by artificially assuming that if a station q becomes congested (i.e. $n_q = N_q$), arrivals are blocked and all other stations $q' \neq q$ are stopped. The product form can then again be used and is justifiable as an approximation.

3.4.3 Marginal waiting times

As a direct consequence of Theorem 1 and for the infinite case ($M = \infty$), we can conclude that the waiting time distribution for each station separately can be computed as a standard multi server queue as if it were in isolation. The proof of this Theorem - Theorem 2 below - is presented in Section 3.9.

Theorem 2. *For the unlimited ($M = \infty$) and exponential case of the system described in Section 3.3, the marginal distribution for the waiting time W_q for each station $q \in \{1, 2, 3\}$ separately is equal to that of an $M/M/s_q$ queue as given by:*

$$\begin{aligned} \mathbb{P}(W_q) &= \mathbb{P}(W_q > 0) e^{-(1-\rho_q)s_q\mu_q t} \\ &= \pi_{q,0} \left[\left(\frac{\lambda}{\mu_q} \right)^{s_q} \frac{1}{s_q!} \right] [1 - \rho_q]^{-1} \, e^{-(1-\rho_q)s_q\mu_q t} \end{aligned} \tag{3.5}$$

Remark. These marginal waiting time distributions, particularly those for the Interview station and the Donation station, are of practical interest for blood collection sites. For example, a frequently encountered perception seems to be that the total waiting time is primarily influenced by the Interview station. While having to wait at the Interview station the donor has already been accepted to the system, but further progress is interrupted and delayed before starting the Donation station. In Section 3.6.3 we will therefore compare waiting time percentiles for the Interview station of the test case and different scenarios based on equation (3.5).

3.4.4 QNA: Approximation of mean waiting time

The product form from Section 3.4.1 shows that a decomposition of the queue length distribution into the individual stations is fully justified in the case of exponentially distributed inter-arrival and service times. However, at collection sites service times are not exponentially distributed. In this Section we will therefore briefly introduce an approximate method, known as QNA, based on Whitt [182]. Basically, this approximation relies on a decomposition of the network into independent queues, adjusting for the non-exponential service and inter-arrival times by:

C_{aq}^2 Squared coefficient of variation of the arrivals at station q

C_{dq}^2 Squared coefficient of variation of the departures from station q

C_{sq}^2 Squared coefficient of variation of the service at station q

Although the product form (3.1) from Theorem 1 is no longer applicable because of this non-exponential distributions, it does provide some justification for the decomposition, as it validates the decomposition for the exponential case. The QNA method, as well as all related ones discussed in Section 3.2, aim to provide an approximation for mean waiting times and not for waiting or delay time percentiles, which will be discussed in Section 3.5. These percentiles are of particular interest for collection sites.

The method works by linking the squared coefficients of variation between stations. The squared coefficient of variation for the external arrivals, in our case to the Registration station, C_{a1}^2, has to be part of the input parameters and could be set to 1 to represent exponential inter-arrival times. For the other stations, $q \in \{1, 2, 3\}$, the squared arrival coefficients C_{aq}^2 can be calculated by:

$$C_{a(q+1)}^2 = C_{dq}^2 = 1 + (1 - \rho_q^2)(C_{aq}^2 - 1) + \frac{\rho_q^2}{\sqrt{s_q}}(C_{sq}^2 - 1) \tag{3.6}$$

Now let $\mathbb{E}_{M/M/s}(W_q)$ denote the expected waiting time of a standard M/M/s queue. Then, with $s = s_q$ and by a simplified version of a formula obtained from Whitt [182], we can approximate the expected waiting time for each station $q \in \{1, 2, 3\}$:

$$\mathbb{E}(W_q) = \left[\frac{(C_{aq}^2 + C_{sq}^2)}{2}\right] \mathbb{E}_{M/M/s}(W_q) \tag{3.7}$$

3.5 Total waiting time distribution

3.5.1 Independent total delay time calculation?

By the product form (3.1) in Section 3.4.1, we have shown that queue lengths are stochastically independent, as if they are independent queues in isolation; an even

more intriguing result is concluded considering that the product form is preserved if we add finite constraints, as argued in Section 3.4.2. Given this product form, one might also expect that the total waiting and total delay time can be computed by simply combining waiting time expressions for each of the stations, as if these stations have stochastically independent waiting time expressions.

However, intuition dictates that queue lengths and waiting times at the consecutive stations are dependent: when an arbitrary donor at the Registration station has a (substantially) higher waiting time than the expected waiting time at this station - due to a higher number of donors at the station than usual - this donor will most likely also experience longer waiting times at the Interview station. As a consequence, this would lead to dependence of waiting times between the stations. Indeed, the intuition is correct: the total waiting time distribution of tandem queues cannot be computed as a convolution of independent waiting time distributions at each queue. These waiting times in tandem queues are not independent, as already proven by Reich [159] and Burke [41].

As an illustrative simulation, we will show that assuming independence is indeed incorrect.

Assume a two-station tandem queue (see Figure 3.2, model A), which exhibits a product form similar to the one from equation (3.1). Assume an exponential inter-arrival time, with an average of one job arrival per time unit. Each station has an exponential service time of 9.4 time units, and 10 servers. This gives an occupancy rate of 94 % at both stations. The system was run five times, in each of the runs generating 100,000 jobs. To show that this gives significantly different delay time distributions than if independence would be assumed, we also simulated the same costumers, experiencing the same service times, arriving independently to the first and second station (Figure 3.2, model B).

Figure 3.2 Simulated two-station tandem model (A) and an independent two-station model (B)

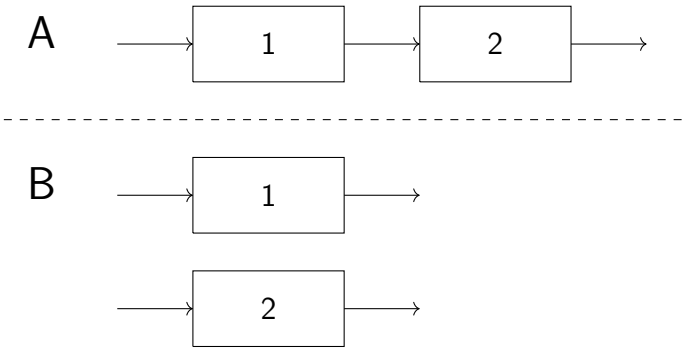


Figure 3.3 shows the simulated probability density function (pdf) of delay time, for both systems. The solid line indicates the pdf for the tandem system (Figure 3.2, A) and the dashed line indicates the pdf for the independent system (Figure 3.2, B). It is clearly visible that these distributions do not coincide. A Kolmogorov-Smirnov

test indicates that the Null hypothesis of identical distributions can be rejected with near certainty ($p-value \ll 0.001$). As could be concluded from the papers of Reich [159] and Burke [41], this numerically shows that we cannot assume that individual queues in a tandem system to be independent in order to calculate the total delay time distribution.

A few points of the cumulative distributions function for both systems have been included in Table 3.2. The maximum difference between these cumulative distributions functions, which is the input for the Kolmogorov-Smirnov test, can be found around 25 time units delay.

Figure 3.3 Probability distribution of the simulated tandem queue and the combination of two independent queues.

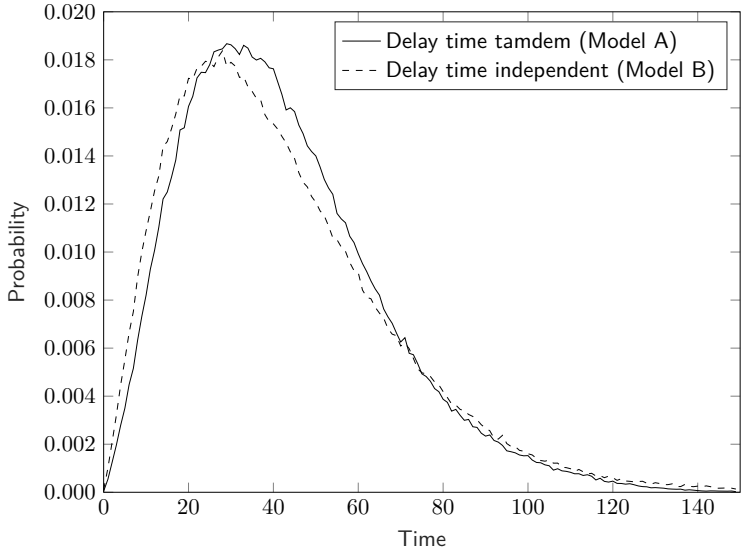


Table 3.2 Cumulative distributions of simulation

Delay time	Cumulative distribution tandem (model A)	Cumulative distribution independent (model B)	Difference (%)
25	0.239	0.278	0.040 (16.7 %)
50	0.670	0.675	0.005 (0.7 %)
75	0.903	0.887	0.017 (1.9 %)

3.5.2 Numerical computation of the total delay time

Although the delay time distributions are dependent, the product form for queue lengths still holds. Based on this product form - and thus under the exponential assumptions, the distribution of the total delay time T can be computed by using the PASTA property (Poisson Arrivals See Time Averages, Wolff [185]) and conditioning

upon the system state upon arrival by

$$\mathbb{P}(T < t) = \sum_n \pi_n \mathbb{P}(T < t|n) \tag{3.8}$$

where π_n can be calculated with the product form expression (3.1).

It is impossible to compute $\mathbb{P}(T < t|n)$ by an explicit expression. Standard equations to derive waiting time or delay expressions such as for M/G/c systems are no longer available. The waiting time at these stations is not independent (as shown in Section 3.5.1) and cannot simply be added as a superposition - except for the expected waiting time and total delay.

In Section 3.10 an algorithm will therefore be presented to numerically compute the total delay. This algorithm roughly distinguishes three parts. First of all, as we need to keep track of a tagged donor, the algorithm needs to expand the Continuous Time Markov Chain (CTMC) from Section 3.4. Next, this CTMC will be approximated arbitrarily closely by using time-discretization. Last and most essentially, the algorithm computes the total sojourn time for an arriving job (donor) until it clears the system. This will be achieved by following the arriving job as a tagged job and by regarding the system as an absorbing Markov chain in which the tagged job passes through each of the stations until it leaves the system, i.e. until it completes its service at the Donation station,

These global steps are worked out and discussed more detailed in Section 3.10.

Remark. To the best of our knowledge, the presented algorithm, based on the exponential assumption, is new. A non-exponential extension, by using phase-type distributions, can be thought of and is certainly of interest from a mathematical point of view. But at this point in time it is likely to become computationally expensive if not prohibitive. This will remain of interest for future research.

3.6 Measurements and computational results

3.6.1 Test case

In this section, we will provide some numerical results to illustrate the use of our methods. Data from a real life collection site has been used as a test case. This site is located in the Dutch city of Zwolle and handles over 30,000 donations annually. The data, gathered in 2012, leads to the parameter settings shown in Table 3.3. The service rates can be considered to be representative for collection sites throughout the Netherlands.

To illustrate the results more effectively, three extra scenarios were compared with the existing, basic scenario:

1. One extra staff member at the Interview station.

2. One extra staff member at the Donation station.

Table 3.3 Input data for the test case

Parameter	Value
Arrival rate, λ	15.0 donors/hour
Service rate per staff member at the Registration station, μ_1	30.0 donors/hour
Staff members (servers) at the Registration station, s_1	1
Service rate per staff member at the Interview station, μ_1	10.2 donors/hour
Staff members (servers) at the Interview station, s_2	2
Service rate per staff member at the Donation station, μ_1	5.0 donors/hour
Staff members (servers) at the Donation station, s_3	4

3. At the Interview station an Hb-test is performed. This test requires roughly 1 minute. We can perform this test directly at the Registration station, changing μ_1 to 20, and μ_2 to 12.3.

3.6.2 Real life measurements and product form computations

Using the product form expression (3.1), in combination with Little's well-known law, we can directly compute the mean delay time by:

$$\begin{aligned} L &= \lambda T \\ L_q &= \lambda T_q \qquad q \in \{1,2,3\} \end{aligned} \tag{3.9}$$

where

- λ Arrival rate of donors (as mentioned before)
- L Mean number of donors in the system (L_q for station $q \in \{1,2,3\}$)
- T Mean total delay time in the system (T_q for station $q \in \{1,2,3\}$)

Note that a similar relation holds for the mean number of donors in the queue and the mean waiting time. It is also possible to directly calculate the expected waiting and delay time W_q and T_q for each station q separately. In Table 3.4, the waiting times that were calculated with the product form result (Theorem 1) are shown together with data from internal reports at Sanquin (Van den Toren et al. [175]). The presented data were collected throughout the Netherlands in 2010.

From Table 3.4 the conclusion can be drawn that the expected waiting times seem to validate quite well with the product form computation. All of the computed waiting times fall within the 95% Confidence Interval - and even the 30% Confidence Interval - of the corresponding real life measurements.

Accordingly, as mentioned in Section 3.3, these results seem to partly justify the assumption of exponential service times, particularly for comparison purposes. Results for the three scenarios from Section 3.6.1 are also included, based on the product form calculations.

The Interview station clearly has more to gain from an extra staff member than the Donation station. Although scenario 3 increases the total waiting time, it might

Table 3.4 Expected waiting times for the three stations in minutes

Scenario	Registration	Interview	Donation	Total
Real waiting time	1.91	8.29	3.67	13.88
Computed waiting time, base scenario	2.00	6.92	6.11	15.04
Scenario 1	2.00	0.87	6.11	8.98
Scenario 2	2.00	6.92	1.42	10.34
Scenario 3	9.00	2.89	6.11	18.00

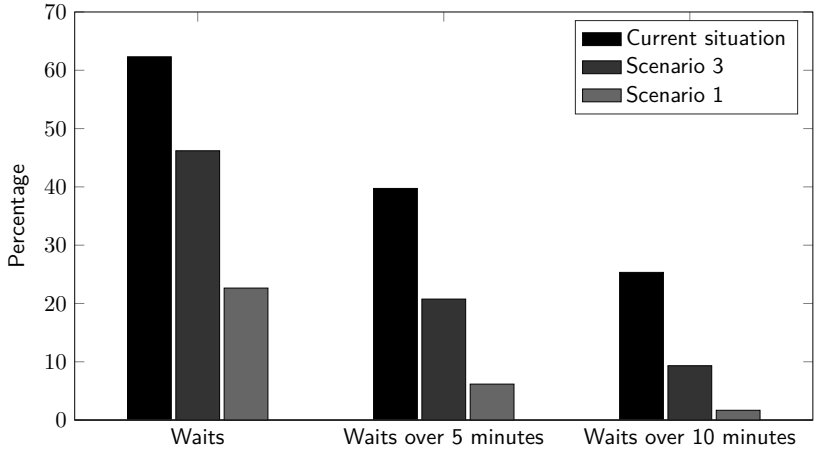
still be interesting. It has been suggested by some people within Sanquin that donors experience waiting time at the Interview station as longer and more annoying than at the two other stations, but there seems to be no evidence to support this theory.

Remark. The validation of the computation, based on exponential service times, seems to be consistent throughout measurements and computations, both at global collection site level and for individual stations. This validation seems to sufficiently justify the use of the product form for global production and waiting time computation.

3.6.3 Computational results of the marginal waiting time distribution

Using equation (3.5), it is possible to calculate percentiles in waiting times for individual stations. In Figure 3.4 we present waiting time percentiles for the Interview station. The results were calculated for a parameter setting based on the current situation and scenarios 1 and 3. Both of these scenarios lower the average workload at the interview station.

Figure 3.4 Percentage of donors that has to wait for the Interview station.



3.6.4 Results of the numerical procedure for total delay time

In addition to the results for expected total waiting times in Section 3.6.2 and the exact results in Section 3.6.3 for the marginal waiting times, it is also possible to compute delay time percentiles by using the computational steps in Section 3.5.2.

To save computational time, the computations just include the Interview and Donation stations. This is justified; as our previous results showed that the Registration station did not lead to high waiting and delay times. As in Section 3.6.2, results were computed for the current situation and two of the scenarios from Section 3.6.1. Scenario 3 was left out as it also concerns the registration station. For calculations we have used the parameter settings from Section 3.6.1. The parameters α, the length of a time interval, and K , the total number of time intervals, were set at twice the maximum of D, the highest rate out of any state, and 150 respectively. In the 'current situation (2)' both were set twice as high to illustrate that there was no considerable effect on any of the outcomes. To keep the computation restricted, the Interview station was given a maximum capacity of 6 (staff members and queue) and the Donation station was given a maximum capacity of 8. Though limited, these values already appeared sufficient to compare the results of the scenarios. For the test case no percentile measurements were available.

Table 3.5 Results from the numerical procedure

	Sojourn time percentile (min)		
	75th	90th	95th
Current situation	33	46	55
Current situation (2)	33	46	56
Scenario 1	29	41	49
Scenario 2	30	42	52

Clearly, the inclusion of an extra staff member at one of either stations has a positive influence on the delay times, especially on the higher percentiles. There is a slight preference to use the extra staff member at the Interview station.

3.7 Discussion

The research presented in the chapter aimed to make a first step to combine production norms for blood collection sites with waiting times on a purely analytical basis. The expressions provided can be regarded as generic - i.e. applicable for different collection sites regardless of arrival and service rates, size and staff numbers - and can also be seen as supportive for approximate methods. In particular, a decomposition result into separate stations appears to be applicable for computation.

By using the decomposition result, waiting time distributions and percentiles could be obtained for individual stations by an analytic, direct expression. The expression is not new, but does not seem to have been justified in a tandem queue setting. This expression is of particular practical interest for general perceptions on waiting time experiences during the donation process. The total waiting and delay

time distribution, however, remains highly intriguing, because the distributions of separate stations are not independent. This dependence meant that, even for the 'simple' exponential case, a computational algorithm had to be developed.

A real life test case, based on a Dutch collection site, was evaluated to illustrate the practical usefulness of both the analytic and computational results to improve the waiting time performance and its perception. Although our simplified three-station tandem queueing model does not capture all aspects for a perfect representation of reality, the results in Section 3.6 seem to support the applicability of the queueing model described. Reasonable approximations as to compare different scenarios for practical application are obtained. We have also shown that, although the exponential assumption on service times is not completely accurate, it does lead to waiting times that closely match those found in real blood collection sites.

From these results, we draw the conclusion that the expected waiting times seem to relate quite well with the product form computation. Accordingly, these results seem to justify the assumption of exponential service times for computational purposes. We emphasize that one of the prime applications for Sanquin is to compare different scenarios in a generic and replicable manner, as presented in Section 3.6.

Next to this useful step in determining and predicting waiting times at blood collection sites, the results are motivating for further research as presented in later chapters of this thesis. An integration and combination with other OR methods, such integer linear programming (see Chapter 5) and Markov decision processes (see Chapter 6) are just some of the options.

3.8 Appendix I: Proof of Theorem 1

First let us introduce some notation. The indicator function $\mathbb{1}_{([condition])}$ is defined as

$$\mathbb{1}_{\{[condition]\}} = \begin{cases} 1 & \text{if } [condition] \text{ is satisfied} \\ 0 & \text{if } [condition] \text{ is not satisfied} \end{cases}$$

Also as standard, the unit vector e_q denotes a vector with value equal to 1 on the q^{th} position and value 0 at all other vector positions. E.g. $e_2 = (0, 1, 0)$. For ease of notation also define:

$$\mu_q(n) = \begin{cases} \mu_q n & n < s_q \\ \mu_q s_q & n \geq s_q \end{cases}$$

The proof will follow by showing that expression (3.1) satisfies the global balance equations. For the description of Section 3.3, these are given by equation (3.10)

below:

$$\left.\begin{array}{l}\pi_n\lambda\mathbb{1}_{\{\sum_{q=1}^{3} n_q<M\}}+\\ \pi_n\mu_1\left(n_1\right)\mathbb{1}_{\{n_1>0\}}+\\ \pi_n\mu_2\left(n_2\right)\mathbb{1}_{\{n_2>0\}}+\\ \pi_n\mu_3\left(n_3\right)\mathbb{1}_{\{n_3>0\}}\end{array}\right\} = \left\{\begin{array}{ll}\pi_{n+e_3}\mu_3\left(n_3+1\right)\mathbb{1}_{\{\sum_{q=1}^{3} n_q<M\}}+ & (3.10.1)\\ \pi_{n-e_1}\lambda\mathbb{1}_{\{n_1>0\}}+ & (3.10.2)\\ \pi_{n+e_1-e_2}\mu_1\left(n_1+1\right)\mathbb{1}_{\{n_2>0\}}+ & (3.10.3)\\ \pi_{n+e_2-e_3}\mu_2\left(n_2+1\right)\mathbb{1}_{\{n_3>0\}} & (3.10.4)\end{array}\right. \quad (3.10)$$

More detailed, we will show that every line (3.10.i) is balanced separately by assuming expression (3.1) to be correct. To show this for (3.10.1) , first note that the indicator functions in the left and right hand side are identical and therefore cancel out. Hence, we only need to show that $\pi_n\lambda = \pi_{n+e_3}[\mu_3(n_3+1)]$. This can be rewritten to show:

$$\frac{\pi_{n+e_3}}{\pi_n} = \frac{\lambda}{\mu_3(n_3+1)}$$

By substituting expression (3.1) in this equation this equation holds. Similarly, by again assuming and substituting expression (3.1), we verify this for (3.10.2) , (3.10.3) and (3.10.4) by:

$$\frac{\pi_{n-e_1}}{\pi_n} = \frac{\mu_1(n_1)}{\lambda}$$

$$\frac{\pi_{n+e_1-e_2}}{\pi_n} = \frac{\mu_2(n_2)}{\mu_1(n_1+1)}$$

$$\frac{\pi_{n+e_2-e_3}}{\pi_n} = \frac{\mu_3(n_3)}{\mu_2(n_2+1)}$$

Hence, we have proven expression (3.1) to satisfy equation (3.10).

3.9 Appendix II: Proof of Theorem 2

Let π^A_{q,n_q} denote the queue length distribution upon arrival at station q, defined as the probability of encountering n_q jobs already present upon arrival at station $q \in \{1,2,3\}$. Then,

$$\mathbb{P}(W_q > t) = \sum_{n_q=0}^{\infty} \sum_{n_{q'}=0;\forall q'\neq q}^{\infty} \mathbb{P}(W_q > t|n)\pi^A_{q,n} \quad (3.11)$$

Using arguments of quasi-reversibility as in Kelly [118], it can be argued that the distribution upon arrival at a station is equal to that as a time average; this is similar to the well-known PASTA (Poisson Arrivals See Time Averages, Wolff [185]) property at the Registration station.

Alternatively, this can also be shown directly by using departure rates from one station to a next one - station q - conditional on the number of donors n_q. Clearly, for $q = 1$, we directly conclude that $\pi^A_{q,n_q} = \pi_{q,n_q}$ due to the well-known PASTA property. For $q \in 2, 3$, these probabilities are obtained as the fraction of all transitions from station $q-1$ to station q that encounter n_q jobs at station q, i.e.:

$$\pi^A_{q,n_q} = \frac{\sum\limits_{q' \neq q} \sum\limits_{n_{q'}=0}^{\infty} \pi_{n+e_q} \mu_{q-1}(n_{q-1}+1)}{\sum\limits_{q'=1}^{3} \sum\limits_{n_{q'}=0}^{\infty} \pi_{n+e_q} \mu_{q-1}(n_{q-1}+1)} \tag{3.12}$$

For notational clarity, first consider $q = 2$. Based on the factorization $\pi_n = \pi_{1,n_1} \cdot \pi_{2,n_2} \cdot \pi_{3,n_3}$, and by summing over all possible options for n_1 and n_3, expression (3.12) can then be rewritten as:

$$\begin{aligned}
\pi^A_{2,n_2} &= \frac{\sum\limits_{n_1=0}^{\infty} \pi_{1,n_1+1} \mu_1(n_1+1) \pi_{2,n_2} \sum\limits_{n_3=0}^{\infty} \pi_{3,n_3}}{\sum\limits_{n_1=0}^{\infty} \pi_{1,n_1+1} \mu_1(n_1+1) \sum\limits_{n_2=0}^{\infty} \pi_{2,n_2} \sum\limits_{n_3=0}^{\infty} \pi_{3,n_3}} \\
&= \frac{\sum\limits_{n_1=0}^{\infty} \pi_{1,n_1+1} \mu_1(n_1+1)}{\sum\limits_{n_1=0}^{\infty} \pi_{1,n_1+1} \mu_1(n_1+1)} \cdot \frac{\pi_{2,n_2} \cdot 1}{1 \cdot 1}
\end{aligned}$$

The same reasoning applies for $q = 3$. Since the waiting time at a station q is independent of the number of donors at other stations and the product form can be written in factorizing terms, $\pi_n = \pi_{1,n_1} \cdot \pi_{2,n_2} \cdot \pi_{3,n_3}$, expression (3.11) can now be rewritten as:

$$\begin{aligned}
\mathbb{P}(W_q > t) &= \sum_{n_q=0}^{\infty} \sum_{n_{q'}=0; \forall q' \neq q}^{\infty} \mathbb{P}(W_q > t | n_q) \pi_{q,n_q} \prod_{q' \neq q} \pi_{q',n_{q'}} \\
&= \sum_{n_q=0}^{\infty} \mathbb{P}(W_q > t | n) \pi_{q,n_q}
\end{aligned}$$

As donors do not have to wait unless there are at least s_q other donors, we can

change the summation to

$$\begin{aligned}&\sum_{n_q=s_q}^{\infty} \mathbb{P}(W_q > t|n)\pi_{q,n_q}\\ =&\mathbb{P}(W_q > t)e^{-(1-\rho_q)s_q\mu_q t}\end{aligned}$$

Here the last equality can be argued by considering that the conditional probability is just a sum of $(n_q - s_q)$ exponential distributions (i.e. an Erlang $((n_q - s_q), \mu_q)$ distribution). A more detailed version of its steps can be found in Literature, e.g. see Cooper [53] (pages 68-69 and 96-97).

3.10 Appendix III: Algorithm to compute total delay

3.10.1 Defining an expanded Continuous Time Markov Chain.

To start the procedure we first need to extend the state description and the state space, in order to keep track of a donor until he/she leaves the system. Therefore, our new state space is:

$$\begin{aligned}S = \{(n_1, n_2, n_3, l, p) \in \mathbb{N}^5 \,|\, 0 \leq n_q \leq N_q, q \in \{1, 2, 3\};\\ 0 \leq l \leq 3; 0 \leq p \leq N^{\max}\}\end{aligned} \tag{3.13}$$

where $N^{\max} = \max\{N_1, N_2, N_3\}$, l is the location of the donors, i.e. the station number, and p is the position of the donor in the queue of station l, including donors in service. Recall that n_q and N_q denote the number of present donors and the maximum number of donors that can be present at station $q \in 1, 2, 3$. Additionally, we define state $(n_1, n_2, n_3, 0, 0)$, which is reached when the donor being tracked exits the system, leaving the system in state (n_1, n_2, n_3) after being absorbed.

Since l is the station where the donor resides, s_l is the number of staff members at the station where the donor is. So, if $p > s_l$ then all staff members are busy, and there are s_l donors in service, and there are $p - s_l - 1$ donors in the queue ahead of the tagged donor. If there are multiple staff members at a station and the donor being tracked is being served, then multiple states have the same associated system. For example, the states $(3, 5, 4, 2, 1)$ and $(3, 5, 4, 2, 2)$ both represent that the tracked donor is in service at the Interview station (assuming the Interview station has at least 2 staff members).

Hence, we construct an absorbing Continuous-Time Markov Chain to measure the time that a tagged donor spends in the system. For ease of notation, we extend the definition of the indicator function $\mathbb{1}_{\{[condition]\}}$ to include multiple conditions. The function returns 1 if all conditions are satisfied, and 0 otherwise. For the same reason, we also define $n_q^{\min} = \min\{n_q, s_q\}$.

Let $Q_{n,n'}$ be the corresponding generator matrix for a transition from $n = (n_1, n_2, n_3, l, p)$ to $n' = (n'_1, n'_2, n'_3, l', p')$ for this Markov Chain:

$$Q_{n,n'} =$$

$$\begin{cases} \lambda \cdot \mathbb{1}_{\{n_1<N_1, n_2<N_2, n_3<N_3\}} & n' = (n_1+1, n_2, n_3, l, p) \\ (n_1^{\min}-1)\mu_1 \cdot \mathbb{1}_{\{l=1, p\leq s_1, n_2<N_2, n_3<N_3\}} & n' = (n_1-1, n_2+1, n_3, l, p) \\ n_1^{\min}\mu_1 \cdot \mathbb{1}_{\{l=1, p>s_1, n_2<N_2, n_3<N_3\}} & n' = (n_1-1, n_2+1, n_3, l, p-1) \\ n_1^{\min}\mu_1 \cdot \mathbb{1}_{\{l\neq 1, n_2<N_2, n_3<N_3\}} & n' = (n_1-1, n_2+1, n_3, l, p) \\ \mu_1 \cdot \mathbb{1}_{\{l=1, p\leq s_1, n_2<N_2, n_3<N_3\}} & n' = (n_1-1, n_2+1, n_3, 2, n_2+1) \\ (n_2^{\min}-1)\mu_2 \cdot \mathbb{1}_{\{l=2, p\leq s_2, n_1<N_1, n_3<N_3\}} & n' = (n_1, n_2-1, n_3+1, l, p) \\ n_2^{\min}\mu_2 \cdot \mathbb{1}_{\{l=2, p>s_2, n_1<N_1, n_3<N_3\}} & n' = (n_1, n_2-1, n_3+1, l, p-1) \\ n_2^{\min}\mu_2 \cdot \mathbb{1}_{\{l\neq 2, n_1<N_1, n_3<N_3\}} & n' = (n_1, n_2-1, n_3+1, l, p) \\ \mu_2 \cdot \mathbb{1}_{\{l=2, p\leq s_2, n_1<N_1, n_3<N_3\}} & n' = (n_1, n_2-1, n_3+1, 3, n_3+1) \\ (n_3^{\min}-1)\mu_3 \cdot \mathbb{1}_{\{l=3, p\leq s_3, n_1<N_1, n_2<N_2\}} & n' = (n_1, n_2, n_3-1, l, p) \\ n_3^{\min}\mu_3 \cdot \mathbb{1}_{\{l=3, p>s_3, n_1<N_1, n_2<N_2\}} & n' = (n_1, n_2, n_3-1, l, p-1) \\ n_3^{\min}\mu_3 \cdot \mathbb{1}_{\{l\neq 3, n_1<N_1, n_2<N_2\}} & n' = (n_1, n_2, n_3-1, l, p) \\ \mu_3 \cdot \mathbb{1}_{\{l=3, p\leq s_3, n_1<N_1, n_2<N_2\}} & n' = (n_1, n_2, n_3-1, 0, 0) \end{cases}$$

with diagonal elements:

$$D_{(n_1,n_2,n_3,l,p)} = -1 \cdot \sum_{(x_1,x_2,x_3,x_4,x_5)\in S\setminus(n_1,n_2,n_3,l,p)} Q_{(n_1,n_2,n_3,l,p),(x_1,x_2,x_3,x_4,x_5)}$$

Clarification of the transition matrix. The first transition represents an arrival. This can only take place if none of the stations is congested. The new state will have one extra donor in the Registration station.

The second transition is a completion at the Registration station of an untagged donor, when the tagged donor is in service at this station. This is only possible if none of the other stations are congested. The position of the donor has to be lowered by one, unless he/she is the only donor left at this station. Furthermore there will be one less donor at the Registration station, and one more at the Interview station.

The third transition is a completion at the Registration station of an untagged donor, when the tagged donor is at this station, but not in service. Again, this can only happen if none of the other stations is congested. After this transition there will be one less donor at the Registration station, one more at the Interview station, and the position of the tagged donor has to be lowered by one, as he/she has moved up one place in the queue.

The fourth transition is a completion at the Registration station of an untagged donor, when the tagged donor is at another station. The same conditions and new state as the previous transition hold, with the exception that the position of the tagged donor doesn't need to be lowered as he/she hasn't moved up in his/her queue.

The fifth transition is a completion of the tagged donor at the Registration station. This is only possible if he/she is actually in service at this station. In the new state the tagged donor will be at the end of the line at the Interview station.

Transitions 6 and 10 are similar to transition 2, transitions 7 and 11 are similar to transition 3, transitions 8 and 12 are similar to transition 4 and transitions 9 and 13 are similar to transition 5.

Note that N_q is conceptually allowed to be infinite. Also note that the last transition, transition 13, into state $(n_1, n_2, n_3 - 1, 0, 0)$ represents that the tagged donor leaves the system. This is to be interpreted as an absorption.

3.10.2 Time-discretization

For computational purposes we now need to transform the continuous-time Markov Chain (CTMC) into a discrete-time Markov Chain (DTMC). We therefore apply the well-known method of Time-discretization - for more details on this and related methods, see Chapter 2. Based on this method, we can then compute the transition matrix P_t for the continuous-time Markov Chain over a time period of length t by:

$$P_t = \sum_{k=0}^{\infty} e^{-\alpha t} \frac{(\alpha t)^k}{k!} P^k \tag{3.14}$$

where I is the identity matrix and P is defined by:

$$P = I + \frac{1}{\alpha} Q \tag{3.15}$$

with

$$\alpha \geq \max_{(n_1,n_2,n_3,l,p) \in S} -1 \cdot D_{(n_1,n_2,n_3,l,p)}$$

The interpretation of this DTMC is that each time-step has an exponential duration with parameter α, hence an average duration of $1/\alpha$. Let π^k be the corresponding state probability vector of the DTMC after k time-steps.

As the DTMC still involves an infinite matrix in its current description, we restrict the system and include finite limitations on the number of donors that can be present in stations 1, 2 and 3 by N_1, N_2 and N_3 respectively. From this exponential DTMC the total delay time distribution can now be computed with the algorithm below.

3.10.3 Numerical Algorithm to track a tagged donor.

Step 1. Definitions.

Let $S' = \{(n_1, n_2, n_3) \in \mathbb{N}^3 | 0 \leq n_q \leq N_q\}$. Notice the difference with S from (3.13)

Let $\pi^{(k)}_{(n_1,n_2,n_3,l,p)}$ be the probability to be in state (n_1, n_2, n_3, l, p) after k time-steps.

Step 2. Initialization.

For all $(x, y, z) \in S'$ (state of the systems at the arrival of the tagged donor):

Set $k = 0$, to indicate no time has passed sice the tagged donor entered.

For $k = 1$ to K (truncation value)

Step 3. Iteration.

Compute:

$$\pi^{(k)}_i = \sum_{j \in S} \pi^{(k-1)}_j P_{j,i} \tag{3.16}$$

Compute the probability that the donors spends less than or equal to k time-steps with exponential length (with parameter α in the systems, given that the donors arrived in state (x, y, z):

$$\mathbb{P}(T_\alpha \leq k | (x, y, z)) = \sum_{(n_1,n_2,n_3) \in S'} \pi^{(k)}_{(n_1,n_2,n_3,0,0)} \tag{3.17}$$

end of both loops, back to step 2 until all $(x, y, z) \in S'$ have been used

Step 4. Back to continuous time.

For any time t, the delay time distribution can be approximated by taking k large and $\alpha = k/t$ (also see Remark 1)

$$\mathbb{P}(T \leq \frac{k}{\alpha} | (x, y, z)) \approx \mathbb{P}(T_\alpha \leq k | (x, y, z)) \tag{3.18}$$

for α large enough

Step 5. Combination with product form.

With $\pi_{(x,y,z)}$, as given by (3.1), compute the unconditional delay time distribution using (3.8)

$$\mathbb{P}(T \leq t) = \sum_{(x,y,z) \in S'} \pi_{(x,y,z)} \mathbb{P}(T \leq t | (x, y, z)) \tag{3.19}$$

Explanatory notes on the Algorithm. The goal of the algorithm is to find the delay time distribution of a blood collection site, by keeping track of the number of time-steps it a tagged donor to pass through the system. The tagged donor arrives at time 0, in state (x, y, z, l, p). Here (x, y, z) represents the number of donors in

the stations 1, 2 and 3 respectively. The l(ocation) label reflects the station where the tagged donor currently resides. The p(osition) label gives the position that the donor currently occupies in the queue at station l. The donor leaves the system as soon as these labels become $p = l = 0$.

For notational purposes, the first computational step introduces an extra, secondary state-space which only consists of the first three components of the total state-space. It also introduces the probability distribution $\pi^{(k)}_{(n1,n2,n3,l,p)}$ of the DTMC after k time-steps.

The second computational step is an initialization step, and sets up the main, iterating step. This has to be run for every initial state (x, y, z) in the secondary state-space S'. First, the index and the probability distribution are initialized. At the start of the algorithm the tagged donor arrives at the system, which is in state (x, y, z). Therefore the probability of being in state (x, y, z) with the tagged donor at the end of the queue in the Registration station is 1, while all other states have probability 0.

In the third computational step, the initial probability distribution $\pi^{(0)}_{(n1,n2,n3,l,p)}$ is multiplied with the transition probability matrix from (3.15). This leads to $\pi^{(1)}_{(n1,n2,n3,l,p)}$, the probability distribution after 1 time-step. After this time-step is completed it is possible to sum over all states that have $l = p = 0$. This summation gives the probability that the tagged donor has left the system in this time-step or earlier. This one-step procedure is repeated until some maximum number of K time-steps has been reached. Then the algorithm will move on to the next initial state (x, y, z), and again start with step 2. In the fourth computational step the probability of leaving the system in k or less time-steps is converted to leaving the system within k/α time units, conditional on arriving in state (x, y, z). Finally, in step 5, as we know the probability of arriving in state (x, y, z) from the product form expression (3.1), and by using PASTA, the conditional expression can be summed over all initial states (x, y, z). We thereby obtain the unconditional total delay time distribution.

Remark 1. By (3.17) we have exactly computed the delay time distribution expressed in k exponential phases with parameter α. The corresponding Erlang distribution with k phases and parameter α has an expectation of k/α and variance k/α^2. Hence, for t fixed, $k = t\alpha$, and α enlarging, the variance will converge to 0 and the expectation will converge to t. By recalling that α can be chosen arbitrarily large, the approximation (3.18) will then be exact for $\alpha \to \infty$. In fact, in Section 3.6.4 we show that doubling α has no effect on the outcome of the algorithm, so the approximation seems to work quite well.

Remark 2. In this algorithm, the maximum delay time for which the probability of absorption is calculated, is $T_{\max} = K/\alpha$. To get a more accurate approximation, it is possible to increase α from the given value. When α is increased, it is recommended to increase K by the same factor to ensure that $T_{\max}$ remains the same.

CHAPTER 4

Queue length computation of time dependent queueing networks

4.1 Introduction

Service systems with free walk-in arrivals rarely have a steady number of arrivals during business hours. Some time intervals will have a high average number of arrivals, and others a low number of arrivals. This happens because people - in general, customers - usually have preferences for certain time intervals. These preferences usually depend on the type of system. Systems with a short service and sojourn time often experience peaks in arrivals just before or after standard office working hours and during lunchtime (e.g. checkout at supermarkets). Systems with longer service and sojourn times usually see preferences for certain days of the week (e.g. hospitals). Due to breaks and part-time employees (servers), service capacities might also not be uniform throughout the day. When standard queueing theory is applied to these kinds of situations, often general queueing expressions are used like Little's law, M/M/s expressions and product forms. However, these methods generally assume a steady-state situation. I.e., these expressions rely on averages over an infinite time-horizon situation. A time-dependent system will usually tend to this average, which is the reason a steady-state approach is well justified in systems that are time-dependent, but only change very slowly. However, if system changes occur relatively often compared to process events - arrivals and service completions - steady-state approaches can give results that do not match reality.

There are several ways to address these steady-state issues. Time-dependent queues have frequently been addressed in literature, as shown in the review paper by Schwarz et al. [164]. Most of the papers that work with time-dependent queues are of a technical nature. These papers generally deal with a single queue, often with just one server. In reality, however, service systems will consist of multiple stations, each handling part of the total demanded service. In addition, each of these stations might have multiple servers. In this chapter, we will show the application of one of the more common methods to deal with transient - i.e. over a finite time horizon - continuous time systems or time-dependent queues: uniformization. As outlined in Chapter 2, uniformization is a method that can be used to analyze a continuous time Markov chain by transforming it into a discrete time Markov chain. As well as being one of the most frequently used methods, uniformization is also considered one

of the methods performing best in literature. An extensive discussion of the theory behind uniformization, and a number of other extensions can be found in Chapter 2. In contrast to the more theoretical and numerical approach in Chapter 2, this chapter will not go into the theory, but focus on the practical application of the method in a time-dependent setting.

The approach presented in this chapter is widely applicable. In this chapter, we will focus on the application that is central to this thesis: the donation process at blood collection sites. A comprehensive description of the process can be found in Section 1.3 in Chapter 1. A large portion of the blood donations, particularly whole blood donations, occurs on a free walk-in basis, and donors prefer certain times during the day for a blood donation, as can be seen in Figure 4.1. Peaks in arrivals are evident early in the morning (around 8.30 am) and during lunchtime (around 12.30 pm). The largest peak in arrivals is after standard office working hours, between 6 pm and 7.30 pm. The difference between the highest and lowest arrival rates can easily reach a factor of 2.5. This clearly indicates that steady-state methods will not be reliable in estimating queue lengths and a transient computation will thus be preferable.

We will start by discussing the relevant literature in Section 4.2, followed by the introduction of the model we will use in Section 4.3. We will then present the steps of the computational method for time-dependent systems in Section 4.4. The technical details of the method will be included in Section 4.7. The results in Section 4.5 are split into two parts. Section 4.5.1 shows an extensive comparison of the time-dependent method and a steady-state approach to demonstrate the benefit of a time-dependent approach. The second part of the results in Section 4.5.2 uses the time-dependent method to analyze several scenarios that aim to decrease queues at blood collections sites. The chapter will be concluded with a discussion in Section 4.6.

4.2 Literature

The study of time-dependent queueing systems is not a new research field. Its first appearance goes back to 1931, in a paper by Kolmogorov [122]. Since then, a large number of papers have appeared, as shown by an extensive and recent literature review by Schwarz et al. [164]. There are two aspects in which papers discussed in the review might be related to our research: the application area and the computational methods.

Blood collection sites have been studied before, as can be seen in [8, 29, 38, 39, 154, 173]. A detailed discussion of these papers can be found in Section 1.4 in Chapter 1. Most of these papers describe a setting that is inherently time dependent. Nevertheless, none of these papers deal with time-dependent aspects of blood collection sites, let alone use uniformization. Therefore, we will first discuss related papers with an application to health care settings in general. Nine papers dealing with time-dependent queues in health care settings are included in the review by Schwarz et al. [2, 26, 36, 40, 52, 90, 167, 176, 189]. None of these papers use uniformization, but two papers [26, 36] use a related approach developed by Brahimi

and Worthington [37]. Their approach also uses a probabilistic one step transition mechanism. However, the method is based on discrete-valued service times, and the transition probabilities are calculated differently. The other health care related papers use a variety of different, less related methods in their time-dependent computations.

The second aspect in which some papers are related, is the method used: uniformization. Jensen [117] introduced uniformization in 1953. Since then, countless papers have used, applied and improved the method (e.g. [139, 157]). The review by Schwarz et al. mentions six papers contributing to the development of uniformization for time-dependent queueing systems. The first of these six papers, by Grassmann [94], looks at the transient behavior of an M/M/1 queue, but allows for constant parameters only. The other five papers [16, 54, 64, 77, 99] all consider time-dependent parameters. [64, 99] provide exact solution methods, but do not include numerical computations, while [16, 54, 77] include approximative computations. Our computational implementation of the algorithm is roughly based on one of these, the paper by Arns et al. [16]. Ingolfsson et al. [111] compare uniformization to six other methods, including an exact one. It is concluded that uniformization is consistently closest to the exact method and often nearly indistinguishable. In most of their examples, though, it is also the slowest approximation method. For our application, and probably many more applications, this is not a major problem. Our Matlab implementation is able to compute results in a couple of seconds (for more information on computational times, see Section 4.4).

There are quite a few papers that apply methods similar to the method described by Brahimi and Worthington [37] - that is, related to uniformization - for real-world problems. However, there only seem to be two papers [112, 138] that use uniformization in a real-world, time-dependent context. Both of these focus on a single-queue call center, in contrast to the multiple station queueing network covered in this chapter.

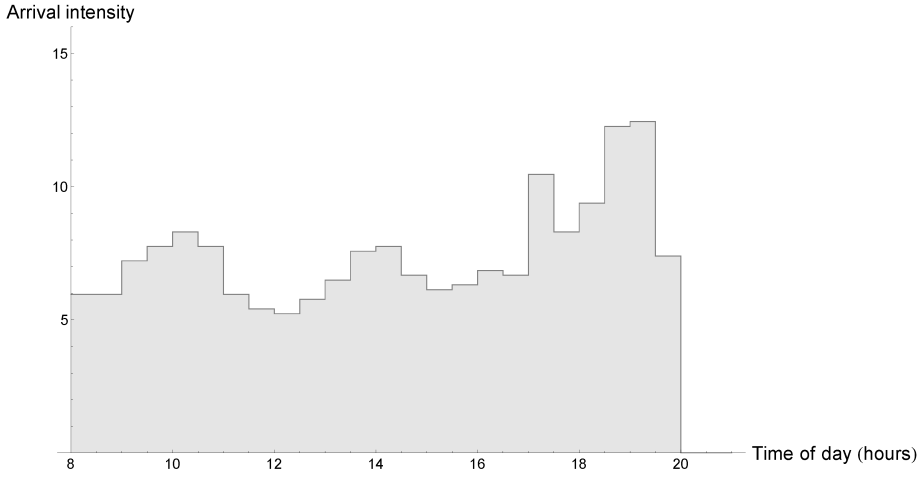
4.3 Model

4.3.1 The blood collection site as a test case

In the remainder of this chapter we will only consider whole blood donations. The most important reason to exclude plasma donations is the appointment system for this kind of donations. This means that the arrival patterns - which are the most important reason for the time-dependent approach - are far less pronounced or even non-existent for plasma donations. The remaining arrival process of whole blood donors is highly time dependent, as can be seen in Figure 4.1. This makes the process of whole blood donation at Dutch blood collection sites a good application of time-dependent queue computation.

As well as being a good test case to show why time-dependent queue estimation is sometimes beneficial, blood collection sites also provide an opportunity to show the effects of several scenarios. Because the process has multiple stations and most staff members are trained to execute all of the different tasks in these stations, there are

Figure 4.1 Average arrival pattern of whole blood donors for a full day's blood collection session in the Netherlands. Arrival rates per half hour are shown.



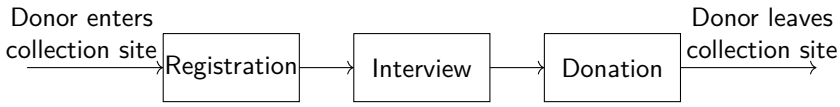
several options for changing the allocation of staff. This will even further increase the time-dependent nature, but will also level the work load and in most cases decrease waiting times. We will, among other scenarios, show the consequences of changing staff allocation just after opening in Section 4.5.2.1, and towards closing time in Section 4.5.2.2. The changes proposed in these sections are quite obvious, but using our method, we can show and predict the effects these changes have on queue length distributions at the blood collection site, while taking time-dependent aspects into account. We will also show the effects of applying the staff scheduling algorithm developed for blood collection sites, and presented in Chapter 5.

In all results in Section 4.5 the arrival rate follows Figure 4.1 - unless mentioned otherwise. This comes down to an average arrival rate of 15 donors per hour. Note that Figure 4.1 shows the arrival rate per half hour. Nine staff members are available throughout the day, one being allocated to the first station, three to the second station and five to the third station. For all results in Section 4.5 we have limited the number of donors who can be present at any of the stations to twelve. This has been done to limit the state-space, a necessity to be able to compute results. Although a limit is necessary, it can easily be increased or decreased.

4.3.2 Technical model

The computational method that will be described in Section 4.4 works for any Jackson network, the well-known type of queueing network introduced by Jackson [114]. A Jackson network contains a number of service stations, each containing a fixed number of servers. Every station can have an outside arrival stream, and arrivals from any other station. When a customer- in this case a donor - completes service at one of the stations, it will instantaneously be routed to a next station by some fixed

Figure 4.2 Model of the blood collection site used throughout the chapter.



probability distribution, independent of the current network state. This next station can be any other station and the "outside world". If the customer selects the outside world, it means that the customer will leave the network.

We will assume all arrival rates and service rates to be Poisson, as is standard in queueing theory. This means that the number of events (either arrivals or service completions) in a given time interval is Poisson distributed. This leads to exponentially distributed times between events, so both the time between two arrivals and the time it takes for one customer to receive service at a station are exponentially distributed. An exponential time between two arrivals leads to memoryless arrivals, meaning that the distribution of the time until the next arrival at any point in time is independent of the time since the last arrival. This is a natural assumption for free walk-in arrivals, as it means that the previous arrival has no influence on the next arrival, i.e. arrivals are independent.

Exponential service times have a similar memoryless property, though interpreted differently. For services the memoryless property means that the time since service began does not influence the probability of completion of service. In most cases this assumption for services does not fully reflect reality, but it is imperative for computational methods like the one presented in this chapter to remain feasible. In most cases the exponential service time assumption leads to an overestimation of waiting and sojourn times, which means that the computations based on these models are usually conservative, i.e. on the safe side. Using this approach, it is also possible to use Erlang distributions or even more general phase-type distributions - a combination of exponential distributions, see Erlang [79] - but this would affect the computational time of the algorithm.

To model the blood donation process, we use a simplified model of a blood collection site. In this model, we distinguish the three main stations of the donation process: the Registration station, the Interview station and the Donation station. The model is visualized in Figure 4.2. This model has been shown to effectively represent a blood collection site in the Netherlands in Chapter 3. Each of the three stations has its own queue. Donors arrive only at the first station, they go sequentially to the second and third stations before leaving the system.

4.4 Methods

As previously mentioned, we present an approach that is based on uniformization, also referred to as randomization, as extensively discussed in Chapter 2. The approach starts by splitting the entire continuous time period in short time intervals during which the system parameters are assumed to be constant. The subsequent

algorithm can be split into two main stages. The first stage of the algorithm, explained in Section 4.4.1, uses uniformization to compute the queueing distribution of the individual intervals. The second stage connects the intervals, and is explained in Section 4.4.2. The technical details of the algorithm can be found in Section 4.7.

4.4.1 Constant intervals

Although it is possible, or even likely, that the system is changing continuously over time, it is usually impossible to estimate arrival and service rates for a single point without some approximation. In contrast, it will usually be possible to get these estimates for a time interval directly from data. The time that has to be evaluated will therefore first be divided into short intervals. These intervals can be arbitrarily short and we will assume that the system is homogeneous during these intervals.

To start the algorithm, we first need an initial vector that contains the probabilities of being in a state - the probability vector - of the system. The state of the system in this case is a combination of all the numbers of people at the different stations. The initial probability vector might be a probability of 1 for an empty system, or any other probability vector based on historical data. As can be seen in Section 4.5.2.1, it is also possible to run instances with different initial vectors.

The algorithm then proceeds to the first interval. For this interval, a so-called generator matrix will be computed. This matrix contains the rates at which the system changes from one state to another, for all possible initial and subsequent states. Of course most of these rates will be 0. With this generator matrix, we can discretize the process. By using the steps described in Section 4.7, we can build a transition matrix from the generator matrix. This matrix contains the probability that the system will end up in some state after one transition, given the previous state. This means that if we multiply the probability vector with this transition matrix, we will end up with a probability vector after one event. If we multiply the vector by the matrix two times, we get the state vector after two events. We can continue this process to get the probability vector after k transitions.

Since we assumed all rates to be Poisson, and the sum of Poisson rates is again a Poisson rate, we know that the number of events k during a given time interval t is Poisson distributed. This means that both the probability of k transitions in a time interval and the probability vector after k transitions are known. If the probability of k events and the probability vector after k events are multiplied, and then summed over all possible k, the probability vector after the time interval t can be calculated. This will work perfectly in theory, but there is one problem in practice. The Poisson distribution gives a non-zero probability for every positive integer value of k, implying that an infinite sum should be used. The probabilities for a very large number of events are very small, so we truncate this sum at some point, i.e. ignoring all numbers of iterations bigger than some K number of events during time interval t. This truncation point has been set such that the Poisson distribution mass that gets ignored is at most 10^{-10}. After these steps, the algorithm has computed the probability vector after one time-step, with a stochastic number of transitions during this time interval.

4.4.2 Interconnected intervals

As long as the system parameters do not change, the time interval described in Section 4.4.1 can be made as long as desired, without losing precision. However, as soon as the system parameters do change, the method runs into problems, as the time at which an event takes place starts to influence the probabilities of ending up in different states. To solve this, we stop the iteration of the uniformization algorithm as soon as one of the system parameters changes. This can be caused by a change in the number of servers, which is usually known in advance, or a change in the rates at which changes occur, which can be derived from historical data.

If the algorithm completes one iteration, it has computed the probability vector after this iteration. This result can be used as an initial probability vector for a new iteration of the uniformization algorithm, with a new generator matrix. At this point, the algorithm described in Section 4.4.1 can start over again. This process of starting a new iteration can be continued until the algorithm reaches the end of the opening hours, or another ending condition occurs. After the algorithm terminates, the queue length distribution can be calculated at the start and end of every uniformization iteration.

4.4.3 Steady-state comparison and numerical implementation

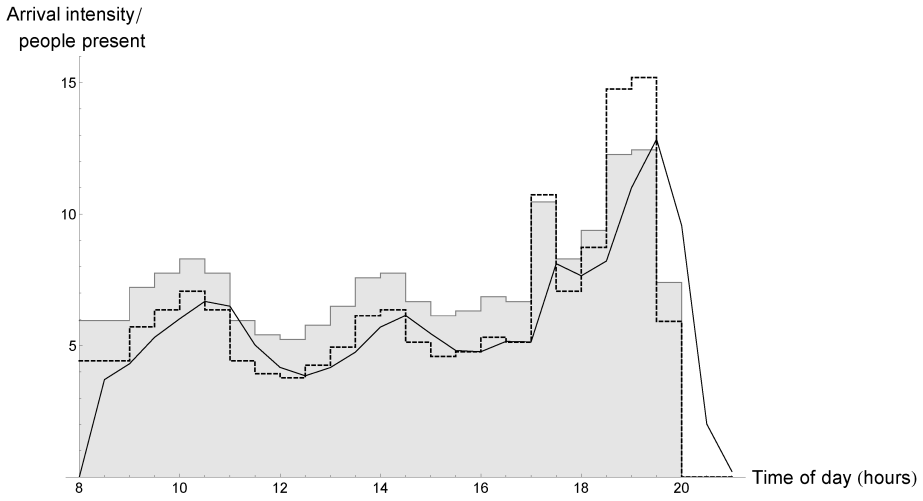
To show the benefit of the time-dependent method, we will compare it to a steady state method. The steady steady state method that will be used is based on the product form expression in Chapter 3. The product form expression solves the birth death equations of a time-homogeneous version of the blood collection site. To compute the results in Section 4.5.1, the system parameters for every interval have been used to compute the steady state for these parameters.

The time-dependent algorithm and the product form have been implemented in both Mathematica 9.0 and Matlab R2015b. Both implementations of the time-dependent algorithm were run on a machine with an Intel Core i5-3437U processor and 8GB of RAM. The Mathematica implementation takes a little less than 10 minutes to compute the queue lengths for a blood collection session of an entire day, if the interval for which calculations are done is set to half an hour. This might be too long for practical purposes. However, the Matlab implementation can do the same computation in a few seconds, making it a very useful tool for a practical application.

4.5 Results

We will first illustrate the benefit of a time-dependent computational queueing method, supported by an example for blood collection sites. Subsequently, based on this application, we will apply the computations to a number of possible improvement scenarios.

Figure 4.3 Difference between the time-dependent method (solid black line) and the steady-state method (dashed black line) for a full collection session. Both methods show the average total number of donors in the collection site. Total in this case means that the number of donors at all of the three stations are added up. The arrival rate is displayed in the background (gray)



4.5.1 Comparison with steady-state method

To show the benefit of time-dependent queueing methods, we will compare the computational method described in Section 4.4 with a steady-state computation. This steady-state computation is based on an exact product form result, shown in Chapter 3.

The output of the computational method described in Section 4.4 is a full probability distribution over all possible states of the system. This means that there is a wide range of performance metrics that could be calculated using this method, such as the probability of exceeding a certain number of people in the system, the probability of exceeding a certain number of people at a particular station, the most likely number of people in the system, etc. To keep the results in this section simple we have chosen to show just the average total number of people in the system, in some cases supplemented with the average number of people at some station of the process.

Figure 4.3 shows the average total number of donors for a typical Dutch blood collection site that has a collection session lasting an entire day. The number of donors is estimated with both the time-dependent method - shown with the solid line - and the steady-state computation - shown with the dashed line. The steady-state value is computed by taking the arrival rate for a specific half hour and assuming this to be constant. In the background, the arrival pattern is shown in gray. Arrivals are allowed from 8 am until 8 pm, after which the collection site operates for one more hour to make sure all donors can donate. It is clearly visible that the methods do not match. If the arrival rate is increasing the steady-state method usually overestimates

the number of people present - e.g. from 8 am till 10.30 am, while when the arrival rate is decreasing, the steady-state method underestimates the total number of people - e.g. from 11 am till 12.30 pm. In general, the steady-state method rarely computes the number of people accurately.

4.5.1.1 Process speed

If both the arrival rate and the service rate are multiplied by a positive number x, the total number of customers present will not change in a steady-state method. So, for a steady-state method, the speed of the services does not influence the number of people present as long as arrivals are adjusted in the same way. In other words: time is scalable.

Suppose we compare two systems. The first system has an arrival rate of 4 customers an hour and a service rate of 5 customers an hour. The second system has an arrival rate of 8 customers an hour and a service rate of 10 customers an hour. In terms of number of people present, these two systems are identical for a steady-state method. In reality, however, the first system will take twice as long to realize the same number of people in the system, if both systems start from the same situation - e.g. empty. With a time-dependent method, this longer 'transition time' will be taken into account.

4.5.1.2 System overload

With a fast changing system, it is likely that at some point in time, the system will get overloaded, i.e. the total arrival rate exceeds the service rate at at least one of the stations in the system. If this happens when a steady-state method is used, two things might happen depending on the other parameters. Results for systems that have some upper bound on the number of people that can be in the system, will become extremely dependent on this upper bound. If the upper bound is increased, the computed average number of people present will increase by a similar percentage. If the system does not have an upper bound, the result for the number of people in the system will simply become infinite. Both of these cases are not an accurate depiction of reality, as the number of people that enter the system during the time it is overloaded, is limited due to time constraints.

The time-dependent method described in this chapter mostly solves this issue by actually tracking the probability that a certain number of people enter the system during the time that the system is overloaded. It does, however, have to include some maximum number of people that can be in the system. Although this will influence results if the system is overloaded for an extended period of time, service systems usually also have a physical limitation on the number of people that can be present in the system, which might make this assumption realistic.

4.5.1.3 System start up

A lot of service systems do not have a continuous period in which service to customers is offered. Most systems will have some down time. This can be during the night,

during the weekend or during holidays. When a system starts up, it is obvious that a sudden change in the system occurs. Before the system starts, there are no services and arrivals will be substantially lower or even non-existent. There are few things that might happen in this situation. Some service systems might have a queue of customers waiting as the system starts, and some might start with empty queues and wait for customers to start coming in.

A steady-state method, however, will assume that the system is in a equilibrium situation right away. This is highly unlikely in systems with arrivals to every queue in the system. If a system contains queues that can only be reached through other queues, and if we assume that the system started empty, it is impossible that the system starts in an equilibrium situation.

The blood collection site always starts empty, although it is possible for donors to line up outside the collection site before it opens. Figure 4.4a shows the average total number of donors in the collection site during the first two hours after the collection site opens. It shows the results for both the steady-state method (dashed line) and the time-dependent uniformization method (solid line). Section 4.5.2.1 discusses a scenario when there are people lined up outside the collection site.

4.5.1.4 System close

When a system closes down, there are multiple actions that can be taken. There are two possible actions on the extreme ends, and some possible combined actions can be thought of. The first of the two extremes is that the arrivals are stopped and the services continue until the system is empty. The other option is to stop both arrivals and services. In this case, usually all customers that were in the queue will leave the system immediately. In this last case, time-dependent methods are less beneficial, because the behavior of the system after it stops can be predicted with absolute certainty. For the case that services continue until the system is empty, however, time-dependent methods are required to predict the behavior of the system after closing. As soon as arrivals stop, a steady-state method would show no more queues. In effect, the difference between both options - to stop all services or continue services - disappears when a steady-state method is used.

In manufacturing systems the extra possibility exists of stopping all arrivals and services and leaving all the queues filled. As customers in service systems are usually people, this option often does not exist in physical service systems.

At blood collection sites, the first option - to continue services until the system is empty - is used, because Sanquin feels the donors should be serviced if they have arrived within opening hours. Figure 4.4b shows the arrival pattern and the average total number of donors in the collection site during the last two hours of the working shift. During the second of these two hours there are no more arrivals.

4.5.1.5 Changing arrival rate

Although the biggest changes in arrival intensity are undoubtedly the opening and closing of a service system, fast changes might also occur during opening hours. On

a very short timescale, changes in arrivals might occur due to the arrival of some form of public transport; a train or a subway. On a slightly longer timescale, changes might occur during lunchtime or after standard working hours. These last type of changes can be predicted easily, and can therefore be used in the prediction of queue lengths. A steady-state method will assume that every change incurs an instant change in queue lengths and number of people present. A time-dependent method will be better able to predict the impact of changes in arrivals

For the blood collection site, changes due to working hours and lunch breaks, are clearly visible in the arrival pattern. Figure 4.4c shows the arrivals from 16.30 to 19.00 hours. During this time, the arrivals to a blood collection site change rapidly. First because people come to the collection site straight from work, then the arrivals go down, most likely due to dinner time, and then go up after dinner time. The steady-state method (dashed line) clearly overestimates the time-dependent method (solid black line) with respect to the number of donors that are in the queues during the busy times, and underestimates during the quiet period from 17.30 till 18.00.

4.5.1.6 Changing number of employees

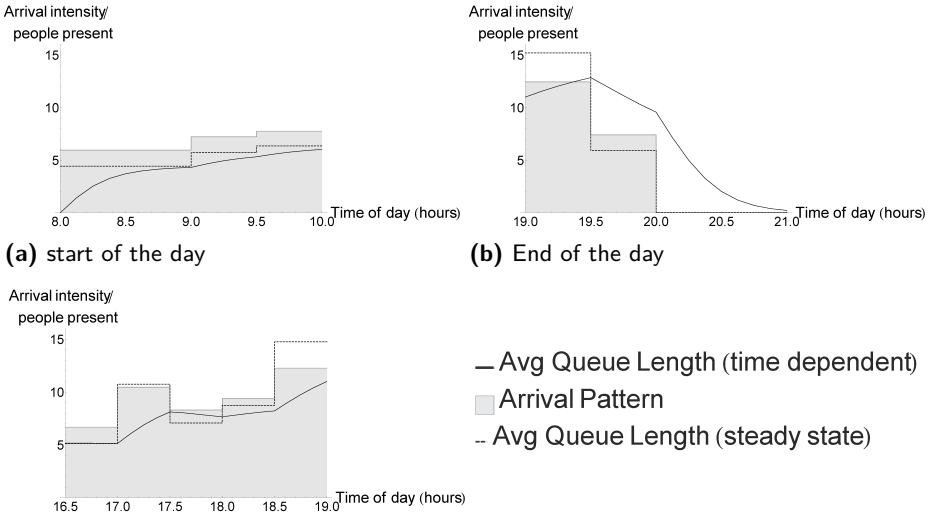
Most of the cases that need time-dependent queueing estimation relate to changes in the arrival rate. However, service rates might also change. On top of the fact that the actual service rate per employee might change due to fatigue or other circumstances, the most obvious change to the service rate is a change in the number of employees that is scheduled. As in previously mentioned situations, a steady-state method would immediately change the number of people at the station where the number of staff members is changed. However, an additional effect might also occur at the subsequent stations of the system. When an extra employee is added to some station, the station starts working through its queue faster, resulting in a short time increase at the subsequent queue. This secondary effect would not happen at all with steady-state estimation.

Changing staff numbers is a common occurrence at blood collection sites. Section 4.5.2.3 shows the result of an algorithm that was developed to optimally change staff numbers throughout the day.

4.5.1.7 Remarks on the difference between steady-state and time-dependent methods

As Figure 4.4 shows, the difference between the time-dependent and steady-state methods can get big enough to cause problems, e.g. if these predictions are used for planning appointments or staff capacity. To emphasize this difference, Table 4.1 shows the difference between the time-dependent and the steady-state method as a percentage of the time-dependent method for two of the three situations in Figure 4.4 - starting the system (opening) and the fast changing arrival rate (changing arrivals). For each of these two instances, the parameters are kept exactly equal to those used for Figure 4.4. Therefore the total time compared for each of the instances is 2 and 2.5 hours respectively. As a table allows for somewhat more details, we not only

Figure 4.4 Difference between the time-dependent method (solid black line) and the steady-state method (dashed black line) for specific times of the day. These figures highlight parts of the results shown in Figure 4.3. For these figures a computational interval of 6 minutes is used to show smoother lines. Both methods show the average total number of donors in the collection site. Total in this case means that the numbers of donors at all of the three stations are added up. The arrival rate is displayed in the background (gray).



(a) start of the day

(b) End of the day

(c) busy time with changes

include the difference between both methods for the total number of people present, but also the differences for each of the three stations of the donation process.

The closing instance could also have been included in Table 4.1, but this would give misleading results. As the steady-state method gives an empty queue as soon as arrivals stop, it would always have a 100 % deviation from the time-dependent method. We have therefore decided to not include these in Table 4.1

For each of the stations and each of the instances, the table shows the difference between both methods as percentages of the time-dependent result. This has been done for three time-points, immediately after a change in the process (start), in the middle of a homogeneous time interval (mid), and just before the next change in the system (end). As the time-dependent queueing method slowly converges to the steady-state estimation, the difference will usually decrease over these three time-points. The results are the absolute differences between the steady-state method and the time-dependent method, as a percentage of the time-dependent method, averaged over the time intervals considered. (i.e. both underestimations and overestimations were counted as positive differences.) For the opening instance the start difference does not include the first time interval, as the time-dependent estimation is 0, and computing the percentage for this time interval would include dividing by 0.

Table 4.1 clearly shows that, as station 1 is the fastest working process, it is

Table 4.1 Differences in total number of donors present between steady-state and the time-dependent method as a percentage of the time-dependent method in different instances.

		Instance	
		Opening	Changing arrivals
Station 1	start	20.1%	70.6%
	mid	5.2%	21.1%
	end	1.3%	10.3%
Station 2	start	16.4%	43.9%
	mid	12.0%	25.0%
	end	3.2%	14.8%
Station 3	start	29.9%	33.9%
	mid	52.9%	28.4%
	end	14.6%	19.9%
Total	start	23.8%	43.2%
	mid	26.8%	25.6%
	end	8.7%	16.1%

fastest to converge to the steady-state estimation, during the three time-points. Conversely, station 3 converges the slowest of each of the stations. It is also visible that the average total difference is always lower than the maximum of the effects for any of the stations. This means that even though the effects are clearly visible in Figure 4.4, the differences for the individual stations might even be bigger. Due to the relatively minor changes in the period after opening - not including the opening itself - and therefore the time-dependent method has time to converge to the steady-state method, resulting in the lowest differences of any of the instances shown.

4.5.2 Scenarios for the blood collection site

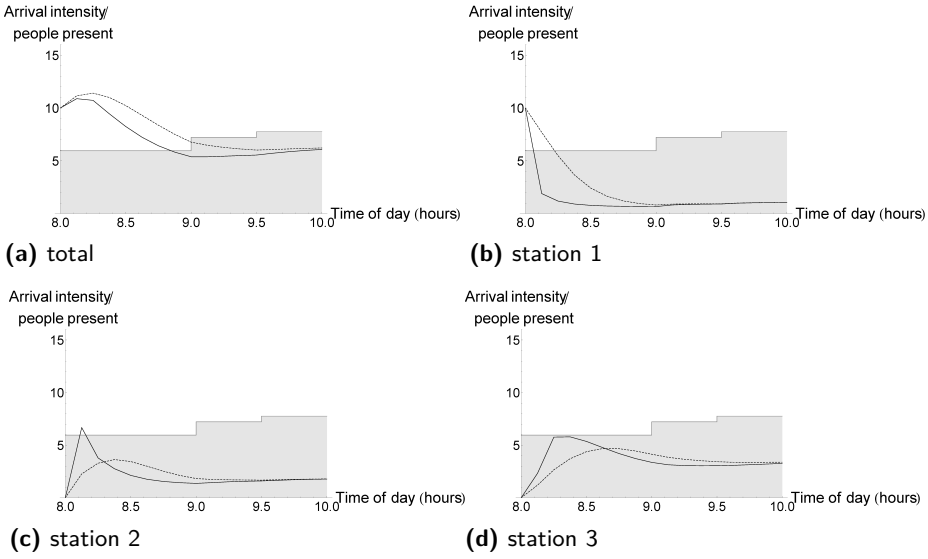
As well as being able to evaluate current performance, the time-dependent method also gives an opportunity to evaluate other scenarios that may improve service. In the next paragraphs we have included some of these scenarios to improve service at blood collection sites, without increasing total capacity. All of the scenarios increase the time-dependent variability of the system, and therefore a time-dependent queueing computation is highly beneficial for the evaluation of these scenarios.

4.5.2.1 Scenario A: Changing allocation after opening

Most staff members at blood collection sites are multi skilled, i.e. they can be allocated to all stations of the donation process. However, in most cases employees will be allocated to just one station for an entire shift. At some times during the day it might be useful to take advantage of the multi skilled employees by changing the allocation of the available employees. Just after opening the collection site is one of these moments, as the second and third stations will be empty until donors have come through the first or the second station respectively.

In the figures and tables shown so far, we have assumed that all queues are empty

Figure 4.5 Average number of people at the different stations of the process. In all figures, the dashed line represents the average number of people for the fixed allocation, while the solid line represents the average for the changed allocation.



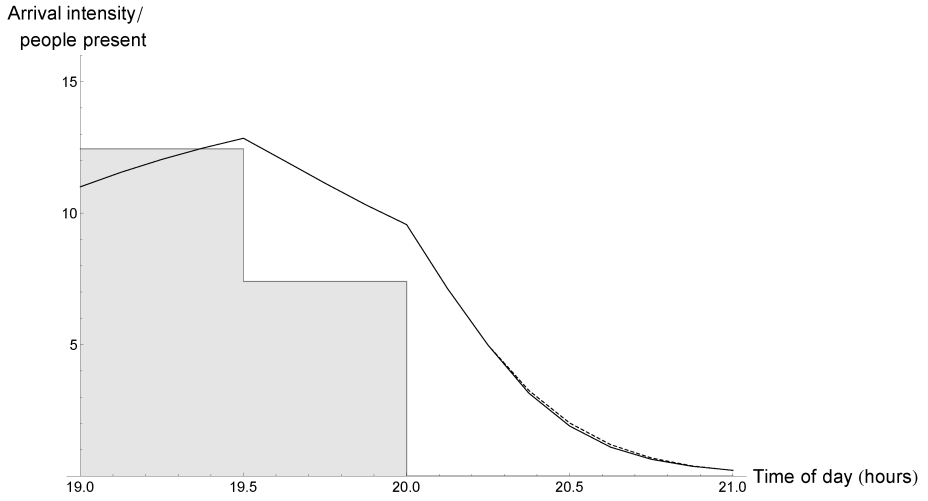
(a) total **(b)** station 1

(c) station 2 **(d)** station 3

when the collection site opens. However, at some collection sites, donors might line up for the Registration station before the collection site opens. In this case it is even more useful to change the allocation, to be able to quickly get these initial donors flowing through the donation process. To illustrate this, Figure 4.5 shows the result of changing the allocation of the staff just after opening (solid line), compared to a fixed allocation (dashed line). In both cases, we assumed that ten donors were waiting when the collection site opens. For the fixed allocation, we take the allocation mentioned in Section 4.3.1: one employee at the first station, three at the second station and five at the third station. For the improved scenario, the allocation was changed for the first 15 minutes. During the first 7.5 minutes, two employees were reallocated from the third to the first station, resulting in three employees for the first station, three for the second station, and three for the last station. During the second 7.5 minutes, the two employees are allocated to the second station, resulting in one employee at the first station, five at the second station and three at the third station. We note that the changes proposed might not always be feasible due to equipment constraints and collection site design. After the first 15 minutes, the allocation is the same as for the fixed staff allocation.

A clear effect of this relatively minor change is visible in Figure 4.5. The average total number of donors in the system shows a reduction of up two donors, and an average decrease of 11.4% during the first two hours. It is important to stress that this improvement is reached without increasing the total staff capacity.

Figure 4.6 Average total number of people at the different stations of the process at the end of the day, compared between a fixed and a changed allocation. The dashed line represents the average average total number of people for the fixed allocation, while the solid line represents the average for the changed allocation.



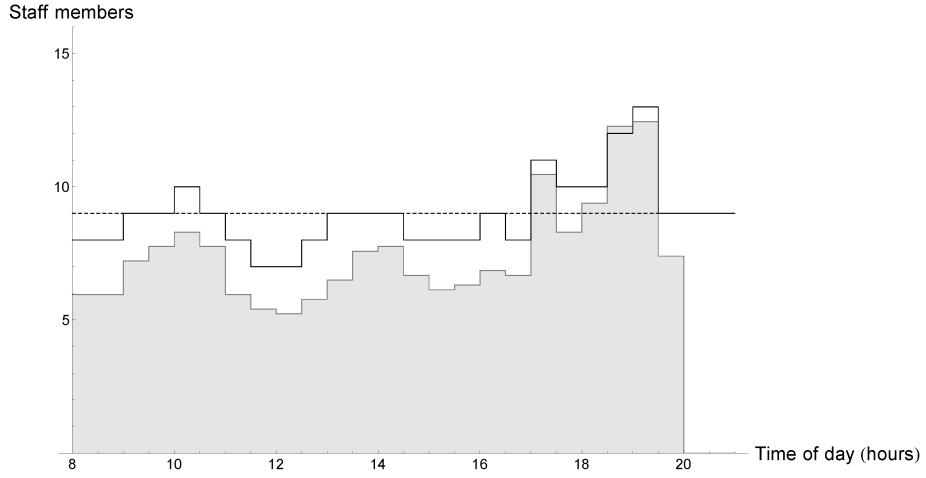
4.5.2.2 Scenario B: Changing allocation after closing

The same thing that happens after opening, also happens after closing, but this time in reverse. After some time, the first station is empty with a very high probability. At this time, it may be very useful to move this employee to a later station of the process. The same happens with employees at the second station of the process somewhat later. If all these employees are moved to the last station of the process, donors will wait less at the last station and leave the system more quickly. This leaves not just the donors better off, with less waiting time, but also the employees, who can close the collection site a little earlier at the end of the day. Although this sounds reasonable, it will be shown that it is much harder to decrease queue lengths in this manner than it is at the start of the day.

As at the start of the day, there are two time-points at which the allocation of the staff members changes. 15 minutes after closing, one employee changes from the second to the third station, resulting in one employee at the first station, two at the second station and six at the last station. 30 minutes after closing, two employees will move to the Donation station, one from the first station and one from the second station. Although it is likely that the queue at the first station is empty long before the 30 minutes after closing mark, it is possible that there is still a donor at the Registration station. If the employee is moved earlier, this donor would never receive service.

A very minor change can be seen in Figure 4.6, but this is negligible. The probabilistic nature at the end of the day that makes it impossible to quickly remove the employee from the first station, and is also the reason that the improvement

Figure 4.7 Number of staff members available using stationary shifts (dashed line) compared with the number of staff members available with optimal shifts (solid line)



is not nearly as high as for the opening of the site. At opening, it is certain that the station where the number of employees is decreased, is empty. This does not exist at the end of the day. To get bigger savings, a dynamic allocation that assigns employees based on the current state of the process, instead of the expected state, would be necessary.

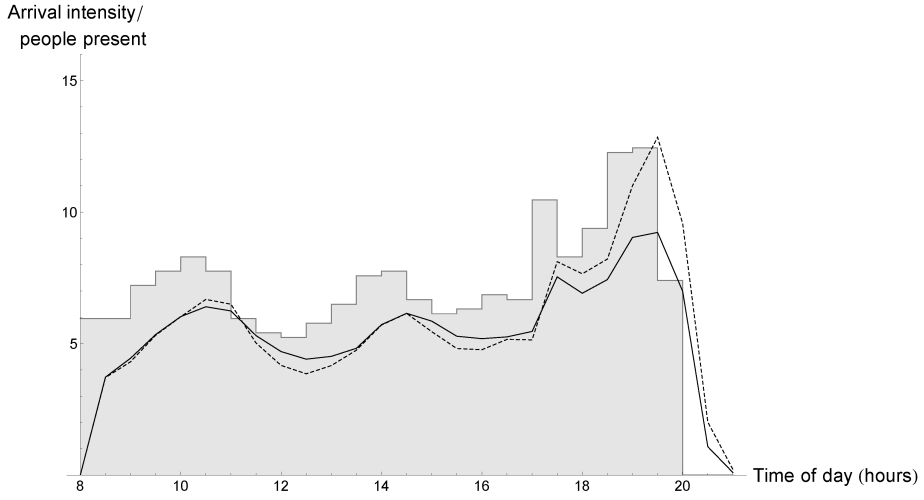
4.5.2.3 Scenario C: Number of staff members

In a Chapter 5, we have developed a methodology to determine the optimal shift schedule for staff members, with regards to arrival patterns. The method mainly focuses on the savings on the number of staff hours required. However, the algorithm can also be used to keep the same number of staff members, and schedule the shifts more efficiently. The goal of the rearrangement of the shifts would then be to reduce and balance queue lengths. Rearranging the shifts would, like the other improvement scenarios mentioned, increase the time-dependent variability of the blood collection process.

Using the algorithm, we found a set of shifts that on average uses nine employees. This will be compared to a stationary scheduling of nine employees throughout the day. The total number of staff members who are working at any point in time is visualized in Figure 4.7. Although peaks in the number of available staff members may only last half an hour, the minimum shift length was set to three hours. Currently, shifts are often shorter than an entire session, but an ending shifts is immediately replaced with another shift.

Figure 4.8 shows the results for the stationary scheduled shifts (dashed line) compared to the optimal shifts from the mentioned algorithm (solid line). For the stationary shifts, the average total number of people present varies greatly during

Figure 4.8 Average total number of people at the different stations of the process. Comparison between stationary shifts (dashed line) and an optimized shift schedule (solid line)



the day. This can be expected when the same number of employees is used to handle varying numbers of arriving donors. Clearly, the algorithm decreases the number of donors present during busy hours. It pays a small price at moments where fewer donors arrive, because fewer employees are available to help donors.

4.6 Discussion

The results in Section 4.5.1 clearly show that ignoring the time-dependent nature of a system results in large discrepancies with reality for queue length and workload computations. Not dealing with time-dependent aspects at all by simply aggregating all arrivals and services throughout the day, and use a steady-state method to compute the queue lengths based on the average rates over the day is clearly very imprecise. But even by computing queue length distributions or averages with some steady-state method for every differentiated time interval clearly leads to incorrect queue estimations.

Uniformization is often mentioned as one of the most accurate methods to compute transient queue length distributions (e.g. Ingolfsson et al [111]), both for stationary and for time-dependent queueing systems. But it is often disregarded because it takes too much computational time. We have shown that for at least some practical situations, such as a blood collection site, it is possible to compute queue length distributions in a matter of seconds. The blood collection site shows that this statement even holds if instead of just a single queue, a small network of queues is considered.

To be able to use uniformization, a few assumptions are required. The assumption of exponential service times is probably the most unrealistic assumption. It is possible to use Erlang or more general phase-type distributions for the service time, but this

comes at a significant computational cost. Another, even more technical and formal option would be to use a stochastic comparison approach. This has been shown for a network somewhat similar to the blood collection site, containing two stations, by van Dijk and Kortbeek [66]. For the queueing network based on the blood collection site, the details would be even more complex. However, most realistic service time distributions will have a coefficients of variation lower than the coefficient of one of the exponential distribution. Therefore, realistic queue lengths can be expected to be lower than the ones computed with exponential service time distributions.

In this chapter, we assumed one staff member can service one donor at the time. However, using uniformization, it is also possible to use some other service discipline, e.g. processor sharing. The total service rate can be any nonlinear function of the number of staff members and donors, as long as the total transition rate remains limited and the generator matrix can be defined. If we were to assume that processor sharing is applied if more donors are present at the Donation station than there are staff members, all results would even remain the same, as this would not influence the transition rates.

In literature, some papers focus on continuously changing parameters in combination with the uniformization method. These run into problems, because the discretized process, and thereby the transition matrix, is continuously changing. The method described in this chapter works with systems for which parameters are assumed to be more or less constant during short time intervals. Although this makes computations easier, it is also motivated by practice. In general, it is not possible to get data on continuous changes in a system, since this would require an infinite set of data points with infinite precision. Usually, one has to estimate the changes of the system by using time intervals, and estimating transition rates based on the historical number of transitions during these time intervals. The more data, given that the data is precise enough, the shorter time intervals can be. Instead of *estimating* a continuous function from these time intervals, we have decided to use the time intervals to our advantage by assuming the system parameters to be constant during a given time interval.

The model also limits the number of donors that can be present at each of the stations of the queueing system. Although this can be considered to be a restriction, most realistic queueing systems, especially service systems, will have a limit on the queue size as well. This is usually a physical limitation - i.e. there is a limited amount of space for the queue to form. It has the added effect that it prevents excessive waiting times, by not allowing too many people to get into the queue.

The results in Section 4.5.1 mainly show that time-dependent queueing methods are beneficial for the analysis of some queueing systems. Section 4.5.2 shows some scenarios that aim to improve service at Dutch blood collection sites. Scenarios A and B reallocate staff members either just after opening or just after closing the collection site. It is interesting to note that whilst the first of these shows clear improvements, the second shows no significant reductions of the total number of people present. Scenario C is based on work presented in Chapter 5. This work focuses on rearranging shifts of staff members based on the arrival pattern of donors. As ex-

pected, this scenario reduces excessive queues and balances the queues throughout the day. Although the number of people present is slightly more balanced, not much is changed during the first hours of the day. However, at the end of the day, the busiest period at blood collection sites, rearranging the shifts shows a clear improvement of service by a significant reduction of the total number of donors present. In all, applying time-dependent queueing methods produced better, more realistic results than the steady-state methods.

The lack of improvement at scenario B - the reallocation of staff at the end of the day - points to an interesting topic for further research. The most likely cause of the lack of improvement is the absence of state dependent reallocation. At the start of the day, the state of the system is known. It is therefore easy to rearrange the staff accordingly. However, at the end of the day, the state of the system is uncertain, and only given by a probability distribution. To get improvements, such as for the total queue length and delays, a state dependent allocation will therefore be required. Off course, the remainder of the day, especially the times during which the arrivals vary most strongly, could also benefit from state-dependent dynamic staff allocation. An extension of the model to incorporate this option might be an interesting topic for further research.

4.7 Appendix: Computational algorithm

The algorithm used for the calculations in this chapter is based on the concept of uniformization. The first thing that we need for the uniformization algorithm is a generator matrix. The generator matrix Q needs to contain all transition rates from one state to another. For ease of notation, we will define the state to be the three dimensional vector

$$n = (n_1, n_2, n_3)$$

with n_q the number of donors at station q. Before formulating the generator matrix for the blood collection site, we first define the function Q^*:

$$Q^*(n, n') = \begin{cases} \lambda & n_1 < N_1,\ n_1' = n_1 + 1,\ n_2' = n_2,\ n_3' = n_3 \\ n_1 \cdot \mu_1 & n_1 < s_1,\ n_2 < N_2,\ n_1' = n_1 - 1,\ n_2' = n_2 + 1,\ n_3' = n_3 \\ s_1 \cdot \mu_1 & n_1 \geq s_1,\ n_2 < N_2,\ n_1' = n_1 - 1,\ n_2' = n_2 + 1,\ n_3' = n_3 \\ n_2 \cdot \mu_2 & n_2 < s_2,\ n_3 < N_3,\ n_1' = n_1,\ n_2' = n_2 - 1,\ n_3' = n_3 + 1 \\ s_2 \cdot \mu_2 & n_2 \geq s_2,\ n_3 < N_3,\ n_1' = n_1,\ n_2' = n_2 - 1,\ n_3' = n_3 + 1 \\ n_3 \cdot \mu_3 & n_3 < s_3,\ n_1' = n_1,\ n_2' = n_2,\ n_3' = n_3 - 1 \\ s_3 \cdot \mu_3 & n_3 \geq s_3,\ n_1' = n_1,\ n_2' = n_2,\ n_3' = n_3 - 1 \\ 0 & \text{else} \end{cases}$$

Here λ is the arrival rate of donors at station 1, and μ_q and s_q are the service rate and the number of staff members at station q respectively. Let N_q be the

maximum number of donors at station q. For the results based on the collection site in Section 4.5, N_q is set to twelve for $q = 1, 2, 3$. The generator matrix should be two-dimensional, instead of the six dimensions of the function Q^*. We will use the following formulas to relabel the states:

$$n_1(i) = \left\lfloor \frac{i-1}{(N_3+1)(N_2+1)} \right\rfloor$$

$$n_2(i) = mod^{(N_2+1)} \left\lfloor \frac{i-1}{N_3+1} \right\rfloor$$

$$n_3(i) = mod^{(N_3+1)}(i-1)$$

where mod^x is the modulo with divisor x. Using these formulas and Q^*, we define the generator matrix Q:

$$Q_{i,j} = \begin{cases} Q^*(n_1(i), n_2(i), n_3(i), n_1(j), n_2(j), n_3(j)) & i \neq j \\ -1 \cdot \sum_{l=1}^{I} Q^*(n_1(i), n_2(i), n_3(i), n_1(l), n_2(l), n_3(l)) & i = j \end{cases}$$

Here I is the total number of states, equal to $(N_1+1)(N_2+1)(N_3+1)$. Let α be the maximum of the absolute values of the diagonal elements:

$$\alpha = \max_i(-1 * Q_{i,i})$$

Now, using α and the generator matrix Q, we can get a transition matrix P for 1 transition of the process:

$$P = \frac{1}{\alpha}Q + I$$

The next step in the algorithm is to compute the state probability vector $\pi(t)$ after some time t, given the start vector $\pi(0)$. Because all the transition rates are assumed to be Poisson, and the sum of Poisson rates is also a Poisson rate, the number of transitions k in a time interval of length t is Poisson distributed. This can be multiplied by the probability vector after k transitions. This in turn can be summed over all possible values for k. This gives to following expression for $\pi(t)$:

$$\pi(t) = \sum_{k=0}^{\infty} \pi(0) P^k \frac{(\alpha t)^k}{k!} e^{-\alpha t}$$

As long as the transition matrix P does not change, this will work for any t. However, since this chapter studies a time-dependent queueing system, the transition matrix P does change over time. Despite the existence of a theoretical exact

Poissonian expression, as in van Dijk [64], we will assume that the transition matrix is not continuously changing for computational purposes. Instead, we will assume that it is piecewise constant during short time intervals $[t_l, t_{l+1})$. Therefore we will add a subscript (t_l, t_{l+1}) to the transition matrix, indicating the stationary transition matrix between times t_l and t_{l+1}. With the distribution at time t_l already computed, and using this transition matrix $P_{[t_l,t_{l+1}]}$, the probability vector at time t_{l+1}, $\pi(t_{l+1})$, can be computed from $\pi(l_l)$ by:

$$\pi(t_{l+1}) = \sum_{k=0}^{\infty} \pi(t_l) P^k_{[t_l,t_{l+1}]} \frac{(\alpha(t_{l+1} - t_l))^k}{k!} e^{-\alpha(t_{l+1}-t_l)}$$

As explained in Section 2.4.2 in Chapter 2, there is a practical limitation: it is impossible to compute an infinite sum numerically. So, for the numerical results in Section 4.5, the sum will be computed for k up to some K. This results in the following practically usable function:

$$\pi(t_{l+1}) = \sum_{k=0}^{K} \pi(t_l) P^k_{[t_l,t_{l+1}]} \frac{(\alpha(t_{l+1} - t_l))^k}{k!} e^{-\alpha(t_{l+1}-t_l)}$$

A suitable K can be found by using a sufficiently small tail of the Poisson distribution. We have used to following truncation:

$$1 - \sum_{k=0}^{K} \frac{(\alpha(t_{l+1} - t_l))^k}{k!} e^{-\alpha(t_{l+1}-t_l)} < 10^{-10}$$

All results in Section 4.5 have been generated using the implementation of these formulas in Mathematica 9.0 and Matlab R2015b.

Part III

Optimization

Chapter 5

S.P.J. van Brummelen, N.M. van Dijk, K. van den Hurk and W.L. de Kort. Waiting time-based staff capacity and shift planning at blood collection sites. *Health Systems*, in press, 2017.

Chapter 6

S.P.J. van Brummelen, K. van den Hurk, W.L. de Kort and N.M. van Dijk. Dynamic staff allocation at blood collection sites. *Submitted*.

Chapter 7

S.P.J. van Brummelen, W.L. de Kort and N.M. van Dijk. Combining appointments and walk-ins at Dutch blood collection sites. *In preparation*.

Chapter 8

J.H.J. van Sambeeck, S.P.J. van Brummelen, and N.M. van Dijk. Blood type specific issuing policies to improve inventory management of red blood cells. *In preparation*.

CHAPTER 5

Waiting time based staff capacity and shift planning

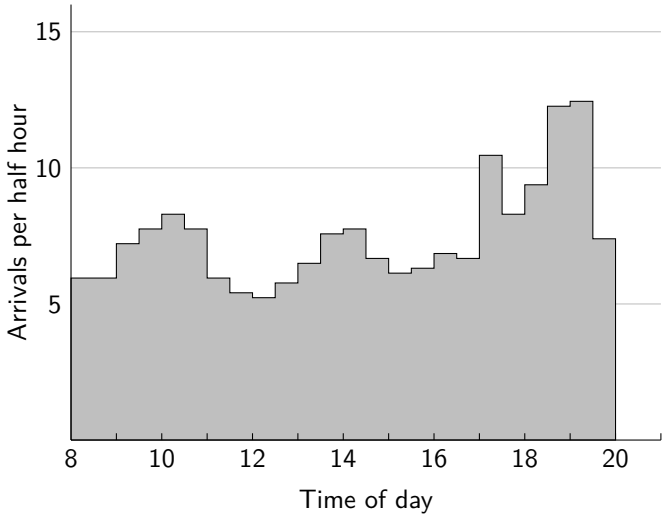
5.1 Introduction

Sanquin currently focuses her collection sites and intake sessions on production. For every session and every hour worked by a staff member the required number of donations has been set in advance, and staff members are scheduled based on these requirements. Waiting times are not consistently taken into account in these scheduling methods. Some managers of collection sites schedule an extra staff member during peak hours to counteract long waiting or sojourn times, but most staff members are scheduled for an entire day or intake session.

Even though whole blood donors can walk in without an appointment, the arrival process is not as random as one might think. Clear patterns show up in the arrival times of donors, and these are mostly independent of day and location (see Van den Toren et al. [175]). Peaks in arrival intensity clearly show up early in the morning, around lunch time, and around dinner time. Additionally, plasma donors make appointments for their donation, which makes the arrival times of these donors more predictable. Much can be gained, both in leveling work pressure and in decreasing waiting times, by adjusting the number of staff members based on the expected arrival pattern of donors. With many part-time employees, as is the case at Sanquin, it might well be possible to combine short and longer shifts to improve the effectiveness of the staff scheduling.

In this chapter we will develop a method using queueing theory and an ILP formulation for this purpose, to take advantage of the patterns in arrival intensities. The proposed method determines starting times and durations of all shifts such that the total number of worked hours is minimized, with certain restrictions on shift lengths. At the same time, the method takes a waiting time restriction into account. A number of ways of implementing these waiting time restrictions are possible. We will propose two methods; the first one is based on sojourn time percentiles (e.g. 95 % of donors should spend less than 60 minutes in the collection site) and the second one is based on an average waiting time, calculated by a slightly more complex, yet also more realistic queueing model. Finally, we use numerical results to show that this method can be implemented without increasing the total number of working hours.

Figure 5.1 A typical arrival pattern for a collection site that is opened the whole day.



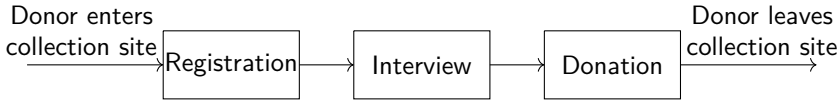
5.1.1 Process description

A comprehensive description of the process at a blood collection site can be found in Section 1.3. In this section we only highlight the aspects of the blood donation process that specifically apply to this Chapter.

From a process point of view, there are two differences between whole blood and plasma donations. The first difference regards the arrivals. Before arriving at a collection site, plasma donors make an appointment. Sanquin aims, and mostly succeeds, to spread out plasma donations over the day. For whole blood donations, there is no appointment system. To be able to control the number of arrivals of whole blood donors, Sanquin sends out invitations to a selection of whole blood donors by post card once a week. Although donors are encouraged to wait for an invitation and to come at their earliest convenience after receiving the invitation, neither is required. Donors may walk in and donate whenever they like, provided they are eligible to donate at that particular time. Arriving whole blood donors also show clear preferences for certain times of the day, as can be seen in figure 5.1. Some days of the week are more popular than others, but the time preferences do not depend on the day of the week. The peaks in arrivals do depend on the opening hours of the collection site. The second major difference between plasma and whole blood donations is the donation itself. All three elements of the donation - starting the donation, collection, and ending the donation - require much more time and equipment for plasma donations, compared to whole blood donations.

Two types of staff members can be distinguished at collection sites: general staff members and physicians. All standard tasks at the collection site are handled by general staff members. The physician has to be present in case of a complication during the donation. The physicians also handle the first interview with a new donor. For these tasks, a collection site always has exactly one physician present. Therefore,

Figure 5.2 Schematic model of a collection site. Every station is modeled as a M/M/s queue.



the described method only focuses on scheduling general staff members.

For this chapter, as in Chapters 3 and 4, we model the collection site by a tandem queue with three stations, as visualized in Figure 5.2: the Registration station, the Interview station and the Donation station. The stations are assumed to have an average service time of 2 minutes, 6 minutes and 12 minutes respectively. So, the combined three stations have a mean service time of 20 minutes for whole blood donors. This would in theory imply that a collection site can handle three donors per staff member per hour. To ensure that waiting time remains acceptable, Sanquin has committed herself to the target that, at every collection site, 85 % of all whole blood donors spend less than 45 minutes in the blood donation process. However, this service level has only implicitly been taken into account when scheduling staff members. Staff is scheduled on the basis that every staff member should help at least two donors per hour. As every staff member could help three donors per hour if they were working at full capacity, Sanquin reckons that waiting times and breaks have been taken into account by using the lower capacity. This, however, has never been formalized. The model in this chapter combines waiting time computation with staff scheduling.

5.1.2 Theoretical background and modeling

From a theoretical perspective, this chapter will combine two different disciplines from the field of Operations Research: Mathematical programming to optimally schedule the staff shifts, and queueing theory to include waiting time targets when scheduling these shifts. This will result in a two-step approach, in line with the terminology used in the extensive review on staff scheduling for service systems by Defraeye and Van Nieuwenhuyse [57].

When faced with an arrival pattern, such as in Figure 5.1, a number of options can be thought of to determine the shifts for staff. The first option is to simply ignore the existence of a pattern, and to ensure that enough staff is available at the peak of the arrival intensity, and scheduling this number of staff members the entire day. This way, excess capacity is available during the remainder of the day. This is common practice at Sanquin. The second option is to break up the day in a few shifts. This way, extra staff can be scheduled only for the peak arrival intensity during a shift, thereby reducing excess capacity. If the intervals are made shorter, the over-capacity is reduced even further.

A combination of overlapping shifts and varying starting times could reduce excess capacity, while preserving viable shift lengths. For example, if we have 3 intervals, and the staff requirements are 1,2,1 respectively, we would be able to cover this with two shifts, both spanning two intervals, one starting the first interval and another

starting the second interval. As this eliminates excess capacity, we can guarantee that this is the optimal solution, where the shortest shift length remains two intervals. However, for large instances, such as the one at Sanquin, it is extremely hard to come up with a solution by hand. And, since there is no way to avoid excess capacity completely, there is no way of knowing how good the solution is. Using mathematical programming, this problem can be formulated as an Integer Linear Program (ILP). Using commercially available solvers, ILP models can usually be solved to optimality.

Before the ILP can be used, the required number of staff members first has to be determined. This can be done in several ways. The most simple is the one currently in use at Sanquin. This computation is based on the presumed number of donors that a staff member should help in an hour. In recent years, Sanquin has set this number to 2.0. This means that for every staff member present, a collection site should collect 2.0 donations per hour. However, even if the staff shifts perfectly match the number of required staff members, waiting times will inevitably occur due to random variations in arrivals and service times. In queueing theory, it is well known that working at full capacity will result in extremely long waiting times. By using queueing theory, the minimum number of staff members can be determined taking waiting times into account, to meet waiting time targets. This will generally increase the required number of working hours, as it will prevent the system from working at full capacity.

Summarizing, there are two competing effects on the total number of working hours. On one hand, the inclusion of flexible staffing could result in a decrease of the number of working hours. The inclusion of waiting times, on the other hand, may require an increase of the number of working hours. This raises the following question: What will happen when both flexible staffing and queueing theory are combined at blood collection sites? The proposed two-step approach in this chapter will be employed to answer this question.

The chapter will be structured as follows. We will start with a literature discussion in section 5.2. A more detailed and technical discussion of the mentioned methods will then be given in section 5.3. Finally, we will provide numerical results for a general approach, in which data from multiple collection sites is combined to give an impression of the average potential of the described method. The chapter will be concluded with a discussion.

5.2 Literature

Although no other literature on staff scheduling at blood collection sites currently exists, the literature on staff scheduling in general is very extensive, as can be seen in the review by Ernst et al. [80]. Most of the papers in the staff scheduling literature cover the same two basic steps used in this chapter: first determine staff requirements within given time intervals, and subsequently determine the optimal shifts to cover these requirements.

The review by Defraeye and Van Nieuwenhuyse [57] focuses on staff scheduling for non-stationary systems. From a technical point of view, that is exactly what we are

trying to achieve for blood collection sites. This review describes a total of 62 papers. Most of these papers use a single queue to calculate their performance indicators. Of the 62 papers, only six use a network of queues. All of these use simulation, while three papers also use an analytical approximation [85, 113, 191]. From a technical point of view, the paper by Izady and Worthington [113] is most closely related to this chapter, as they use similar performance indicators: sojourn time percentiles and average waiting times. The other three papers that use a network of queues [4, 45, 168] only use simulation for performance evaluation. Two of these papers, the papers by Ahmed and Alkhamis [4] and by Centeno et al. [45], use a sojourn time percentile for performance evaluation, similar to this chapter.

None of the papers discussed in the review by Defraeye and Van Nieuwenhuyse [57] cover blood collection sites or blood banks in general, but eight papers discuss a health care setting. Of those papers, seven are applied to an emergency department [4, 45, 56, 98, 113, 168, 191] and one is applied to ambulance services [78]. The paper by Defraeye and Van Nieuwenhuyse uses a similar performance metric as this chapter - the expected waiting time. However, it does not use a network of queues. Although they evaluate the average waiting time, it is not included in the performance goals. The only paper that does use a network of queues, but which is not applied to health care, is the paper by Fukunaga et al. [85]. This paper is based on a call center. Although call centers seem by far the most frequent application of time dependent staff scheduling methods in literature, these systems mostly use a single queue for performance evaluation.

As can be seen in the mentioned papers, most recent papers surrounding staff scheduling have used simulation as a tool to estimate waiting times. This has the advantage that it can handle very large and complex systems, but because our system is limited in size, an analytic model is a faster and more consistent way to calculate waiting times.

Some papers have been written on other logistical issues at blood collection sites. For a detailed discussion of these papers, please see Section 1.4. We will, however, highlight one of the papers here. From a practical point of view, the paper by Blake and Shimla [29] is closely related to the research contained in this chapter. In the paper, a blood collection site is modeled as a flow shop, and then the results for the required number of staff members for each station are adjusted for uncertainty by describing every station as an M / M / s queueing model. The minimum number of required staff members is then computed by setting a waiting time restriction for each of the stations. This is close to how we will model the blood collection site.

The main contribution of the research presented in this chapter is the combination of exact methods from two fields of research in Operations Research - queueing theory and Integer Linear Programming (ILP) - to incorporate waiting time estimation in the determination and planning of staff capacity at blood collection sites. First, we expand the waiting time estimation of Blake and Shimla [29] to be able to include waiting and sojourn time restrictions on the total blood collection process. Different queueing computations will be used for this purpose. Second, to actually minimize the number of staff working hours, we will use an ILP model to schedule shifts based

on the required number of staff members. These required numbers can either be based on a production standard (as currently used in practice) or on these waiting time restrictions. The ILP is able to incorporate fluctuating arrivals to the blood collection site by allowing shorter and more flexible shifts.

We note that the methods to compute the waiting time and to determine the optimal shifts are not new methods from a mathematical perspective (e.g. see the review by Defraeye and Van Nieuwenhuyse [57]). However, the combination of the methods at blood collection sites, or even health care systems in general, has not been reported on before. Additionally, in Chapter 3, we have shown that modeling the blood collection site as a tandem queue gives a good approximation of the waiting times. The combination of this queueing model and a small ILP model, results in a fast computation of good shift options for practical purposes.

5.3 Queueing methods

In this section three methods to calculate the minimum number of required staff members will be discussed. These methods will be used for the first step in the two-step procedure. In contrast to the first method - a production standard - from section 5.3.1, the second two methods will take waiting time into account by using queueing theory. Like the production standard, these methods will determine the minimum number of staff members required for every half hour during the opening hours of the collection site.

For both queueing theory methods discussed in this section, we will assume exponential distributions for both inter-arrival times and service times.For arrivals, this seems like a natural assumption, as it implies that arrivals are independent, a likely situation because there are no appointments. For services, we have also assumed exponential times. There are two justifications for this. The first is a lack of reliable data on service times. The second, more important, reason is that exponential service times reasonably predict waiting times in Dutch blood collection sites, as shown in Chapter 3. Nevertheless, for theoretical purpose, a method without the exponential requirement will also be included in Section 5.3.3.

Before discussing the specific methods, it is important to note the difference between:

- A production standard η (used for the production standard method, section 5.3.1). The production standard entails the number of donors that should at least be served by a staff member every hour.
- A service capacity μ (used for the methods M/M/s and network model, sections 5.3.2 and 5.3.3 respectively). A service capacity entails the number of donors that could be helped by a staff member every hour, assuming continuous production.

5.3.1 Current situation: Production standard

Currently Sanquin uses a production standard when scheduling their staff. This means that for every staff member, a fixed number of donations η should be completed every hour. Currently η is set at 2.0. From a utilization standpoint it can be argued that this production standard can be increased, as the average time a staff member is needed during the donation process is less than 30 minutes. However, we can conclude from basic queueing theory that increasing the production standard would undoubtedly lead to longer waiting times. This argument is also used against a production standard of 2.5 or even 3.0; numbers that imply an average service time closer to the actual average service time of 20 minutes. Although this argument is valid, the exact implications of increasing the production standard are unknown, as this method does not include waiting time estimation.

To model time dependent arrivals, an arrival pattern has been included, an example of which is shown in Figure 5.1. The arrival pattern is based on van Mechelen and Zonneveld [133]. This arrival pattern specifies the arrivals expected in each half hour interval. This implies that the minimum number of staff members will also be calculated for every half hour during the opening hours of the system. A uniform and user specified arrival pattern are also included in the tool for collection sites.

With λ_h the arrival rate in half hour h, the minimal required number of staff member to be present B_h can be calculated by

B_h using a Production Standard

$$B_h^{(1)} = \left\lceil \frac{\lambda_h}{\eta} \right\rceil \tag{5.1}$$

This will be used as a basedline input for the ILP model, to allow for the optimization of staff shifts, as outlined in Section 5.4.

5.3.2 M/M/s model

As a first simple option, we could model the collection site as a standard M/M/s multi server system, with a service time equal to the sum of the service times of the individual stations of the process. This can be justified if it is assumed that a staff member follows the donor throughout the system. Although this is not applied at Dutch blood collection sites, it is used in blood collection. This practice is commonly referred to as "go with the flow."

Exact formulas are known to calculate the average waiting time, the average delay and even the waiting time and the delay distribution (e.g. Winston [183, Chapter 20]). As previously mentioned, the official Sanquin policy is that 85 % of the whole blood donors should spend less than 45 minutes in the collection site. This means that the 85th percentile of the delay distribution should be lower than 45 minutes. By using equation 5.2, we can check this, and possibly other, service goals for a given staff level.

The main drawback of this model is that it is not possible to take the system's multiple stations into account. The model simply takes an average occupancy for the entire system. This is a problem because the relation between occupancy and queue lengths is not linear. In reality the process steps are interrupted and not all stations have the same service time or number of staff members. Therefore, the occupancy will not be the same for each of the stations. If the variations in occupancy between stations are large, the M/M/s model might give approximations for the waiting time that are too optimistic.

The most appealing feature of the M/M/s computation is the option to base decisions on a percentile in the waiting time distribution. When a percentage α of the donors has to have a delay lower than t hours, the minimal required number of staff members B_h can be calculated using equation 5.2.

B_h by M/M/s computation:

$$\begin{aligned} B_h^{(2)} = \quad & \text{minimize} \quad s \\ & \text{subject to} \\ & e^{-\mu t}\left(1+\mathbb{P}\left(j \geq s\right)\frac{1-e^{-\mu t(s-1-s\rho)}}{s-1-s\rho}\right) < 1-\alpha \end{aligned} \tag{5.2}$$

Here $\rho = \lambda/(s*\mu)$ with λ and μ the arrival rate and service rate respectively. λ and μ should use the same time unit as t. $\mathbb{P}\left(j \geq s\right)$ represents the probability that there are as many or more donors than there are staff members available. This probability can be calculated using standard M/M/s formulas.

5.3.3 Network model

The second, more complicated, but also more realistic modeling option, could be to use some form of a queueing network. These kind of models incorporate the fact that the system has multiple stations and multiple servers working at each station. This allows us to use the full model, as depicted in Figure 5.2. Although it is still possible to calculate sojourn time distributions, as shown in Chapter 3, this is a very time-consuming process. Therefore, for network models this chapter will only deal with the average waiting time.

The Queuing Network Analyzer (QNA) [182] will be used to calculate average waiting times in a queueing network. QNA is based on a set of approximative expressions using the coefficients of variation of the external arrivals and coefficients of variation of preceding stations. Due to the serial nature of the system at the Dutch blood bank, the original expressions can be slightly simplified. The expression below describes how the coefficients of variations of departures depend on the coefficients of variation of the arrivals and services, and the parameters of the station in question. Since there is no splitting and superposition of donor flows in a collection site, the coefficients of variation of the departures are the same as the coefficients of variation

Table 5.1 Parameters and variables used

Parameters	
η	Production standard
τ	Total expected service time
τ_q	Expected service time at station i
μ	Service rate ($= 1/\tau$)
μ_q	Service rate at station i ($= 1/\tau_q$)
λ	Arrival rate
s_q	Number of servers at station i
ρ_q	Occupancy, $= \lambda\tau_q/s_q$
C_{sq}^2	Squared coefficient of variation of services at station q
C_{dq}^2	Squared coefficient of variation of departures at station q
C_{aq}^2	Squared coefficient of variation of arrivals at station q
Variables	
W	Total waiting time
W_q	Waiting time at station q
T	Total sojourn time (delay)
T_q	Delay at station i

of the arrivals at the next station. The description of the parameters and variables used can be found in Table 5.1.

$$C_{a(q+1)}^2 = C_{dq}^2 = 1 + (1-\rho_q^2)(C_{aq}^2 - 1) + \frac{\rho_q^2}{\sqrt{s_q}}(C_{sq}^2 - 1).$$

Further, let $\mathbb{E}_{M/M/s_q}(W_q)$ denote the expected waiting time for an $M/M/s_q$ queue. This can be computed standardly by:

$$\mathbb{E}_{M/M/s_q}(W_q) = \frac{(s_q\rho_q)^{s_q}}{s_q!\left(\sum_{n=0}^{s_q-1}\frac{(s_q\rho_q)^n}{n!} + \frac{(s_q\rho_q)^{s_q}}{(1-\rho_q)s_q!}\right)(1-\rho_q)^2 s_q}.$$

Then $\mathbb{E}_{s_q}(W_q)$ for the non-exponential case can be calculated by:

$$\mathbb{E}_{s_q}(W_q) = \frac{C_{aq}^2 + C_{sq}^2}{2}\mathbb{E}_{M/M/s_q}(W_q).$$

Note that for the exponential case, i.e. if all coefficients of variation are equal to 1, the computation by QNA is equal to the exact expressions that are available for the exponential case.

Because a donor can only visit a station once, the expected delay $\mathbb{E}(T_q)$ at station q can be calculated by:

$$\mathbb{E}_{s_q}(T_q) = \mathbb{E}_{s_q}(W_q) + \tau_q.$$

For an average total delay less than t minutes, B_h can be computed by solving equation 5.3.

B_h for QNA method:

$$\begin{aligned} B_h^{(3)} = \quad & \text{minimize} \quad \sum_{q=1}^{3} s_q \\ & \text{subject to} \quad \sum_{q=1}^{3} \mathbb{E}_{s_q}(T_q) < t \end{aligned} \tag{5.3}$$

This is an integer, non-linear optimization problem, so in general it is very hard to solve. But, since there are only a finite number of configurations of the staff in each half hour interval - for a typical blood donor center this could be 1 or 2 staff members at the registration station, between 2 and 4 staff members at the interview station and 3 to 6 staff members at the donation station, we could solve this by applying brute force, i.e. checking every possible combination of staff members between some lower bound and upper bound. It is possible to do this for every interval that has to be scheduled, and then these numbers can be used as input for the ILP model of section 5.4.

QNA also allows the use of coefficients of variation of the inter arrival times and service times, meaning that these are not required to be exponential. Although this is very useful in most systems, for the numerical results in this chapter, these coefficients are set to 1 for the blood collection site, resulting in exponential service times, as discussed previously.

5.4 ILP model

The second step is the use of an Integer Linear Program (ILP) to schedule shifts. Once the minimum number of staff members $B_h^{(i)}$ has been determined by either eq. (5.1), (5.2) or (5.3), an ILP model can be formulated to determine optimal shifts lengths and starting times. This ILP model is given in Box 1. The parameters and variables are explained in Table 5.2.

Each of the restrictions in the ILP model in Box 1 has their own interpretation, as specified below:

Interpretation of restrictions

1. This restriction ensures that there are at least as many staff members present as the restrictions calculated by any of the three options discussed in sections 5.3.1, 5.3.2 and 5.3.3.

Table 5.2 Parameters and variables for the ILP model

Indices	
h, h'	Half hours
t	Shift length
Parameters	
k_t	Cost of a staff member for shift duration t
$B_h^{(i)}$	Required number of staff members present at half hour h (calculated by method i)
mn_h	Minimum number of staff members at half hour h
$q_{t,h,h'}$	1 if $h \leq h' < t + h$ and a shift of length t, starting at half hour h is allowed, 0 otherwise
Variables	
$x_{t,h}$	Starting shifts at half hour h of length t of
y_h	Staff members present at half hour h
$z_{t,h,h'}$	Number of breaks at half hour h' of a staff member that has (a shift length t and started at half hour h)

Box 1: The ILP model

Minimize

$$\sum_t \sum_h x_{t,h} \cdot k_t$$

Subject to:

(1) $$y_h - \sum_{t=12}^{18} \sum_{h'} z_{t,h',h} \geq B_h^{(i)} \qquad \forall(h),\ B_h^{(i)} \text{ from equation } (i)$$

(2) $$y_h - \sum_{t=12}^{18} \sum_{h'} z_{t,h',h} \geq mn_h \qquad \forall(h)$$

(3) $$\sum_t \sum_h x_{t,h} \cdot q_{t,h,h'} = y_{h'} \qquad \forall(h')$$

(4) $$x_{t,h} \leq \sum_{h'=h+1}^{h+t-1} z_{t,h,h'} \qquad \forall(h), t \geq 12$$

(5) $$x_{t,h} \in \mathbb{N} \qquad \forall(t,h)$$

(6) $$z_{t,h,h'} \in \mathbb{N} \qquad \forall(t,h,h')$$

2. This restriction ensures that there are at least as many staff members present as the minimum number required. This is not a value that has been calculated,

Table 5.3 The costs associated with the various shift lengths

Shift duration	Costs
2	2
3	3
4	3.99
5	4.99
6	5.98
7	6.98
8	7.97
9	8.97

but some value that has been set as an absolute minimum by the user of the algorithm. This is to ensure that some minimum number of staff members is always present. E.g. the M/M/s model is based on a single station, which could require only one staff member. However, all stations have to be manned at all times, requiring at least three staff members.

3. This restriction converts $x_{t,h}$, the starting shifts for staff members, to y_h, the number of staff members present. It also makes sure that shift lengths that are not allowed, do not convert to staff members that are working.

4. This restriction ensures that there is enough slack in the schedule to give everyone who is entitled to a break can get a break.

5. This restriction ensures that the solution is integer, i.e. no fractions of staff members.

6. This restriction ensures that the solution is integer, i.e. no fractions of breaks.

The costs k_t can be seen in Table 5.3. The costs are set such that the model will always select one longer shift rather than a combination of two sequential shorter shifts, by making a longer shift slightly cheaper than the combined cost of two shorter shifts. The difference is small enough that longer shifts will not be selected if a combination of two shorter shifts results in fewer working hours.

Given the calculated minimum staff levels and the ILP model, we will use commercially available packages to compute the optimal solution. We have used modeling tool AIMMS 4.5.2 to build the ILP model and its restrictions and have used CPLEX 12.6.1 to solve the ILP. Even for the largest Sanquin cases - collection sessions of 12 hours, the CPLEX reaches the optimal solution within a second.

5.5 Results

5.5.1 Current situation (base scenario)

The exact method that Sanquin uses to schedule staff has not been formalized. Based on discussions with employees and team leaders, we may conclude that the method

Table 5.4 Session types at Sanquin and their opening hours

Session name	Opening hours
M1	8.00 - 11.00
M2	8.00 - 12.00
MA	8.00 - 15.30
AE	12.30 - 20.00
E1	16.00 - 20.00
E2	17.00 - 20.00
MAE	8.00 - 20.00

Table 5.5 Average changes in staff hours based on all session types and multiple collection site sizes, compared to the current situation (*). The methods in the first column are based on sections 5.3.1, 5.3.2 and 5.3.3 respectively. In case a result shows NP, it is not possible to meet the waiting time restriction with this service capacity, irrespective of the capacity used. [a] Note that in the case of a production standard method the production standard is equal to the service capacity

	Possible shifts	Waiting time restriction	Service capacity [a] 2.0	2.5	3.0
Production Standard	Session shifts	N/A	*	-18.6%	-32.0%
	Flexible shifts	N/A	-26.2%	-40.1%	-49.5%
M/M/s	Session shifts	$\mathbb{P}(T > 45 \text{ min }) < 0.12$	NP	NP	-10.2 %
		$\mathbb{P}(T > 45 \text{ min }) < 0.15$	NP	NP	-17.8 %
		$\mathbb{P}(T > 60 \text{ min }) < 0.15$	27.4%	-7.9%	-23.9%
	Flexible shifts	$\mathbb{P}(T > 45 \text{ min }) < 0.12$	NP	NP	-29.5 %
		$\mathbb{P}(T > 45 \text{ min }) < 0.15$	NP	NP	-36.3 %
		$\mathbb{P}(T > 60 \text{ min }) < 0.15$	-1.8%	-29.5%	-42.0%
Network Model	Session shifts	$\mathbb{E}(W) < 5$ min	43.3%	17.9%	3.2%
		$\mathbb{E}(W) < 10$ min	31.6%	8.4%	-6.6%
		$\mathbb{E}(W) < 15$ min	26.6%	3.9%	-11.5%
	Flexible shifts	$\mathbb{E}(W) < 5$ min	12.6%	-6.1%	-18.5%
		$\mathbb{E}(W) < 10$ min	2.8%	-15.4%	-27.4%
		$\mathbb{E}(W) < 15$ min	-2.0%	-19.8%	-31.6%

that is closest to reality - which will therefore be used as a base scenario in this section - is the production standard method that has been presented in section 5.3.1. A production standard of 2.0 is used to determine the minimum required number of staff members.

Staff members are scheduled for an entire session, except for long sessions, which are split into two shifts, but these two shifts usually have the same number of assigned staff members. This means that Sanquin will usually staff the number of employees that are required during peak hours for the entire day. Employees will get a shift length equal to either the total or half of the session length plus some additional time before opening and after closing the collection site. This extra time is required

to set up and shut down equipment respectively [1]. In table 5.5 this method of shift planning will be called "session shifts". As it is closest to the current situation, it will be referred to and used as the base scenario, indicated with * in table 5.5.

5.5.2 Alternative scenarios

Table 5.5 shows the three different methods to calculate the minimum number of required staff members, B_h, that were presented in this chapter: production standard, M/M/s and network modeling. The last two are accompanied by a waiting time restriction. For M/M/s this is the probability that the delay time, i.e. the total time spent in the system, will exceed a certain threshold. For the network model this is a restriction on the total mean waiting time. These restrictions should hold for every half hour, meaning that a busy period with long waiting times can not be compensated for by a quiet period with very short waiting times. Note that the individual restrictions of the M/M/s and network models are not linked. E.g. we do not claim that an expected waiting time below 5 minutes implies that less than 12 % of donors spend longer than 45 minutes at the collection site.

Table 5.5 also includes a distinction between scenarios that only allow session shifts, as explained in section 5.5.1 and scenarios that allow "flexible shifts". Flexible shifts, in this case, allows for shifts that start at any half hour during the day (e.g. 9.00, 9.30, 10.00 etc.) and last a whole number of hours between 3 and 9 hours. Finally, Table 5.5 includes results for a production standard/service capacity of 2.0, 2.5 and 3.0. It is important again to note the difference between the production standard (used for the production standard method) and the service capacity (used for the M/M/s model and network model). This means that a service capacity of 3.0 seems reasonable, as the total process has a service time of approximately 20 minutes, but a production standard of 3.0 results in extremely long waiting times.

To get an impression of the results that can be achieved by the proposed combination of queueing and ILP, 35 instances will be used for every scenario. Table 5.5 shows the average result of all these instances for every scenario. The instances are a combination of 5 arrival rates for all of the 7 session types that Sanquin distinguishes. These 7 session types are shown in Table 5.4. The average donor arrival rates per hour that were used range from 12 to 20, with increments of 2.

Some scenarios for the M/M/s computation are shown to be not possible (NP). In these cases the tail of the service time distribution exceeds the required probabilities. This means that even without any waiting time, the delay time restriction still cannot be met due to the assumed stochasticity of the exponential service times. This cannot happen with the network model, as it places a restriction on the waiting time. The waiting time can be arbitrarily close to 0 if enough staff is added.

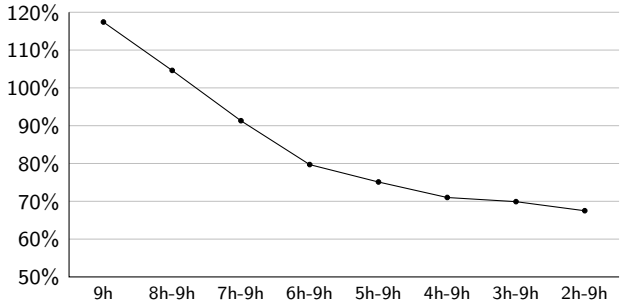
As a first observation, it can be seen that a waiting time restriction increases the required staff hours. By just introducing a waiting time restriction, while still assuming a service capacity of 2.0, staff hours increase by up to 43.3% if waiting

[1] as this extra time is required, it is included in all scenarios for employees that work the first or last shift.

times are only allowed to be 5 minutes. However, it is safe to assume a higher service capacity for these queueing methods. Even a very safe service capacity increase to 2.5 decreases the extra staff hours required to at most 17.9%, and is able to completely negate the increase for the M/M/s method for the $\mathbb{P}(T > 60 \text{ min }) < 0.15$ case. However, the lower waiting time restrictions are still impossible. When increasing the service capacity to a realistic 3.0, all but one scenario show a decrease in the number of staff hours.

The second main observation is, as expected, that flexible staffing results in significant savings on staff hours. By just introducing flexible staffing, i.e. comparing session shifts and flexible scenarios with the same further settings, savings are around 20 %, ranging from 20.4 % for the network model with a waiting time restriction of 5 minutes and a service capacity of 2.5 to 26.5 % for the production standard method with a production standard of 2.0.

Figure 5.3 Effects of adding extra shift length possibilities on number of working hours for the MAE session (see Table 5.4. Required number of staff members based on the network model. Results are in number of working hours as a percentage of session shifts and are based on an average of the 9 different scenarios for the network model included in Table 5.5.



The benefits of flexible shifts are again shown in Figure 5.3. This shows the effect of additional shift length options. It is based on an average of the 9 scenarios for the network model from Table 5.5, and results are expressed as a percentage of the session shift option. The first data point is the number of hours that are needed to staff the collection site if only 9 hour shifts are allowed, the second data point adds shifts of 8 hours, etc. Only the data from MAE sessions (see table 5.4) was taken into account, because the other sessions are not opened for 9 hours, making the 9 hour shifts redundant in these sessions. If only 9 hour shifts are allowed, flexible shifts are worse than session shifts. This has to do with the fact that two 9 hour shifts cover more that the total session, while one is not enough. A combination of 8 and 9 hour shifts still shows the same effect, but it is significantly reduced. Also note that the marginal effect decreases; the additional effect of adding 6 hour shifts is much larger that the additional effect of adding 2 hour shifts. This means that a large portion of the beneficial effects of flexible shifts can already be achieved without very short shifts.

Finally, a combination of flexible staffing and a waiting time restriction almost

exclusively results in savings of staff hours. Even for a safe service capacity assumption of 2.5, savings are substantial for all included waiting time requirements. For a realistic service capacity assumption of 3.0, savings are at least 18.5% compared to the current situation, and savings go as high as 42.0%, while still guaranteeing that 85% of all donors spend at most 60 minutes at the collection site.

5.6 Discussion

With the presented combined approach, savings on personnel are a possibility - assuming the results can be followed exactly with regard to employment contracts. At the same time, by aligning employee shifts and arrival patterns, it is possible to include waiting or delay time restrictions. Generally, three observations can be obtained from the results in section 5.5:

1. By including waiting time restrictions, an increase in staff working hours will be required.

2. By using flexible shift planning, substantial savings on working hours by staff members can be obtained.

3. By combining flexible shift planning and waiting time restrictions, no extra staff is needed, and generally a small saving on staff hours remains a possibility.

Most of these savings originate from a more flexible way of scheduling the shifts of staff members, in which shorter shifts are possible. In the flexible staffing in our Results section we allowed for all shifts lengths from 3 to 9 hours, but other shift possibilities and restrictions can easily be incorporated, depending on specific requirements from certain blood collection sites.

If we recall from section 5.1.1 that the production standard of 2.0 was set to include waiting times, it is worthwhile to compare the production standard of 2.0 with some of the waiting time restrictions with the realistic service capacity of 3.0, while maintaining the session shift assumption. We then observe that the result closest to a 0 % increase is for an expected waiting time restriction of 5 minutes, with all other restrictions resulting in a decrease of the number of staff hours. This means that the 2.0 production standard is probably quite low in most cases, and that an increase would most likely be possible in these cases. However, it is important to note that this might not hold for all collection sites. Especially small centers might still see a sharp increase in waiting and delay times if the production standard is increased.

Clearly, more advanced approximate results for queueing networks could be beneficial to the method described in this chapter. However, these results would most likely require more computational time, which might affect the applicability of the method in practice.

If the approach were to be applied at Sanquin collection sites, two issues can be thought of that could cause a difference between the computational results and the

results in reality after implementation. First, as the method does not take actual, individual employment contracts into account, and does not assign employees to shifts, realistic savings would probably be a bit lower. However, the savings would still be expected to be substantial enough to implement the approach, especially for long sessions. A second possible discrepancy might be caused by the waiting time computation. Although a large difference between the results from the model and reality is unlikely, especially for the network model, some differences might still occur. Some methods exist that would likely lead to smaller differences between reality and the model, but these methods come with their own downsides. Simulation could be an alternative method to give a more realistic approximation. However, this would most likely slow down calculations substantially and it would eliminate the possibility for an exact answer. Most importantly, though, simulation would not be generic and it would require adapting the simulation model to each individual collection site.

Since we can combine significant savings with waiting time guarantees and fast calculations - individual cases are solved in a matter of (milli)seconds, Sanquin investigated practical consequences of implementing the proposed approach with favorable results. The next step will be to actually apply the approach.

CHAPTER 6

Dynamic staff allocation

6.1 Introduction

A Sanquin blood collection site, like most other blood collection sites, has three stations that require staff members: Registration, Interview and testing, and the Donation. Staff members of collection sites are trained to do work at any of these stations. Currently, staff members generally work at just one of these stations for an entire shift or session. This means that there is a considerable amount of flexibility that is not being utilized, even though the arrival process of donors is highly time-dependent, as shown in Chapter 4. Staff members could change stations every time they finish a task or, if this is not desirable due to the number of changes this might cause, a few times during a shift. This would lead to more staff members working at stations where donors are available, thereby decreasing waiting times and leveling work pressure. This chapter will present an algorithm to optimize the reallocation of these staff members, based on a Markov Decision Process (MDP).

In Chapter 4, we have shown a computational method to compute and evaluate queueing distributions of time dependent queueing networks. Some of the presented numerical examples in Chapter 4 concern the allocation of staff members. The reallocations covered in Chapter 4 were all pre-determined, i.e. they are not state dependent. Nevertheless, some of these already showed decreasing queue lengths. Further improvements might be expected if state dependence is taken into account. To be able to use the full potential of reallocating the staff members, we further developed the model from chapter 4 into an MDP. Given the number of staff members that are working, the MDP is able to compute the optimal staff allocation of the staff members. This allocation is both state-dependent and time-dependent.

This chapter we will present this MDP model, which is able to compute the staff allocation that optimizes the expected number of waiting donors, present donors, or some other function based on the state of the collection site. We are able to numerically solve the presented MDP for a realistic, time-dependent problem size in reasonable time. Since the MDP requires some assumptions that might not fully reflect reality, we will also include a simulation to test the solutions computed by the MDP. The simulation will offer the opportunity to verify that the results computed by the MDP can be achieved in an even more realistic setting.

6.2 Literature

The problem studied in this chapter is often referred to as the server assignment problem. Early work mainly focuses on the static version of this problem. Papers that discuss this problem are generally concerned with designing an optimal production line, by changing buffer sizes and server assignments. After this design phase, the production line cannot be easily changed. The work by Yamazaki et al. [187] and the references therein show early examples of this work. More recent examples of this work can be found in van Woensel et al. [184].

More recently, there has also been work on the dynamic server assignment problem. The first paper discussing dynamic server assignment, seems to be by Ostolaza et al. [150]. Although this paper was unavailable to us, a continuation of the research is discussed by McClain et al. [130]. Both discuss a version of the dynamic assignment problem in which a server at a station can help the server at the next station if this server has fallen behind on its work. Gel et al. [86] expand this work by including the system architecture in their model and using an MDP to solve the problem. Ahn and Righter [5] study the dynamic server assignment problem if servers are trained in a subset of consecutive tasks. They show that often either a *last buffer first served* or a *first buffer first served* policy is optimal. Kirkizlar et al. [119] study the robustness of dynamic assignment policies for systems with non-exponential service time distributions and finite buffers. The paper also includes a heuristic policy that preforms near-optimal for these non-exponential systems. This work has inspired many papers that focus on throughput maximization; some examples of this work can be found in [10, 11, 13, 17, 18, 48, 106, 120, 177].

Zavadlav et al. [190] uses the work of Ostolaza et al. for a so-called Toyota Sewn-Products Management System (TSS). This is a system in which the server moves downstream with the job until it is handed of to another server, at which point the server moves upstream to pick up another job. Bischak [28] compares the TSS system to a static approach where every server handles one station. Bartholdi and Eisenstein [22] show an optimal sequence of the servers for a TSS system. Bartholdi et al. [23] show that this optimal sequence also works in a stochastic setting. McClain et al. [131] evaluate TSS and related techniques in a wider variety of situations.

Duenyas et al. [74] study the server assignment problem in a tandem queueing system with one server and setup time. The problem of what job to work on next is described as an MDP, and both an exact and a heuristic policy are tested with simulation. Sennott et al. [165] study a system in which every station has its own server, with one additional server that can change between stations. Next to setup time, their work also includes setup cost and holding cost. Andradottir et al. [12] study the dynamic server assignment problem in a system with two sequential stations, two servers and setup cost.

Our work differs from this research in a number of important ways. Most research that formulates the server assignment problem as an MDP either have difficulties getting a numerical solution for larger systems or do not include a numerical study at all. These papers focus on analytical results instead. Generally, these rely on

average reward results over an infinite horizon and are limited to a fixed number of servers. In this chapter, in contrast, we formulate the problem as an MDP, allow for a generic number of servers and are able to numerically compute the optimal policy for a time-dependent system in reasonable time. These numerical policies show similar performance if we simulate the realistic setting, as shown in section 6.6.

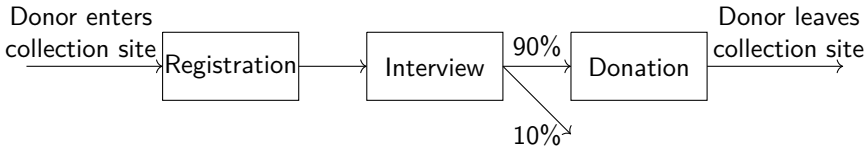
Moreover, earlier research of the server assignment problem is exclusively applied to production systems. The application at a blood collection site requires several differences in modeling. First of all, jobs in production systems are inanimate objects, which validates the focus on throughput taken by previous research. At blood collection sites, in contrast, we deal with donors and accordingly also deal with perceptions of delays and waiting times by these "jobs". Therefore, our objective is to minimize queue lengths. A blood collection site also forces us to work with a number of factors that have never been combined for the server assignment problem. Firstly, as a blood collection site is only open for a specific interval, the problem is intrinsically finite time. In addition, the arrivals are generally non homogeneous. A blood collection site therefore has to be modeled as a time-dependent, multi-server, tandem queue. All of these three factors - time-dependent, multi-server and tandem queues - have been studied before separately, but have never been combined.

There is only paper that considers flexible assignment of staff members at collection sites, by Brennan et al. [38]. The paper uses a simulation model to study a blood collection site and test several strategies to improve service. Among their tested strategies is the idea to use a 'floater', a staff member who can help at multiple stations. They, however, do not optimize the use of this floater, and don't allow all staff members to change their assignment. As the remainder of literature on blood collection sites does not specifically apply to this chapter, the reader is referred to section 1.4 for a discussion of the this literature.

To cover all aspects of the blood collection site, we model the assignment problem at the blood collection site as a Markov Decision Process. For an extensive description of Markov Decision Processes, the reader is refered to [155].

6.3 Queueing model

Figure 6.1 Schematic representation of queueing model of the collection site



An extensive description of a blood donation from a process point of view can be found in Section 1.3. For the queueing model, the complexity of the described process has been reduced down to just three stations: the Registration station, the Interview station and the Donation station, as is schematically shown in Figure 6.1.

All of these queues will be modelled as an M/M/s queue. After the Interview station, approximately 10% of the donors leave the system due to inelegebility at that visit. The average service times that will be used throughout the chapter are 2 minutes, 6 minutes and 12 minutes for the three stations respectively. In this queueing model, we will assume that all service times and inter-arrival times are exponentially distributed. This assumption is required for numerical experiments. As a consequence, the total state of the system can be described by just the number of people at the stations. This keeps the state-space small enough for the MDP to be numerically solvable. Although this model does not completely describe the reality of a blood collection site, this relatively simple queueing model can be used to compute quite accurate waiting times and queue lengths, as is shown in Chapter 3.

The combination of this model with an MDP formulation to compute the optimal staff allocation, requires two additional assumptions. The first assumption is a maximum number of donors that can be present at any of the stations, such that the state-space is finite. If we do not limit the state-space by imposing a maximum number of donors, the MDP model would be required to compute an optimal decision for an infinite number of states, which is numerically infeasible.

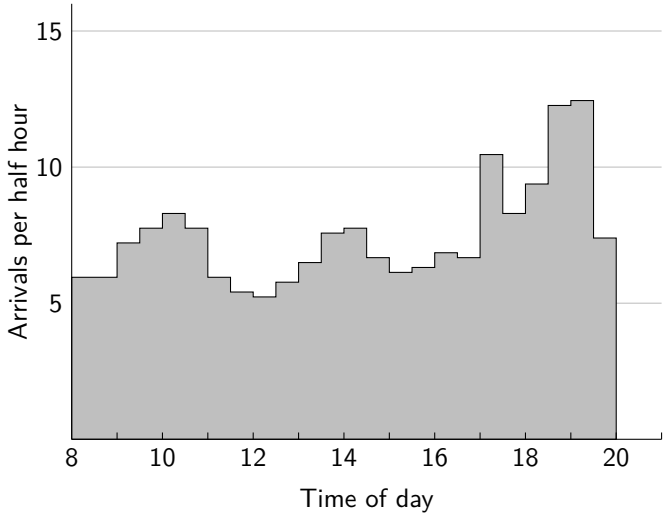
The second assumption is that decisions about reallocations are taken at fixed time points with regular intervals, at which point services are preempted when a staff member is reallocated. If this assumption had not been included, a decision would have to be taken every time a staff member finishes his task, which could be any time during the day, as exponential distributions are continuous. But, since we deal with a finite-horizon time dependent system, this would mean that the time would have to be included in the state, which would again lead to an infinite size of the state-space. With these two additional assumptions, the queueing model can be used to describe an MDP model that determines the optimal allocation of the staff members throughout the day.

These assumptions make the model less realistic, but are required for the MDP model to be numerically tractable. Although it has been shown in Chapter 3 that this model is able to predict waiting times and queue lengths reasonably accurately, we will test the results of the MDP model with a simulation. The simulation will not require most unrealistic assumptions such as exponential service times, preemption and a maximum number of donors. The simulation will also include actions such as filling out the questionnaire and splits the Donation station in three separated stations. The simulation model will be described in detail in section 6.6, followed by results form the simulation model.

6.4 Method: Markov Decision Process

We will model the staff allocation decision as a Markov Decision Process (MDP). There is no collection site that is continuously opened, so the decision problem has a natural, finite horizon. The time dependent aspects of the collection site cause two parameters of the MDP to change during a collection session. The possibly changing number of available staff members induces a changing action-space. When

Figure 6.2 Arrival pattern of whole blood donors for a full day blood collection session in the Netherlands.



this is combined with the arrival pattern from Figure 6.2, it also induces transition matrices that change over time. These time dependent parameters force us to solve the problem as an MDP model that changes over time.

6.4.1 Structure of the Markov decision process

An MDP is generally defined by five aspects: the states, the actions, the transition probabilities, the costs (or rewards) and the optimization criterion. Since the collection site will be modeled using a finite horizon MDP, we will also include horizon cost h^K. Whenever a superscript k is included, the parameter is dependent on the time step k.

6.4.1.1 States

The states are the first of the five aspects that define an MDP. The state for our MDP is defined as (n_1, n_2, n_3), where n_q is the number of donors at station q. To limit the size of the state-space, the maximum number of donors that can be present at station q is limited by N_q. The realistic state-space is therefore defined as:

$$S' = \left\{(n_1, n_2, n_3) \in \mathbb{N}^3 \mid 0 \leq n_1 \leq N_1; 0 \leq n_2 \leq N_2; 0 \leq n_3 \leq N_3\right\} \qquad (6.1)$$

Using this definition, the state-space would have 3 dimensions. To make computations easier, we have chosen to convert this to a one-dimensional state-space. The

Table 6.1 Parameters, variables and functions used in the MDP.

Parameters	
t	Duration of a time step in hours
T	Duration of the collection session in hours
K	Number of time steps in T: $K = T/t$
A^k	Set of possible actions at time step k
a_i^k	Action taken in state i at time step k
a^k	Vector of length I containing a_i^k's
λ^k	Arrival rate at time step k
$\bar{\lambda}$	Average arrival rate of donors
μ_q	Service rate of donors at station q
N_q	Maximum number of donors at station q
C_q	Maximum number of staff members at station q
C^k	Number of staff members working at time step k
S	state-space $\{1, 2...(N_1+1)(N_2+1)(N_3+1)\}$
I	Size of the state-space $\lvert S\rvert = (N_1+1)(N_2+1)(N_3+1)$
p^k_{i,j,a_i^k}	Transition probability from state i to state j at time step k when action a_i^k is taken
$P^k_{a^k}$	$I \times I$ transition matrix containing p^k_{i,j,a_i^k}'s
$r^{(1)}_{i,a_i^k}$	Costs based on total donors in state i when action a_i^k is selected
$R^{(1)}_{a^k}$	Vector of length I containing $r^{(1)}_{i,a_i^k}$'s
$r^{(2)}_{i,a_i^k}$	Costs based on waiting donors in state i when action a_i^k is selected
$R^{(2)}_{a^k}$	Vector of length I containing $r^{(2)}_{i,a_i^k}$'s
w_q	Weight for cost of having a queue at station q
h^k	Horizon cost at the end of time-step k
Functions	
$n_1(i)$	Number of donors at the Registration station in state i
$n_2(i)$	Number of donors at the Interview station in state i
$n_3(i)$	Number of donors at the Donation station in state i
$c_q(a_i^k)$	Element q of a_i^k, i.e. the number of staff members at station q when action a_i^k in taken.
Variables	
a_i^{*k}	Optimal action in state i at time step k
a^{*k}	Vector of length I containing a_i^{*k}'s
V^k	Value vector for the start of time-step k
π^k	Distribution of donors present after time step k
M^k	Average number of staff members reallocated after time step k

formulas to convert this one-dimensional state-space back to three dimensions can be found in section 6.8.1. The state-space that will be used for computations is:

$$S = \big\{1, 2, 3, ..., (N_1+1)(N_2+1)(N_3+1)\big\} \tag{6.2}$$

6.4.1.2 Actions

The second aspect of an MDP are the actions. An action is given by a tuple (A_1, A_2, A_3), where A_q is the number of staff members that is allocated to station q. To make sure that the staff allocation is feasible in a real collection site, two restrictions apply to the actions that are allowed. The first is that all staff members have to be allocated to one of the stations. Secondly, there is a restriction on the maximum number of staff members that can be allocated to one station. In reality, this would mean that there is a physical capacity constraint. For the Interview station this could for example be a finite number of interview rooms. Since an interview room can only be used by one staff member, allocating more staff members to the interview station than there are interview rooms is not allowed. This means that the action-space A^k at time step k is defined by:

$$A^k = \{(A_1, A_2, A_3) \in \mathbb{N}^3 |\ 0 \leq A_1 \leq C_1; 0 \leq A_2 \leq C_2; 0 \leq A_3 \leq C_3; \\ A_1 + A_2 + A_3 = C^k\} \tag{6.3}$$

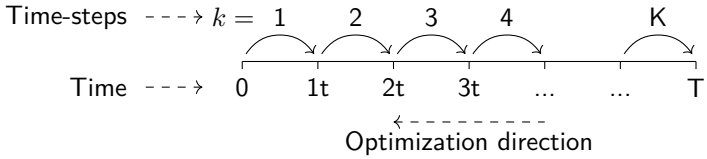
6.4.1.3 Transitions

The third aspect of an MDP are the transition probabilities from state i to state j given by p^k_{i,j,a^k_i}. These are dependent on the selected action a^k_i. The transition probabilities are combined in transition matrices $P^k_{a^k}$, dependent on the vector of actions $a^k = \{a^k_1, a^k_2, a^k_3 ... a^k_I\}$. To compute the transition matrices, we have used the method described in chapter 4. This method is based on the concept of uniformization, as described in Chapter 2. This requires exponential times between events and uniformly bounded transition rates. Uniformization converts the continuous time Markov chain (CTMC) to a discrete time Markov chain (DTMC) by first converting the generator matrix of a CTMC to a one step transition matrix. This can be combined with a Poisson distribution which determines the number of transitions in a time period to get a transition matrix for one time-step. The technical details of the method can be found in Chapter 4, and a detailed description of uniformization can be found in Chapter 2.

6.4.1.4 Cost

The fourth aspect of an MDP are the cost R_a. The cost in our model is based on the number of donors that are currently in the collection site. The number of donors is used as a cost, instead of waiting time, as the number of donors in the system or in queues can directly be deduced from the state combined with the action. We have defined two different cost structures, both have been defined as a function below. The function in equation 6.4 infers cost for every donor that is in the collection site, whether in service or waiting. The second cost structure, defined in equation 6.5, only infers a cost for every donor that is waiting for their service. Both structures will be used for numerical results in section 6.5.

Figure 6.3 The difference between time, time-steps and the direction of optimization of the dynamic program.



For the number of present donors:

$$r^{(1)}_{i,a^k_i} = \sum_{q=1}^{3} w_q \cdot n_q(i) \tag{6.4}$$

For the number of waiting donors:

$$r^{(2)}_{i,a^k_i} = \sum_{q=1}^{3} w_q \cdot \left[n_q(i) - c_q(a^k_i)\right]^+ \tag{6.5}$$

where $x^+ = \max(x, 0)$.

6.4.1.5 Optimization criterion

This definition of the costs also relates to the fifth aspect of an MDP, the optimization criterion. The costs are based on the number of donors in the system. Since we want to minimize the total time that donors have spent either in the system or the time they have spent waiting, and the process has a clear end time, we will use a finite horizon MDP minimizing the total cost. The discount factor has been set to 1.

6.4.2 Solving the Markov decision process

The MDP description from section 6.4.1 does not include the time-dependent nature of the collection site. However, as we will use dynamic programming to solve the finite horizon MDP, it is possible to change the parameters for different time-steps of the MDP. Figure 6.3 shows the difference between time, time-steps and the direction of optimization. The last time-step K will be the only one where the horizon cost are not pre-defined. To make sure the MDP does its best to clear the system, a large cost should be incurred if donors are left in the collection site after closing time. Therefore, for every donor still in the collection site, a cost will be incurred that is equal to having the donor wait for an hour at the end of the time horizon. If t is the step length in hours, this can be defined using the cost function from equation (6.4):

$$h^K = \frac{1}{t} \cdot R^{(1)}_{a^K_i} \tag{6.6}$$

Note that $r^{(1)}_{i,a^k_i}$, and therefore $R^{(1)}_{a^k}$, is independent of a^k. With these horizon cost, and the other MDP parameters given in section 6.4.1, we will solve the MDP backwards for $k = K-1, K-2, \ldots$ by iteratively solving the following two equations:

$$V^k_i = \min_{a^k_i \in A^k} \left[r^{(b)}_{i,a^k_i} + \sum_{j=1}^{I} p^k_{i,j,a^k_i} \cdot V^{k+1}_j \right] \tag{6.7}$$

$$a^{*k}_i = \arg\min_{a^k_i \in A^k} \left[r^{(b)}_{i,a^k_i} + \sum_{j=1}^{I} p^k_{i,j,a^k_i} \cdot V^{k+1}_j \right] \tag{6.8}$$

Here, $b \in \{1, 2\}$ represents the cost function that is chosen. The algorithm will keep decreasing k by 1 until it terminates as soon as $k = 0$. At this point, the algorithm has computed the optimal value V^k_i and optimal allocation of staff members at any time-step k and state i.

6.4.3 Number of staff reallocations

The probability distribution π^k over the states after every time-step k can be computed using the decisions and transition matrices described in the previous sections. The starting distribution π^0 has to be given as well. It is most natural to assume the collection site is empty when it opens, but it is also possible to use some other distribution, to reflect a possible initial queue when the collection site opens. When the optimal actions a^{*k}_i for all of the time steps k and states i have been computed, the queueing distribution can be computed by iteratively computing

$$\pi^k = \pi^{k-1} \cdot P^k_{a^{*k}} \tag{6.9}$$

Using this, we can compute the average queue length at the starting times of all decision intervals. Additionally, by using the $P^k_{a^k}$ of some strategy, it is possible to compute the queue length distribution and average queue length for any strategy.

An important factor in the applicability of the model, is the number of staff members that have to change their position when a new decision interval starts. By using the π^k's, the average number of staff changes per decision interval can be computed. The number of reallocations between time step k and $k+1$ can be computed by:

$$\begin{aligned} M^k = \sum_{i=1}^{I} \Bigg[\pi^k_i \cdot \sum_{j=1}^{I} \sum_{q=1}^{3} \Big[\big(c_q(a^{*k}_i) - \\ c_q(a^{*k+1}_j)\big)^+ \cdot p^k_{i,j,a^{*k}} \Big] - \big(C^{k+1} - C^k\big)^+ \Bigg] \end{aligned} \tag{6.10}$$

Note that by using this equation we assume that only the minimally required number of changes occurs, i.e. staff members only change position if they are superfluous at their current station, and directly move to the station that should have more staff members than it currently has. In case of shift changes, some extra assumptions apply. As we do not include shifts in our model, two consecutive shifts by different staff members cannot be distinguished, and will be regarded as one staff member for reallocations. When the start or end of a shift induces a change in the number of staff members present, this will not be counted as a reallocation.

6.5 Numerical results

6.5.1 Implementation and computational times

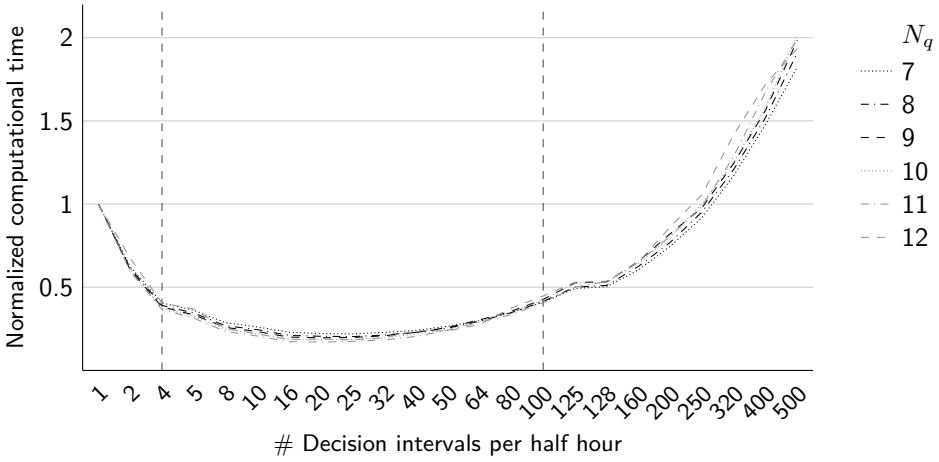
We have implemented the approach from section 6.4 in Matlab R2015b. This application has been chosen for two reasons. First, because it is known for its speed with matrix computations. The second reason was the existence of the Markov Decision Process toolbox, which has been used to compute the optimal decision for the individual MDPs.

The total computational time is dependent on the parameter settings. For most parameters the direction of this dependency can be easily predicted - e.g. if N_q or C_q increases, respectively the state-space or the action-space increases, and therefore the computational time will increase. However, the relation with the number of decision intervals is intriguing. If the number of decision intervals increases, the length of a time-step decreases, which makes the transition matrices easier to compute. At the same time, if the number of decision intervals increases, the number of optimal decisions that have to be computed also increases.

Figure 6.4 shows the computational time of the MDP algorithm for different numbers of decision intervals. The horizontal axis shows the number of decision intervals per half hour. The Figure shows the normalized computational times for different numbers of N_q ($N_1 = N_2 = N_3$), ranging from 7 to 12, as shown in the legend. For every scenario in the Figure - i.e. every distinct number of N_q and number of decision intervals - the optimal allocation has been computed for an MAE session. The computational times have been normalized by setting the computational time for 1 decision interval to 1, so the graphs for the different numbers of N_q can be compared. For completeness, the total computational time in seconds for 1 decision interval per half hour are included in 6.8.2.

Figure 6.4 clearly and consistently shows that the computational times follow a convex structure. The relative computational times of different values for N_q are indistinguishable, and do not differ much between 4 and 100 decision intervals per half hour. For 100 decision intervals, the optimal decision is computed for every 18 seconds. This is low relative to the service times, so there is no need for higher settings of this parameter. On the other end, four time intervals correspond to a decision interval of 7.5 minutes, which is in the same order of magnitude as the service times, and therefore an interesting setting to use for the results in section 6.5.2.

Figure 6.4 Relationship between the number of decision intervals and the normalized computational time (one decision interval = 1, absolute times can be found in 6.8.2). The different lines correspond to different values for N_q ($N_1 = N_2 = N_3$) that were used (as shown in the legend). The normalized computational time is lowest and not changing very much between the dashed lines at 4 and 100 decision intervals.



6.5.2 Queue length reductions

Currently, staff members do not change their allocation throughout the day. To measure the improvement of the MDP method described in this chapter, we will use two outcome measures, both of which measure the improvement compared to the best possible current situation. For the results in this chapter, we have chosen to show the reductions for eight, nine and ten staff members. The best possible static allocation of these staff members is shown in Table 6.2. Table 6.4 shows the reduction of the expected number of donors that are present at the collection site if the method is used, and Table 6.5 shows the reductions of the expected number of waiting donors at the collection site.

Table 6.2 Best possible static allocation of staff members. These allocations have been used as baseline for the results in Tables 6.4 and 6.5

Number of staff members available	Number allocated to station 1	station 2	station 3
8	1	2	5
9	1	3	5
10	1	3	6

We have included a variety of parameters settings. Next to three values for the number of staff members, three values will also be used for the length of the decision interval: 450 seconds, 90 seconds and 36 seconds. This corresponds to 4, 20 and 50 decision intervals per half hour respectively, all of which are in the range mentioned in Section 6.5.1. For the average arrival rate 5 values will be used: the average

number of donors arriving at the collection site will be varied between 12.5 and 22.5, with increments of 2.5. The average arrival rate has been multiplied with the arrival pattern shown in Figure 6.2.

Table 6.3 Scenarios for optimization and evaluation in Tables 6.4 and 6.5.

		Optimized for:	
		Donors present (cost function 1)	Donors waiting (cost function 2)
Evaluated for	Donors present	Columns with L in Table 6.4	Columns with L_q in Table 6.4
	Donors waiting	Columns with L in Table 6.5	Columns with L_q in Table 6.5

We have also used two different cost functions for the optimization in the MDP algorithm. The first cost function is based on the total number of donors at the collection site - indicated with L in Tables 6.4 and 6.5. The second cost function is based on the number of donors waiting - indicated with L_q in Tables 6.4 and 6.5. Staff members are scheduled for up to one hour after closing time to clean the equipment and help donors arriving shortly before closing time. All donors that are present one hour after closing time will incur a cost equivalent to one hour of waiting or being present for the first and second cost functions respectively. Note that this means that there are four possibilities for the combination of optimization and evaluation, as shown in Table 6.3. Also note that the difference between optimizing for the number of present donors and the number of waiting donors is very small, so the differences are expected to be small. However, the improvement should be larger if the optimization and evaluation are both based on the same cost function.

A number of things stand out in Tables 6.4 and 6.5. First, and above all, the possible improvements are very large. For most scenarios, a reduction of over 50% of the expected number of waiting donors seems achievable, and in the best scenario that was included in the numerical experiments, a reduction of 83.4% in the expected number of waiting donors was achieved. For the total number of donors present at the collection site, the reductions are naturally lower since time spent in service cannot be reduced with the methods in this chapter. The lowest reduction of the expected number of donors present is 8.0%, but this already includes a reduction of 50.5% of waiting donors. From this, we can conclude that the maximum possible reduction of the expected number of donors present for this scenario is around 16%.

We can also see that the differences between optimizing for the expected number of waiting donors and the expected number of donors present is very small. This is to be expected. It is possible to use the method to get more people into service, which leads to a reduction of the expected number of waiting donors. However, to get to the optimal allocation, the best strategy to both decrease the expected number of waiting donors and donors present, is to get the donors to exit the collection site as fast as possible. With long reallocation intervals, the optimization of waiting donors might choose to guarantee that during the interval as many donors as possible are in service. However, for shorter intervals, both methods will do essentially the same.

Table 6.4 Reductions on the expected number of present donors compared to the static allocation of staff members. Results directly from the MDP model.

Arrival rate	Decision interval	8 staff members optimized for L	8 staff members optimized for L_q	9 staff members optimized for L	9 staff members optimized for L_q	10 staff members optimized for L	10 staff members optimized for L_q
$\bar{\lambda} = 12.5$	450 seconds	-18.6%	-18.3%	-8.2%	-8.0%	-8.6%	-8.6%
	90 seconds	-22.1%	-22.1%	-10.4%	-10.4%	-9.7%	-9.7%
	36 seconds	-22.7%	-22.7%	-10.8%	-10.8%	-9.8%	-9.8%
$\bar{\lambda} = 15$	450 seconds	-24.5%	-24.3%	-10.8%	-10.6%	-11.4%	-11.3%
	90 seconds	-28.9%	-28.9%	-14.1%	-14.1%	-13.1%	-13.1%
	36 seconds	-29.8%	-29.8%	-14.7%	-14.7%	-13.3%	-13.3%
$\bar{\lambda} = 17.5$	450 seconds	-28.6%	-28.4%	-13.0%	-12.8%	-13.8%	-13.7%
	90 seconds	-33.6%	-33.6%	-17.3%	-17.3%	-16.1%	-16.1%
	36 seconds	-34.6%	-34.6%	-18.1%	-18.1%	-16.5%	-16.5%
$\bar{\lambda} = 20$	450 seconds	-30.8%	-30.6%	-14.3%	-14.2%	-15.2%	-15.1%
	90 seconds	-36.0%	-36.0%	-19.3%	-19.3%	-18.1%	-18.0%
	36 seconds	-37.1%	-37.1%	-20.3%	-20.3%	-18.6%	-18.5%
$\bar{\lambda} = 22.5$	450 seconds	-30.7%	-30.5%	-15.2%	-15.0%	-15.6%	-15.5%
	90 seconds	-36.0%	-35.9%	-20.5%	-20.4%	-18.8%	-18.8%
	36 seconds	-37.1%	-37.1%	-21.6%	-21.5%	-19.4%	-19.4%

Table 6.5 Reductions on the expected number of waiting donors compared to the static allocation of staff members. Results directly from the MDP model.

Arrival rate	Decision interval	8 staff members optimized for L	8 staff members optimized for L_q	9 staff members optimized for L	9 staff members optimized for L_q	10 staff members optimized for L	10 staff members optimized for L_q
$\bar{\lambda} = 12.5$	450 seconds	-62.0%	-62.8%	-49.4%	-50.5%	-65.6%	-66.0%
	90 seconds	-79.7%	-79.8%	-73.1%	-73.2%	-78.6%	-78.6%
	36 seconds	-83.4%	-83.4%	-77.7%	-77.7%	-80.7%	-80.7%
$\bar{\lambda} = 15$	450 seconds	-60.1%	-60.6%	-44.5%	-45.2%	-59.3%	-59.6%
	90 seconds	-75.3%	-75.4%	-66.4%	-66.5%	-72.5%	-72.6%
	36 seconds	-78.6%	-78.6%	-70.9%	-70.9%	-74.9%	-75.0%
$\bar{\lambda} = 17.5$	450 seconds	-57.4%	-57.8%	-40.0%	-40.5%	-53.4%	-53.7%
	90 seconds	-70.5%	-70.5%	-59.6%	-59.7%	-66.2%	-66.2%
	36 seconds	-73.3%	-73.3%	-63.8%	-63.8%	-68.6%	-68.7%
$\bar{\lambda} = 20$	450 seconds	-54.3%	-54.6%	-35.7%	-36.1%	-47.3%	-47.6%
	90 seconds	-65.6%	-65.7%	-53.0%	-53.1%	-59.3%	-59.4%
	36 seconds	-68.1%	-68.2%	-56.7%	-56.8%	-61.7%	-61.8%
$\bar{\lambda} = 22.5$	450 seconds	-50.0%	-50.5%	-32.2%	-32.5%	-41.5%	-41.8%
	90 seconds	-60.2%	-60.4%	-47.3%	-47.5%	-52.5%	-52.7%
	36 seconds	-62.4%	-62.7%	-50.6%	-50.8%	-54.8%	-55.0%

If all three stations are merged, there is a service capacity of 3 donors per staff member per hour. So, the total average load on the system $\rho = \bar{\lambda}/(3C)$, where C is the number of staff members. The average load of the scenarios included in the numerical experiments is between 0.42 and 0.94. However, since the number of arrivals is not constant, but instead follows the structure of Figure 6.2, the arrival

rate varies between 69.7% and 165.9% of the average arrival rate. This causes nine out of fifteen scenarios to contain at least one half hour period where the system is 'unstable', i.e. the arrival rate is higher than the maximum service rate. For an arrival rate of 22.5 and 8 staff members, 6 half hour periods have a load on the system larger than 1. Since the model contains a restriction on the total number of donors that can be present, one could assume that the difference between the static and flexible allocation would decrease. But even in these scenarios, the MDP performs very well.

Reductions are always higher when a shorter decision interval is used. This should be the case, as by definition it introduces more flexibility. It is also visible that the difference between 450 and 90 seconds is much larger that the difference between 90 and 36 seconds because the situation at the blood collection site cannot change very much in 90 seconds, let alone 36 seconds. This makes the additional flexibility largely redundant.

The reductions for eight staff members are larger than those of nine and ten staff members. This is mainly caused by the inefficiency of the static allocation for eight staff members. There is only a small difference between allocating the last available staff member to the Interview or Donation station. With the flexible allocation, the staff member can alternate between the Interview and Donation station. The results in section 6.5.3 support this, as the expected number of staff reallocations for eight staff members is significantly higher than for nine and ten staff members.

A noticeable difference between Tables 6.4 and 6.5 can be found when the average arrival rate $\bar{\lambda}$ is increased. The reductions in Table 6.4 are consistently larger for a higher arrival rate, while the reductions in Table 6.5 are decreasing. As the arrival rate increases, both the expected number of waiting donors and the expected number of donors present increase. With very few waiting donors, for the low arrival rates, the flexibility offered by the reallocations can almost entirely eliminate waiting donors, especially with frequent reallocations. However, the savings on the total number of present donors are clearly limited by the fact that only the number of waiting donors can be reduced, not the number of donors in service. When the collection site gets busier, the MDP can effectively do more with the offered flexibility, as shown in the reductions of the expected number of present donors. However, the method cannot completely negate the increase in the expected number of waiting donors, which leads to decreasing relative savings on the expected number of waiting donors. The decreasing gap between the reductions on the expected number of waiting donors and present donors also shows that the expected number of waiting donors is increasing.

Summarizing, the MDP model can accomplish significant reductions on numbers of waiting donors and numbers of donors present. Compared to the current static allocation, the best results are achieved for 8 staff members. The MDP profits from a modest number of reallocation intervals, but the difference between reallocating every 90 and every 36 seconds is very small.

6.5.3 Reallocated staff members

A major factor in the possible implementation of the described method will be the number of reallocations that are actually necessary. If staff members have to change every few minutes, the method will not be adopted, as staff members and their superiors will not cooperate with the implementation. The time a reallocation takes has not been included in the MDP optimization. For a low number of reallocations, the time a reallocation takes is negligible, and can be safely excluded from the model. However, if the number of reallocations increases, the time a reallocation takes starts to add up. We have therefore kept track of the average number of reallocations the optimal strategy requires. To do this, we have used the method described in section 6.4.3. The results are shown in Table 6.6 per half hour and Table 6.7 per decision interval. The structure of these is the same as for Tables 6.4 and 6.5, and the same scenarios have been included. It shows both the average number of reallocations.

Table 6.6 Total number of staff reallocations per half hour. Results directly from the MDP model.

		8 staff members		9 staff members		10 staff members	
Arrival	Decision	optimized for		optimized for		optimized for	
rate	interval	L	L_q	L	L_q	L	L_q
	450 seconds	1.6	1.6	1.3	1.3	1.1	1.1
$\bar{\lambda} = 12.5$	90 seconds	7.0	7.2	7.0	7.1	6.3	6.3
	36 seconds	18.8	18.7	17.8	17.7	21.6	17.3
	450 seconds	2.2	2.1	1.8	1.7	1.4	1.5
$\bar{\lambda} = 15$	90 seconds	9.9	10.1	9.6	9.7	8.2	8.2
	36 seconds	25.9	25.8	24.3	24.2	24.3	22.3
	450 seconds	2.9	2.7	2.3	2.1	1.8	1.8
$\bar{\lambda} = 17.5$	90 seconds	13.2	13.2	12.2	12.2	9.9	9.9
	36 seconds	33.8	33.3	30.9	30.5	27.1	26.9
	450 seconds	3.5	3.4	2.7	2.6	2.1	2.1
$\bar{\lambda} = 20$	90 seconds	16.5	16.4	14.5	14.6	11.3	11.3
	36 seconds	41.4	41.0	36.7	36.4	30.9	30.6
	450 seconds	4.2	4.0	3.1	3.0	2.3	2.3
$\bar{\lambda} = 22.5$	90 seconds	19.7	19.4	16.6	16.7	12.7	12.6
	36 seconds	48.5	48.5	42.1	41.8	34.3	33.8

Table 6.7 shows that the expected number of reallocations per decision interval does not seem to change very much. In general there are more reallocations if the arrival rate increases and if there are less staff members. Essentially, this is the same effect, as both of these increase the load on the system. The intuition behind this is most likely as follows: as the load increases, the average number of donors at the collection site also increases. This means that there is a higher probability that all staff members are currently working, which in turn increases the total number of events influencing the state of the system - arrivals and service completions. As a consequence the optimal allocation also changes at a higher rate, which induces more staff changes. The lack of an obvious optimal static strategy for eight staff

Table 6.7 Total number of staff reallocations per decision interval. Results directly from the MDP model.

Arrival rate	Decision interval	8 staff members optimized for L	8 staff members optimized for L_q	9 staff members optimized for L	9 staff members optimized for L_q	10 staff members optimized for L	10 staff members optimized for L_q
$\bar{\lambda} = 12.5$	450 seconds	0.40	0.40	0.33	0.32	0.27	0.28
	90 seconds	0.35	0.36	0.35	0.36	0.31	0.31
	36 seconds	0.38	0.37	0.35	0.35	0.43	0.35
$\bar{\lambda} = 15$	450 seconds	0.55	0.52	0.45	0.43	0.36	0.37
	90 seconds	0.49	0.50	0.48	0.48	0.41	0.41
	36 seconds	0.52	0.52	0.49	0.48	0.49	0.45
$\bar{\lambda} = 17.5$	450 seconds	0.72	0.68	0.56	0.54	0.44	0.45
	90 seconds	0.66	0.66	0.61	0.61	0.50	0.50
	36 seconds	0.68	0.67	0.62	0.61	0.54	0.54
$\bar{\lambda} = 20$	450 seconds	0.89	0.85	0.67	0.65	0.52	0.52
	90 seconds	0.83	0.82	0.73	0.73	0.57	0.57
	36 seconds	0.83	0.82	0.73	0.73	0.62	0.61
$\bar{\lambda} = 22.5$	450 seconds	1.04	1.01	0.77	0.76	0.58	0.58
	90 seconds	0.98	0.97	0.83	0.83	0.63	0.63
	36 seconds	0.97	0.97	0.84	0.84	0.69	0.68

members, as mentioned in section 6.5.2, also plays a role in the higher number of reallocations for eight staff members.

Although not very surprising, it is important to note the difference between 90 and 36 second intervals. The expected number of reallocations per decision interval is almost identical. Thus, the total number of reallocations per half hour is higher for the 36 second intervals. Combined with the fact that this does not cause a large decrease in the expected number of donors that are either waiting or present in the collection site, the 36 second option does not seem very appealing. However, most of these changes will likely include a preemption, so the simulation results will be required to get a more realistic indication of the expected number of reallocations.

Finally, it is important to note that for the 450 second intervals, the expected number of reallocations is reasonable. The expected number of reallocations for eight staff members and 450 second intervals is between 1.6 and 4.2 per half hour. This means that every staff member would, even in the worst case, only change its station every hour on average.

6.6 Simulation

6.6.1 Outline of the simulation model

The queueing model is a very useful way to model the system, as it allows for the use of methods like an MDP to optimize the allocation of staff members. However, it does require assumptions not completely reflecting reality. To assess the performance of the optimal allocation, a simulation model has been developed. The simulation model can be used to test the impact of the new allocation in a setting closer to

reality. It is important to note that the simulation will not be used to search for a better allocation, but only to evaluate the optimized policy computed by the MDP. The simulation model was built in Arena Simulation Software version 14.70.

Table 6.8 Schematic overview of the difference between the simulation model and the queueing model used for the MDP.

	Queueing model		Simulation model	
	Distribution	Value/ $\mu(\sigma)$ (minutes)	Distribution	Value/ $\mu(\sigma)$ (minutes)
Arrivals	Markov process	See Figure 6.2	Markov process	See Figure 6.2
Registration	Exponential	2	Lognormal	2 (2.53)
Questionnaire	*Assumed to be done while waiting for Interview*		Lognormal	3.5 (1)
Interview	Exponential	6	Lognormal	6 (2.91)
Deferral	Bernoulli	0.1	Bernoulli	0.1
Donation start	Exponential	12	Lognormal	7 (4.19)
Donation idle	Exponential	12	Lognormal	8 (1)
Donation end	Exponential	12	Lognormal	5 (4.19)
State restriction (N_q)	*Maximum of 12 donors can be present at each stations*		*No state restriction*	
Staff restriction (C_q)	*Maximum of 2, 4 and 6 staff members can work at the Registration, Interview and Donation station respectively.*			

The detailed differences between the simulation model and the queueing model on which the MDP was based can be found in Table 6.8. These differences can roughly be separated into four categories. The first entails the service time distributions. The queueing model has to use exponential distributions to remain numerically solvable. In the simulation model, however, we can use a more realistic service time distribution. After testing on data from Sanquin, a lognormal distribution resulted in the best fit for the service time distributions, so these will be used in the simulation.

The second and third major category of differences between the queueing model and the simulation model are the number of stations and the number of donors allowed to be at any of these stations. These variables determine the size of the state-space. This size is directly related to the numerical complexity of the MDP, and is therefore limited in the queueing model. There is no practical limit on the number of donors that can be present in the simulation model. In the case that there are more than twelve donors at one of the stations in the simulation model, we will use the optimal decision that is associated with twelve donors at this station, since the MDP has only computed the optimal decision for up to twelve donors per station.

The number of stations in the simulation model is also increased from the queueing model. The Donation station of the queueing model is split into three different stations in the simulation model. The first is the start of the donation. This is the moment when a staff member starts the donation by setting up the equipment and placing the needle. The second station of the donation - in Table 6.8 referred to as the 'donation idle' station - is the time when the blood is drawn from the donor.

Generally, no staff member is strictly required for this station, although staff members will check on the donor regularly in between tasks. The final station of the donation - the 'donation end' station in Table 6.8 - is the station where the staff member ends the donation. Note that the time where a staff member is required during the donation in the simulation model is equal to the total time required for the donation in the queueing model. The simulation model also includes the time spent by donors to complete their questionnaire. Like the donation idle station, this does not require staff members, and is therefore not included in the queueing model.

The fourth and last major difference between the queueing and the simulation model does not show up in Table 6.8 and has to do with preemption. The MDP model from section 6.4 reallocates all staff members after a time interval, including staff members that are currently servicing a donor at one of the stations. Although this might be a 'dummy reallocation' - i.e. they are allocated to the same station they are currently working at, it is also possible that they are allocated to a different station. In the queueing model, the staff member will then stop working on its current task, and immediately switch to its new task. The simulation model doesn't require this assumption.

6.6.2 Results from simulation evaluation

6.6.2.1 Queue length reductions

With the simulation, we acquired the same results as in section 6.5.2, and have also used the same structure for the Tables 6.9 and 6.10. So, for a detailed description of the structure of these tables, see the first paragraphs of section 6.5.2.

To get the results for Tables 6.9 and 6.10, we have used the simulation model to simulate 100 collection sessions for each scenario. After the 100 runs, the seed was reset, such that all scenarios with the same arrival intensity have dealt with exactly the same donors - i.e., we have used common random numbers. Wherever significance is mentioned, a 5% significance level is used.

The first thing to note in Tables 6.9 and 6.10, is that all but one scenario still show a significant improvement over the static allocation. The only scenario that does not show a significant result, both in the expected number of waiting donors and the expected number of present donors, is the scenario with $\bar{\lambda} = 22.5$, 36 second decision interval, 9 staff members and optimized for number present. The reductions are the highest for 8 staff members, followed by 10 staff members and 9 staff members. This effect is also visible with the numerical experiments, and has been explained in section 6.5.2, but it stands out more in the simulated results.

There are of course a few differences. It is interesting to note that although the results for a decision interval of 450 seconds look similar to the results of the MDP model in section 6.5.2, the results generally deteriorate for 90 and 36 seconds. For these short decision intervals, the MDP model assumes an amount of flexibility that does not exist in reality. The MDP only looks forward 36 or 90 seconds, and assumes the staff member can be reallocated after this time. In reality, the staff member will be fixed to the station he or she was sent to until he or she finishes at least one

Table 6.9 Reductions on the expected number of present donors compared to the static allocation of staff members. Results from the simulation.

Arrival rate	Decision interval	8 staff members optimized for L	8 staff members optimized for L_q	9 staff members optimized for L	9 staff members optimized for L_q	10 staff members optimized for L	10 staff members optimized for L_q
$\bar{\lambda} = 12.5$	450 seconds	-11.8%	-11.2%	-6.4%	-5.8%	-7.1%	-6.6%
	90 seconds	-11.2%	-11.3%	-6.6%	-6.7%	-7.3%	-7.3%
	36 seconds	-10.5%	-11.0%	-6.4%	-6.5%	-7.4%	-7.4%
$\bar{\lambda} = 15$	450 seconds	-16.7%	-16.5%	-8.0%	-7.6%	-9.1%	-8.7%
	90 seconds	-14.8%	-15.3%	-8.0%	-8.1%	-9.5%	-9.6%
	36 seconds	-13.6%	-14.3%	-7.6%	-7.8%	-9.5%	-9.6%
$\bar{\lambda} = 17.5$	450 seconds	-21.1%	-21.3%	-8.9%	-8.8%	-11.1%	-10.8%
	90 seconds	-17.9%	-18.9%	-8.5%	-8.7%	-11.5%	-11.8%
	36 seconds	-16.3%	-17.0%	-7.9%	-8.3%	-11.5%	-11.6%
$\bar{\lambda} = 20$	450 seconds	-25.3%	-26.3%	-8.5%	-9.0%	-12.2%	-12.2%
	90 seconds	-20.4%	-22.3%	-7.4%	-8.0%	-12.8%	-13.4%
	36 seconds	-17.9%	-19.0%	-6.6%	-7.4%	-12.8%	-13.2%
$\bar{\lambda} = 22.5$	450 seconds	-29.6%	-32.5%	-6.2%	-7.7%	-12.9%	-13.6%
	90 seconds	-23.5%	-26.6%	-4.4%	-5.6%	-13.6%	-14.5%
	36 seconds	-18.2%	-20.1%	-3.2%	-4.6%	-13.6%	-14.2%

Table 6.10 Reductions on the expected number of waiting donors compared to the static allocation of staff members. Results from the simulation.

Arrival rate	Decision interval	8 staff members optimized for L	8 staff members optimized for L_q	9 staff members optimized for L	9 staff members optimized for L_q	10 staff members optimized for L	10 staff members optimized for L_q
$\bar{\lambda} = 12.5$	450 seconds	-66.7%	-63.0%	-63.2%	-57.6%	-76.1%	-71.0%
	90 seconds	-63.5%	-64.2%	-65.1%	-65.7%	-79.0%	-79.0%
	36 seconds	-59.1%	-62.5%	-63.2%	-64.4%	-79.3%	-79.2%
$\bar{\lambda} = 15$	450 seconds	-63.8%	-63.1%	-55.7%	-53.3%	-71.0%	-67.7%
	90 seconds	-56.7%	-58.6%	-56.2%	-56.8%	-74.1%	-74.4%
	36 seconds	-52.0%	-54.7%	-53.2%	-54.9%	-74.1%	-74.2%
$\bar{\lambda} = 17.5$	450 seconds	-59.4%	-59.9%	-43.9%	-43.5%	-61.4%	-59.5%
	90 seconds	-50.5%	-53.2%	-41.9%	-43.0%	-63.4%	-65.0%
	36 seconds	-45.9%	-47.9%	-39.0%	-40.9%	-63.2%	-64.1%
$\bar{\lambda} = 20$	450 seconds	-54.4%	-56.7%	-31.4%	-33.3%	-50.4%	-50.8%
	90 seconds	-43.8%	-48.0%	-27.4%	-29.7%	-53.1%	-55.4%
	36 seconds	-38.3%	-41.0%	-24.1%	-27.4%	-52.9%	-54.6%
$\bar{\lambda} = 22.5$	450 seconds	-49.8%	-55.2%	-17.0%	-21.6%	-41.2%	-43.6%
	90 seconds	-39.2%	-44.9%	-11.4%	-15.2%	-43.6%	-46.5%
	36 seconds	-30.2%	-33.9%	-8.0%	-12.2%	-43.5%	-45.4%

service at this station, which almost always takes longer than 90 seconds. The fact that the only exceptions to this are scenarios with a very low load confirms this. If the load is very low, staff members have a lot more idle time where the staff member can be reallocated after 36 or 90 seconds.

Summarizing, the results from the simulation support the MDP results. Although

results are not identical, which is to be expected as the simulated model is much closer to reality, both show a large potential for improvement, with the highest reductions for 8 staff members.

6.6.2.2 Reallocated staff members

We have also used the simulation model to check the expected number of reallocations that were determined by the numerical experiments. Table 6.11 has the same structure as Table 6.6, but reallocations per decision interval are not included. In the simulation, staff is not reallocated at the start of decision intervals, but only when a staff member finishes their task. Reallocations per decision interval therefore no longer have a clear interpretation.

Table 6.11 Number of staff reallocations per half hour. Results from the simulation.

		8 staff members		9 staff members		10 staff members	
Arrival	Decision	Optimized for		Optimized for		Optimized for	
rate	interval	L	L_q	L	L_q	L	L_q
	450 seconds	3.0	2.5	2.3	2.0	1.5	1.3
$\bar{\lambda} = 12.5$	90 seconds	3.7	3.7	2.6	2.6	1.6	1.6
	36 seconds	3.7	3.7	2.6	2.6	1.6	1.6
	450 seconds	4.1	3.4	3.2	2.8	2.2	1.9
$\bar{\lambda} = 15$	90 seconds	5.1	5.0	3.7	3.7	2.4	2.4
	36 seconds	5.1	5.1	3.6	3.6	2.4	2.4
	450 seconds	5.0	4.4	4.1	3.7	2.8	2.5
$\bar{\lambda} = 17.5$	90 seconds	6.4	6.3	4.6	4.6	3.1	3.1
	36 seconds	6.4	6.4	4.5	4.6	3.0	3.1
	450 seconds	5.8	5.2	4.7	4.4	3.2	3.0
$\bar{\lambda} = 20$	90 seconds	7.4	7.3	5.3	5.3	3.5	3.5
	36 seconds	7.3	7.4	5.3	5.3	3.4	3.5
	450 seconds	6.2	5.9	5.3	5.1	3.4	3.3
$\bar{\lambda} = 22.5$	90 seconds	8.0	8.1	5.8	5.9	3.7	3.8
	36 seconds	7.9	8.3	5.8	5.9	3.7	3.7

The expected number of reallocations for decision intervals of 450 seconds are in the same order of magnitude as the same results from the numerical experiments, although the expected number of reallocations is slightly higher for the simulation. This confirms that for this decision interval, the simulation is able to reasonably closely follow the optimal decisions of the MDP model, which also results in similar results for the expected number of waiting and present donors, as can be seen in Tables 6.9 and 6.10.

In contrast, the expected number of reallocations for decision intervals of 90 and 36 seconds are much lower than the corresponding results from the numerical experiments. This also matches with the conclusions in section 6.6.2.1: the MDP uses flexibility that does not exist in reality. To reach the reductions of waiting and present donors that are shown by the numerical results, the much higher number of reallocations that go along with these savings turn out to be really necessary.

Table 6.11 shows that the expected number of reallocations for the decision intervals of 90 and 36 seconds are almost identical. Although these similarities are not as clear for the expected number of donors, the reductions shown for 90 and 36 seconds in Tables 6.9 and 6.10 are often not significant as well.

6.7 Discussion

The flexibility offered by the possible reallocation of staff members can be utilized effectively by using the MDP model outlined in section 6.4. The numerical results in section 6.5.2 show substantial reductions of up to 83.4% on the expected number of waiting donors for a low average arrival rate. A simulation study confirms that large, but slightly lower reductions of around 60% on the expected number of waiting donors are possible if the optimal policy is used.

The MDP is constructed such that it can reallocate all staff members after a fixed decision interval. Shortening this interval therefore creates more flexibility for the MDP method, which should result in more savings on waiting and present donors. This is indeed supported by larger reductions in the numerical experiments of section 6.5. However, the simulations in section 6.6.2 show that this greater flexibility does not actually reduce numbers of waiting and present donors. The flexibility that is assumed by the MDP, namely that staff members can be reallocated extremely quickly, is not actually present in reality, where staff members first have to finish their task before switching. The longer decision interval, which has to take more future changes into account with its decision, performs better in the simulation. This effect is strengthened if the load on the system increases, because more staff members are fixed to their current station performing service. Due to this effect, the largest reductions suggested by the numerical results are not achievable in practice, but reductions of approximately 60% seem plausible.

If a staff member changes its allocation, this will require some time. Some equipment might have to be shut down, and physically changing to another position in the system will also require some time. Although adding these setup times to the MDP model would make it more realistic, it would also require an expansion of the state-space, making the problem harder, if not impossible to solve. Additionally, the expected number of reallocations is low for longer decision intervals and in the simulation model, it is safe to assume that the addition of these setup times will not significantly influence the results.

As could have been expected, the savings on the expected number of waiting and present donors are higher if the best static allocation of the staff members is less obvious. This is supported by both the numerical experiments and the simulation, where the reductions for eight staff members are clearly superior to the reductions for nine and ten staff members.

Numerical results from the MDP model show that the expected number of times a staff member has to reallocate during a session greatly increases if the decision interval shortens. The numerical results even show that the expected number of reallocations per decision interval is more or less constant for a given arrival rate and

number of staff members. This implies that the expected number of reallocations between a decision interval of 450 seconds and one of 90 seconds differs by approximately a factor 5. The simulation does not show this result. Although the expected number of reallocations for 90 and 36 second decision intervals are somewhat higher than those of 450 second intervals, the differences are not nearly as large as for the numerical results. However, given the results on the reductions of waiting and present donors, it still seems advisable to use a decision interval of 450 seconds.

The main challenge remaining is the implementation of a reallocation policy. Roughly, two directions can be taken for the implementation. The first option is to deduce simple rules from the extensive optimal strategies, such that staff members can use these as a rule of thumb on when to reallocate. The second option is to develop a digital system which keeps track of donors at the collection site and notifies staff members on when to change their allocation. In this case, the entire optimal strategy from the MDP can be used.

Summarizing, the MDP based model is able to significantly reduce waiting in systems with a number of sequential servers. In some cases queue lengths can be reduced by over 60%. For the blood collection site, a further indication of these reductions are given in sections 6.5 and 6.6.2. Although for other similar systems, more detailed studies might be necessary to validate the presented model, these indicated results are highly promising.

6.8 Appendix

6.8.1 Appendix I: Formulas to compute three dimensional state

$$n_1(i) = \left\lfloor \frac{i-1}{(N_3+1)(N_2+1)} \right\rfloor \tag{6.11}$$

$$n_2(i) = mod^{(N_2+1)} \left\lfloor \frac{i-1}{N_3+1} \right\rfloor \tag{6.12}$$

$$n_3(i) = mod^{(N_3+1)}(i-1) \tag{6.13}$$

6.8.2 Appendix II: Absolute computational times for Figure 6.4

Table 6.12 Computational times in seconds for the MAE session with a decision every half hour

Maximum donors per station	7	8	9	10	11	12
computational time (seconds)	184.2	349.3	636.9	1097.0	1825.6	2995.1

CHAPTER 7

Combining appointments and walk in donors

7.1 Introduction

Previous chapters discussed methods to compute and decrease waiting times by using the available capacity at the collection site as efficiently as possible. Chapter 5 specifically dealt with the problem of assigning staff such that the number of available staff members matches the number of expected arrivals. This chapter will address essentially the same problem, but will do this by altering the arrival process instead of the service process.

By introducing appointments, it is possible to direct arrivals of donors to times when more staff capacity is available. However, a substantial number of whole blood donors prefer the flexibility of walk-in arrivals. Sanquin therefore plans to introduce appointments as an option for whole blood donors, while donors that prefer walk-in will not be forced to plan appointments. This chapter will deal with planning these appointments, while taking the plasma appointments and walk-in appointments into account. The eventual goal of Sanquin is for 80% of whole blood arrivals to have an appointment. The underlying notion exists that these appointments smoothen the arrival process and thereby reduce waiting times. This is to be investigated in this chapter.

Arriving donors first enter a queue at the registration desk, where the type of donor is not yet known. In the remainder of the collection process, plasma donors receive priority over whole blood donors at the collection site. To encourage donors to make appointments, a new priority level will be introduced for whole blood appointments. Whole blood donors with an appointment will have priority over whole blood donors without one, but will still have lower priority than plasma donors.

These priorities are set by Sanquin management, but it is worth noting that they are not always strictly followed in practice. Staff members might sometimes choose to serve a whole blood donor that has been waiting for a while, even if a plasma donor is present in the same queue. As this is dependent on staff members, location and other factors, and the exact situation of queues is not kept track off, it is not possible to take this into account. For this reason, the remainder of the chapter will use strict priority rules.

Although the computational model and planning approach introduced in this chapter is based on blood collection sites, processes with a registration desk and

priority rules after this initial queue, exist in many more situations. As examples one might think of an outpatient phlebotomy laboratory, where the priority is based on urgency of patients, a bank office, or a city hall. For several of these situations, similar models would be applicable.

We will start this chapter by discussing the relevant literature on appointment scheduling in Section 7.2 and the contribution of this paper to the literature. As an extensive survey of the literature on blood collection sites can already be found in Section 1.4, this will not be repeated here. Subsequently, we will discuss the method developed for the scheduling of appointments in Section 7.3. We will then introduce the test case for the approach based on the blood collection site in the city of Enschede in the Netherlands in Section 7.4.1. Based on some numbers from this collection site, random arrival patterns will be generated, and the approach will also be tested on these patterns. Results from these random arrival patterns are discussed in Section 7.4.3. The chapter will then be concluded with a discussion in Section 7.5

7.2 Literature

Although appointment scheduling at blood collection sites has rarely been studied, extensive research has been done on appointment scheduling in a broader healthcare setting. Reviews of this literature can be found in the surveys by Cayirli and Veral [43], Gupta and Denton [102], Hulshof et al. [110] and Ahmadi-Javid et al. [3]. We will refer the reader to these papers for a full overview of appointment scheduling in Healthcare. As our main goal is to limit waiting times by combining appointments with walk-ins, we will elaborate on this area of research.

7.2.1 Appointment scheduling in combination with walk-in

Reilly et al. [161] propose a method called "delay scheduling" by which walk-in patients to a clinic can be assigned a delay time, i.e. invited to return at a later time, if the clinic is very busy.

Su and Shih [171] study an outpatient clinic with a very high number of walk-in patients. They use a simulation model to test several appointment scheduling policies. Some of these policies significantly improve service.

Green et al. [97] use a dynamic programming approach to schedule outpatient appointments for diagnostic facilities, while taking walk-in inpatients and urgent patients into account. Their dynamic program simultaneously provides priority rules for admitting new patients into service.

Sickinger and Klisch [166] present an extension of the Bailey-Welch rule combined with a neighborhood search to schedule outpatient appointments. Their aim is to maximize the total reward, which incorporates costs for waiting patients.

Dobson et al. [70] study the effects of reserving appointment slots for urgent patients in primary care.

Koeleman and Koole [121] study the scheduling of non-emergency patients when arrivals of emergency patients are time-dependent. Their objective is to balance

waiting times, idle times and overtime. Using a local search algorithm, they are able to show significant improvements.

Luo et al. [127] study an outpatient clinic, for which a framework is developed to schedule outpatients. A balance has to be found between waiting times and server utilization, while taking unscheduled emergency patients into account.

Peng et al. [151] present a hybrid simulation and genetic algorithm to develop scheduling templates for clinics, if walk-in patients are also allowed. Their method is able to reduce cost significantly.

Kortbeek et al. [123] use an iterative procedure consisting of an access time model and a waiting time model to plan appointments for a CT scanner. The method is used to balance waiting time at the facility and access time by planning the appointment during times when a low number of walk-in patients is expected.

In most of this research, walk-in patients are emergency or urgent patients, whom are given priority over scheduled patients. In the situation at the blood collection site, however, these priorities are reversed. This may result in different decisions.

Compared with the most closely related paper in the existing literature, the one by Alfonso et al. [6], our method has important advantages, especially for the Dutch situation. First of all, our method takes transient behavior into account, compared to the steady state performance of the petri net model used in Alfonso et al. Transient analysis is important both in the case that the previous appointments have not yet cleared the system as new appointment donors arrive, and as the arrival process of the donors without appointment is random and continuously changing. Both of these imply that the system never reaches a steady state situation. The presented model can easily be extended to incorporate changing numbers of staff members throughout the day, which would increase the necessity for a transient analysis even more.

Secondly, our model includes priorities. In reality, donors with appointments receive priority over donors without appointments, and plasma donors receive priority over whole blood donors. Ignoring these priorities would vastly underestimate the queues for donors without appointments, which in turn could lead to losing these donors for future donations.

Finally, a third aspect of our contribution in this chapter is the combination of donors with and without appointments. Although healthcare systems often have to deal with patients without an appointment (e.g. urgent patients), only a limited number of papers (17) appear in the most recent literature review [3] that include walk-in patients.

7.3 Method

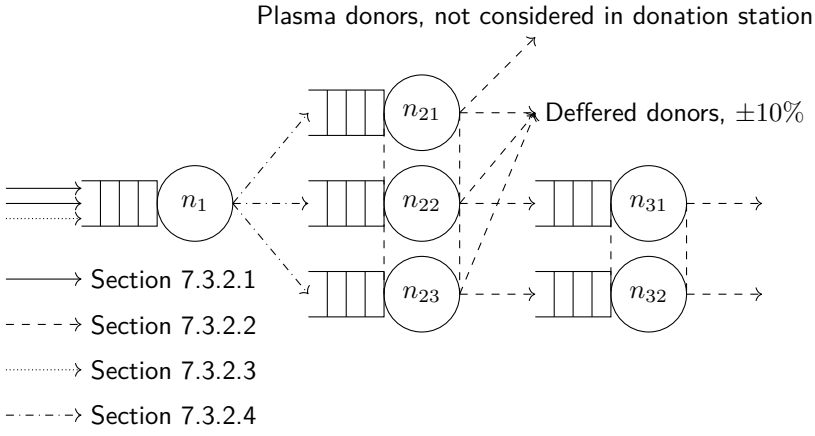
In this section, we present a description of the proposed approach for the computation and evaluation of appointment schedules for whole blood donors. First, Section 7.3.1 presents the modeling of a blood collection site including plasma donors and three required priority classes. Section 7.3.2 then present the corresponding transition structure and probabilities for computation of queueing distributions of this model

of a blood collection site. Finally, section 7.3.3 presents how appointment schedules are generated and improved.

7.3.1 Model

We will model the blood collection site as a tandem queue with priorities, as shown in Figure 7.1. Donors first enter the queue for the registration station. At this station, no priority rules apply, as the type of donor is not yet known in reality at this point. After having received service at the registration station, the donors move on to one of the queues for the interview station, depending on their type. Similarly to reality, we will distinguish three types of donors: plasma donors, whole blood donors with appointment and whole blood donors without appointment. From this point in the system, plasma donors are in the highest priority class, followed by whole blood donors with an appointment and then whole blood donors without an appointment. We assume that 10 % of donors are rejected for a donation at that visit and leave the system after the interview station, which corresponds to data obtained from Sanquin. The plasma donors are also discraded in the model after the interview station, as their donation station is seperated and will not be considered in this model. The whole blood donors move on to the donation station, again seperated based on their type. After the donation station, the remaining donors also leave the system.

Figure 7.1 Schematic representation of the model.



Inter-arrival times will be exponential for whole blood donors without appointments, which is a natural assumption for walk-in arrivals. However, the exponential distribution for the inter-arrival times can easily be replaced by some other discrete distribution (see Section 7.3.2.3). Appointment arrivals are assumed to either not show up with probability $p^{(\mathsf{ns})}$, or to show up exactly at the time of their appointment, at which time they will join the queue of the registration desk. Services at the registration station have a deterministic service time of 2 minutes. Services at the registration desk do not have a great variance, which justifies this assumption.

However the most important reason for choosing a deterministic service time is the possibility to track which donor is in what spot in the line for the registration desk, without having to include it in the state. To keep the problem tractable, services at the interview and donation station have exponentially distributed service times with an average of 6 and 12 minutes respectively.

To be able to keep the problem tractable, we need two additional assumptions. First, services at the interview and donation station will be preempted if a higher priority donor comes to the station. I.e. a plasma donor can preempt any whole blood donor in service, and a whole blood donor with an appointment can preempt a donor without an appointment.

Second, we will limit the number of donors that can be present in the registration station by N_1, the interview station by N_2 and at the donation station by N_3. This results in the following state-space:

$$\begin{aligned} S = \{(n_1, n_{21}, n_{22}, n_{23}, n_{31}, n_{32}) \in \mathbb{N}_0^6 \mid \\ n_1 \leq N_1, n_{21} + n_{22} + n_{23} \leq N_2, n_{31} + n_{32} \leq N_3\}, \end{aligned} \tag{7.1}$$

where n_1 is the number of donors at the Registration station. n_{21} refers to the number of plasma donors at the interview station, n_{22} and n_{23} refer to the number of whole blood donors with and without appointments at the Interview station respectively. n_{31} and n_{32} refer to the number of whole blood donors with and without appointment at the donation station respectively. This is also visualized in Figure 7.1. For notational convenience, we will use i and j to represent states in this state-space:

$$\begin{aligned} i &= (n_1, n_{21}, n_{22}, n_{23}, n_{31}, n_{32}) \quad \in S \\ j &= (n_1', n_{21}', n_{22}', n_{23}', n_{31}', n_{32}') \quad \in S, \end{aligned} \tag{7.2}$$

where i will refer to the current state, and j to the next state.

7.3.2 Transitions

The time between the start and end of the donation session will be divided into short time-steps. Each of these time-steps will have length τ_1, the constant, deterministic service time at the registration desk. The queueing distribution will be computed in a forward manner. We therefore assume that the queueing distribution π_{t_k} at time step t_k is known. During every time-step, four different transition types will be distinguished. With these the queueing distribution $\pi_{t_{k+1}}$ will be computed. Methodologically, the first transitions to take place are the arrivals of appointments. Then, services take place at the second and third station, directly followed by the arrivals of walk-in donors. Lastly, at most one service is completed at the first station. We will distinguish four different types of transitions: Arrivals of donors with an appointment in Section 7.3.2.1, services at the Interview and Donations stations in Section 7.3.2.2, Walk-in arrivals of donors in Section 7.3.2.3, and finally services at the Registration station in Section 7.3.2.4. Section 7.3.2.5 will then briefly outline how the four transition types are combined.

7.3.2.1 Appointment arrivals

The first of the transitions to occur during a time-step are the appointment arrivals. We will deal with these arrivals one by one. With $p^{(\mathsf{ns})}$ the no-show probability of a donor with appointment, i.e. the probability that a donor with an appointment does not show up to the collection site, the one-step transition matrix for these arrivals is given by:

$$P^{(\mathsf{aa})}_{i,j} = \begin{cases} (1-p^{(\mathsf{ns})}) & j = i + e_1, n_1 < N_1 \\ p^{(\mathsf{ns})} & j = i, n_1 < N_1 \\ 1 & j = i, n_1 = N_1 \\ 0 & \text{else.} \end{cases} \tag{7.3}$$

After the first queue, appointment donors get priority over other donors. However, if we were to explicitly track these donors, the state space would become very large. Therefore, we will track the location of appointment donors independently of the state space. Let $\sigma^{(\mathsf{p})}_{t_k} = (\sigma^{(\mathsf{p})}_{t_k,1}, \ldots, \sigma^{(\mathsf{p})}_{t_k,N_1})$ be a vector with elements $\sigma^{(\mathsf{p})}_{t_k,n_1}$ containing the probability that a plasma donor is in the queue at the registration station, with $n_1 - 1$ other donors waiting before this plasma donor, and let $\sigma^{(\mathsf{wb})}_{t_k}$ with elements $\sigma^{(\mathsf{wb})}_{t_k,n_1}$ be the same vector for whole blood donors.

Let $a^{(\mathsf{p})}_{t_k}$ and $a^{(\mathsf{wb})}_{t_k}$ be the number of plasma and whole blood appointments scheduled for time t_k respectively. We will use $\pi^*_{t_k}$ to indicate the temporary queueing distributions that track progress in between π_{t_k} and $\pi_{t_{k+1}}$, and $\pi^{1*}_{t_k}$ to indicate the queueing distribution aggregated for all stations except for station 1.

$$\begin{aligned} \sigma^{(\mathsf{wb})}_{t_k,n_1} &= \sigma^{(\mathsf{wb})}_{t_k,n_1} + \pi^{1*}_{t_k,n_1-1} \frac{a^{(\mathsf{wb})}_{t_k}}{a^{(\mathsf{wb})}_{t_k} + a^{(\mathsf{p})}_{t_k}} (1-p^{(\mathsf{ns})}), \qquad n_1 = 1,2,\ldots,N_1. \\ \sigma^{(\mathsf{p})}_{t_k,n_1} &= \sigma^{(\mathsf{p})}_{t_k,n_1} + \pi^{1*}_{t_k,n_1-1} \frac{a^{(\mathsf{p})}_{t_k}}{a^{(\mathsf{wb})}_{t_k} + a^{(\mathsf{p})}_{t_k}} (1-p^{(\mathsf{ns})}), \qquad n_1 = 1,2,\ldots,N_1. \end{aligned} \tag{7.4}$$

We will use $\pi^*_{t_k}$ to store the intermediate queueing distribution. Before the first step, this is initialized as $\pi^*_{t_k} = \pi_{t_k}$. We then update $\pi^*_{t_k}$ by:

$$\pi^*_{t_k} := \pi^*_{t_k} P^{(\mathsf{aa})}. \tag{7.5}$$

To compute the queue location vectors $\sigma^{(\mathsf{p})}_{t_k}$ and $\sigma^{(\mathsf{wb})}_{t_k}$ and the queueing distribution after the arrival of appointments, iterate equations (7.4) and (7.5) $a^{(\mathsf{p})}_{t_k} + a^{(\mathsf{wb})}_{t_k}$ times.

7.3.2.2 Services at Interview and Donation stations

The second transition type to take place, will be the services at the interview and donation station. These services will be computed based on uniformization (as extensively described in Chapter 2). To use uniformization, we first have to specify

the transition rates from service completions at the Interview and Donation stations. These are given by:

$$q_{i,j,t_k} = \begin{cases} \min\{s_2, n_{21}\} \cdot \mu_2 & j = i - e_2 \\ \min\{\max\{s_2 - n_{21}, 0\}, n_{22}\} \cdot \mu_2 & j = i - e_3 + e_5, n_3 < N_3 \\ \min\{\max\{s_2 - n_{21} - n_{22}, 0\}, n_{23}\} \cdot \mu_2 & j = i - e_4 + e_6, n_3 < N_3 \\ \min\{s_3, n_{31}\} \cdot \mu_3 & j = i - e_5 \\ \min\{\max\{s_3 - n_{31}, 0\}, n_{32}\} \cdot \mu_3 & j = i - e_6 \\ 0 & \text{else,} \end{cases} \tag{7.6}$$

where $n_3 = n_{31} + n_{32}$ and e_i is a vector with all zeros except for a one in position i. These transition rates form the generator matrix, total rates:

$$q_{i,t_k} = \sum_{j \in S} q_{i,j,t_k}. \tag{7.7}$$

Let the maximum of the total rates be $\alpha_{t_k} = \max_{i \in S} q_{i,t_k}$. The one step transition matrix for the services at the second and third station is given by:

$$P^{(\text{s23})}_{i,j,t_k} = \begin{cases} q_{i,j,t_k}/\alpha_{t_k} & j \neq i \\ 1 - q_{i,t_k}/\alpha_{t_k} & j = i. \end{cases} \tag{7.8}$$

With these definitions, the transitions can be computed similar to Section 2.4.2 from Chapter 2. To this end, initialize the queueing distribution $\pi^{(l)}$ after exactly l transitions. We then start rebuilding the intermediate queueing distribution $\pi^*_{t_k}$ by adding the queueing distribution after 0 transitions $\pi^{(0)}$ multiplied by the probability of 0 transitions.

$$\begin{aligned} \pi^{(0)} &= \pi^*_{t_k} \\ \pi^*_{t_k} &= \frac{((t_{k+1} - t_k)\alpha_{t_k})^0}{0!} e^{-(t_{k+1} - t_k)\alpha_{t_k}} \pi^{(0)}. \end{aligned} \tag{7.9}$$

We may now iteratively compute the queueing distribution after $l = 1, 2, \ldots$ transitions and update the intermediate distribution $\pi^*_{t_k}$ by:

$$\pi^{(l)} = \pi^{(l-1)} P^{(\text{s23})}_{t_k} \tag{7.10}$$

$$\pi^*_{t_k} := \pi^*_{t_k} + \frac{((t_{k+1} - t_k)\alpha_{t_k})^l}{l!} e^{-(t_{k+1} - t_k)\alpha_{t_k}} \pi^{(l)}. \tag{7.11}$$

Note that := indicates a variable gets updated. In equation (7.11), = would only hold if everything after the plus sign equals zero. This iterative computation is exact for $l \to \infty$, as explained in Chapter 2. However, as this is numerically infeasible, the iteration will be terminated if

$$1 - \sum_{l'=0}^{l} \frac{((t_{k+1} - t_k) * \alpha_{t_k})^{l'}}{l'!} e^{(t_{k+1}-t_k)*\alpha_{t_k}} < 10^{-6}. \tag{7.12}$$

7.3.2.3 Walk in arrivals

The third transition type to occur are the walk-in donors. For the walk-in donors, an exponential inter-arrival time will be assumed. The number of arrivals in a fixed time period will therefore be Poisson distributed, resulting in the following transition matrix:

$$P_{i,j}^{(\mathsf{aw})} = \begin{cases} \frac{(\lambda t)^l}{l!} e^{-(\lambda t)} + \\ \qquad \mathbb{1}_{\{l=N_1-n_1\}} \sum_{l'=l+1}^{\infty} \frac{(\lambda t)^{l'}}{l'!} e^{-(\lambda t)} & j = i + l \cdot e_1, n_1 + l \leq N_1 \\ 0 & \text{else,} \end{cases} \tag{7.13}$$

where $t = t_{k+1} - t_k$. Here, we recall that i and j represent state vectors from expression (7.2). As the maximum number of present donors at the registration station is limited by N_1, the tail of the Poisson distribution is added to the maximum number of allowed arrivals, assuming the remainder of the possible arrivals to be blocked. Then, $\pi^*_{t_k}$ is updated by:

$$\pi^*_{t_k} := \pi^*_{t_k} P^{(\mathsf{aw})}. \tag{7.14}$$

7.3.2.4 Services at station 1

The final transition to take place is exactly one service at the Registration station, if at least one donor is present here. As stated, the length of a time-interval is based on the service time at this station, and the service time at the first station is assumed to be deterministic. The transition matrix for the services at the first station is given

by:

$$P^{(\mathsf{s1})}_{i,j} = \begin{cases} \sigma^{(\mathsf{p})}_{t_k,1}/(1-\pi^{1*}_{t_k,0}) & j = i - e_1 + e_2, n_2 < N_2, n_1 > 0 \\ \sigma^{(\mathsf{wb})}_{t_k,1}/(1-\pi^{1*}_{t_k,0}) & j = i - e_1 + e_3, n_2 < N_2, n_1 > 0 \\ 1 - (\sigma^{(\mathsf{wb})}_{t_k,1} + \sigma^{(\mathsf{p})}_{t_k,1})/(1-\pi^{1*}_{t_k,0}) & j = i - e_1 + e_4, n_2 < N_2, n_1 > 0 \\ 1 & j = i, n_2 < N_2, n_1 = 0 \\ 1 & j = i, n_2 = N_2 \\ 0 & \text{else.} \end{cases} \tag{7.15}$$

where $n_2 = n_{21}+n_{22}+n_{23}$Since $\sigma^{(\mathsf{p})}_{n_1,t_k}$ and $\sigma^{(\mathsf{wb})}_{n_1,t_k}$ contain the unconditional probabilities that a plasma or whole blood appointment donor respectively stands at position n_1 in the queue at the registration station, we need to divide by $(1-\pi^{1*}_{t_k,0})$ to get the probability that a plasma or whole blood donor stands at position n_1, given that at least one donor is in this queue.

The queueing distribution after all transitions, the distribution at time t_{k+1} can now be computed by:

$$\pi_{t_{k+1}} := \pi^*_{t_k} P^{(\mathsf{s1})}. \tag{7.16}$$

After this transition, update the probabilities of appointment donors being in position n_1 in the queue, for $n_1 = 1, 2, \ldots, N_1$,

$$\begin{aligned} \sigma^{(\mathsf{wb})}_{n_1,t_{k+1}} &= \sigma^{(\mathsf{wb})}_{n_1+1,t_k}(1-\pi^{2*}_{t_k,N_2}) + \sigma^{(\mathsf{wb})}_{n_1,t_k}(\pi^{2*}_{t_k,N_2}) \\ \sigma^{(\mathsf{p})}_{n_1,t_{k+1}} &= \sigma^{(\mathsf{p})}_{n_1+1,t_k}(1-\pi^{2*}_{t_k,N_2}) + \sigma^{(\mathsf{p})}_{n_1,t_k}(\pi^{2*}_{t_k,N_2}). \end{aligned} \tag{7.17}$$

7.3.2.5 Combining the transitions

The computations outlined in Sections 7.3.2.1 until 7.3.2.4 are applied in this sequence to get from π_{t_k} to $\pi_{t_{k+1}}$. The time between these two distributions is exactly equal to the constant, deterministic service time at the Registration station. This is required to be able to track which donor is in which position in the queue for the Registration station. The assumption of a constant service time at the Registration station is not very unrealistic, as the registration consists of a few actions that do not contain a lot of stochasticity.

Summarizing, the queueing distribution $\pi_{t_{k+1}}$ at the start of time step t_{k+1}, is computed by:

$$\pi_{t_{k+1}} = \pi_{t_k} \prod_{m=1}^{a^{\mathsf{p}}_{t_k}+a^{\mathsf{wb}}_{t_k}} [P^{\mathsf{aa}}]\, P^{\mathsf{s23*}} P^{\mathsf{aa}} P^{\mathsf{aa}}, \tag{7.18}$$

where

$$P^{\text{s23*}} = \sum_{l=0}^{\infty} \frac{((t_{k+1} - t_k)\alpha_{t_k})^l}{l!} e^{-(t_{k+1}-t_k)\alpha_{t_k}} P^{\text{s23}l}. \tag{7.19}$$

7.3.3 Finding a schedule through binary search

Section 7.3.2 describes how $\pi_{t_{k+1}}$ is determined from π_{t_k}. Now, we would like to find an optimal schedule for the whole blood appointments. Optimal, in this case, could be with respect to multiple criteria or goal functions. We are interested in delivering the best possible service to the donors, given the structure of the process and the restrictions on the capacity of the service stations. The goal functions $f(\pi_{t_k})$ used in this chapter will therefore be aimed at minimizing the queues at the collection site.

However, finding an optimal schedule without extreme computational cost does not seem possible. Although the described process is close to a process that could be optimized by a Markov Decision Process (MDP), like the problem solved in Chapter 6, this is not possible because of the dependency on $\sigma_{t_k}^{(\text{wb})}$. Since $\sigma_{n_1,t_k}^{(\text{wb})}$ is dependent on the decision that have to be made, which makes that the problem can not be solved backwards, as would be required for a time-dependent, finite horizon MDP. So, to find a good schedule, we have to use heuristics. The heuristic that will be used is based on binary search.

First, we will have to determine when appointments can be planned. The length of a time step is equal to the mean service time at the registration station. As this is most likely a very short, it is likely that appointments should not be planned every time step, but in a subset of the time-steps. For this purpose, introduce the set of time-steps where appointments can be planned ζ,

$$\zeta = \{t_k | \text{appointments can be planned at time } t_k\}.$$

Next, introduce a vector with thresholds L. Without loss of generality, assume the goal function, queue lengths in this chapter, should be minimized. In this case, the values in L should be be non-increasing. At every time step $t_k \in \zeta$, we plan as many appointments as there are thresholds above the current value of the goal function $f(\pi_{t_k})$. For a goal function that should be maximized, we plan as many appointments as there are thresholds lower than the current value of the goal function. E.g., if $f(\pi_{t_k}) < L_3$ and $f(\pi_{t_k}) \geq L_4$, we plan 3 appointments in time-slot t_k.

With this method, whole blood appointments can be planned for the entire day or collection session. Binary search can then be applied to find the appropriate levels of the thresholds to plan the required number of appointments. Note that this required number of appointments could also be a range in which the planned number of appointments should lie.

The following procedure was used as an implementation of binary search. After the iteration, with some vector L, it is possible that the required number of

appointments, is planned. In this case the procedure can immediately be stopped. Otherwise, the number of planned appointments is either too high, in which case we add some vector C, or the number of planned appointments is too low, in which case we subtract the same vector C.

$$\begin{aligned} \text{too many appointments:} \quad & L_{\text{new}} = L_{\text{old}} + C \\ \text{too few appointments:} \quad & L_{\text{new}} = L_{\text{old}} - C, \end{aligned} \tag{7.20}$$

where C should have the same non-increasing property as the threshold vector L. If after the iteration both an upper and a lower bound on the thresholds have been found, divide C by half. Now apply the new vector L to find a new schedule until a schedule has been found with the appropriate number of appointments.

Figure 7.2 shows an example of the binary search method. Figure 7.2a shows the initial values for the thresholds, and the resulting expected number of present donors, which will be used as the goal function. These thresholds do not plan enough appointments, so the thresholds are increased in Figure 7.2b, and then decreased twice in Figures 7.2c and 7.2d to reach the required number of appointments.

The initial values of C and L of course influence the schedule that is found at the end of the procedure. Although more and less appropriate values can be thought of, the authors do not believe 'correct' initial values exist. These depend on preferences on the required schedule, and some trial and error might be involved in finding the most appropriate values.

7.4 Results

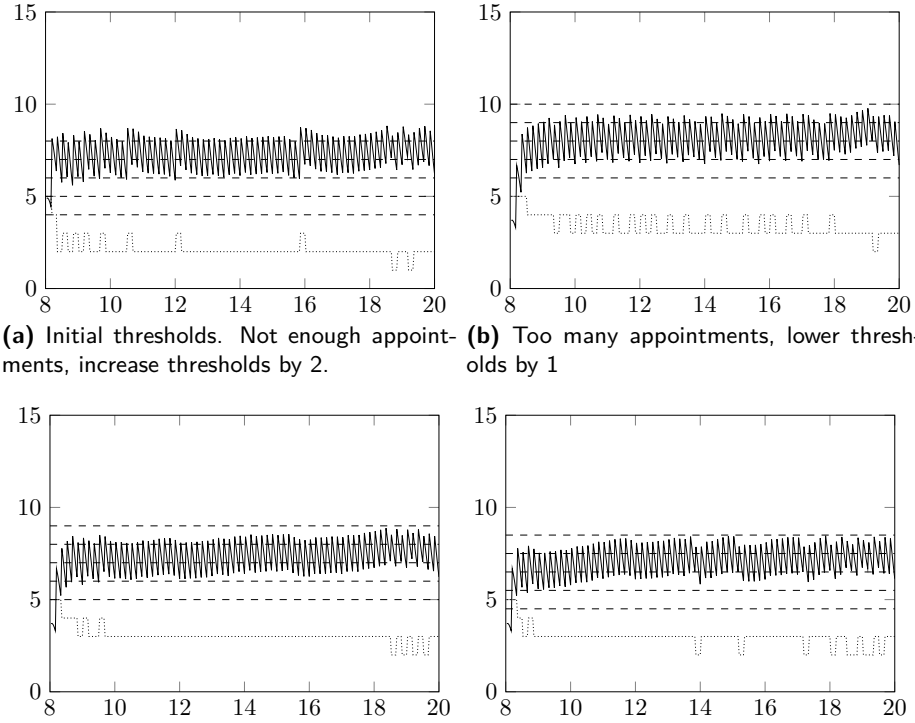
In this section, we will show some results of the proposed approach. We will first introduce our test case, based on the collection site in the Dutch city of Enschede, and show the application of the method to this collection site. Subsequently, we will also apply the method to some randomly generated arrival patterns.

All results in this Section are based on the model presented in Section 7.3.1. The total number of donors at the Registration station, Interview station and Donation station has been limited to 10 per station. This already results in 207,636 possible states for the system. The mean service time at the stations is 2 minutes, 6 minutes and 12 minutes respectively. All results were computed using Matlab 2015b, run on a laptop with an Intel Core i5-3437U CPU and 8GB of Ram.

7.4.1 Test case: Enschede

We were supplied with the arrival data from the collection site in the Dutch city of Enschede for 2015. This is a small fixed collection site. We have chosen to show the application of our approach with collection sessions lasting an entire day, from 8.00 AM until 8.00 PM. This session has the largest variation in arrivals, which makes it a case where improvements can be achieved, while also being a non-trivial case. The

Figure 7.2 Example for the binary search algorithm. thresholds are dashed, the number of planned appointments are dotted and the resulting expected number of donors, as computed by Section 7.3.2, is shown as the solid line.



(a) Initial thresholds. Not enough appointments, increase thresholds by 2.

(b) Too many appointments, lower thresholds by 1

(c) Too many appointments, lower thresholds by 0.5

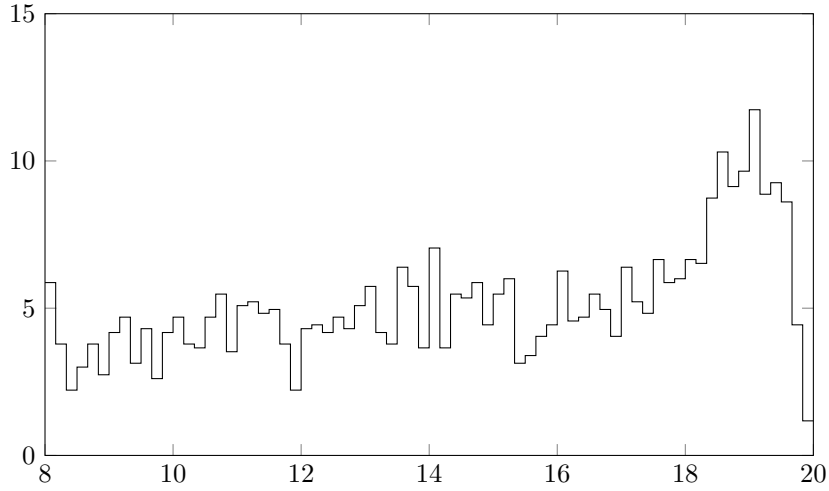
(d) Enough appointments. Binary search can be stopped

collection site in Enschede had one of these sessions per week, on Mondays. The average arrivals for all Monday sessions in Enschede in 2015 is shown in Figure 7.3.

The utilization at the collection site is relatively low. With 6 staff members, the current assignment including plasma donotations at the collection site, the average utilization throughout the day at the busiest station is 0.56, with a maximum utilization of 0.89 in the evening. To also show the application of the approach for larger centers and higher occupancies, we have created three main test cases. The first is based on the current collection site in Enschede, with 6 staff members. Additionally, we have also multiplied the arrivals by a factor of 2.5 (i.e. an increase of 150%) and 3.5 (i.e. an increase of 250%) for the second and third test case respectively. The second and third test case will have 9 staff members. These two test cases are included to replicate the situation at larger collection sites.

Appointments for plasma donors in Enschede, similarly to the rest of the Netherlands, can be planned every 10 minutes. Currently, on average 0.67 plasma appointments arrive during every 10 minute interval. Combined with the knowledge of a

Figure 7.3 Arrival pattern for Enschede. Average number of arriving whole blood donors, in number of donors per hour. The average number of arriving donors is shown per 10 minute interval, based on data from 2015.



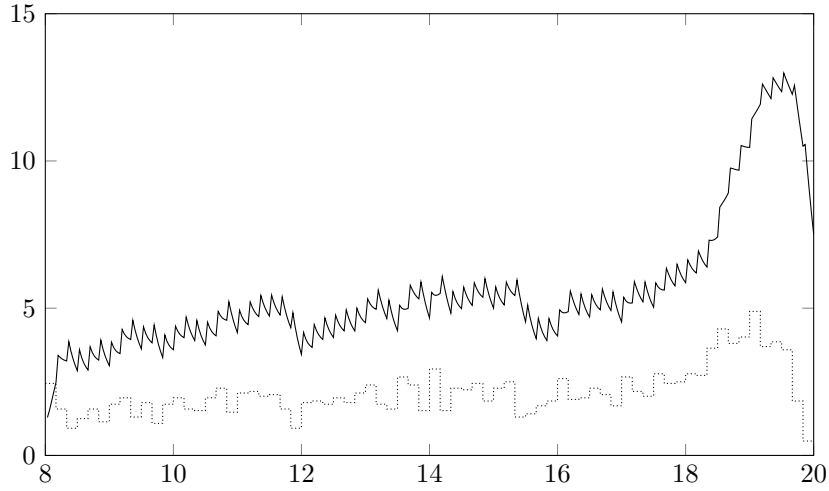
current no-show probability of approximately 20%, we have chosen to include one plasma appointment during every 10 minute interval for all scenarios. For the new whole blood appointments the same 10 minutes interval will be used. The same 20% no-show probability will also be used. A better estimation than the no-show probability for plasma donors is not available, as the whole blood appointments are not introduced yet.

We will start by highlighting one test case. We have chosen to include the test case with 150% increased arrivals, as the utilization in the base test case is too low to show clear effects. Figure 7.4 shows the situation without whole blood appointments and with 1 plasma appointment per 10 minutes. The dashed line shows the arrival rate in arriving donors per 10 minutes - Note that this is different from Figure 7.3, which shows the arrivals in arriving donors per hour. The solid line shows the average total number of donors present, computed by the approach from Section 7.3.2. The average total number of donors present is the average number of donors present, summed over all six (sub)stations.

Before looking at the results in more detail, we would first like to make a note about small variations visible in Figure 7.4. At the beginning of a 10 minute interval, the average number of donors present goes up, to then slowly decrease untill the start of the next interval. This zigzag behaviour is caused by the arrival of appoinment donors - plasma donors and in subsequent figures also whole blood donors, whom we assumed to always arrive exactly on time. In reality, this behaviour would most likely be drowned out as donors are early or late by at least a few minutes.

Figure 7.4 also shows the effect we want to counteract with the proposed approach. During prolonged high arrival rates, in this case during the evening, we can clearly see that the total number of donors in the collection site grows strongly, and

Figure 7.4 Expected number of donors at the collection site (solid line) and the arrival pattern in arrivals per 10 minutes (dotted line) for the collection site in Enschede without whole blood appointments. 150% extra arrivals test case.



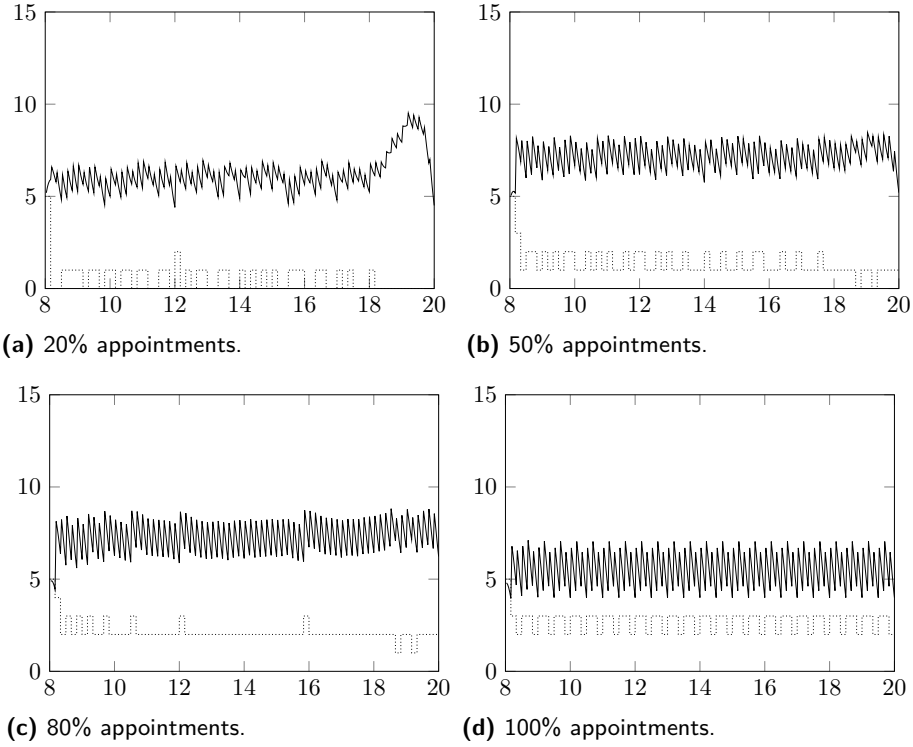
in sharp contrast to the remainder of the day. By introducing appointments arrivals should be spread out over the day. As a result, peaks could be decreased or even removed.

Figure 7.5 shows the effects of introducing appointments. For every Figure, a different percentage of whole blood arrivals is transferred from walk-in to appointments, varying from 20% to 100%. We assumed that all arrival rates are reduced by the applicable percentage, thereby retaining the same arrival pattern for whole blood donors. The number of appointments that has to be planned has been determined by first taking the expected reduction in walk-in arrivals. This is then divided by the no-show percentage, to take into account that not every appointment donor will show up. The approach then starts searching for the thresholds described in Section 7.3.3 that plan this number of appointments, plus or minus two appointments.

If 20% of arrivals are transferred to appointments, the peak during the evening is significantly reduced, but still visible. The graphs for the expected number of present donors with 50% or 80% show very little difference. If the number of arrivals that is transferred is increased to 100%, the expected number of present donors is significantly reduced. This can easily be explained by the fact that most queues are caused by variability in the system, and appointment arrivals remove a lot of the variability in the arrival process.

To be able to combine results from far more cases we will show 7 main performance metrics in the remaining results. The first three performance metrics deal with the number of donors in the collection site: the expected number of donors present, the expected number of queued donors and the difference between the lowest number of present donors and the highest number of present donors throughout the day. These three performance metrics are shown in Table 7.1. This last per-

Figure 7.5 Expected number of donors at the collection site (solid line) and the planned number of appointments (dotted line) for the collection site in Enschede for indicated percentage of whole blood arrivals with an appointment. 150% extra arrivals test case.



(a) 20% appointments. **(b)** 50% appointments.

(c) 80% appointments. **(d)** 100% appointments.

formance metric is included to give an indication of fluctuations and possible peaks present in the system. As the earlier described zigzag behavior will be present in all of our results, we have created two subsets of measurements: the highs - just after the arrival of appointments, and the lows - just before the arrival of appointments. The collection site always starts empty, so the first few measurements will also not give an accurate representation of the results. We have therefore excluded the first hour after opening from all results. Summarizing: For every session, we have results for 11 hours, all containing 6 intervals: a total of 66 measurements per session per subset. The results in Table 7.1 are shown based on these to subsets, and based on all measurements in a session.

Additionally, we have tracked a few blocking probabilities. The first of these probabilities is the probability that an arrival is blocked due to N_1 present donors in the first queue, and is shown in Table 7.2. This is tracked to make sure all appointment scenarios, from 0% to 100% appointment arrivals are compared fairly, with the same expected number of donors going through the collection site. We have also tracked the probability that a type of donor is being blocked from service, i.e. that all staff members are working for higher priority classes at the station in

Table 7.1 Expected number of present donors, queued donors and the difference between the highest and lowest number of present donors throughout the day respectively. For all three performance indicators, the average has been taken of all low measurements (just before arrival of appointments), all measurements, and all high measurements (just after the arrival of appointments).

		Expected present			Expected queued			Maximum difference		
		lows	all	highs	lows	all	highs	lows	all	highs
Appointments	0%	5.47	5.76	6.12	0.91	1.00	1.09	9.03	9.66	9.14
	20%	5.78	6.25	6.78	0.87	1.03	1.16	4.45	5.12	3.60
	50%	6.44	7.18	7.97	1.04	1.34	1.58	2.52	3.26	1.03
	80%	6.33	7.32	8.33	0.92	1.33	1.76	1.37	3.04	0.84
	100%	4.33	5.53	6.73	0.23	0.63	1.31	0.62	3.06	0.59

question. These probabilities are shown in Table 7.3, and are included to track if all priority classes receive enough service. For all of these probabilities, for figures will be shown: the average blocking probability over all arrival scenarios[1] and time-intervals, the average of the maximum blocking probabilities in every arrival scenario and the maximum blocking probability over all time-intervals and arrival scenarios. All of these 7 performance metrics can be directly computed from the results computed by Section 7.3.2.

Table 7.2 Probability of blocking an arrival for the arrival pattern of the collection site in Enschede with 150% increased arrivals. Columns indicate respectively: average probability, average of maximum probability per arrival scenario, maximum probability over all arrival scenarios.

		Avg	Avg Max	Max
Appointments	0%	0.000	0.002	0.002
	20%	0.000	0.000	0.000
	50%	0.000	0.000	0.000
	80%	0.000	0.000	0.000
	100%	0.000	0.000	0.000

It is important to note that the probability that arrivals to the Registration station are blocked due to capacity limit N_1 is negligible. The probabilities is Table 7.3 do give an indication of a potential problem. Although the introduction of appointments is able to decrease the peak in the number of present donors, as is visible in Table 7.1, the probability that donors without an appointment will have to wait does increase significantly. Especially if 50% or more of the whole blood arrivals are transferred to appointments, the probability that none of the whole blood donors without an appointment is receiving service at the Interview station is high. The other two probabilities shown in Table 7.3 will most likely not cause issues at the collection site.

[1] This section only considers one arrival scenario. However, in Section 7.4.3, multiple arrival scenarios will be considered. Therefore, the tables shown already allow for the analysis of multiple arrival scenarios, even though it is not relevant for this section.

Table 7.3 Probability that all staff members at the indicated station are occupied by donors with a higher priority class than the one indicated, for the arrival pattern at the collection site of Enschede, with arrivals increased by 150%. For all three combinations of station and priority class, the following columns are included: average probability, average of maximum probability per arrival scenario, maximum probability over all arrival scenarios.

		Station 2 with appointment			Station 2 without appointment			Station 3 without appointment		
		Avg	Avg Max	Max	Avg	Avg Max	Max	Avg	Avg Max	Max
Appointments	0%	0.005	0.025	0.025	0.005	0.025	0.025	0.000	0.000	0.000
	20%	0.007	0.031	0.031	0.072	0.556	0.556	0.000	0.005	0.005
	50%	0.010	0.029	0.029	0.215	0.564	0.564	0.006	0.048	0.048
	80%	0.011	0.026	0.026	0.369	0.712	0.712	0.032	0.114	0.114
	100%	0.012	0.025	0.025	-	-	-	-	-	-

7.4.2 Generating random arrival patterns

Besides the results based on the collection site in Enschede, random arrival patterns were generated to test the approach. Like the arrival patterns previously discussed and shown in Figure 7.3, these patterns are 12 hours long, with a different arrival rate every 10 minutes. The easiest way to generate these new arrival patterns would be to simply estimate a normal distribution based on the 72 arrival rates of the arrival pattern of Enschede. New arrival patterns could then be generated by drawing 72 arrival rates from this estimated distribution. However, this would result in a arrival pattern were subsequent arrival rates are unconnected, and the probability for prolonged periods with high or low arrival rates is low. If high arrival rates only last a few intervals, the system can always recover from high arrival rates quickly, and a problem with large queues never appears. This makes the problem fairly easy.

Therefore, we have decided to generate the arrival patterns based on a random walk, with a drift towards the average arrival rate. To do this, we have first computed the average arrival rate $\bar{\lambda}$, and the standard deviation of the difference between subsequent arrival rates σ_λ. The service time at the registration station, by definition the time length of a time-step, is 2 minutes. The arrival rate - changing every 10 minutes - therefore changes every 5 time-steps. If the arrival rate for a time step λ_{t_k} is known, the subsequent arrival rate is drawn from a normal distribution by:

$$\lambda_{t_{k+5}} \sim N(\nu\bar{\lambda} + (1-\nu)\lambda_{t_k}, \sigma_\lambda^2) \tag{7.21}$$

The average of the normal distribution is a weighted average between the current arrival rate and the average arrival rate. For $\nu = 0$, the arrival pattern is a random walk without a drift, and for $\nu = 1$, the subsequent arrival rates are independent. A drift has to be included, as a pure random walk could result in very high arrival rates, resulting in extreme queue lengths. These cases, with occupancies much higher than 1, are not interesting, as no substantial improvement can be achieved without increasing capacity. After some trials with different values of ν, $\nu = 0.2$ has been used for the results in Section 7.4.3. If a value lower than 0 is drawn, $\lambda_{t_k} = 0$ is used

instead, as negative arrival rates are not possible.

7.4.3 Results from random arrival patterns

For the results in this section, 100 arrival scenarios are generated using the method described in Section 7.4.2. The arrival scenarios are then multiplied by 2.5 and 3.5 to create a total of 300 arrival scenarios, similar to the three test cases based on the collection site in Enschede. Appointments will be planned based on all of the arrival scenarios. These will be combined with the arrival scenario of Enschede for a total of 101 arrival scenarios per arrival rate level.

Table 7.4 Summary of the generated arrival patterns, compared to the base test case of Enschede. St. dev. = standard deviation.

	Base arrivals	150% extra arrivals	250% extra arrivals
Staff members Registration	1	1	1
Staff members Interview	2	3	3
Staff members Donation	3	5	5
Average utilization (Enschede)	0.558	0.630	0.804
St. dev. utilization (Enschede)	0.097	0.162	0.234
Maximum utilization (Enschede)	0.887	1.178	1.643
Average utilization (all)	0.560	0.634	0.808
Average st. dev. utilization (all)	0.115	0.191	0.272
Maximum st. dev. utilization (all)	0.163	0.273	0.397
Average maximum utilization (all)	0.834	1.093	1.493
Maximum utilization (all)	1.031	1.463	2.048

Table 7.4 shows some summarizing numbers of the arrival scenarios that were generated. All of the figures are based on the utilization. However, all parameters influencing the utilization except for the arrival rate are constant within an arrival rate level, so these values also give a direct indication of the generated arrival rates. Note that all values are based on the scenario without appointments. The standard deviations in Table 7.4 show that the generated arrival patterns have a bit more fluctuation than the Enschede case, but the average utilization of all scenarios is almost identical to the Enschede case.

7.4.3.1 Base level arrivals

The average blocking probabilities of an arrival shown in Table 7.5 are negligible, so all appointment scenarios had to deal with approximately the same expected number of arriving donors, the only variation coming from the small range allowed for the number of appointments that had to be scheduled.

As can be expected with a very low utilization, the average expected number of donors present and average expected number of donors in queue shown in Table 7.6 are very low. Depending on the measurement, the numbers of donors is the highest

Table 7.5 Probability of blocking an arrival, as an average over the 101 arrival scenarios without increased arrivals (100 random arrival patterns + Enschede). Columns indicate respectively: average probability, average of maximum probability per arrival scenario, maximum probability over all arrival scenarios.

		Avg	Avg Max	Max
Appointments	0%	0.000	0.000	0.000
	20%	0.000	0.000	0.000
	50%	0.000	0.000	0.000
	80%	0.000	0.000	0.000
	100%	0.000	0.000	0.000

Table 7.6 Expected number of present donors, queued donors and the difference between the highest and lowest number of present donors throughout the day respectively, all as an average over the 101 arrival scenarios, without increased arrivals. For all three performance indicators, the average has been taken of all low measurements (just before arrival of appointments), all measurements, and all high measurements (just after the arrival of appointments).

		Expected present			Expected queued			Maximum difference		
		lows	all	highs	lows	all	highs	lows	all	highs
Appointments	0%	2.31	2.61	3.02	0.11	0.14	0.17	3.15	3.82	3.06
	20%	2.39	2.76	3.23	0.10	0.14	0.18	1.72	2.41	1.31
	50%	2.55	3.03	3.59	0.11	0.18	0.22	1.20	2.04	0.84
	80%	2.43	3.02	3.68	0.08	0.15	0.18	0.86	2.07	0.80
	100%	1.91	2.59	3.31	0.02	0.07	0.11	0.57	1.99	0.63

for the appointment scenario with 50% or 80% appointments, but the differences are miniscule. The maximum difference, also shown in Table 7.6, shows a clear reduction going from 0% to 100% appointments, but all average differences are small enough not to be problematic.

The blocking probability of a donor with appointment at the Interview station, as shown in Table 7.7 is very small. It does show an increasing trend in the appointment

Table 7.7 Probability that all staff members at the indicated station are occupied by donors with a higher priority class than the one indicated, as an average over the 101 arrival scenarios without increased arrivals. For all three combinations of station and priority class, the following columns are included: average probability, average of maximum probability per arrival scenario, maximum probability over all arrival scenarios.

		Station 2 with appointment			Station 2 without appointment			Station 3 without appointment		
		Avg	Avg Max	Max	Avg	Avg Max	Max	Avg	Avg Max	Max
Appointments	0%	0.062	0.147	0.149	0.062	0.147	0.149	0.000	0.000	0.000
	20%	0.068	0.179	0.184	0.157	0.692	0.841	0.002	0.027	0.083
	50%	0.079	0.184	0.192	0.284	0.761	0.841	0.011	0.055	0.142
	80%	0.090	0.188	0.201	0.391	0.774	0.887	0.038	0.080	0.208
	100%	0.095	0.192	0.207	-	-	-	-	-	-

scenarios that can partly be explained by increasing number of present donors, but can not be completely explained by intuition. The trend is not clearly visible with 150% and 250% extra arrivals, and might therefore be a coincidence. The blocking probability of a donor without an appointment in Table 7.7 is, even for the low utilization in this arrival rate level, rising to problematic levels, especially for the 50% and 80% appointment scenarios. The blocking probability at the Donation station is also increasing, but remains acceptable. The increasing trend of both can simply be explained by the fact that the number of donors with appointment, whom have priority, is increasing.

7.4.3.2 150% extra arrivals

The arrival blocking probabilities in Table 7.8 are decreasing with the introduction of appointments. This makes sense as the method distributes arrivals over the day, and thereby decreases peaks that can cause arrivals to be blocked. However, the probabilities are still negligible and can safely be ignored with the analysis of the remaining results.

Table 7.8 Probability of blocking an arrival, as an average over the 101 arrival scenarios with arrivals increased by 150%. Columns indicate respectively: average probability, average of maximum probability per arrival scenario, maximum probability over all arrival scenarios.

		Avg	Avg Max	Max
Appointments	0%	0.000	0.002	0.013
	20%	0.000	0.000	0.003
	50%	0.000	0.000	0.001
	80%	0.000	0.000	0.001
	100%	0.000	0.000	0.000

Table 7.9 Expected number of present donors, queued donors and the difference between the highest and lowest number of present donors throughout the day respectively, all as an average over the 101 arrival scenarios, with arrivals increased by 150%. For all three performance indicators, the average has been taken of all low measurements (just before arrival of appointments), all measurements, and all high measurements (just after the arrival of appointments).

		Expected present			Expected queued			Maximum difference		
		lows	all	highs	lows	all	highs	lows	all	highs
Appointments	0%	5.61	5.91	6.30	1.09	1.19	1.30	9.11	9.68	8.87
	20%	5.76	6.24	6.79	0.91	1.07	1.22	4.11	4.74	2.99
	50%	6.36	7.11	7.90	1.07	1.36	1.63	2.13	3.13	1.12
	80%	6.47	7.48	8.51	1.05	1.48	1.95	1.36	3.18	0.91
	100%	4.40	5.60	6.81	0.29	0.70	1.38	0.58	3.00	0.58

The expected number of present donors in Table 7.9 is again highest for the 50% and 80% appointment scenarios, but differences are small. The number of queued donors increases from 0% to 80% appointments. It is however, intresting

Table 7.10 Probability that all staff members at the indicated station are occupied by donors with a higher priority class than the one indicated, as an average over the 101 arrival scenarios with arrivals increased by 150%. For all three combinations of station and priority class, the following columns are included: average probability, average of maximum probability per arrival scenario, maximum probability over all arrival scenarios.

		Station 2 with appointment			Station 2 without appointment			Station 3 without appointment		
		Avg	Avg Max	Max	Avg	Avg Max	Max	Avg	Avg Max	Max
Appointments	0%	0.006	0.022	0.040	0.006	0.022	0.040	0.000	0.000	0.000
	20%	0.008	0.026	0.035	0.080	0.544	0.589	0.001	0.010	0.053
	50%	0.010	0.029	0.034	0.215	0.581	0.710	0.007	0.044	0.154
	80%	0.011	0.027	0.035	0.377	0.707	0.830	0.039	0.112	0.315
	100%	0.012	0.025	0.028	-	-	-	-	-	-

to note that both the expected number of present and the expected number of queued donors is lowest for a scenario with 100% appointments. The main goal of introducing the appointments, decreasing peaks in the number of present donors, is clearly succesfull if we look at the maximum difference in Table 7.9. The relatively high difference if all measurements are considered is caused by simultanious arrival of multiple appointments, causing the number of present donors to immediately go up.

Although a high percentage of appointments seems a good idea if only Table 7.9 is considered, Table 7.10 might very well weaken this position. The probability that a donor without an appointment is blocked at the Interview station is again high, with an average highest blocking probability of 0.707, and a scenario where the probability even increases to 0.830. The other two blocking probabilities shown in Table 7.10 are low enough not to cause a problem, and show not further interesting behavior.

7.4.3.3 250% extra arrivals

With 250% aditional arrivals, the blocking probabilities are still negligible, but do start to be visible. Table 7.11 does not show a clear pattern, and seems to be independent of the appointment scenario.

Table 7.11 Probability of blocking an arrival, as an average over the 101 arrival scenarios with arrivals increased by 250%. Columns indicate respectively: average probability, average of maximum probability per arrival scenario, maximum probability over all arrival scenarios.

		Avg	Avg Max	Max
Appointments	0%	0.002	0.014	0.036
	20%	0.001	0.005	0.023
	50%	0.000	0.001	0.009
	80%	0.001	0.004	0.035
	100%	0.000	0.002	0.019

Although the values for the average expected number of present donors, the average expected number of queued donors and the maximum differences are higher,

Table 7.12 Expected number of present donors, queued donors and the difference between the highest and lowest number of present donors throughout the day respectively, all as an average over the 101 arrival scenarios, with arrivals increased by 250%. For all three performance indicators, the average has been taken of all low measurements (just before arrival of appointments), all measurements, and all high measurements (just after the arrival of appointments).

		Expected present			Expected queued			Maximum difference		
		lows	all	highs	lows	all	highs	lows	all	highs
Appointments	0%	9.13	9.43	9.78	3.22	3.37	3.54	14.14	14.56	13.72
	20%	9.07	9.61	10.19	2.73	3.01	3.30	6.97	7.53	5.25
	50%	9.68	10.59	11.52	2.92	3.45	4.00	2.96	3.97	1.37
	80%	10.31	11.59	12.86	3.34	4.13	5.01	1.71	3.93	1.10
	100%	7.03	8.57	10.10	1.40	2.22	3.38	0.82	3.88	0.82

Table 7.13 Probability that all staff members at the indicated station are occupied by donors with a higher priority class than the one indicated, as an average over the 101 arrival scenarios with arrivals increased by 250%. For all three combinations of station and priority class, the following columns are included: average probability, average of maximum probability per arrival scenario, maximum probability over all arrival scenarios.

		Station 2 with appointment			Station 2 without appointment			Station 3 without appointment		
		Avg	Avg Max	Max	Avg	Avg Max	Max	Avg	Avg Max	Max
Appointments	0%	0.009	0.037	0.081	0.009	0.037	0.081	0.000	0.000	0.000
	20%	0.011	0.030	0.059	0.126	0.590	0.808	0.003	0.040	0.196
	50%	0.012	0.027	0.037	0.335	0.678	0.874	0.028	0.114	0.288
	80%	0.012	0.023	0.032	0.576	0.785	0.948	0.130	0.253	0.517
	100%	0.013	0.022	0.028	-	-	-	-	-	-

the same general trends can be seen in Table 7.12 that were visible in Table 7.9 for 150% extra arrivals. Increasing the percentage of appointments from 0% to 80% will increase the number of present and to a lesser extent queued donors, but the drastic decrease of the maximum difference seems to be the more important result.

The blocking probabilities in Table 7.13 also show similar trends to the 150% extra arrivals case. The blocking probability for a donor without an appointment at the Interview station is now even higher, as can be expected with the higher utilization. The blocking probability for the same donor without an appointment now also starts to be noticable at the Donation station.

7.5 Discussion

The proposed approach to plan appointments for whole blood donors at Sanquin blood collection sites is able to accomplish its goal. Even with just a small fraction of the arrivals having an appointment, the highest queues during the day can be reduced substantially. If 50% or more of the whole blood arrivals have an appointment, the

queues become almost equally spread and constant during the day. The average number of donors at the collection site does increase slightly, but this will most likely not cause any problems.

The probability that the donors without an appointment do not receive service will cause problems. Over all scenarios, the average maximum probability is almost 0.8 in the scenario with the highest utilization. On average, over half the time no donor without an appointment is receiving service at the Interview station if 80% of whole blood arrivals have an appointment, the percentage that Sanquin strives for.

The priority rules as currently proposed by Sanquin, might very well introduce problems if they are implemented in this way. The priority rules were intended to tempt donors to make an appointment. This will certainly work for a portion of the donors, but the donors without an appointment will get such a disadvantage that they might not be willing to remain blood donors. As losing a substantial part of the donors might be disastrous for Sanquin, careful consideration is required before implementing these priority rules.

On a more technical note, the queueing model in Section 7.3.1 seems to work quite well. Even for over 200,000 states, the distribution over all these states, for every 2 minutes during the collection session, can still be computed in approximately 30 seconds. Although some of the exponential assumptions and the preemptive priorities might not give a completely accurate representation of reality, similar models have performed well in Chapters 3 and 6.

Although using a binary search approach for the scheduling of appointments might seem fairly simple, it does perform very well. However, no guarantees for optimality can be made. The approach can easily be extended by some local search that shifts an appointment slot to another time. We have chosen not to include this for two reasons. First of all, the schedules resulting directly from the binary search initial schedule already accomplish the goal of this research. Secondarily, although 30 seconds per iteration is no problem for the binary search approach, computing the queueing distribution for hundreds or thousands of alternative schedules would take too long for practical purposes.

Summarizing, the proposed approach to schedule appointments appears to be very effective at reducing peak loads on the system. However, the priority rules suggested by Sanquin require serious reconsideration. These might cause donors without an appointment to experience excessive waiting times. A possible solution would be to create one priority class for all whole blood donors, without a distinction for donors with an appointment. Although this will have a significant influence on the blocking probabilities, it is unlikely that the results with respect to total number of donors present and queue lengths will change.

CHAPTER 8

Blood type specific issuing policies to improve inventory management

8.1 Introduction

Hospitals in the Netherlands use approximately 420,000 red blood cell (RBC) units every year [163]. These units are donated by voluntary non-remunerated donors at collection sites of the Dutch blood supplier, Sanquin, throughout the country. After collection, the donated blood units are processed, typed for more than fifteen different antigens, tested for several infectious diseases and finally stored in one of the distribution centers until they are requested by hospitals. Although describing the process as such makes it sound trivial, several complications arise, making inventory management of RBC units an important and interesting topic for research.

First of all, RBC units are used during major surgeries or as a treatment for leukemia, anemia, and blood disorders. Not being able to satisfy requests from hospitals comes at a very high cost, since this may lead to delays and therefore poses transfusion recipients at risk. Hence, an adequate and timely availability of RBC units is essential.

Second, RBCs are perishable products. After 35 days of storage, the unit has to be discarded. In the inventory management for perishable products, always a balance has to be found between the probability of outdating and the probability of shortage, as decreasing one usually increases the other. RBCs are produced from donations by voluntary, non-remunerated donors. Donors are substantially motivated by the fact that their blood donation is necessary and saves lives. Increasing outdating would affect the motivation by donors and, additionally, is ethically undesirable. So, for RBCs and other blood products, both outdating and shortages should be minimized at the least, and preferably totally prevented.

Minimizing outdating and shortages simultaneously is a challenge by itself, but different blood-types pose another challenge. The ABO blood-types are relatively known, but many more blood-types exist. A request for a unit of RBCs is always accompanied by a requested blood type. This consists of the antigens for which the bag of blood should be typed negative. Such a request can be fulfilled by any unit that is negative for at least these antigens. If a unit is not tested for certain antigens, it has to be regarded as positive for these antigens. Although compatible issuing can

be done, identical issuing is preferred, as compatible issuing can cause a shortages of rare blood types. If the recipient is positive for some antigen, it does not matter if the RBC unit is positive or negative for this antigen. However, the reverse is not true. If the recipient is negative, a positive unit can not be used for this recipient. Besides minimizing outdating and shortages, we will also take the rarity of an issued unit of RBCs into account, and try to increase the percentage of exact matches. To be clear, in this paper we only focus on the issuing of RBC units stored in the general RBC inventory. For extremely rare blood types there exists a separate frozen inventory.

Finally, inventory management of blood products has to deal with stochasticicy. Most inventory management systems deal with stochastic demand. However, the supply side of blood inventory also contains stochasticity. As donors are donating voluntarily and are non-remunerated, the probability of a no-show for a donation is substantial. Additionally, as Sanquin does not use appointments for blood donors, the lead-time between inviting a donor and the donor showing up is highly uncertain, ranging from a week up to a month.

These four complications make inventory management of RBCs non-standard an non-trivial. In this paper, we will present an integrated method for the complete inventory management of RBCs at Sanquin. The method shows the number of donors that should be invited to control the supply side of the inventory. Next, it uses a min-cost max-flow algorithm to determine which units should be issued to which request. Finally, by simulation, we investigate the successive application of the matching (ILP) solutions to combine supply and issuing in one approach. In this way, a much better performance with respect to shortage and outdating can be achieved.

Medical literature has not reached a consensus on the benefit of using blood blood that has been stored for a shorter time [46, 82, 137, 178]. From an inventory management perspective, using a FIFO (first in, first out) policy is preferable, so this will be the base for our model. We will, however, report the average storage time before a RBC unit was issued to a hospital.

Let us briefly sketch the outline of this paper. Section 8.2 provides an overview on the existing relevant literature on the inventory management of blood products in which we focus on issuing policies. Next, in Section 8.3 we introduce a mathematical model for the daily allocation of the RBC inventory. In this model we assume that the supply and demand of RBCs is deterministic. To incorporate the stochasticity in the demand and supply of RBCs we develop in section 8.4 a simulation model. We will then discuss the data that was used to run the model and the simulation experiments in Section 8.5, followed by Section 8.6, which discusses the results from the simulation. The paper will be concluded with a discussion of the method and results, and will indicate fertile directions for further research.

8.2 Literature review

The inventory management of perishable products is an important topic in Operations Research. [144], [145] gave a review of the available literature with respect to this topic. A part of this literature consists of the inventory management of blood products. The management of these specific products also garners much attention itself. [25] and [149] are two recent review papers. Beliën and Forcé [25] classified the papers according to the blood component under study: platelets, RBCs, plasma, whole blood, frozen blood, other/unclear. Whereas we are interested in the inventory management of perishable products we focus on the literature about platelets and RBCs. The most important difference with respect to the inventory management of these two blood components is their maximum shelf life. Platelets are considered to be expired after 7 days and RBCs after 35 days.

In Section 8.2.1 we will review the recent literature on the inventory management of platelets, and continue afterwards with the literature on the inventory management of RBC in Section 8.2.2. Finally, in Section 8.2.3 we will discuss the extension of our research on the existing literature.

8.2.1 Platelets

Whereas platelets expire after 7 days most papers about platelet inventory management consider the percentage of outdated units as main performance measure and apply a FIFO issuing policy [169], [143], [83]. Though, the inventory size should be sufficiently large to prevented or maintain shortages. One way to maintain shortages is by using a predetermined maximum shortage level [72]. Another way to maintain both outdating and shortages is by including both performance measures into the objective function [1], [51].

A FIFO issuing policy seemed to be the optimal issuing policy with respect to the inventory management of platelets. However, for some patient groups fresh platelets increase survival rates [104], [68] and therefore some papers use different issuing policies. One way to this is to make a difference between 'young' and 'old'/'any' platelets [50], [100]. Moreover, whereas Gunpinar and Centeno [100] looked at the inventory management of hospitals they incorporated a cross-match-to-transfusion ratio and the cross-match release period in their models. Civelek et al. [51] shortened the maximum shelf life of the platelets to three days and classified them as 'young', 'mature', and 'old'. Moreover, they included protection levels and substitution costs to limit the amount of 'young' platelets issued to satisfy requests where the age of the platelets did not matter.

With respect to replenishment policies Duan and Liao [72] considered an old inventory ratio policy to avoid shortages. This policy states that if the proportion of old units in stock exceeded a certain threshold, making it likely that some units get outdated, then some extra donors should be invited. A similarly approach was proposed by Haijema et al. [104], [105]. They investigated 1D and 2D order-up-to-rules, where in the 2D order-up-to-rule donors where invited based on both, the total amount of platelet units in stock and the amount of young platelet units in stock.

Recently, [156] analyzed several ordering policies for platelets based on hospital size and demand variation.

8.2.2 Red blood cells

Also for the inventory management of RBCs the percentage outdating and shortage remain important performance measures. Though, some papers suggest that the maximum shelf life of 35 days may be reduced without excessive increases of outdating or shortage rates Blake et al. [30], [73].

[69] developed a two stage inventory control model, where in the first stage decisions about the review period and order-up-to-levels are made and in the second stage decision about the daily operation of the system are considered. Moreover, they investigate the difference between an exact issuing policy and a compatible issuing policy with respect to ABO, RhD compatibility.

Atkinson et al. [19] apply a single threshold policy in which blood younger than the threshold is issued according to an FIFO-policy and blood older than the threshold is issued according to a LIFO policy. For a threshold of 14 days they show that the mean age of transfused blood decreases by 10 to 20 days.

8.2.3 Literature extension

So far, the existing literature on the inventory management of blood products only considers the eight common ABO-RhD blood groups. However, since the beginning of the 21th century, hospitals are extending their matching strategies for some patient categories [44]. As a result, transfusion recipients belonging to one of these patient categories are not only matched for the standard ABO-RhD blood groups, but also for some additional blood groups. The extensive matching strategies therefore lead to a growing demand for more specific blood types. To support real world decision making better, it is necessary to include these extended blood types into the mathematical models.

Clearly, by including more antigen both the diversity among the blood units in inventory and the diversity among the blood units requested increases. Hence, the likelihood of finding an exact match between a unit requested and a unit issued decreases. Therefore, we extend the existing performing measures such as outdating, shortage, and issuing age by a fourth performance measure: the quality of a match.

In the primary objective of the (mathematical) model that will be presented in this paper we aim to prioritize between the quality of a match and the age of an issued unit (FIFO). Whereas maximizing the quality of a match will correspond to saving rare units to prevent shortages and applying a FIFO policy corresponds to the prevention of outdating, these performance measures are not included in the objective of the mathematic model. However, the simulation that evaluates the quality of a match and age of issued units over time, also keeps track of these performance measures.

8.3 Daily inventory allocation problem

First, in order to achieve an optimal issuing policy, which prioritizes between the rarity and age of the RBC units in stock, we show that the daily inventory allocation problem can be modeled as a transportation problem. Second, to investigate the impact of these daily decisions in the long run (i.e., the composition of the inventory, the average issuing age of particular blood types, the number of outdated units, and the number of units short) a simulation study is conducted (see Section 8.4). This simulation study does not only evaluate the proposed issuing policy, but also incorporates the stochasticity in the supply and demand of red blood cells.

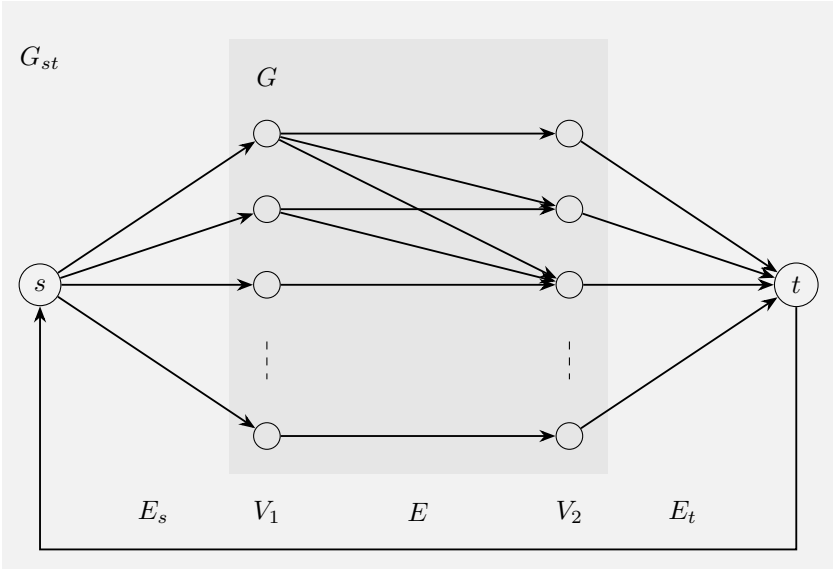
The inventory allocation problem is in essence equivalent to a transportation problem. Therefore, we will first mathematically describe transportation problem in Section 8.3.1 and demonstrate how it can be solved by reformulating it as a circulation flow problem (CFP). Next, in Section 8.3.2, we show how the vertex and edge set of the transportation problem should be modified to make the translation to blood inventory management. Finally, in Section 8.3.3, we will incorporate the age of blood units by extending the vertex set.

8.3.1 Circulation flow problem

Let $G = (V, E)$ be a directed bipartite graph with vertex set $V = V_1 \cup V_2$ ($V_1 \cap V_2 = \emptyset$) and edge set E, such that $e = \{i, j\} \in E$ is an edge from $i \in V_1$ to $j \in V_2$ (see Figure 8.1). Interpret the set V_1 as a set of sources and the set V_2 as a set of destinations. Each vertex $i \in V_1$ has a non-positive demand $d_i \leq 0$, which can be interpreted as the supply of source i. Similarly, each vertex $j \in V_2$ has a non-negative demand $d_j \geq 0$, which is the demand of destination j. An edge $e = \{i, j\}$ between vertices i and j implies that source i can be used to satisfy the demand of destination j. The maximum flow that can be transported over this edge is equal to the minimum of the amount supplied by i and the amount demanded by j. Therefore, we say that the capacity of edge e is equal to $u_e = \min\{-d_i, d_j\}$. Define the set of decision variables, $x_e \in \mathbb{N}_0$ ($e = \{i, j\}$), as the amount distributed from source i to destination j. The transportation cost for distributing one unit over edge e is given by the parameter c_e. So, the total distribution cost are equal to $\sum_{e \in E} c_e x_e$.

To find the maximum flow that can be transported over G we convert this graph to a flow graph $G_{st} = (V_{st}, E_{st})$, which can be used to solve a circulation flow problem. First, the vertex set of G is extended with a source node s and a sink node t. Second, two new edge sets are introduced: E_s and E_t. $E_s = \{\{s, i\} : i \in V_1\}$ consists of the edges from s to all vertices in V_1. The edges $e \in E_s$ have capacity $u_e = -d_i$ and cost $c_e = 0$. $E_t = \{\{j, t\} : j \in V_2\}$ consists of the edges from all vertices in V_2 to t. The edges $e \in E_t$ have capacity $u_e = d_j$ and cost $c_e = 0$. Third, an edge $\{t, s\}$ from the sink to the source is constructed. This edge has infinite capacity, but for computational convenience the capacity of this edge can be restricted by the total flow leaving s: $u_{\{t,s\}} = -\sum_{i \in V_1} d_i$. The cost of edge $\{t, s\}$ should obtain a value, such that every cycle $s - i - j - t - s$ ($i \in V_1$, $j \in V_2$) has negative cost. This implies that the cost of edge $\{t, s\}$ could be equal to $c_{\{t,s\}} = -\max_{e \in E}\{c_e\} - 1$. Hence, the

Figure 8.1 Flow graph $G_{st} = (V_{st}, E_{st})$ with vertex set $V_{st} = s \cup V_1 \cup V_2 \cup t$ and edge set $E_{st} = E_s \cup E \cup E_t \cup \{t, s\}$. The edges $e \in E_{st}$ have capacity u_e and costs c_e.



flow graph $G_{st} = (V_{st}, E_{st})$, depicted in Figure 8.1, has vertex set $V_{st} = \{s\} \cup V \cup \{t\}$ and edge set $E_{st} = E_s \cup E \cup E_t \cup \{t, s\}$.

To find a solution to the transportation problem the following LP-formulation of the circulation flow problem could be solved:

Circulation flow problem

$$\min \quad \sum_{e \in E_{st}} c_e x_e \tag{8.1}$$

$$\text{s.t.} \quad \sum_{e=\{i,j\}:\ j \in V_{st}} x_e - \sum_{e=\{j,i\}:\ j \in V_{st}} x_e = 0 \qquad \forall i \in V_{st} \tag{8.2}$$

$$0 \leq x_e \leq u_e \qquad \forall e \in E_{st} \tag{8.3}$$

The objective function (8.1) minimizes the total distribution costs. Constraints (8.2) are called the flow balance equations and state that the amount of flow that enters a vertex $i \in V_{st}$ should be equal to the amount of flow that leaves this vertex. Finally, constraints (8.3) restrict the total flow that can be send over an edge $e \in E_{st}$.

Finding a maximum flow with minimal cost for the transportation problem is equivalent to solving the corresponding circulation flow problem. The maximum flow is ensured, because of the negative cost cycles in the circulation flow problem. Let $\bar{x}$ be the optimal solution of the circulation problem with objective value $f(\bar{x})$. Then $x^* = \{x_e^* = \bar{x}_e \ : \ e \in E\}$ is the optimal solution of the transportation problem with objective value $f(x^*) = \sum_{e \in E} c_e x_e^*$.

8.3.2 Blood inventory model

Let $\mathcal{A}$ be the set of antigens taken into consideration ($|\mathcal{A}| = n$). Let $\mathcal{B}$ be the set of different blood types ($|\mathcal{B}| = m$), where a blood type is defined as a unique combination of antigens that are either present or absent on the surface of an RBC. An individual with blood type $i \in \mathcal{B}$ is called a-positive if antigens $a \in \mathcal{A}$ are present on the individuals RBCs and a-negative if the antigens are absent. For computational convenience we will represent a blood type $i \in \mathcal{B}$ by a binary vector $b(i)$, where b is a bijective function that maps $\mathcal{B}$ to $\{0,1\}^n$:

$$b_a = \begin{cases} 1 & \text{if an individual is } a\text{-positive} \\ 0 & \text{if an individual is } a\text{-negative} \end{cases}, \qquad \forall a \in \mathcal{A} \tag{8.4}$$

For the sake of simplicity we use in the remainder of the paper that $\mathcal{B} = \{0,1\}^n$, implying that $i \in \mathcal{B}$ is equal to $b(i) \in \{0,1\}^n$.

An important aspect of the issuance of blood units is the concept of compatibility. This concept states which blood types $i \in \mathcal{B}$ can be issued to fulfill a request for blood type j. If a hospital requests a blood unit they indicate for which antigens $a \in \mathcal{A}' \subseteq \mathcal{A}$ the issued unit should be negative, i.e. $i_a = 0$ for all $a \in \mathcal{A}'$. However, they do not indicate whether the other antigens should be positive or negative. Hence, i_a could be either 0 or 1 for all $a \in \mathcal{A} \backslash \mathcal{A}'$. Suppose we define j as follows: $j_a = 0$ for all $a \in \mathcal{A}'$ and $j_a = 1$ for all $a \in \mathcal{A} \backslash \mathcal{A}'$. Then, all blood types i that could be issued to fulfill a request for blood type j, satisfy the relation $i \leq j$. We say that blood type i is a compatible substitution for j if $i \leq j$. Therefore, we define the compatibility matrix $C \in \{0,1\}^{m \times m}$, which indicates whether blood type i is a compatible substitution for j, in the following way:

$$C_{ij} = \begin{cases} 1 & \text{if } i \leq j \\ 0 & \text{otherwise} \end{cases}, \qquad \forall i,j \in \mathcal{B} \tag{8.5}$$

Let V_1 and V_2, previously interpreted as the sets of sources and destinations, consist of all possible combinations of blood types, i.e. $V_1 = \{i \ : \ i \in \mathcal{B}\}$ and $V_2 = \{j \ : \ j \in \mathcal{B}\}$. The vertices $i \in V_1$ and $j \in V_2$ have demand $d_i \leq 0$ and $d_j \geq 0$ respectively. d_i represents the inventory level of blood type i in a blood distribution center and d_j shows the total demand for blood type j at the time the decision is taken. We say that there exists an edge $e = \{i,j\} \in E$ between $i \in V_1$ and $j \in V_2$ if blood type i is a compatible substitution for blood type j, i.e. $C_{ij} = 1$.

In practice, it turns out that the frequency at which different blood types occur in the inventory is not equal. This frequency depends on two factors, namely the distribution of antigens in the donor population and the fraction of donors typed for particular antigens. Suppose that coverage rate of a blood type $i \in \mathcal{B}$ is equal to p_i. Define the cumulative coverage rate as $\tilde{p}_i = \sum_{j \in \mathcal{B}} C_{ij} p_j$. Then the costs for using

blood type i to fulfill a request for blood type j equals:

$$c_e = 1 - \frac{\tilde{p}_i}{\tilde{p}_j} \qquad e = \{i, j\} \in E, i \in V_1, j \in V_2. \tag{8.6}$$

In this paragraph the daily allocation of a blood inventory is described as a transportation problem. The best allocation, with respect to the rarity of particular blood types, could be found by solving a circulation problem as described in Section 8.3.1. Though, it may happen that the inventory levels are not sufficient to meet all requests, leading to shortages for particular blood types. A shortage for blood type j is defined as the difference between the amount demand for this particular blood type and the amount issued:

$$s_j = d_j - \sum_{e \in E:\ e = \{i,j\}} x_e, \qquad \forall j \in \mathcal{B}. \tag{8.7}$$

8.3.3 Blood inventory model with residual shelf lives

A special feature of issuing policies for perishable product is that often old items have to be issued first to maintain outdating, a so-called first-in-first-out (FIFO). To add this feature to blood inventory model some adaptions have to be made. However, first a clear overview of the different sets, indices, parameters, and variables is given.

Sets
$\mathcal{A}$ – Set of antigens, $|\mathcal{A}| = n$
$\mathcal{B}$ – Set of blood types, $|\mathcal{B}| = 2^n = m$
$\mathcal{R}$ – Set of residual shelf lives, $|\mathcal{R}| = R_{\text{max}}$

Indices
a – Antigen, $a \in \mathcal{A}$
i^r – Blood type, $i \in \mathcal{B}$ with residual shelf life $r \in \mathcal{R}$ (in inventory)
j – Blood type, $j \in \mathcal{B}$ (demand)
r – Residual shelf life of an item on stock, $r \in \mathcal{R}$

Parameters
d_i^r – Number of units with a residual shelf life of $r \in \mathcal{R}$ days and blood type $i \in \mathcal{B}$ in stock
d_j – Number of units with blood type $j \in \mathcal{B}$ demanded
l^r – Residual shelf life of an an item in stock
s_j – Number of units with blood type $j \in \mathcal{B}$ short
o_i – Number of units with blood type $i \in \mathcal{B}$ outdated
C_{ij} – Binary compatibility matrix with $C_{ij} = 1$ if blood type $i \in \mathcal{B}$ can be substituted to satisfy demand for blood type $j \in \mathcal{B}$

Variables
x_{ij}^r – Quantity of blood type $i \in \mathcal{B}$ with residual shelf life $r \in \mathcal{R}$ that is used to satisfy the demand of blood type $j \in \mathcal{B}$

Adaptations made to the flow graph are:

- the vertex set V_1 is extended by adding to every blood type a residual shelf life r: $V_1^r = \{i^r \ : \ i \in V_1, r \in R\}$.
- the demand of each vertex $i^r \in V_1^r$ is equal to $d_i^r \leq 0$, such that $\sum_{r \in R} d_i^r = d_i$.
- the edge set $E_s^r = \{\{s, i^r\} \ : \ i^r \in V_1^r\}$ consists of the edges from s to all vertices in V_1^r, where an edge $e \in E_s^r$ has capacity $u_e = -d_i^r$ and cost $c_e = 0$.
- the edge set $E^r = \{\{i^r, j\} \ : \ i^r \in V_1^r, j \in V_2\}$, which implies that blood type i is a compatible substitution for blood type j. An edge $e \in E^r$ has capacity $u_e = \min\{-d_i^r, d_j\}$ and cost $c_e = 1 - \tilde{p}_i/\tilde{p}_j$.

In this way the flow graph G_{st} is transformed into the flow graph $G_{st}^r = (V_{st}^r, E_{st}^r)$ with vertex set $V_{st}^r = \{s\} \cup V_1^r \cup V_2 \cup \{t\}$ and edge set $E_{st}^r = E_s^r \cup E^r \cup E_t \cup \{t, s\}$.

To add the FIFO issuing policies to LP-formulation of the circulation problem an extra term is added to the objective function:

$$\beta \sum_{e \in E^r} l^r x_e \tag{8.8}$$

Second, constraints (8.3) of the LP formulation have to be adapted. This leading to the following LP-formulation of the circulation flow model, which incorporates the FIFO policy:

Circulation flow problem FIFO

$$\min \quad \sum_{e \in E^r} (c_e + \beta l^r) x_e \tag{8.9}$$

$$\text{s.t.} \quad \sum_{e=\{i,j\}:\ j \in V_{st}^r} x_e - \sum_{e=\{j,i\}:\ j \in V_{st}^r} x_e = 0 \qquad \forall i \in V_{st}^r \tag{8.10}$$

$$0 \leq x_e \leq u_e \qquad \forall e \in E_{st}^r \tag{8.11}$$

8.4 Simulation

In the previous section we have seen that the decisions about which units to issue to satisfy requests is based on three aspects. The first aspect is a hard constraint and states that a blood type issued should be compatible with a blood type requested. The second aspect is the age of the issued unit. To reduce wastage, old units should be issued first. Finally, the third aspect states that the rarity of the blood type issued should be relatively close to the rarity of the blood type requested. However, with respect to the second and third aspect some trade-off should be made. Dependent

on the value of the parameter β more weight is put on the second or third aspect. To investigate the impact of this weight as well as the impact of other parameters, such as the average age of the inventory, we apply a simulation study.

The goal of the simulation study is to mimic the inventory management of a blood bank. The daily decisions about which units to issue are made by the linear programming model discussed in the previous section. It could be that there are multiple decision moment ($DMom$) in a day, which means that the inventory is allocated multiple times per day. For each day the demand is simulated based on historical data (see Section 8.4.1). The supply of RBC units is a bit more complicated, so we will first look at and discuss a small example in which only one blood type is considered. Thereafter, we extend the example to multiple blood types.

8.4.1 Simulating the supply of blood units

To explain the supply side of the simulation model we will first look at a small example. Suppose that we have a single blood type, which has an expiration date of three days. The daily demand (D) for this blood type follows discrete distribution with expectation $\mathbb{E}[D] = 2$. The supply side of the simulation model can be controlled by inviting donors to donate their blood. Whether a donor does or does not responds on an invitation is modeled by a Bernoulli variable Y, which has expectation $\mathbb{E}[Y] = p$. If a donor responds to an invitation ($Y = 1$) then there is a fixed lead time (L_f) of one day and a variable lead time (L_v) of one or two days, both occurring with probability 0.5. Finally, to avoid stockouts a safety stock of four units is kept ($SS = 4$).

In this example we do not start with an empty system. At the end of day $t = 0$ the inventory on hand has the following composition $I_0^e = (3, 1, 0)$, which means that 3 units have age 1, 1 unit has age 2, and 0 units have age 3. Furthermore, the expected number of donors that will show up after respectively 1, 2, and 3 days equals $\mathbb{E}[Z_0^1] = 1.5$, $\mathbb{E}[Z_0^2] = 1$, and $\mathbb{E}[Z_0^3] = 0.5$.

At time $t = 1$ it turns out that $Z_0^1 = 2$, which implies that the inventory at the beginning of the day is equal to $I_1^s = (2, 3, 1)$. During the day the blood supplier receives two requests ($D_1 = 2$). Since this blood supplier issues the blood units according to a FIFO-policy the composition of the inventory at the end of the day becomes $I_1^e = (2, 2, 0)$. At the end of the day the blood supplier has to decide about the number of donors that should be invited to donate A_1 (the so-called reorder quantity). This number depends on the safety stock (SS), the expected demand during lead time ($\mathbb{E}[D_L] = \mathbb{E}[D] \cdot \mathbb{E}[L]$), the inventory on-hand (I_1^e), the inventory in transit ($\mathbb{E}[Z_0^2] + \mathbb{E}[Z_0^3]$), and the expected attendance rate (p):

$$\begin{aligned} A_1 &= \left\lfloor \frac{SS + \mathbb{E}[D_L] - \sum_{i=1}^{2} \left(I_1^e(i)\right) - \left(\mathbb{E}[Z_0^2] + \mathbb{E}[Z_0^3]\right)}{p} \right\rfloor \\ &= \left\lfloor \frac{4 + 2 \cdot 1.5 - 4 - (1 + 0.5)}{0.5} \right\rfloor \\ &= 3, \end{aligned}$$

where $\lfloor . \rceil$ means that the number between these brackets is rounded to the nearest integer. Finally, the expected number of donors that will show up after 1, 2, or 3 days have to be updated: $\mathbb{E}[Z_1^1] = \mathbb{E}[Z_0^2] = 1$, $\mathbb{E}[Z_1^2] = \mathbb{E}[Z_0^3] + \mathbb{P}[L = 2] \cdot \mathbb{E}[p] \cdot A_1 = 1.25$, and $\mathbb{E}[Z_1^3] = \mathbb{P}[L = 3] \cdot \mathbb{E}[p] \cdot A_1 = 0.75$. Note that there was no wastage or shortage this day, because $W_1 = \max\{0, I_1^s(3) - D_1\} = \max\{0, 1 - 2\} = 0$ and $S_1 = \max\left\{0, D_1 - \sum_{i=1}^{3} I_1^s(i)\right\} = \max\{0, 2 - 6\} = 0$.

In Table 8.1 for a time period of five days a possible realization of the variables on the demand and supply side of the simulation model is given (In the table the realization of the variable is shown in *italics*). Moreover, we computed the number of donors that should to be invited every period.

Table 8.1 Different values of the parameters for the small example to explain how the supply side of the simulation model works. Realizations of uncertain variables are shown in *italics*

Time (t)		0	1	2	3	4	5
Inventory begin (I_t^b)	1	-	2	1	0	3	4
	2	-	3	2	1	0	1
	3	-	1	2	0	0	0
Demand (D_t)		-	*2*	*4*	*2*	*2*	*2*
Inventory end (I_t^e)	1	3	2	1	0	1	3
	2	1	2	0	0	0	0
	3	0	0	0	0	0	0
Invited (A_t)		-	3	10	6	4	2
Expected Supply ($\mathbb{E}[Z_t]$)	1	1.5	1	1.25	3.25	4	2.5
	2	1	1.25	3.25	4	2.5	1.5
	3	0.5	0.75	2.5	1.5	1	0.5
Supply	*1*	*2*	*1*	*0*	*3*	*4*	*2*
	2	*1*	*0*	*3*	*4*	*2*	*2*
	3	*0*	*2*	*2*	*1*	*2*	*1*
Wastage (W_t)		-	0	0	0	0	0
Shortage (S_t)		-	0	0	1	0	0

In the example we considered a single blood type and donors were invited based on the safety stock, the inventory on hand and the inventory in transit. In practice however, donors are invited based on their ABO, RhD blood type. The intuition in the example still holds if we look at the entire inventory as eight sub-inventories, namely one for each blood type. Each sub-inventory has its own safety stock, inventory on hand, and inventory in transit. Given that $R_{\max}$ is the maximal shelf life the amount of donors with ABO, RhD blood type j that should be invited to donate at time t is equal to

$$A_t(j) = \left\lfloor \frac{SS(j) + \mathbb{E}[D_L] - \sum_{i=1}^{R_{\max}-1} \left(I_t^e(j^i)\right) - \sum_{i=2}^{L_{\max}} \mathbb{E}[Z_t^i(j)]}{p(j)} \right\rceil. \tag{8.12}$$

Here $SS(j)$ is the safety stock for ABO, RhD blood type j. This safety stock depends

on the expected demand during lead time and the input parameter k. This parameter states what the ideal average shelf life is of the items in stock. To prevent outdating the value of k should be smaller than half of the maximum shelf life ($k < \frac{R_{\max}}{2}$).

Note that we can control the supply of the ABO, RhD blood types by inviting donors accordingly. However, the supply for the extended blood types can only be controlled indirectly via the ABO, RhD blood types and the percentage of donors typed for different antigens. Hence, a donor is invited based on its ABO, RhD blood group, but not based on its entire extended blood type. The replenishment of the inventory is based on three factor 1) the ABO, RhD blood groups of the donors invited 2) The distribution of non- ABO, RhD antigens in the donor population [160] and 3) the percentage of donors typed for these antigens.

8.5 Data acquisition

The simulation model discussed in the previous section incorporates random data streams concerning the demand for and supply of RBC units. The distributions regarding the requests for particular blood types, the lead time of donors, and the percentage of donors responding to an invitation were estimated based on data gathered from Sanquin. This section discusses how these data were were transformed and incorporated :

1. data about the requests for particular blood types,
2. data about the percentage of donors responding to an invitation,
3. data about the lead time of donors,
4. data about the percentage of donors typed for particular antigens,
5. data about the distribution of antigens in the donor population.

These data streams were obtained from the central software system of Sanquin.

8.5.1 Demand data

We extracted data about the demand for RBC units (01-01-2014 up to 04-12-2017) from E-Progesa, a database containing information about all RBC units requested by hospitals. A record in this database contains information about the date of the request, the number of units requested, the number of units delivered, and the typing of the requested units. We excluded records for which no ABO-RhD blood group was known (1.31%), records containing requests for biologically impossible typings (0.02%), and records that were requested during holidays or maintenance of the database system (1.29%).

Dependent on the antigens that were taken into consideration the remaining records (97.28%) were divided into groups based on the requested typing and the day of the week (i.e. Monday, Tuesday, ..., Sunday). Subsequently, within each group, all units requested on the same date were aggregated.

To investigate whether the data of two groups (where these groups had an identical typing but different weekday) were drawn from different distributions we applied log rank test. Finally, for each combination of typing and weekday a negative binomial distribution was fitted.

8.5.2 Supply data

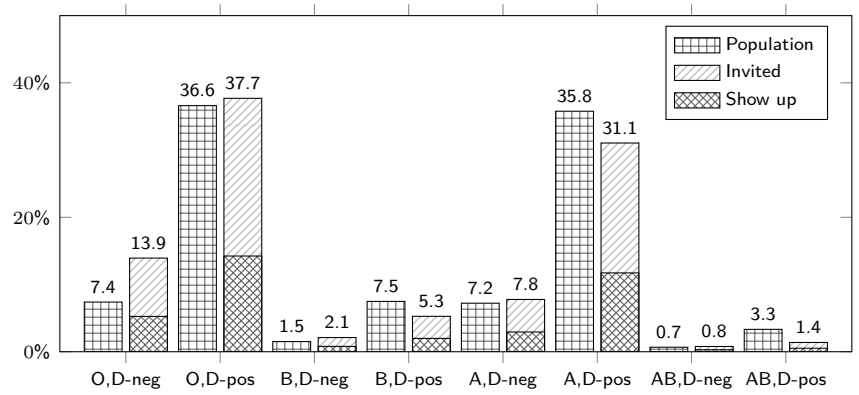
Donors are invited to donate blood based on their ABO, RhD blood group. We extracted data from E-Progesa (01-01-2014 up to 31-03-2017) and investigated whether the distribution of ABO, RhD blood groups of the invited donors were in line with population numbers (see Figure 8.2a). It turns out that if a donor has blood group O and/or RhD-neg the donor is more likely to be invited. Moreover, we also investigated the attendance rate of donors. It turns out that the probability a donor responds on an invitation is 0.3778. We investigate whether there was a difference in respond rate between ABO, RhD blood groups by computing 95% confidence intervals. It turned out that if we look at the daily response rate the differences are negligible (see Figure 8.2b).

We also collected data about the leadtime of a donor. We excluded donation during weekend days (0.03%), as collection sites are not opened on weekend days. Based on the day the invitation was send we estimated a leadtime distribution.

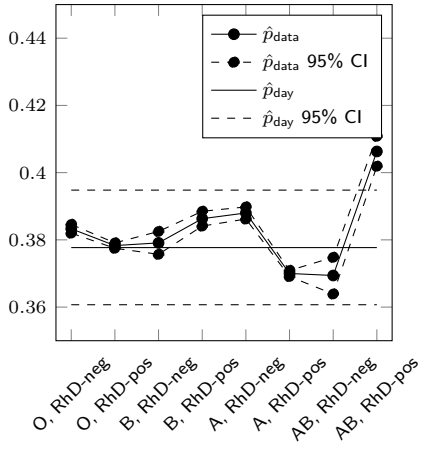
8.5.3 Typing data

To be able to issue extensively typed blood products, we have to look at the probability that a particular typing will occur in our inventory. This probability depends on several factors including the distribution of antigen profiles in the donor population, the inviting policy (currently donor are invited based on their ABO,RhD blood group), the amount of donors typed for different antigens. For estimating the distribution of antigens in the donor population we used the Bloodgroup Antigen Factbook [160].

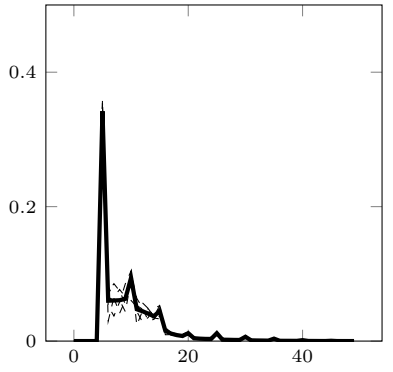
Figure 8.2 Donor-data figures.



(a) The left columns show the percentage of individuals with a specific ABO, RhD blood group in the population. The right columns show the percentage of individuals with a specific ABO, RhD blood group that are invited to donate and the percentage of donors that show up after receiving an invitation.



(b) Show up probability p based on the data for particular ABO, RhD blood groups. Based on these numbers a daily show up probability is computed.



(c) Lead time distribution.

8.6 Computational experiments and results

To compare the proposed approach to the current issuing policy, and compare different parameter settings, several simulations have been run. Every simulation run is one year (365 days) long. This is preceded by 56 days to warm up the model. At the start of the warm up, it always starts with a uniformly distributed inventory from 0 days to $2k$ days old, where k is the average shelf life parameter. For the first and

second antigen sets, 100 runs will be used to compute the average performance and the confidence intervals, for the third antigen set, 10 runs will be used to reduce computational complexity. This results from the fact that thousands of decisions have to be taken, the decision at one decision moment, as would be required in practice, still only requires a few seconds.

For the main parameters of the model, results will be shown for three different values. The number of decision moments per day DMom is one, three or five. Currently, every time a request comes in from a hospital, RBC units are assigned to this request, but it is not unreasonable to save up requests and assign the units a few times per day, as most requests are not shipped immediately. For the average shelf life k, on which the order up to quantity depends, the values 5, 10 and 15 have been used. For β, the parameter indicating the trade off between age and rarity of blood products, values .1, 1 and 10 have been used. Finally, we have also used three different antigen sets (see Table 8.3 for all antigens and their system). The first series of simulations for antigen set $\mathcal{A}_1$ only include the ABO system and RhD, the second antigen set $\mathcal{A}_2$ also includes the remainder of the Rhesus system and K. The third, most extended set $\mathcal{A}_3$ also includes S, s and the Duffy and Kidd systems.

Shortage and out-dating are very small in all scenarios. There is no shortage for all runs with antigen sets $\mathcal{A}_1$ and $\mathcal{A}_2$. For the runs with antigen set $\mathcal{A}_3$ there is some shortage, but this is at most 0.1%. Interestingly, shortage is independent of the average shelf life k, even for $k = 5$, problems do not occur with the antigen sets $\mathcal{A}_1$ and $\mathcal{A}_2$. Out-dating is dependent on k, but is very low. For $k = 5$ or $k = 10$ outdating is non-existent, and for $k = 15$, out-dating is at most 0.15%. With current policies, out-dating is approximately 2%, so even 0.15% is a significant improvement. It is hard to compare shortages with current values, as it is not registered when insufficient inventory levels are the reason to not supply a hospital with a requested RBC unit.

Table 8.2 shows the percentage of requests that can be supplied with an exactly matching unit instead of a compatible unit. For the first antigen set $\mathcal{A}_1$, the percentages are shown in table 8.2a. Although differences can be observed between parameter settings, some of them also significant, all fractions are extremely high. If all further antigens are dropped, volumes are very high, and reaching almost 100% exact matches is not very hard with a smart issuing policy. If β is increased, a little more weight is given to the age of products, and it makes sense that the number of exact matches slightly decreases. The effects of increasing the number of decision moments is negligible. The same holds for the average shelf life k when the number of exact matches is concerned.

Tables 8.2b and 8.2c show the fraction of exact matches for the second and third antigen set respectively. Note that requests for all standard ABO and RhD typed blood have been excluded from these results. Most donors in the Netherlands have been typed for a number of extra antigens, making a donor that is only negative for ABO RhD antigens very rare. On the other hand, the number of requests for these types is very high. So, if these requests would be included, results would be very distorted, and reaching high exact matching fractions would be impossible. Comparing Tables 8.2b and 8.2c to Table 8.2a therefore does not make sense.

Table 8.2 Percentage of blood units issued that were an exact match between the blood type issued and the blood type supplied.

(a) Antigen set $\mathcal{A}_1$.

DMom	k	$\beta = 0.1$	$\beta = 1$	$\beta = 10$
1	5	0.995 (0.993, 0.996)	0.990 (0.987, 0.992)	0.990 (0.987, 0.992)
	10	0.994 (0.992, 0.996)	0.990 (0.988, 0.993)	0.990 (0.988, 0.993)
	15	0.994 (0.992, 0.996)	0.991 (0.988, 0.993)	0.991 (0.988, 0.993)
3	5	0.993 (0.990, 0.995)	0.987 (0.984, 0.989)	0.987 (0.984, 0.989)
	10	0.993 (0.991, 0.996)	0.988 (0.985, 0.991)	0.988 (0.985, 0.991)
	15	0.993 (0.991, 0.995)	0.988 (0.986, 0.991)	0.988 (0.986, 0.991)
5	5	0.992 (0.990, 0.995)	0.987 (0.984, 0.990)	0.987 (0.984, 0.990)
	10	0.993 (0.991, 0.995)	0.988 (0.985, 0.991)	0.988 (0.985, 0.991)
	15	0.993 (0.990, 0.995)	0.989 (0.986, 0.991)	0.989 (0.986, 0.991)

(b) Antigen set $\mathcal{A}_2$, standard ABO-D blood types excluded.

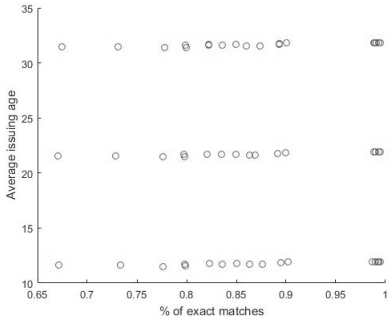
DMom	k	$\beta = 0.1$	$\beta = 1$	$\beta = 10$
1	5	0.902 (0.896, 0.908)	0.894 (0.889, 0.901)	0.894 (0.889, 0.901)
	10	0.899 (0.894, 0.904)	0.891 (0.885, 0.897)	0.891 (0.885, 0.897)
	15	0.901 (0.896, 0.906)	0.893 (0.887, 0.900)	0.893 (0.887, 0.900)
3	5	0.850 (0.845, 0.856)	0.823 (0.815, 0.830)	0.823 (0.815, 0.830)
	10	0.850 (0.843, 0.856)	0.820 (0.814, 0.827)	0.820 (0.814, 0.827)
	15	0.850 (0.844, 0.856)	0.822 (0.816, 0.829)	0.822 (0.816, 0.829)
5	5	0.836 (0.830, 0.841)	0.798 (0.792, 0.806)	0.798 (0.792, 0.806)
	10	0.835 (0.830, 0.842)	0.797 (0.791, 0.804)	0.797 (0.791, 0.804)
	15	0.836 (0.828, 0.843)	0.799 (0.792, 0.805)	0.799 (0.792, 0.805)

(c) Antigen set $\mathcal{A}_3$, standard ABO-D blood types excluded.

DMom	k	$\beta = 0.1$	$\beta = 1$	$\beta = 10$
1	5	0.876 (0.873, 0.878)	0.863 (0.858, 0.866)	0.863 (0.858, 0.866)
	10	0.869 (0.866, 0.875)	0.863 (0.857, 0.866)	0.863 (0.857, 0.866)
	15	0.874 (0.872, 0.878)	0.860 (0.856, 0.867)	0.860 (0.856, 0.867)
3	5	0.798 (0.796, 0.801)	0.733 (0.726, 0.742)	0.733 (0.726, 0.742)
	10	0.798 (0.796, 0.801)	0.728 (0.725, 0.736)	0.728 (0.725, 0.736)
	15	0.800 (0.797, 0.804)	0.731 (0.724, 0.738)	0.731 (0.724, 0.738)
5	5	0.776 (0.770, 0.782)	0.671 (0.664, 0.677)	0.671 (0.664, 0.677)
	10	0.777 (0.771, 0.784)	0.670 (0.662, 0.677)	0.670 (0.662, 0.677)
	15	0.778 (0.772, 0.781)	0.674 (0.666, 0.680)	0.674 (0.666, 0.680)

In Table 8.2b differences between the scenarios are slightly higher than for the first antigen set, so a few more observations can be made. The number of decision moments is now significantly influencing results. This is not surprising, as increasing

Figure 8.3 Scatter plot of the average issuing age and percentage of exact matches for all scenarios considered.



the number of decision moment directly increases the difficulty level of the problem. It is important to note that only having one decision moment per day would most likely increase delivery time to hospitals too much. Differences for different values of k are still negligible.

The fraction of exact matches with the third antigen set $\mathcal{A}_3$ shown in Table 8.2c are off course lower than those in Table 8.2b, as increasing the number of included antigens decrease the probability of an exact match significantly. With this in mind, the ratio of exact matches that can be achieved with this approach is very high. Table 8.2c further shows similar results to those that can be found in Table 8.2b.

Results from $\beta = 1$ are equal to $\beta = 10$, indicating that decisions and results most likely will not change for any value of $\beta > 1$. It is also interesting to note that different values for k never result in significant results, and it therefore seems that the value for k can be based on other factors, without the need to take the percentage of exact matches into account.

As Table 8.2 includes a lot of scenarios, we have included Figure 8.3 to visualize some of the results. Every point represents the average issuing age (on the vertical axis) and the average fraction of exact matches (on the horizontal axis) of one of the scenarios in Table 8.2. Ignoring very small (non significant) differences, the average issuing age is completely dependent on the value chosen for k. As also shown in Table 8.2, the fraction of exact matches depends on most major parameters of the approach, except for k. this can be nicely seen as every point corresponds to two other points below or above it, as if on a vertical line.

8.7 Discussion

A blood type is defined by the antigens for which someone is negative. If a blood type is requested by a hospital, Sanquin can supply the hospital with blood that is negative for an antigen that is positive in the request, but not the other way around. Blood that is negative for a large number of antigens is rare, and a well-designed inventory management system for blood collection sites should incorporate as many

antigens as possible. However all models presented for this purpose in literature only incorporate the 3 ABO-RhD antigens, resulting in 8 blood types. The model presented in this paper is generalized, and can incorporate any antigen. The most extensive numerical study included in the paper includes 14 antigens, resulting in 16384 different blood types.

The model does not take a decision every time a request comes in, but instead batches the requests, and takes a decision a few times a day. This decision than allocates a unit of RBCs from the inventory to all requests that have been batched. Even for the most extensive numerical study with 16384 different blood types and 35 days maximal storage time Ű all days of which are distinguished as different products by the model, an average laptop computer is able to take this decision in approximately 5 seconds. This is well within the requirements for practical implementation.

By incorporating a lot of different blood types, we deal with (very) rare blood types. It is not possible to supply a hospital with a unit less rare than the requested unit, making it important to keep rare units in our inventory. In the ideal situation, therefore, we would be supplying every request with a unit of RBCs that exactly matches the requested unit. So, one of the most important performance measures for our model, is the percentage of requests that were supplied with an exactly matching unit of RBCs. This performance measure has also not been reported before in literature, as this gets less important if less antigens are incorporated in the model.

Currently, issuing units of RBCs is a manual process. This has serious disadvantages, as the inventory of RBCs is far too large and complex for a human to keep track of. This inability without a doubt leads to inefficiencies in the issuing of RBC units.

Using our model, a very high percentage of requests can be fulfilled with an exactly matching unit of RBCs. This is combined with a significant reduction in the percentage of units outdated, and a very low shortage. The percentage of exact matches and shortage seem independent on the average shelf life of blood products within the range tested in our numerical experiments. Outdating is dependent on the average shelf life, as outdating is only present for an average shelf life of 15 days, and not for 10 and 5 days. A short average shelf life therefore seems recommendable.

Appendix

Table 8.3 Binary representation of the blood groups and their prevalence in the Caucasian, African, and Asian population [160].

Blood group system	Name	Special name	Binary representation	% Caucasian	% Blacks	% Asian
ABO		AB	(1,1)	4	4	5
	B-	A	(1,0)	43	27	27
	A-	B	(0,1)	9	20	25
	A-, B-	O	(0,0)	44	49	43
Rhesus		R_1R_2	(1,1,1,1,1)	13.3	4.0	30.0
	e-	R_2R_z	(1,1,1,1,0)	0.1	rare	0.4
	E-	R_1r	(1,1,1,0,1)	34.9	21.0	8.5
	c-	R_1R_z	(1,1,0,1,1)	0.2	rare	1.4
	c-, e-	R_zR_z	(1,1,0,1,0)	rare	rare	rare
	c-, E-	R_1R_1	(1,1,0,0,1)	18.5	2.0	51.8
	C-	R_2r	(1,0,1,1,1)	11.8	18.6	2.5
	C-, e-	R_2R_2	(1,0,1,1,0)	2.3	0.2	4.4
	C-, E-	R_0r	(1,0,1,0,1)	2.1	45.8	0.3
	D-	r'r''	(0,1,1,1,1)	rare	rare	rare
	D-, E-	r''r^y	(0,1,1,1,0)	rare	rare	rare
	D-, e-	r'r	(0,1,1,0,1)	0.8	0.5	0.1
	D-, c-	r'r^y	(0,1,0,1,1)	rare	rare	rare
	D-, c-, e-	$r^y r^y$	(0,1,0,1,0)	rare	rare	rare
	D-, c-, E-	r'r'	(0,1,0,0,1)	rare	rare	0.1
	D-, C-	r''r	(0,0,1,1,1)	0.9	rare	rare
	D-, C-, e-	r''r''	(0,0,1,1,0)	rare	rare	rare
	D-, C-, E-	rr	(0,0,1,0,1)	15.1	6.8	0.1
Kell			(1,1)	8.8	2	*
	k-		(1,0)	0.2	rare	*
	K-		(0,1)	91	98	*
	K-, k-		(0,0)	rare	rare	rare
Kidd			(1,1)	50.3	40.8	49.1
	Jk^a-		(1,0)	26.3	51.1	23.2
	Jk^b-		(0,1)	23.4	8.1	26.8
	Jk^a-, Jk^b-		(0,0)	rare	rare	0.9
Duffy			(1,1)	49	1	8.9
	Fy^a-		(1,0)	22	9	90.8
	Fy^b-		(0,1)	34	22	0.3
	Fy^a-, Fy^b-		(0,0)	rare	68	0
MNS			(1,1,1,1)	24	13	*
	s-		(1,1,1,0)	4	2	*
	S-		(1,1,0,1)	22	33	*
	S-, s-		(1,1,0,0)	0	0.4	*
	N-		(1,0,1,1)	14	7	*
	N-, s-		(1,0,1,0)	6	2	*
	N-, S-		(1,0,0,1)	8	16	*
	N-, S-, s-		(1,0,0,0)	0	0.4	*
	M-		(0,1,1,1)	6	5	*
	M-, s-		(0,1,1,0)	1	2	*
	M-, S-		(0,1,0,1)	15	19	*
	M-, S-, s-		(0,1,0,0)	0	0.7	*

Part IV

Practice and Outlook

Chapter 9

S.P.J. van Brummelen, N.M. van Dijk, W.L. de Kort and K. van den Hurk. Combining optimal shift scheduling and reallocation policies at Dutch blood collection sites. *In preparation*.

CHAPTER 9

Application of staff scheduling and reallocation: Case studies

9.1 Introduction

Blood services across the world have to deal with pressure to reduce health care costs, while many also face another challenge: the decreasing demand for blood (e.g. see the annual reports of Sanquin [163]). As the cost of staff and other resources at blood collection sites are a major part of the cost of blood, the efficient use of these resources is important and requires research.

Blood donors value their time and long waiting times are an important reason to stop donating (Mckeever et al. [132]). Simple methods to decrease waiting times without increasing the number of staff members do not exist, as, in general, the amount of unused capacity in the system is inversely related to waiting times. The most simple M|M|1 queueing model already shows a direct relation between unused capacity $(1 - \text{occupancy})$ and the total expected time spent in the system (sojourn time). If, for explanatory purposes, we were to assume that an M|M|1 model can describe a blood collection site, the total expected time a donor spends in this collection site can be estimated by (e.g. see Winston [183, Chapter 20]):

$$\text{Sojourn time} = \frac{1}{1 - \text{occupancy}} \times \text{Mean service time} \tag{9.1}$$

Here, the occupancy is the fraction of the capacity that is required to service all incoming clients. Although simply adding more staff is not an option, we can change the allocation of capacity. This can both be done by changing working times of staff members, or by changing the task that is performed by a staff member. In this way, so to speak, additional capacity can be created where it is necessary, by shifting this capacity from moments where it is not required. The first of these options, changing the working times, has already been adressed for other service systems (e.g. see the review by Defraeye and Van Nieuwenhuyse [57]), but has never been applied to a blood collection site. Changing the allocation of staff members has never been studied in a process as stochastic as a blood collection site. This stochasticity originates in time-dependent random arrivals and several random service times.

In the Netherlands, every year approximately 800,000 blood donation attempts are made, and similar systems exist throughout the world. In spite of this high volume, research into the logistics of blood collection sites apprears to be scarce (Baş et al. [20]). The first study to investigate the logistics of blood collection sites was done by Pratt et al. [154]. In this study, a simulation model was used to investigate queues at blood collection sites, and investigate improvements with respect to arrivals of donors. As can be predicted, they find that equally spread arrivals result in less queues than uneven arrivals, and that staff utilization is optimal if the system is at or near to its maximal capacity. The most interesting suggestion is that, as donations by men are quicker, scheduling more men at the start of a session can be beneficial. Subsequently, simulation studies have also been performed by Brennan et al. [38], De Angelis et al. [15] and Alfonso et al. [8]. In essence, simulation is a tool that can be used for evaluation of scenarios and to achieve better performance by choosing the best scenario. However, no guarantees can be made that the chosen scenario is optimal or even near optimal. Another disadvantage of simulation is that it has to be adapted for each specific set-up of a blood collection site. Although some general conclusions can be drawn from these papers, such as to use staff as flexible as possible and that service can be increased with an increased budget, more specific results only apply to the collection site that has been simulated.

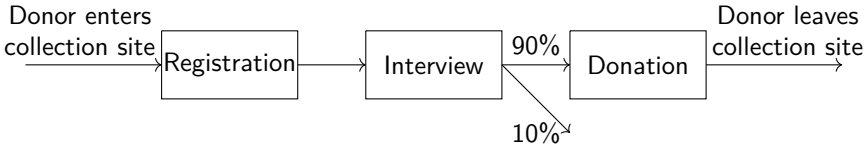
Another option to study waiting times and efficiency at blood collection sites, is the use of analytical methods. Although some assumptions are usually required, these methods produce exact and replicable results, in contrast to simulation. Bretthauer and Côté [39] were the first to apply such methods to a blood collection site. Their paper first presents a general framework to determine required resources for a new or redesigned health care system. Subsequently, they apply their method to a blood collection site. Based on expected arrivals and a required service level, they compute the number of required staff members and equipment. This model is aimed at a longer time frame, and is not built for staff scheduling or other short term decisions at day, session or hourly level.

Chapters 3 and 4 show multiple results for the analytical computation of expected waiting times, waiting time distributions and queue length distributions at blood collection sites. Blake and Shimla [29] also use an analytical queueing model. Their paper describes the use of a relatively standard queueing model to determine the required number of simultaneously working staff members for a blood collection site, depending on the number of expected donor arrivals. Testik et al. [173] use data mining to determine arrival patterns and then use methods similar to those of Blake and Shimla to determine the required number of staff members.

Although it is useful to know the required number of staff members, this number might fluctuate wildly throughout the day as at many blood collection sites donors do not come to the collection site uniformly throughout the day. In Chapter 5 we have shown that an Integer Linear Program (ILP) can be used to compute optimal shifts for staff members based on the required number of staff members, given arrival patterns of donors and the size of the specific blood collection site.

Subsequently, we will use the results from this model by computing the optimal

Figure 9.1 Model of a blood collection site. The first station is the Registration, the second is the Interview station and the third is the Donation station.



allocations and reallocations of the staff members to the different stations during a collection session at a blood collection site to minimize waiting time. This is of course dependent on the number of queueing donors at the different stations of the donation process. Chapter 6 provides a Markov Decision Process (MDP) that can be used for this purpose.

In this chapter, we combine the ILP technique for optimal shift planning from Chapter 5 and the MDP technique for optimal staff allocation from Chapter 6 during a collection session to use the capacity as effectively as possible. We will apply the two techniques to data from three blood collection sites of different sizes. By allocating staff members optimally both in time and location, we will decrease waiting times without increasing cost.

9.2 Methods

The combined approach used for the results in this paper is based on two distinct approaches:

- Shift scheduling with an ILP based on arrival patterns, presented in Chapter 5 and shortly explained in Section 9.2.2.
- Staff reallocation based on an MDP formulation, presented in Chapter 6 and shoftly explained in Section 9.2.3.

9.2.1 Model

An extensive description of the processes at a blood collection site can be found in Section 1.3 in Chapter 1. As in previous chapters, we will model a blood collection site as three sequential queues, as shown in figure 9.1. The stations are Registration, Interview and Donation respectively. Staff members are always assigned to one of these three stations for a period of time, and each of the stations always has at least one staff member present. The only exception is that, after the collection site has closed, a station can be abandoned if there are no more donors that have to go through this station.

All service times are assumed to be exponentially distributed. We will assume the average average service times to be 2 minutes for the Registration station, 6 minutes for the Interview station and 12 minutes for the Donation station. These values are based on data gathered throughout the Netherlands and model validation

in Chapter 3. The time between two arrivals is also assumed to be exponentially distributed. The arrivals are based on the arrivals of whole blood donors for one day of the week at a collection site. We have chosen to include the Wednesday session of the location in Nijmegen, the Thursday session of the location in Leiden and the Thursday session of the mobile location in Almelo. For every half hour during the selected sessions, the average number of donors in 2015 has been used to establish an arrival pattern.

9.2.2 Shift scheduling: ILP

The shift scheduling algorithm is used to compute the shifts that minimize the number of worked hours, given that enough staff members are present at any time. The approach is based work presented in Chapter 5. The method consists of two basic steps. The first step is to determine the minimum required number of staff members for every half hour. This can be done by a variety of ways, and in this chapter, we will use a M|M|c model (e.g. see Winston [183, Chapter 20]). By using this model, we compute the required number of staff members for a time interval based on the arrivals that are expected for this interval, total service time for a donor and some sojourn time requirement. This method does not distinguish between the three stations, but models one station, with one queue and a total service time of 20 minutes.

The second step is to use the minima from the first step to compute the optimal shifts. This is done by minimizing the number of worked hours, with two basic restrictions: 1) At any time, at least the computed minimum number of staff members have to be present and 2) shifts can only have a duration of 3, 4, 5, 6, 7, 8 or 9 hours. This second step is solved with a so-called Integer Linear Program (ILP) (e.g. see Winston [183, Chapter 9]).

9.2.3 Staff allocation and reallocation: MDP

The staff (re)allocation algorithm is used to determine the optimal assignment of staff members to the three stations of the blood donation process: Registration, Interview, and Donation. The approach is based on work presented in Chapter 6. The algorithm uses a Markov Decision Process (MDP, e.g. see Puterman [155]), which is used to compute the optimal staff allocation at fixed intervals. The number of present donors, their location in the process, and the expected number of arrivals are all taken into account to compute the optimal allocation of each staff member at each interval.

9.3 Results

To compute the minimum number of staff members with the M|M|s model, we have used the requirement that at most 12% of donors should have a sojourn time of more than 45 minutes. Note that, as the aggregate station does not accurately

reflect reality, this requirement is not automatically satisfied if only the minimum number of staff members is scheduled. This minimum number of staff members has been computed for every half hour, and then the ILP has been used to compute shifts using the minimum total number of staff members. The average number of staff members scheduled is shown in the first column (first three rows) of Table 9.1.

This scenario is compared with two other scenarios: One where we schedule the (rounded up) average number from the ILP the entire day, and one where we schedule the peak number scheduled by the ILP for the entire day. The first of these has slightly more staff members, scheduled less efficiently, and the third has a lot more staff members. The first of these, the "mean ILP rounded up" scenario, will be used as a reference. Currently, the number of present staff members rarely changes during the day. As this scenario will use slightly more but a comparable number of staff hours, it is fairest to use this as a comparison to show the improvements from the ILP model.

In the first three rows, we have assumed that a staff member can help 3 donors per hour ($\mu = 60$ minutes/20 minutes per donor), corresponding to the 20 minutes total service time mentioned in Section 9.2.1. For rows 4-6, μ has been set to 3.5, and for rows 7-9, μ has been set to 4. The remaining procedure is equal to that previously described.

Table 9.1 Average number of working staff members per blood collection site per half hour. For the first column the M|M|s method has been used to compute the required number of staff members, and then the staff scheduling algorithm has been run. For the second column, the average number of present staff members from the ILP has been rounded up and scheduled every time interval, for the third column, the highest number of staff members scheduled in some time interval by the ILP has been scheduled the entire day. μ is the number of serviced donors per staff member per hour.

			Mean ILP	
		ILP	rounded up	Max ILP
$\mu = 3$	Nijmegen	7.06	8	10
	Leiden	5.60	6	8
	Almelo	8.71	9	12
$\mu = 3.5$	Nijmegen	5.35	6	8
	Leiden	4.20	5	5
	Almelo	6.41	7	9
$\mu = 4$	Nijmegen	4.76	5	7
	Leiden	3.64	4	4
	Almelo	5.53	6	8

After determining the number of staff members, we used the MDP to compute the optimal staff allocation. For every staff schedule, we ran the MDP 4 times. The first time, the MDP reallocated staff members every 7.5 minutes, the second time every 15 minutes and the third time every 30 minutes. The fourth time, the MDP modeled a fixed allocation of staff members. The reallocations do not imply that all staff members had to be reassigned every time interval, just that some staff

members might have to change their allocation. The output of the MDP is a complete distribution over all possible states, i.e. numbers of donors at each of the stations, at all times. Given the assumptions, this is an exact computation, and is not subject to confidence intervals. From this, the expected number of present and waiting donors was computed for every half hour. Table 9.2a shows the average expected and the highest expected (in parentheses) number of donors present during the day in Nijmegen. Table 9.2b shows the average and highest expected (in parentheses) number of waiting donors during the day in Nijmegen. Table 9.3 shows the same results for Leiden, and Table 9.4 shows the results for Almelo.

Tables 9.2, 9.3 and 9.4 show that the ILP results in negligible reductions when it comes to average expected number of present and waiting donors compared to just scheduling the mean number of staff members the entire day. It should be noted that the "mean ILP" column always has more total staff hours, and often quite a lot more, than the ILP scenario. However, the main goal of the ILP is not to decrease the average number of present or waiting donors, but to reduce the peaks of these variables by scheduling more staff at peak arrival times. Compared to scheduling the mean number of staff members the entire day, the schedule from the ILP is able to decrease the maximum expected number of donors present and waiting at the blood collection site in almost all cases. A notable exception is the collection site in Leiden, when $\mu = 3.5$ or $\mu = 4$ is used. The collection site in Leiden has quite evenly spread out arrivals, resulting in the (rounded up) average number of staff members being equal to the maximum number of staff members. In this case, the ILP will always perform worse in terms of waiting times than the other two scenarios, as the ILP always has equal or less staff members available.

The MDP method decreases both the average and the maximum of both the number of present and the number of waiting donors. The reductions on the number of waiting donors are higher. The reductions also increase if the interval in which the allocations are changed decreases. As the intervals decrease, the flexibility increases, so it makes sense that this influences the performance of the algorithm. This of course has to be weighted with the number of times a staff member has to reallocate, as shown in Table 9.5.

9.4 Discussion

We can conclude that combining the ILP model to compute optimal shifts for staff members with the MDP model to allocate and reallocate the staff members to the three stations at a blood collection site is effective in decreasing queues at equal or lower costs. The ILP distributes the staff members to the periods of the day where they are most necessary. The implicit goal of this approach is to reduce the peak queues at blood collection sites, a goal that is achieved in the numerical experiments.

Additionally, the MDP is constantly optimizing the allocation of staff members to the three stations to reduce the average number of queued donors. By doing this, both the peak queues and the average queues are decreased. This of course comes at the cost of staff members having to change their assignment a few times per day,

Table 9.2 The number of donors present (Table 9.2a) and waiting (Table 9.2b) at the collection site in Nijmegen. The average expected number of donors present/waiting in the collection site and the highest expected number of donors present/waiting (in parentheses) are shown for a number of scenarios. Rows correspond to a combination of number of serviced donors per staff member per hour, $\mu = 3$, $\mu = 3.5$ or $\mu = 4$ and the MDP that can change the allocation of the staff members every 7.5 minutes, 15 minutes, 30 minutes or not at all. The columns correspond to those in Table 9.1. Columns and rows indicated with * are closest the the current practice at Sanquin, and should be considered as reference values.

		ILP	Mean ILP rounded up*	Max ILP
$\mu = 3$	MDP 7.5 minutes	3.18 (5.53)	3.11 (5.87)	2.97 (5.53)
	MDP 15 minutes	3.27 (5.28)	3.19 (5.82)	2.99 (5.28)
	MDP 30 minutes	3.35 (5.94)	3.27 (6.32)	3.02 (5.94)
	No reallocations*	3.39 (6.39)	3.44 (6.79)	3.02 (6.39)
$\mu = 3.5$	MDP 7.5 minutes	4.26 (5.87)	3.74 (6.74)	3.11 (5.87)
	MDP 15 minutes	4.49 (5.82)	3.95 (7.20)	3.19 (5.82)
	MDP 30 minutes	4.78 (6.53)	4.18 (7.79)	3.27 (6.32)
	No reallocations*	5.22 (7.71)	4.08 (8.06)	3.44 (6.79)
$\mu = 4$	MDP 7.5 minutes	5.47 (7.57)	4.95 (8.73)	3.30 (6.20)
	MDP 15 minutes	5.78 (8.00)	5.26 (8.98)	3.43 (6.28)
	MDP 30 minutes	6.15 (9.08)	5.68 (9.93)	3.60 (6.64)
	No reallocations*	6.33 (9.94)	9.34 (13.44)	3.55 (6.95)

(a) Present donors at the collection site.

		ILP	Mean ILP rounded up*	Max ILP
$\mu = 3$	MDP 7.5 minutes	0.23 (0.55)	0.16 (0.78)	0.05 (0.37)
	MDP 15 minutes	0.33 (0.67)	0.25 (1.02)	0.08 (0.45)
	MDP 30 minutes	0.44 (0.83)	0.36 (1.19)	0.12 (0.55)
	No reallocations*ă	0.70 (1.57)	0.74 (2.37)	0.30 (1.42)
$\mu = 3.5$	MDP 7.5 minutes	1.26 (2.40)	0.74 (2.34)	0.16 (0.78)
	MDP 15 minutes	1.49 (2.68)	0.99 (2.84)	0.25 (1.02)
	MDP 30 minutes	1.82 (3.11)	1.25 (3.40)	0.36 (1.19)
	No reallocations*	2.70 (3.89)	1.40 (3.68)	0.74 (2.37)
$\mu = 4$	MDP 7.5 minutes	2.48 (5.44)	1.93 (4.78)	0.33 (1.23)
	MDP 15 minutes	2.77 (5.41)	2.23 (5.24)	0.47 (1.51)
	MDP 30 minutes	3.20 (5.41)	2.67 (5.82)	0.66 (1.90)
	No reallocations*	3.96 (6.60)	6.98 (10.00)	0.86 (2.47)

(b) Waiting donors at the collection site.

but as the number of reallocations is at most 1 staff member per half hour, this seems realistically feasible.

The initial parameter settings ($\mu = 3$) of the M|M|s method to compute the required number of staff members clearly scheduled more staff members than neces-

Table 9.3 The number of donors present (Table 9.3a) and waiting (Table 9.3b) at the collection site in Leiden. The average expected number of donors present/waiting in the collection site and the highest expected number of donors present/waiting (in parentheses) are shown for a number of scenarios. Rows correspond to a combination of number of serviced donors per staff member per hour, $\mu = 3$, $\mu = 3.5$ or $\mu = 4$ and the MDP that can change the allocation of the staff members every 7.5 minutes, 15 minutes, 30 minutes or not at all. The columns correspond to those in Table 9.1. Columns and rows indicated with * are closest the the current practice at Sanquin, and should be considered as reference values.

		ILP	Mean ILP rounded up*	Max ILP
$\mu = 3$	MDP 7.5 minutes	2.75 (3.66)	2.68 (3.73)	2.46 (3.35)
	MDP 15 minutes	2.87 (3.88)	2.79 (3.88)	2.49 (3.42)
	MDP 30 minutes	2.96 (4.18)	2.85 (4.18)	2.53 (3.63)
	No reallocations*	3.03 (4.21)	2.76 (4.21)	2.56 (3.84)
$\mu = 3.5$	MDP 7.5 minutes	3.97 (5.14)	3.10 (4.28)	3.10 (4.28)
	MDP 15 minutes	4.20 (5.56)	3.26 (4.55)	3.26 (4.55)
	MDP 30 minutes	4.50 (6.25)	3.43 (5.12)	3.43 (5.12)
	No reallocations*	5.49 (8.29)	5.00 (8.19)	5.00 (8.19)
$\mu = 4$	MDP 7.5 minutes	6.13 (8.63)	4.85 (7.35)	4.85 (7.35)
	MDP 15 minutes	6.43 (9.29)	5.15 (7.58)	5.15 (7.58)
	MDP 30 minutes	6.68 (10.47)	5.52 (8.64)	5.52 (8.64)
	No reallocations*	6.80 (11.45)	5.92 (9.90)	5.92 (9.90)

(a) Present donors at the collection site.

		ILP	Mean ILP rounded up*	Max ILP
$\mu = 3$	MDP 7.5 minutes	0.30 (0.83)	0.23 (0.56)	0.04 (0.13)
	MDP 15 minutes	0.41 (1.00)	0.36 (0.80)	0.08 (0.20)
	MDP 30 minutes	0.54 (1.21)	0.46 (0.97)	0.12 (0.29)
	No reallocations*	0.79 (1.63)	0.48 (1.00)	0.28 (0.60)
$\mu = 3.5$	MDP 7.5 minutes	1.50 (2.63)	0.62 (1.33)	0.62 (1.33)
	MDP 15 minutes	1.70 (2.86)	0.76 (1.57)	0.76 (1.57)
	MDP 30 minutes	2.02 (3.13)	0.97 (1.89)	0.97 (1.89)
	No reallocations*	3.42 (5.62)	2.90 (5.22)	2.90 (5.22)
$\mu = 4$	MDP 7.5 minutes	3.75 (7.08)	2.41 (4.51)	2.41 (4.51)
	MDP 15 minutes	3.96 (7.20)	2.64 (4.90)	2.64 (4.90)
	MDP 30 minutes	4.28 (7.69)	3.05 (5.59)	3.05 (5.59)
	No reallocations*	4.89 (9.21)	3.90 (6.96)	3.90 (6.96)

(b) Waiting donors at the collection site.

sary. If the proposed approach would be implemented in some other blood collection site or even another process entirely, it therefore remains important to calibrate the approach for the specific application.

The ILP model can schedule different numbers of staff members per half hour,

Table 9.4 The number of donors present (Table 9.4a) and waiting (Table 9.4b) at the collection site in Almelo. The average expected number of donors present/waiting in the collection site and the highest expected number of donors present/waiting (in parentheses) are shown for a number of scenarios. Rows correspond to a combination of number of serviced donors per staff member per hour, $\mu = 3$, $\mu = 3.5$ or $\mu = 4$ and the MDP that can change the allocation of the staff members every 7.5 minutes, 15 minutes, 30 minutes or not at all. The columns correspond to those in Table 9.1. Columns and rows indicated with * are closest the the current practice at Sanquin, and should be considered as reference values.

		ILP	Mean ILP rounded up*	Max ILP
$\mu = 3$	MDP 7.5 minutes	4.11 (6.84)	4.12 (7.27)	3.90 (6.84)
	MDP 15 minutes	4.21 (6.53)	4.23 (7.26)	3.91 (6.53)
	MDP 30 minutes	4.31 (7.18)	4.35 (7.67)	3.92 (7.18)
	No reallocations*ă	4.41 (7.86)	4.30 (8.08)	4.02 (7.86)
$\mu = 3.5$	MDP 7.5 minutes	5.10 (7.33)	4.87 (8.13)	4.12 (7.27)
	MDP 15 minutes	5.42 (7.81)	5.15 (8.51)	4.23 (7.26)
	MDP 30 minutes	5.76 (8.52)	5.52 (9.27)	4.35 (7.67)
	No reallocations*	6.26 (9.95)	5.59 (9.93)	4.30 (8.08)
$\mu = 4$	MDP 7.5 minutes	7.03 (10.7)	6.04 (9.62)	4.37 (7.61)
	MDP 15 minutes	7.53 (11.49)	6.48 (10.21)	4.55 (7.79)
	MDP 30 minutes	8.04 (12.79)	7.06 (11.11)	4.74 (8.01)
	No reallocations*	9.43 (15.09)	7.13 (11.58)	5.26 (9.47)

(a) Present donors at the collection site.

		ILP	Mean ILP rounded up*	Max ILP
$\mu = 3$	MDP 7.5 minutes	0.25 (0.73)	0.25 (1.06)	0.09 (0.55)
	MDP 15 minutes	0.36 (1.00)	0.38 (1.33)	0.10 (0.57)
	MDP 30 minutes	0.49 (1.27)	0.54 (1.66)	0.12 (0.61)
	No reallocations*	0.89 (1.93)	0.75 (2.49)	0.47 (1.93)
$\mu = 3.5$	MDP 7.5 minutes	1.18 (2.64)	0.94 (2.61)	0.25 (1.06)
	MDP 15 minutes	1.51 (3.22)	1.24 (3.08)	0.38 (1.33)
	MDP 30 minutes	1.95 (3.87)	1.62 (3.74)	0.54 (1.66)
	No reallocations*	2.95 (5.24)	2.10 (4.71)	0.75 (2.49)
$\mu = 4$	MDP 7.5 minutes	3.15 (6.60)	2.09 (4.71)	0.47 (1.56)
	MDP 15 minutes	3.61 (7.33)	2.54 (5.33)	0.66 (1.94)
	MDP 30 minutes	4.26 (8.45)	3.20 (6.13)	0.90 (2.33)
	No reallocations*	6.48 (11.01)	3.67 (6.60)	1.77 (4.14)

(b) Waiting donors at the collection site.

which will almost always result in an average number of staff members that is not a whole number. It is therefore hard to compare to a situation with the same average number of staff members scheduled throughout the day, as it is impossible to schedule a non-whole number of staff members. If the ILP works well, it schedules just enough

Table 9.5 The average number of reallocations of staff members per half hour, with the total number of working staff members between parentheses as a reference, for all three collection sites. Rows correspond to a combination of number of serviced donors per staff member per hour, $\mu = 3$, $\mu = 3.5$ or $\mu = 4$ and the MDP that can change the allocation of the staff members every 7.5 minutes, 15 minutes or 30 minutes. Columns correspond to those in Table 9.1

		ILP	Mean ILP rounded up	Max ILP
$\mu = 3$	MDP 7.5 minutes	0.81 (7.06)	1.17 (8)	0.89 (10)
	MDP 15 minutes	0.55 (7.06)	0.59 (8)	0.58 (10)
	MDP 30 minutes	0.12 (7.06)	0.04 (8)	0.13 (10)
$\mu = 3.5$	MDP 7.5 minutes	0.70 (5.35)	0.40 (6)	1.17 (8)
	MDP 15 minutes	0.33 (5.35)	0.09 (6)	0.59 (8)
	MDP 30 minutes	0.02 (5.35)	0.09 (6)	0.04 (8)
$\mu = 4$	MDP 7.5 minutes	0.77 (4.76)	0.10 (5)	1.35 (7)
	MDP 15 minutes	0.16 (4.76)	0.09 (5)	0.15 (7)
	MDP 30 minutes	0.02 (4.76)	0.04 (5)	0.06 (7)

(a) Reallocations For Nijmegen

		ILP	Mean ILP rounded up	Max ILP
$\mu = 3$	MDP 7.5 minutes	0.29 (5.60)	0.19 (6)	0.70 (8)
	MDP 15 minutes	0.07 (5.60)	0.05 (6)	0.30 (8)
	MDP 30 minutes	0.00 (5.60)	0.00 (6)	0.00 (8)
$\mu = 3.5$	MDP 7.5 minutes	0.07 (4.20)	0.05 (5)	0.05 (5)
	MDP 15 minutes	0.03 (4.20)	0.04 (5)	0.04 (5)
	MDP 30 minutes	0.00 (4.20)	0.00 (5)	0 (5)
$\mu = 4$	MDP 7.5 minutes	0.02 (3.64)	0.08 (4)	0.08 (4)
	MDP 15 minutes	0.01 (3.64)	0.02 (4)	0.02 (4)
	MDP 30 minutes	0.00 (3.64)	0.00 (4)	0.00 (4)

(b) Reallocations For Leiden

		ILP	Mean ILP rounded up	Max ILP
$\mu = 3$	MDP 7.5 minutes	1.14 (8.71)	0.93 (9)	0.40 (12)
	MDP 15 minutes	0.61 (8.71)	0.49 (9)	0.16 (12)
	MDP 30 minutes	0.36 (8.71)	0.06 (9)	0.08 (12)
$\mu = 3.5$	MDP 7.5 minutes	0.81 (6.41)	0.72 (7)	0.93 (9)
	MDP 15 minutes	0.24 (6.41)	0.13 (7)	0.49 (9)
	MDP 30 minutes	0.02 (6.41)	0.10 (7)	0.06 (9)
$\mu = 4$	MDP 7.5 minutes	0.86 (5.53)	1.14 (6)	1.79 (8)
	MDP 15 minutes	0.21 (5.53)	0.10 (6)	0.47 (8)
	MDP 30 minutes	0.02 (5.53)	0.10 (6)	0.04 (8)

(c) Reallocations For Almelo

staff members. Scheduling less staff members could result in excessive queues. We have therefore chosen to round up the number of staff members for the comparison scenario "rounded up mean". Even though this does mean that this comparison scenario often has an advantage, the maximum queue is lower with the ILP in most cases, combined with lower personnel cost.

The other comparison scenario is based on the maximum number of staff members scheduled by the ILP, which results in a lot more scheduled staff members. It is clearly visible that the queue lengths do not decrease proportionally, and scheduling this many staff members seems pointless.

In summary, it seems possible to decrease queues without increasing the hours worked by staff, by combining optimal shift planning with flexible staff allocation at blood collection site stations.

CHAPTER 10

Conclusion and outlook

As donors donate blood voluntarily and are non-remunerated (i.e. not compensated with money), the natural intuition is to give donors the best possible service. Additionally, there is a financial incentive for a high quality of service to donors, as donations by returning donors are less costly than recruiting new donors. A frequently mentioned reason for blood donors to stop donating is the experience of or at least perception of excessive waiting times at collection sites. This is supported by scientific research showing a negative association between return behavior of donors and waiting times [81, 132]. This thesis presents a number of approaches, based on methods from the field of Operations Research, to compute and decrease waiting times at blood collections sites, without the need for additional capacity.

Modeling blood collection sites with mathematical methods has only been addressed by a small number of references. Structured methods to analyze potential improvements are still of considerable interest. Part II of this thesis aims to provide methods for this purpose. Chapter 3 is focused on distributions of waiting time and delay time. A closed form expression is provided for the waiting time distribution of each individual station of the process, while a numerical approach is provided to compute the delay time for the entire process. The total delay time computation could be of interest for optimization approaches, but is too slow for this purpose. Chapter 4, therefore, provides an iterative approach, based on uniformization, to efficiently compute queueing distributions for blood collection sites. As this approach is much faster, it is used for the optimization approaches in subsequent chapters.

One of the main problems at Dutch blood collection sites, is the misalignment of staff and the expected arrivals of donors. As shown in our test cases, the number of available staff members is more or less uniform throughout the day, while donors show clear preferences for particular periods, sometimes resulting in a 150% higher arrival rate during the busiest periods compared to the quietest. This misalignment leads to a considerable amount wasted capacity. Simply lowering the staff capacity without a thorough look at the problem may cause waiting times to explode. Attempts can be made to tackle this problem using different approaches. One way is to use flexible shifts to align shifts of staff members such that these fit the arrival pattern, as is done in Chapter 5. Another option could be to strive for spreading the arrivals of donors over the day. This is done in Chapter 7, by proposing the introduction of appointments for whole blood donors. Both approaches are able to accomplish less waiting times without adding more staff. However the current proposal by Sanquin

for the priorities of appointments might cause serious problems.

If service to donors is the main concern of Sanquin, forcing its donors into a new, appointment based, arrival system might not be a desirable option, and might even lose Sanquin some of its donors. A combination of appointments and more effective shift planning is another an option. In Chapter 7, the number of slots that is made available is equal to the number of donations that is wanted for plus some percentage of no-shows. An interesting scenario for further research is the option to make more slots available, such that even the last donor to make an appointment has a choice for the time of his/her appointment, and is not forced into the last available slot. Subsequently, methods like the one proposed in Chapter 5 could be used to fit the shifts of staff members to the appointments that have been made. This might improve the results of the staff scheduling, as it will then be based on a far more fixed arrival pattern. A disadvantage will be that staff members should then be scheduled after most, or all, appointments have been made.

While Chapters 5 and 7 are both focused on optimizing the blood collection process before it takes place, on the tactical to off-line operational level, Chapter 6 discusses an approach to optimize the assignment of staff during the collection process, on the on-line operational level. Based on the actual number of present donors, the algorithm presented in Chapter 6 may determine that it is optimal to change the allocation of one or more staff members during a session. Although the algorithm is based on exponential and preemptive assumptions, simulation results of a blood collection site show similar results: reductions of up to 70% on queue lengths.

Chapter 8 follows the route of blood products from donor to recipient one step further to the inventory management. Chapter 8 discusses an entirely new way of allocating blood products in the inventory of the blood bank to requests from hospitals. Based on the rarity and age of a unit of red blood cells, an artificial cost is determined for the match between units of red blood cells and the request. A subsequent min-cost-max-flow algorithm determines the matches between units of cells and requests. The approach is able to significantly improve the ratio of requests that is given an exact match, while simultaneously decreasing out-dating and shortages. As the algorithm also determines the donors that should be sent a request to donate, this comes back to the process at the blood collection site.

Chapter 9 concludes the research based chapters of this thesis by the application and combination of the approaches discussed in Chapters 5 and 6. The approaches were applied to data from three collection sites in the Netherlands: the fixed collection sites in Leiden and Nijmegen, and the mobile collection site in Almelo. The numerical experiments with this data show that the approach from Chapter 5 is able to reduce the peak queues at blood collection sites, while the method from Chapter 6 decreases both the average and maximum queue lengths at blood collection sites.

Most approaches in this thesis assume exponential inter-arrival and service times. While exponential inter-arrival times are not disputed, exponential service times do not fully represent reality, as realistic service times are closer to a log-normal distribution. Although a direct insertion of log-normal distributions in the presented methods is impossible, such distributions could be approximated by phase-type dis-

tributions, which can be inserted into the methods. However, this would come at very high computational costs. The methods in Part III of this thesis are fast enough to be implemented in practice. This would not remain the case with the inclusion of phase-type distributions, unless significant improvements are made to CPU speeds or optimization algorithms. Additionally, the simulation in Chapter 6 shows that the exponential assumption does not seem to have a large effect on the results in terms of queue reductions.

In all chapters of this thesis, the demand of blood products from hospitals is taken as a given, and some of the methods need to be adapted slightly if the demand suddenly rises or falls. The demand will always remain stochastic, but it can be partly predicted. Taking a demand prediction into account could improve the approaches presented in this thesis, most notably the inventory management and staff scheduling.

The research in this thesis focused on the operational and tactical levels. Next to the major challenge of implementation, the long term strategic level also offers opportunities for further research. The demand for red blood cells is decreasing, while the demand for plasma is increasing. This shift in demand requires Sanquin to adapt to new circumstances, most importantly because plasma can currently not be collected in mobile collection sites. Issues like the optimal location of collection sites and opening times of collection sites will likely be affected by this shift in demand, and have not been studied in this thesis. Some neighboring countries have created specialized plasma collection centers, another option that could be studied by Sanquin.

Entirely different fields of research can even be used for the issues studied in this thesis. There are motivations for donors to visit a blood collection site at at a particular time are not well understood. This thesis may have speculated at times to what these motivations might be, but they are essentially unknown. The motivations might even be influenced, resulting in more favorable arrival patterns. Studying the reasons behind the arrival patterns could be very interesting, and remains an open question for psychologists and statisticians to study.

In summary, this thesis provides a number of approaches to compute and decrease queues at blood collection sites. The chapters in Part II first develop methods to evaluate waiting times and queues, followed by optimization and improvement techniques in Part III. The approaches in this final part could all be implemented at blood collection sites, although the required effort differs. The staff scheduling is already done centrally, and could therefore easily be improved by the approach in Chapter 5. On the other side, the reallocation of staff members from Chapter 6 would require real-time tracking of donors in the process, which will require substantial investment. Supported by the final results in Chapter 9, this thesis shows that substantial reductions of queues are possible without additional capacity or staff.

Bibliography

[1] U. Abdulwahab and M. I. M. Wahab. Approximate dynamic programming modeling for a typical blood platelet bank. *Computers & Industrial Engineering*, 78:259–270, 2014.

[2] S. R. Agnihothri and P. F. Taylor. Staffing a centralized appointment scheduling department in Lourdes-hospital. *Interfaces*, 21(5):1–11, 1991.

[3] A. Ahmadi-Javid, Z. Jalali, and K. J. Klassen. Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research*, 258(1):3 – 34, 2017. ISSN 0377-2217.

[4] M. A. Ahmed and T. M. Alkhamis. Simulation optimization for an emergency department healthcare unit in Kuwait. *European Journal of Operational Research*, 198 (3):936–942, 2009.

[5] H.-S. Ahn and R. Righter. Dynamic load balancing with flexible workers. *Advances in Applied Probability*, 38(3):621–642, 2006.

[6] E. Alfonso, X. Xiaolan, and V. Augusto. A queueing network approach for appointment scheduling of blood donors. *IFAC Proceedings Volumes*, 45(6):303–308, 2012.

[7] E. Alfonso, X. Xie, V. Augusto, and O. Garraud. Modeling and simulation of blood collection systems. *Health Care Management Science*, 15(1):63–78, 2012.

[8] E. Alfonso, X. Xie, V. Augusto, and O. Garraud. Modelling and simulation of blood collection systems: improvement of human resources allocation for better cost-effectiveness and reduction of candidate donor abandonment. *Vox Sanguinis*, 104(3): 225–233, 2013.

[9] E. Alfonso, X. Xie, and V. Augusto. A simulation-optimization approach for capacity planning and appointment scheduling of blood donors based on mathematical programming representation of event dynamics. In *2015 IEEE International Conference on Automation Science and Engineering*, pages 728–733. IEEE, 2015.

[10] S. Andradottir and H. Ayhan. Throughput maximization for tandem lines with two stations and flexible servers. *Operations Research*, 53(3):516–531, 2005.

[11] S. Andradottir, H. Ayhan, and D. G. Down. Server assignment policies for maximizing the steady-state throughput of finite queueing systems. *Management Science*, 47(10): 1421–1439, 2001.

[12] S. Andradottir, H. Ayhan, and E. Kirkizlar. Flexible servers in tandem lines with setup costs. *Queueing Systems*, 70(2):165–186, 2012.

[13] S. Andradottir, H. Ayhan, and D. G. Down. Optimal assignment of servers to tasks when collaboration is inefficient. *Queueing Systems*, 75(1):79–110, 2013.

[14] A. Andreychenko, P. Crouzen, L. Mikeev, and V. Wolf. On-the-fly uniformization of time-inhomogeneous infinite Markov population models. *arXiv preprint arXiv:1006.4425*, 2010.

[15] V. de Angelis, G. Felici, and P. Impelluso. Integrating simulation and optimisation in health care centre management. *European Journal of Operational Research*, 150(1): 101–114, 2003.

[16] M. Arns, P. Buchholz, and A. Panchenko. On the numerical analysis of inhomogeneous continuous-time Markov chains. *Informs Journal on Computing*, 22(3):416–432, 2010.

[17] R. Arumugam, M. E. Mayorga, and K. M. Taaffe. Inventory based allocation policies for flexible servers in serial systems. *Annals of Operations Research*, 172(1):1–23, 2008.

[18] R. G. Askin and J. Chen. Dynamic task assignment for throughput maximization with worksharing. *European Journal of Operational Research*, 168(3):853–869, 2006.

[19] M. P. Atkinson, M. J. Fontaine, L. T. Goodnough, and L. M. Wein. A novel allocation strategy for blood transfusions: investigating the tradeoff between the age and availability of transfused blood. *Transfusion*, 52(1):108–117, 2012.

[20] S. Baş, G. Carello, E. Lanzarone, Z. Ocak, and S. YalÃğÄśndağ. *Management of Blood Donation System: Literature Review and Research Perspectives*, pages 121–132. Springer, 2016.

[21] C. Baier, B. Haverkort, H. Hermanns, and J.-P. Katoen. Model checking continuous-time Markov chains by transient analysis. In *International Conference on Computer Aided Verification*, pages 358–372. Springer, 2000.

[22] J. J. Bartholdi and D. D. Eisenstein. A production line that balances itself. *Operations Research*, 44(1):21–34, 1996.

[23] J. J. Bartholdi, D. D. Eisenstein, and R. D. Foley. Performance of bucket brigades when work is stochastic. *Operations Research*, 49(5):710–719, 2001.

[24] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *Journal of the ACM*, 22(2): 248–260, 1975.

[25] J. Beliën and H. Forcé. Supply chain management of blood products: A literature review. *European Journal of Operational Research*, 217(1):1–16, 2012.

[26] J. C. Bennett and D. J. Worthington. An example of a good but partially successful OR engagement: Improving outpatient clinic operations. *Interfaces*, 28(5):56–69, 1998.

[27] S. Bhulai, A. C. Brooms, and F. M. Spieksma. On structural properties of the value function for an unbounded jump Markov process with an application to a processor sharing retrial queue. *Queueing Systems*, 76(4):425–446, 2014.

[28] D. P. Bischak. Performance of a manufacturing module with moving workers. *IIE transactions*, 28(9):723–734, 1996.

[29] J. T. Blake and S. Shimla. Determining staffing requirements for blood donor clinics: the Canadian Blood Services experience. *Transfusion*, 54(3pt2):814–820, 2014.

[30] J. T. Blake, M. Hardy, G. Delage, and G. Myhal. Déjà-vu all over again: using simulation to evaluate the impact of shorter shelf life for red blood cells at Héma-Québec. *Transfusion*, 53(7):1544–1558, 2013.

[31] H. Blok and E.M. Spieksma. Structures of optimal policies in Markov decision processes with unbounded jumps: the state of our art. *Markov Decision Processes in Practice*, pages 139–196, 2015.

[32] V. Bosnes, M. Aldrin, and H. E. Heier. Predicting blood donor arrival. *Transfusion*, 45(2):162–170, 2005.

[33] R. J. Boucherie and N. M. van Dijk. On a queueing network model for cellular mobile telecommunications networks. *Operations Research*, 48(1):38–49, 2000.

[34] R. J. Boucherie and N. M. van Dijk. *Queueing networks: a fundamental approach*, volume 154. Springer Science & Business Media, 2010.

[35] O. J. Boxma and H. Daduna. *Sojourn Times in Queueing Networks*. Centre for Mathematics and Computer Science, 1989. (Amsterdam, Netherlands).

[36] M. Brahimi and D. J. Worthington. Queuing models for outpatient appointment systems - a case-study. *Journal of the Operational Research Society*, 42(9):733–746, 1991.

[37] M. Brahimi and D. J. Worthington. The finite capacity multi-server queue with inhomogeneous arrival rate and discrete service time distribution âĂŤ and its application to continuous service time problems. *European Journal of Operational Research*, 50 (3):310–324, 1991.

[38] J. E. Brennan, B. L. Golden, and H. K. Rappoport. Go with the flow: Improving red cross bloodmobiles using simulation analysis. *Interfaces*, 22(5):1–13, 1992.

[39] K. M. Bretthauer and M. J. Cote. A model for planning resource requirements in health care organizations. *Decision Sciences*, 29(1):243–270, 1998.

[40] A. M. de Bruin, A. C. van Rossum, M. C. Visser, and G. M. Koole. Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science*, 10(2):125–137, 2007.

[41] P. J. Burke. The dependence of delays in tandem queues. *The Annals of Mathematical Statistics*, pages 874–875, 1964.

[42] J. A. Buzacott. Automatic transfer lines with buffer stocks. *International Journal of Production Research*, 5(3):183–200, 1967.

[43] T. Cayirli and E. Veral. Outpatient scheduling in health care: a review of literature. *Production and operations management*, 12(4):519–549, 2003.

[44] CBO. Blood transfusion guidelines, 2011. `https://www.sanquin.nl/repository/documenten/en/prod-en-dienst/287294/blood-transfusion-guideline.pdf` [Accessed: September 2017].

[45] M. A. Centeno, R. Giachetti, R. Linn, and A. M. Ismail. Emergency departments II: a simulation-ILP based tool for scheduling ER staff. In *Proceedings of the 35th conference on Winter simulation: driving innovation*, pages 1930–1938. Winter Simulation Conference, 2003.

[46] C. Chai-Adisaksopha, P. E. Alexander, G. Guyatt, M. A. Crowther, N. M. Heddle, P. J. Devereaux, M. Ellis, D. Roxby, D. I. Sessler, and J. W. Eikelboom. Mortality outcomes in patients transfused with fresher versus older red blood cells: a meta-analysis. *Vox Sanguinis*, 112(3):268–278, 2017.

[47] K. M. Chandy and A. J. Martin. A characterization of product-form queuing networks. *Journal of the ACM*, 30(2):286–299, 1983.

[48] J. Chen and R. G. Askin. Throughput maximization in serial production lines with worksharing. *International Journal of Production Economics*, 99(1âĂŞ2):88–101, 2006.

[49] K.L. Chung. *Markov chains with stationary transition probabilities*. Springer-Verlag, 1967.

[50] Y. T. Chung and F. Erhun. Designing supply contracts for perishable goods with two periods of shelf life. *IIE Transactions*, 45(1):53–67, 2013.

[51] I. Civelek, I. Karaesmen, and A. Scheller-Wolf. Blood platelet inventory management with protection levels. *European Journal of Operational Research*, 243(3):826–838, 2015.

[52] T. Collings and C. Stoneman. $m|m|\infty$ queue with varying arrival and departure rates. *Operations Research*, 24(4):760–773, 1976.

[53] R. B. Cooper. *Introduction to queueing theory*. Macmillan, 1972.

[54] S. Creemers, M. Defraeye, and I. van Nieuwenhuyse. G-RAND: A phase-type approximation for the nonstationary $G(t)|G(t)|s(t) + G(t)$ queue. *Performance Evaluation*, 80:102 – 123, 2014. ISSN 0166-5316. 'SI: Service Science of Queues.

[55] Y. Dallery and S. B. Gershwin. Manufacturing flow line systems: a review of models and analytical results. *Queueing Systems*, 12(1-2):3–94, 1992.

[56] M. Defraeye and I. van Nieuwenhuyse. Controlling excessive waiting times in small service systems with time-varying demand: An extension of the ISA algorithm. *Decision Support Systems*, 54(4):1558–1567, 2013.

[57] M. Defraeye and I. van Nieuwenhuyse. Staffing and scheduling under nonstationary demand for service: A literature review. *Omega*, 58:4–25, 2016. ISSN 0305-0483.

[58] F. Didier, T. A. Henzinger, M. Mateescu, and V. Wolf. Fast adaptive uniformization of the chemical master equation. In *High Performance Computational Systems Biology, 2009. HIBI'09. International Workshop on*, pages 118–127. IEEE, 2009.

[59] J. D. Diener and W. H. Sanders. Empirical comparison of uniformization methods for continuous-time Markov chains. In *Computations with Markov chains*, pages 547–570. Springer, 1995.

[60] N. M. van Dijk. *Controlled Markov processes: time discretization*, volume 1CWI Tract 11. Centrum voor Wiskunde en Informatica, 1984.

[61] N. M. van Dijk. A note on extended uniformization for non-exponential stochastic networks. *Journal of applied probability*, pages 955–961, 1991.

[62] N. M. van Dijk. Transient error bound analysis for continuous-time Markov reward structures. *Performance evaluation*, 13(3):147–158, 1991.

[63] N. M. van Dijk. Approximate uniformization for continuous-time Markov chains with an application to performability analysis. *Stochastic processes and their applications*, 40(2):339–357, 1992.

[64] N. M. van Dijk. Uniformization for nonhomogeneous Markov chains. *Operations research letters*, 12(5):283–291, 1992.

[65] N. M. van Dijk. *Queueing networks and product forms: a systems approach*, volume 4. John Wiley & Son Limited, 1993.

[66] N. M. van Dijk and N. Kortbeek. Erlang loss bounds for OTâĂŞICU systems. *Queueing Systems*, 63(1-4):253–280, 2009.

[67] N. M. van Dijk and M. L. Puterman. Perturbation theory for Markov reward processes with applications to queueing systems. *Advances in applied probability*, 20(1):79–98, 1988.

[68] N. M. van Dijk, R. Haijema, J. van der Wal, and C. S. Sibinga. Blood platelet production: a novel approach for practical optimization. *Transfusion*, 49(3):411–420, 2009.

[69] M. Dillon, F. Oliveira, and B. Abbasi. A two-stage stochastic programming model for inventory management in the blood supply chain. *International Journal of Production Economics*, 187:27–41, 2017.

[70] G. Dobson, S. Hasija, and E. J. Pinker. Reserving capacity for urgent patients in primary care. *Production and Operations Management*, 20(3):456–473, 2011. ISSN 1937-5956.

[71] D. W. Dormuth and A. S. Alfa. Two finite-difference methods for solving $map(t)|ph(t)|1|k$ queueing models. *Queueing Systems*, 27(1):55–78, 1997. ISSN 1572-9443.

[72] Q. Duan and T. W. Liao. A new age-based replenishment policy for supply chain inventory optimization of highly perishable products. *International journal of production economics*, 145(2):658–671, 2013.

[73] Q. Duan and T. W. Liao. Optimization of blood supply chain with shortened shelf lives and ABO compatibility. *International Journal of Production Economics*, 153:113 – 129, 2014.

[74] I. Duenyas, D. Gupta, and T. L. Olsen. Control of a single-server tandem queueing system with setups. *Operations Research*, 46(2):218–230, 1998.

[75] E. B. Dynkin. *Markov processes*. Springer, 1965.

[76] W. K. Ehrlich, R. Hariharan, P. K. Reeser, and R. D. van der Mei. *Performance of Web servers in a distributed computing environment*, volume Volume 4, pages 137–148. Elsevier, 2001.

[77] S. G. Eick, W. A. Massey, and W. Whitt. $m_t|g|\infty$ queues with sinusoidal arrival rates. *Management Science*, 39(2):241–252, 1993.

[78] G. Erdogan, E. Erkut, A. Ingolfsson, and G. Laporte. Scheduling ambulance crews for maximum coverage. *Journal of the Operational Research Society*, 61(4):543–550, 2010.

[79] A. K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *The Post Office Electrical Engineers Journal*, 10: 189–197, 1917.

[80] A. T. Ernst, H. Jiang, M. Krishnamoorthy, and D. Sier. Staff scheduling and rostering: A review of applications, methods and models. *European Journal of Operational Research*, 153(1):3–27, 2004.

[81] E. Ferguson. Predictors of future behaviour: a review of the psychological literature on blood donation. *British Journal of Health Psychology*, 1(4):287–308, 1996.

[82] W. A. Flegel, C. Natanson, and H. G. Klein. Does prolonged storage of red blood cells cause harm? *British journal of haematology*, 165(1):3–16, 2014.

[83] M. J. Fontaine, Y. T Chung, W. M. Rogers, H. D. Sussmann, P. Quach, S. A. Galel, L. T. Goodnough, and F. Erhun. Improving platelet supply chains through collaborations between blood centers and transfusion services. *Transfusion*, 49(10): 2040–2047, 2009.

[84] B. L. Fox and P. W. Glynn. Computing Poisson probabilities. *Communications of the ACM*, 31(4):440–445, 1988.

[85] A. Fukunaga, E. Hamilton, J. Fama, D. Andre, O. Matan, and I. Nourbakhsh. Staff scheduling for inbound call centers and customer contact centers. *AI Magazine*, 23 (4):30–40, 2002.

[86] E. S. Gel, W. J. Hopp, and M. P. van Oyen. Factors affecting opportunity of worksharing as a dynamic line balancing mechanism. *IIE Transactions*, 34(10):847–863, 2002.

[87] Paul L. F. Giangrande. The history of blood transfusion. *British Journal of Haematology*, 110(4):758–767, 2000. ISSN 1365-2141.

[88] I. I. Gihman and A. V. Skorohod. *Controlled stochastic processes*. Springer Science & Business Media, 2012.

[89] I. I. Gihman and A. V. Skorokhod. *Introduction to the theory of random processes*. WB Saunders Philadelphia, 1969.

[90] J. Gillard and V. Knight. Using singular spectrum analysis to obtain staffing level requirements in emergency units. *Journal of the Operational Research Society*, 65 (5):735–746, 2014.

[91] W. J. Gordon and G. F. Newell. Closed queuing systems with exponential servers. *Operations Research*, 15(2):254–265, 1967.

[92] W. J. Gordon and G. F. Newell. Cyclic queuing systems with restricted length queues. *Operations Research*, 15(2):266–277, 1967.

[93] A. Goyal, S. S. Lavenberg, and K. S. Trivedi. Probabilistic modeling of computer system availability. *Annals of Operations Research*, 8(1):285–306, 1987.

[94] W. K. Grassmann. Transient solutions in Markovian queues. *European Journal of Operational Research*, 1(6):396–402, 1977.

[95] W. K Grassmann. Transient solutions in Markovian queueing systems. *Computers & Operations Research*, 4(1):47–53, 1977.

[96] W. K. Grassmann. Finding transient solutions in Markovian event systems through randomization. *Numerical solution of Markov chains*, 8:37–61, 1991.

[97] L. V. Green, S. Savin, and B. Wang. Managing patient service in a diagnostic medical facility. *Operations Research*, 54(1):11–25, 2006.

[98] L. V. Green, J. Soares, J. F. Giglio, and R. A. Green. Using queueing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine*, 13(1):61–68, 2006.

[99] D. Gross and D. R. Miller. The randomization technique as a modeling tool and solution procedure for transient Markov processes. *Operations Research*, 32(2):343–361, 1984.

[100] S. Gunpinar and G. Centeno. Stochastic integer programming models for reducing wastages and shortages of blood products at hospitals. *Computers & Operations Research*, 54:129–141, 2015.

[101] X. Guo, O. Hernández-Lerma, T. Prieto-Rumeau, X.-R. Cao, J. Zhang, Q. Hu, M. E. Lewis, and R. Vélez. A survey of recent results on continuous-time Markov decision processes. *Top*, 14(2):177–261, 2006.

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

[103] X. Haifeng, T. Chaussalet, and M. Rees. A semi-open queueing network approach to the analysis of patient flow in healthcare systems, 20-22 June 2007 2007.

[104] R. Haijema, J. van der Wal, and N. M. van Dijk. Blood platelet production: Optimization by dynamic programming and simulation. *Computers & Operations Research*, 34(3):760–779, 2007.

[105] R. Haijema, N. M. van Dijk, J. van der Wal, and C. S. Sibinga. Blood platelet production with breaks: optimization by SDP and simulation. *International Journal of Production Economics*, 121(2):464–473, 2009.

[106] J. J. Hasenbein and B. Kim. Throughput maximization for two station tandem systems: a proof of the Andradottir-Ayhan conjecture. *Queueing Systems*, 67(4): 365–386, 2011.

[107] H. Hermanns, U. Herzog, and J.-P. Katoen. Process algebra for performance evaluation. *Theoretical computer science*, 274(1):43–87, 2002.

[108] J. Hillston. *A compositional approach to performance modelling*, volume 12. Cambridge University Press, 2005.

[109] A. Hordijk and R. Schassberger. Weak convergence for generalized semi-Markov processes. *Stochastic Processes and their Applications*, 12(3):271–291, 1982.

[110] P. J. H. Hulshof, N. Kortbeek, R. J. Boucherie, E. W. Hans, and P. J. M. Bakker. Taxonomic classification of planning decisions in health care: a structured review of the state of the art in OR/MS. *Health Systems*, 1(2):129–175, 2012.

[111] A. Ingolfsson, E. Akhmetshina, S. Budge, Y. Li, and X. Wu. A survey and experimental comparison of service-level-approximation methods for nonstationary $m(t)|m|s(t)$ queueing systems with exhaustive discipline. *INFORMS Journal on Computing*, 19 (2):201–214, 2007.

[112] A. Ingolfsson, F. Campello, X. Wu, and E. Cabral. Combining integer programming and the randomization method to schedule employees. *European Journal of Operational Research*, 202(1):153–163, 2010.

[113] N. Izady and D. J. Worthington. Setting staffing requirements for time dependent queueing networks: The case of accident and emergency departments. *European Journal of Operational Research*, 219(3):531–540, 2012.

[114] J. R. Jackson. Networks of waiting lines. *Operations Research*, 5(4):518–521, 1957.

[115] J. R. Jackson. Jobshop-like queueing systems. *Management Science*, 10(1):131–142, 1963.

[116] D. L. Jagerman. Nonstationary blocking in telephone traffic. *Bell Labs Technical Journal*, 54(3):625–661, 1975.

[117] A. Jensen. Markoff chains as an aid in the study of Markoff processes. *Scandinavian Actuarial Journal*, 1953(sup1):87–91, 1953.

[118] F. P. Kelly. *Reversibility and Stochastic Networks*. Cambridge University Press, 2011.

[119] E. Kirkizlar, S. Andradottir, and H. Ayhan. Robustness of efficient server assignment policies to service time distributions in finiteâĂŘbuffered lines. *Naval Research Logistics*, 57(6):563–582, 2010.

[120] E. Kirkizlar, S. Andradottir, and H. Ayhan. Flexible servers in understaffed tandem lines. *Production and Operations Management*, 21(4):761–777, 2012.

[121] P. M. Koeleman and G. M. Koole. Optimal outpatient appointment scheduling with emergency arrivals and general service times. *IIE Transactions on Healthcare Systems Engineering*, 2(1):14–30, 2012.

[122] A. Kolmogorov. Sur le probleme d'attente. *Matematicheskii Sbornik*, 38(1-2):101–106, 1931.

[123] N. Kortbeek, M. E. Zonderland, A. Braaksma, I. M. H. Vliegen, R. J. Boucherie, N. Litvak, and E. W. Hans. Designing cyclic appointment schedules for outpatient clinics with scheduled and unscheduled patient arrivals. *Performance Evaluation*, 80: 5–26, 2014.

[124] M. Kwiatkowska, G. Norman, and D. Parker. Stochastic model checking. In *International School on Formal Methods for the Design of Computer, Communication and Software Systems*, pages 220–270. Springer, 2007.

[125] G. Latouche and V. Ramaswami. *Introduction to matrix analytic methods in stochastic modeling*. SIAM, 1999.

[126] M. Loève. Probability theory, volume II. *Graduate texts in mathematics*, 46:311, 1978.

[127] J. Luo, V. G. Kulkarni, and S. Ziya. Appointment scheduling under patient no-shows and service interruptions. *Manufacturing & Service Operations Management*, 14(4): 670–684, 2012.

[128] M. Malhotra, J. K. Muppala, and K. S. Trivedi. Stiffness-tolerant methods for transient analysis of stiff Markov chains. *Microelectronics Reliability*, 34(11):1825–1841, 1994.

[129] N. S. R. Maluf. History of blood transfusion. *Journal of the History of Medicine and Allied Sciences*, 9(1):59–107, 1954. ISSN 00225045, 14684373. URL `http://www.jstor.org/stable/24619834`.

[130] J. O. McClain, L. J. Thomas, and C. Sox. "on-the-fly" line balancing with very little WIP. *International Journal of Production Economics*, 27(3):283–289, 1992.

[131] J. O. McClain, K. L. Schultz, and L. J. Thomas. Management of worksharing systems. *Manufacturing and Service Operations Management*, 2(1):49–67, 2000.

[132] T. McKeever, M. R. Sweeney, and A. Staines. An investigation of the impact of prolonged waiting times on blood donors in Ireland. *Vox Sanguinis*, 90(2):113–8, 2006.

[133] I van Mechelen and P. L. M. Zonneveld. Capaciteitplanning en wachttijdbepaling, 2013.

[134] B. Melamed and M. Yadin. Randomization procedures in the computation of cumulative-time distributions over discrete state Markov processes. *Operations Research*, 32(4):926–944, 1984.

[135] A. J. Meulenbroek and A. N. C. J. de Regt. *Van bloed tot geneesmiddel; Van donor tot patiÃńnt*. Amsterdam, 2007.

[136] J. D. Michaels, J. E. Brennan, B. L. Golden, and M. C. Fu. A simulation study of donor scheduling systems for the American red cross. *Computers & Operations Research*, 20(2):199–213, 1993.

[137] R. A. Middelburg, L. M. G. van de Watering, E. Briët, and J. G. van der Bom. Storage time of red blood cells and mortality of transfusion recipients. *Transfusion medicine reviews*, 27(1):36–43, 2013.

[138] S. K. Mok and J. G. Shanthikumar. A transient queueing model for business office with standby servers. *European Journal of Operational Research*, 28(2):158–174, 1987.

[139] A. P. A. van Moorsel and W. H. Sanders. Adaptive uniformization. *Stochastic Models*, 10(3):619–647, 1994.

[140] A. P. A. van Moorsel and W. H. Sanders. Transient solution of Markov models by combining adaptive and standard uniformization. *IEEE Transactions on Reliability*, 46(3):430–440, 1997.

[141] A. P. A. van Moorsel and K. Wolter. Numerical solution of non-homogeneous Markov processes through uniformization. In *ESM*, pages 710–717, 1998.

[142] J. K. Muppala and K. S. Trivedi. Numerical transient solution of finite Markovian queueing systems. In *Queueing and Related Models*. 1992.

[143] A. Nagurney, A. H. Masoumi, and M. Yu. Supply chain network operations management of a blood banking system with cost and risk minimization. *Computational Management Science*, 9(2):205–231, 2012.

[144] S. Nahmias. Perishable inventory theory: A review. *Operations research*, 30(4): 680–708, 1982.

[145] S. Nahmias. *Perishable inventory systems*, volume 160. Springer Science & Business Media, 2011.

[146] M. F. Neuts. *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Courier Corporation, 1981.

[147] M. F. Neuts and J.-M. Li. An algorithm for the $p(n,t)$ matrices of a continuous BMAP, 1996.

[148] A. R. Odoni. On finding the maximal gain for Markov decision processes. *Operations Research*, 17(5):857–860, 1969.

[149] A. F. Osorio, S. C. Brailsford, and H. K. Smith. A structured review of quantitative models in the blood supply chain: a taxonomic framework for decision-making. *International Journal of Production Research*, 53(24):7191–7212, 2015.

[150] Jose Ostolaza, L. J. Thomas, and J. O. McClain. The use of dynamic (state-dependent) assembly-line balancing to improve throughput. *Journal of Manufacturing Operations Management*, 3:105–133, 1990.

[151] Y. Peng, X. Qu, and J. Shi. A hybrid simulation and genetic algorithm approach to determine the optimal scheduling templates for open access clinics admitting walk-in patients. *Computers and Industrial Engineering*, 72:282 – 296, 2014. ISSN 0360-8352.

[152] H. G. Perros. *Queueing networks with blocking*. Oxford University Press, Inc., 1994.

[153] B. Pittel. Closed exponential networks of queues with saturation: The Jackson-type stationary distribution and its asymptotic analysis. *Mathematics of Operations Research*, 4(4):357–378, 1979.

[154] M. L. Pratt and A. J. Grindon. Computer simulation analysis of blood donor queueing problems. *Transfusion*, 22(3):234–7, 1982.

[155] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley and Sons, 2005.

[156] S. Rajendran and A. R. Ravindran. Platelet ordering policies at hospitals using stochastic integer programming model and heuristic approaches to reduce wastage. *Computers & Industrial Engineering*, 2017.

[157] A. Reibman and K. Trivedi. Numerical transient analysis of Markov models. *Computers & Operations Research*, 15(1):19–36, 1988.

[158] A. Reibman, R. Smith, and K. Trivedi. Markov and Markov reward model transient analysis: An overview of numerical approaches. *European Journal of Operational Research*, 40(2):257–267, 1989.

[159] E. Reich. Waiting times when queues are in tandem. *The Annals of Mathematical Statistics*, 28(3):768–773, 1957.

[160] M. E. Reid, C. Lomas-Francis, and M. L. Olsson. *The blood group antigen factsbook*. Academic Press, 2012.

[161] T. A. Reilly, V. P. Marathe, and B. E. Fries. A delay-scheduling model for patients using a walk-in clinic. *Journal of Medical Systems*, 2(4):303–313, 1978.

[162] A. Rindos, S. Woolet, I. Viniotis, and K. Trivedi. Exact methods for the transient analysis of nonhomogeneous continuous time Markov chains. In *Computations with Markov chains*, pages 121–133. Springer, 1995.

[163] Sanquin Blood Supply Foundation. Annual report, 2010-2016. `https://www.sanquin.nl/en/about/about-sanquin/annual-reports/` [Accessed: September 2017].

[164] J. A. Schwarz, G. Selinka, and R. Stolletz. Performance analysis of time-dependent queueing systems: survey and classification. *Omega*, 63:170–189, 2016.

[165] L. I. Sennott, M. P. van Oyen, and S. M. R. Iravani. Optimal dynamic assignment of a flexible worker on an open production line with specialists. *European Journal of Operational Research*, 170(2):541–566, 2006.

[166] S. Sickinger and R. 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, 2009.

[167] M. Singer and P. Donoso. Assessing an ambulance service with queuing theory. *Computers and Operations Research*, 35(8):2549–2560, 2008.

[168] D. Sinreich and O. Jabali. Staggered work shifts: a way to downsize and restructure an emergency department workforce yet maintain current operational performance. *Health Care Management Science*, 10(3):293–308, 2007.

[169] S. H. W. Stanger, R. Wilding, N. Yates, and S. Cotton. What drives perishable inventory management performance? lessons learnt from the UK blood supply chain. *Supply Chain Management: An International Journal*, 17(2):107–123, 2012.

[170] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Springer-Verlag, 2002.

[171] S. Su and C.-L. Shih. Managing a mixed-registration-type appointment system in outpatient clinics. *International Journal of Medical Informatics*, 70(1):31–40, 2003.

[172] P. G. Taylor. *Insensitivity in Stochastic Models*, pages 121–140. Springer, 2010.

[173] M. C. Testik, B. Y. Ozkaya, S. Aksu, and O. I. Ozcebe. Discovering blood donor arrival patterns using data mining: a method to investigate service quality at blood centers. *Journal of Medical Systems*, 36(2):579–94, 2012.

[174] H. C. Tijms. *Stochastic modelling and analysis: a computational approach*. John Wiley & Sons, Inc., 1986.

[175] K. van den Toren, K. Habets, and F. Atsma. Wacht- en doorlooptijden noordoost. 2010.

[176] P. T. Vanberkel, R. J. Boucherie, E. W. Hans, and J. L. Hurink. Optimizing the strategic patient mix combining queueing theory and dynamic programming. *Computers and Operations Research*, 43:271–279, 2014.

[177] X. C. Wang, S. Andradottir, and H. Ayhan. Dynamic server assignment with task-dependent server synergy. *IEEE Transactions on Automatic Control*, 60(2):570–575, 2015.

[178] L. M. G. van de Watering. Age of blood: does older blood yield poorer outcomes? *Current opinion in hematology*, 20(6):526–532, 2013.

[179] W. van der Weij, R. van Der Mei, and B. G. F. Phillipson. Optimal server assignment in a two-layered tandem of multi-server queues. In *Proceedings 3rd International Working Conference on Performance Modelling and Evaluation of Heterogeneous Networks*, volume 1. Citeseer, 2005.

[180] W. van der Weij, N. M. van Dijk, and R. D. van der Mei. Product-form results for two-station networks with shared resources. *Performance Evaluation*, 69(12):662–683, 2012.

[181] W. Whitt. Approximations of dynamic programs. *Mathematics of Operations Research*, 3(3):231–243, 1978.

[182] W. Whitt. The queueing network analyzer. *Bell System Technical Journal*, 62(9): 2779–2815, 1983.

[183] W. L. Winston. *Operations Research: Applications and Algorithms*. Thomson Brooks/Cole, 2004.

[184] T. van Woensel, R. Andriansyah, F. R. B. Cruz, J. M. Smith, and L. Kerbache. Buffer and server allocation in general multi-server queueing networks. *International Transactions in Operational Research*, 17(2):257–286, 2010.

[185] R. W. Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2):223–231, 1982.

[186] E. Yale. First blood transfusion: A history, 2015. URL `https://daily.jstor.org/first-blood-transfusion/`.

[187] G. Yamazaki, H. Sakasegawa, and J. G. Shanthikumar. On optimal arrangement of stations in a tandem queueing system with blocking. *Management Science*, 38(1): 137–153, 1992.

[188] G. B. Yom-Tov. *Queues in hospitals: Semi-open queueing networks in the QED regime*. PhD thesis, 2007.

[189] G. B. Yom-Tov and A. Mandelbaum. Erlang-R: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing and Service Operations Management*, 16(2):283–299, 2014.

[190] E. Zavadlav, J. O. McClain, and L. J. Thomas. Self-buffering, self-balancing, self-flushing production lines. *Management Science*, 42(8):1151–1164, 1996.

[191] S. Zeltyn, Y. N. Marmor, A. Mandelbaum, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, A. Shtub, T. Lauterman, D. Schwartz, K. Moskovitch, S. Tzafrir, and F. Basis. Simulation-based models of emergency departments: Operational, tactical, and strategic staffing. *ACM Transactions on Modeling and computer simulation*, 21(4):1–25, 2011.

Summary

This thesis consists of four parts: The first part contains an introduction, the second presents approaches for the evalutation of waiting times at blood collecction sites, the third uses these to present approaches that improve waiting times at blood collection sites. The final part shows the application of two of the approaches to data from real blood collection sites, followed by the conclusions that can be drawn from this thesis.

Part I: Introduction, contains two chapters. **Chapter 1** introduces the context for this thesis: blood banks in general, the Dutch blood bank Sanquin and blood collection sites. The chapter sketches some of the challenges faced with respect to blood collection sites. As blood donors are voluntary and non-remunerated, delays and waiting times within blood collection sites should be kept at acceptable levels. However, waiting times are currently not incorporated in staff planning or in other decisions with respect to blood collection sites. These blood collection sites will be the primary focus of this thesis. This thesis provides methods that do take waiting times into account, aiming to decrease waiting times at blood collection sites and leveling work pressure for staff members, without the need for additional staff.

Chapter 2 then presents a technical methods that will be used most of the chapters in this thesis: uniformization. Uniformization can be used to transform Continuous Time Markov Chains (CTMCs) — that are very hard to analyze — into Discrete Time Markov Chains (DTMCs) — that are much easier to analyze. The chapter shows how the method works, provides an extensive overview of the literature related to the method, the (technical) intuition behind the method as well as several extensions and applications. Although not all of the extensions and applications are necessary for this thesis, it does provide an overview of one of the most valuable methods for this thesis.

Part II: Evaluation, contains two chapters that propose and adapt several methods to compute waiting times and queues at blood collection sites. A blood collection site is best modeled as a time-dependent queueing network, requiring non-standard approaches.

Chapter 3 considers a stationary, i.e. not time-dependent model of blood collection sites as a first step. A blood collection site consists of three main stations: Registration, Interview and Donation. All three of the stations can have their own queue. This means that even the stationary model is non-trivial for some computations. However, for the stationary model, an analytic so-called product form expression is derived. Based on this product form, two more results are shown. The first result is that the standard waiting time distributions from M|M|s queues are

applicable, as if the queue is in isolation. It is then concluded that no closed form expression exist for the total waiting or delay time distribution, as the distributions of the three stations in tandem are not independent. Therefore a numerical approach is presented to compute the total delay time distribution of a collection site. All of the results are supported by numerical examples based on a Dutch blood collection site. The approach for the computation of the total delay time distribution can also be combined with the approach from Chapter 4 for an extension to a time-dependent setting.

Chapter 4 shows an approach to deal with these time-dependent aspects in queueing systems, as often experienced by blood collection sites and other service systems, typically due to time-dependent arrivals and capacities. Easy and quick to use queueing expressions generally do not apply to time-dependent situations. A large number of computational papers has been written about queue length distributions for time-dependent queues, but these are mostly theoretical and based on single queues. This chapter aims to combine computational methods with more realistic time-dependent queueing networks, with an approach based on uniformization. Although uniformization is generally perceived to be too computationally prohibitive, we show that our method is very effective for practical instances, as shown with an example of a Dutch blood collection site. The objective of the results is twofold: to show that a time-dependent queueing network approach can be beneficial and to evaluate possible improvements for Dutch blood collection sites that can only be properly assessed with a time-dependent queueing method.

Part III: Optimization, contains four chapters that aim to improve service levels at Sanquin. The first three chapters focus on three different methods to decrease queues at blood collection sites. Chapters 5 and 6 focus on improving the service by optimizing staff allocation to shifts and stations. Chapter 7 focuses on improving the arrival process with the same goal. Chapter 8 is focused at improving inventory management of red blood cells.

Donors do not arrive to blood collection sites uniformly throughout the day, but show clear preferences for certain times of the day. However, the arrival patterns that are shown by historical data, are not used for scheduling staff members at blood collection sites. As a first significant step to shorten waiting times we can align staff capacity and shifts with walk-in arrivals. **Chapter 5** aims to optimize shift scheduling for blood collection sites. The chapter proposes a two-step procedure. First, the arrival patterns and methods from queueing theory are used to determine the required number of staff members for every half hour. Second, an integer linear program is used to compute optimal shift lengths and starting times, based on the required number of staff members. The chapter is concluded with numerical experiments that show, depending on the scenario, a reduction of waiting times, a reduction of staff members or a combination of both.

At a blood collection site three stations (Registration, Interview and Donation) can roughly be distinguished. Staff members at Dutch blood collection sites are often trained to work at any of these stations, but are usually allocated to one of the stations for large fractions of a shift. If staff members change their allocation this is

based on an ad hoc decision. **Chapter 6** aims to take advantage of this mostly unused allocation flexibility to reduce queues at blood collection sites. As a collection site is a highly stochastic process, both in arrivals and services, an optimal allocation of staff members to the three stations is unknown, constantly changing and a challenge to determine. Chapter 6 provides and applies a so-called Markov Decision Process (MDP) to compute optimal staff assignments. Extensive numerical and simulation experiments show the potential reductions of queues when the reallocation algorithm would be implemented. Based on Dutch blood collection sites, reductions of 40 to 80% on the number of waiting donors seem attainable, depending on the scenario.

Chapter 7 also aims to align the arrival of donors with scheduled staff, similarly to Chapter 5. Chapter 7 tries to achieve this by changing the arrivals of donors. By introducing appointments for an additional part of donors, arrivals can be redirected from the busiest times of the day to quiet times. An extended numerical queueing model with priorities is introduced for blood collection sites, as Sanquin wants to incentive donors to make appointments by prioritizing donors with appointments over donors without appointments. Appointment slots are added if the average queue drops below certain limits. The correct values for these limits, i.e. the values that plan the correct number of appointments, are then determined by binary search. Numerical results show that the method succeeds in decreasing excessive queues. However, the proposed priorities might result in unacceptably high waiting times for donors without appointments, and caution is therefore required before implementation.

Although this thesis mainly focuses on blood collection sites, many more logistical challenges are present at a blood bank. One of these challenges arises from the expectation that Sanquin can supply hospitals with extensively typed red blood cell units directly from stock. **Chapter 8** deals with this challenge. Currently, all units are issued according to the first-in-first-out principle, irrespective of their specific typing. These kind of issuing policies lead to shortages for rare blood units. Shortages for rare units could be avoided by keeping them in stock for longer, but this could also lead to unnecessary wastage. Therefore, to avoid both wastage and shortages, a trade-off between the age and rarity of a specific unit in stock should be made. For this purpose, we modeled the allocation of the inventory as a circulation flow problem, in which decisions about which units to issue are based on both the age and rarity of the units in stock. We evaluated the model for several settings of the input parameters. It turns out that, especially if only a few donors are typed for some combinations of antigens, shortages can be avoided by saving rare blood products. Moreover, the average issuing age remains unchanged.

Part IV: Practice and Outlook concludes this thesis. The first of two chapters in this part shows the combined application of two approaches from this thesis to data from three collection sites in the Netherlands. The final chapter of this thesis presents the conclusions that can be drawn from this thesis and discusses an outlook for further research.

Chapter 9 shows the combined application of the methods in Chapters 5 and 6 to three real collection sites in Dutch cities: Nijmegen, Leiden and Almelo. The collection sites in Nijmegen and Leiden are both large fixed collection sites. The

collection site in Almelo is a mobile collection site. The application of each one of the methods individually reduce waiting times significantly, and the combined application of the methods reduces waiting times even further. Simultaneously, small reductions in the number of staff hours are attainable.

The results from Chapter 9 summarize the main message of this thesis: waiting time for blood donors at blood collection sites can be reduced without the need for more staff members when the working times of staff members are used more effectively and efficiently, and controlling the arrival process of donors. The approaches presented in this thesis can be used for this purpose. This is not only beneficial for blood donors, but will also result in more balanced workload for staff members, as fluctuations in this workload are reduced significantly.

Samenvatting

Dit proefschrift bestaat uit vier delen: het eerste deel is een introductie, het tweede behandeld methoden om wachttijden te bepalen van bloedafnamelocaties, het derde behandeld methoden om het proces te verbeteren of optimaliseren. Het laatste deel laat de toepassing van twee methoden uit deel 3 zien op werkelijke data, gevolgd door de conclusies die uit dit proefschrift getrokken kunnen worden.

Deel I:Introductie bestaat uit twee hoodstukken. **Hoofdstuk 1** beschrijft de context van dit proefschrift: Bloeddonatie in Nederland; De meeste hoofdstukken van dit proefschrift gaan over afnamelocaties van Sanquin, de Nederlandse bloedbank. Hoofdstuk 1 schetst een aantal van de uitdagingen die Sanquin tegenkomt bij haar afnamelocaties. Bloeddonoren zijn vrijwillig en krijgen geen geldelijke vergoeding voor hun donaties. Wachttijden voor deze donoren moeten daarom zo veel mogelijk worden beperkt. Wachttijden worden momenteel echter niet of nauwelijks meegenomen bij onder andere de planning van personeel of andere beslissingen bij de afnamelocaties. Het verminderen van deze wachttijden, zo mogelijk zonder dat extra werkuren nodig zijn, zal daarom de primaire focus zijn van dit proefschrift.

Hoofdstuk 2 laat een overzicht zien van een technische methode die in veel hoofdstukken word gebruikt: Uniformizatie. Univormizatie wordt gebruikt om een zogenaamde "Continuous Time Markov Chain" (CTMC) - die erg ingewikkeld zijn om te analyseren - om te zetten in een zogenaamde "Discrete Time Markov Chain" (DTMC) - die veel eenvoudiger zijn om te analyseren. Het hoofdstuk beschrijft hoe de methode werkt, laat een uitgebreid overzicht zien van de relevante literatuur, beschrijft de (technische) intuitie achter de methode en toont enkele uitbreidingen en toepassingen. Veel van deze uitbreidingen en toepassingen zijn niet benodigd voor dit proefschrift, maar laten wel de veelzijdigheid zien van Uniformizatie.

Deel II: Evaluatie bestaat uit twee hoofdstukken die een aantal methoden beschrijven aan aanpassen om wachttijden en -rijen te berekenen voor bloedafnamelocaties. Deze methoden moesten grotendeels nieuw worden ontwikkeld, omdat een afnamelocatie enkele kenmerken heeft die standaard methdoen uitsluiten.

Hoofdstuk 3 beschrijft eerst het eenvoudigste model van een afnamelocaties: een stationair tandem-model met drie stations: Registratie, Interview en Donatie. Ieder van deze stations heeft haar eigen wachtrij. Zelfs dit eenvoudigste model is daarmee niet triviaal. Voor dit model is wel een analytische, zogenaamde "product form" afgeleid. Van deze product form kan weer expliciete uitdrukking voor de marginale wachttijdsverdeling van één van de stations worden afgeleid. Vervolgens word in dit hoofdstuk beargumenteerd waarom een expliciete uitdrukking voor de verdeling van de totale doorlooptijd niet kan bestaan. Daarom word een numerieke procedure

getoond om deze totale verdeling te bepalen. Ook deze numerieke procedure is gebaseerd op de "product form". De procedure kan worden gecombineerd met het werk uit hoofdstuk 4 voor een tijdsafhankelijke bepaling van de doorlooptijdverdeling. Alle resultaten worden ondersteund met numerieke voorbeelden.

Hoofdstuk 4 beschrijft de tijdsafhankelijke versie van het in hoofdstuk 3 geïntroduceerde model. Tijdsafhankelijkheid komt voor in veel processen, meestal door variërende aankomsten en capaciteiten. Eenvoudige methoden om wachttijden te schatten zijn onder tijdsafhankelijkheid niet meer van toepassing. Verschillende onderzoeken zijn gedaan naar tijdsafhankelijke wachtrijbepalingen, maar deze zijn veelal erg theoretisch. Hoofdstuk 4 combineert de methoden uit de theorie met een realistischere modelering van de praktijk in de vorm van wachtrijnetwerken. Er wordt een methode beschreven op basis van uniformizatie. Hoewel dit vaak wordt beschouwd als een methode die te veel rekenkracht nodig heeft, toont het hoofdstuk dat uniformizatie goed bruikbaar is voor praktische toepassingen zoals een bloedafnamelocatie. Numerieke resultaten tonen vervolgens zowel dat een tijdsafhankelijke bepaling van wachtrijen voordelig is, als het evalueren van verschillende scenario's voor bloedafnamelocaties.

Deel III: Optimalisatie bestaat uit vier hoofdstukken die service van Sanquin verbeteren. De eerste drie hoofdstukken beschrijven methoden om wachttijden op afnamelocaties te verminderen. Hoofdstukken 5 en 6 doen dit door de inzet van medewerkers te optimaliseren. Hoofdstuk 7 beschrijft een methode om de aankomsten te beïnvloeden met hetzelfde doel. Het vierde hoofdstuk in dit deel, Hoofdstuk 8 focust op het verbeteren van voorraadbeheer van rode bloedcellen.

Donors komen niet gelijkmatig verdeeld over de dag aan bij een afnamelocatie, maar tonen duidelijke voorkeur voor bepaalde tijden. De hieruit af te leiden patronen worden echter niet gebruikt bij de planning van medewerkers. Wachttijden van donors kunnen significant worden verlaagd door de diensten van medewerkers af te stemmen op verwachtte aankomsten. **Hoofdstuk 5** zorgt voor deze afstemming door gebruik te maken van een twee-staps procedure. De eerste stap bestaat uit het bepalen van de minimaal benodigde medewerkers per half uur, om de in dat half uur aankomende donors te kunnen bedienen met een korte wachttijd. Vervolgens word een zogenaamd "Integer Linear Program" (ILP) gebruikt om diensten voor medewerkers te bepalen. De diensten worden zodanig bepaald dat aan elk van de eisen wat betref benodigde medewerkers word voldaan met zo min mogelijk werkuren. Afhankelijk van het scenario laten de resultaten een verlaging van wachttijd, een verlaging van het aantal gewerkte uren, of een combinatie van beiden zien.

Bij een bloedafnamelocatie kunnen drie stations worden onderscheiden (Registratie, Interview en Donatie). Medewerkers van Nederlandse afnamelocaties zijn geschoold om op elk van deze drie stations te werken, maar worden over het algemeen toegewezen aan een van de stations voor een groot deel van of een gehele dienst. Als een medewerker wisselt gebeurd is dit niet op basis van gestructureerde beslissingen. **Hoofdstuk 6** poogt de beslissingen om medewerkers aan stations toe te wijzen te optimaliseren. Een afnamelocatie is een erg stochastisch proces, zowel door tijdsafhankelijk aankomsten als ongelijke serviceduren. De optimale toewijzing van

medewerkers veranderd daarom continu en is een uitdaging om te bepalen. Hoofdstuk 6 beschrijft een zogenaamd "Markov Decision Process" (MDP) om de optimale toewijzing van medewerkers te bepalen met regelmatige intervallen. Numerieke en simulatie resultaten laten zien dat de wachtrijen bij Nederlandse afnamelocaties kunnen worden verkort met 40% tot 80%.

Hoofdstuk 7 richt zich op dezelfde uitdaging als Hoofdstuk 5: het afstemmen van aankomsten en de aanwezigheid van medewerkers. Hoofdstuk 7 pakt dit echter aan door de aankomsten van donors te veranderen. Door een deel van de volbloeddonors een afspraak te laten maken, kunnen de aankomsten worden gestuurd richting tijden dat het rustig is op de afnamelocatie. Donors worden aangemoedigd een afspraak te maken door ze met voorrang te bedienen op een afnamelocatie. Een uitgebreid wachtrijmodel voor de afnamelocatie met voorrangsregels is daarom ontwikkeld om goede afsprakenschema's te bepalen. Als in verwachting minder donors aanwezig zijn dan een nader te bepalen limiet, word een afspraak toegevoegd. De juiste limieten, de limieten die in precies genoeg afspraken resulteren, worden bepaald met een zoekstrategie die bekend staat als "binary search". Numerieke resultaten laten zien dat uitschietende wachtrijen met succes worden verminderd. De voorgestelde voorrangsregels zouden echter erg slecht kunnen uitpakken voor donors zonder afspraak. Voor implementatie zal over deze voorrangsregels dus nog moeten worden nagedacht.

Hoewel dit proefschrift zich hoofdzakelijk focust op afnamelocaties, zijn er nog veel meer logistieke uitdagingen bij de bloedbank. Een van deze uitdagingen komt voort uit de verwachting dat Sanquin uitgebreid getypeerde bloedproducten direct uit voorraad kan leveren aan ziekenhuizen. **Hoofdstuk 8** beschrijft een potentieel nieuw systeem voor voorraadbeheer om hier veel vaker aan te kunnen voldoen. Op dit moment worden alle rode bloedcellen geleverd volgens het "first-in-first-out" principe, onafhankelijk van typering. Dit leidt tot tekorten van uitgebreid getypeerde producten. Door uitgebreid getypeerde producten langer in voorraad te houden zouden tekorten teruglopen, maar kan verspilling van deze producten ook toenemen door een maximale houdbaarheid. Om tekorten en verspilling tegen te gaan, wordt daarom een afweging gemaakt tussen zeldzaamheid en leeftijd van producten. Het ontwikkelde systeem maakt deze afweging met behulp van een zogenaamd "circulation flow problem". Uit gesimuleerde resultaten blijkt dat het nieuwe systeem veel vaker een product levert dat exact overeenkomt met de vraag, waardoor uitgebreid getypeerde producten alleen worden uitgegeven wanneer ze benodigd zijn. Tekorten en verspilling nemen hierbij ook nog eens af.

Deel IV: Praktijk en Conclusies sluit dit proefschrift af met de laatste twee hoofdstukken. Het eerste hoofdstuk laat de toepassing van twee methoden zien op data van drie afnamelocaties in Nederland. Het laatste hoofdstuk toont de conclusies die uit dit proefschrift kunnen worden getrokken en de mogelijkheden voor verder onderzoek.

Hoofdstuk 9 laat de gezamelijk toepassing van de methoden uit hoofdstukken 5 en 6 zien op data van de afnamelocaties in Nijmegen, Leiden en Almelo. De locaties in Nijmegen en Leiden zijn beide grote, vaste afnamelocaties, terwijl Almelo wordt

bezocht door een mobiele afnamelocatie. De afzonderlijke methoden laten verbeterde prestatie zien met betrekking tot wachtrijen, terwijl een gezamenlijke toepassing de wachtrijen nog verder verkort. Tegelijkertijd lijkt een kleine verlaging van het aantal gewerkte uren haalbaar.

De resultaten van hoofdstuk 9 vatten de belangrijkste boodschap van dit proefschrift samen: wachttijden en -rijen bij bloedafname kunnen worden verkort zonder dat meer medewerkers nodig zijn. Dit kan worden bereikt door medewerkers efficiënter en effectiever in te zetten en de aankomsten van donors meer te controleren. De methoden beschreven in dit proefschrift zijn daarvoor een goed uitgangspunt, die kunnen worden geïmplementeerd als eerste stap en verder kunnen worden ontwikkeld. Dit is niet alleen voordelig voor donors, maar ook voor de medewerkers door de werkdruk te verspreiden over de werkdag.

About the author

Sem van Brummelen was born in Uithoorn on November 23rd, 1990. He grew up in Amstelveen, and went to high school at Keizer Karel College. After finishing high school in 2009, he started studying Econometrics and Operations Research at the University of Amsterdam. He obtained a bachelor, with specialization in Operations Research, with a thesis on scheduling balancing periods for staff working hours, minimizing average deviations from contractual hours. He then started a master in Operations Research at the University of Amsterdam, where he first got to know Prof. dr. Nico van Dijk, whom got him interested in doing PhD research and supervised his master thesis. In his master thesis, which was written in cooperation with the Dutch Transplant Association, he studied waiting time for corneas transplantations, and proposed several interventions to reduce waiting time and introduce more equality between recipients of corneas.

In November of 2013 he started his PhD, again under supervision of Nico van Dijk. The project, which resulted in this thesis, was called "Waiting time and personnel capacity optimization". It is a cooperation between the department Donor Studies at Sanquin Research and the research group CHOIR at the university of Twente. The research and developed methods try to reduce waiting times for donors at blood collection sites without increasing the total capacity of staff members and other equipment.

List of publications

S.P.J. van Brummelen, M.B.A. Heemskerk, H.A. van Leiden and N.M. van Dijk. Met het oog op wachttijd. [Dutch] *STAtOR* 15 (1):18-23, 2014.

L.P. Besselink, S.P.J. van Brummelen, W.L. de Kort, P.L.M. Zonneveld and N.M. van Dijk. OR voor betere bloeddoorstroming. [Dutch] *STAtOR* 15 (3):16-21, 2014.

S.P.J. van Brummelen, W.L. de Kort and N.M. van Dijk. Waiting time computation for blood collection sites. *Operation Research for Health Care* 7:70-80, 2015.
(Basis for Chapter 3.)

S.P.J. van Brummelen, N.M. van Dijk, K. van den Hurk and W.L. de Kort. Waiting time based staff capacity and shift planning at blood collection sites. *Health Systems, in press*
(Basis for Chapter 5.)

S.P.J. van Brummelen, W.L. de Kort and N.M. van Dijk. Queue length Computation of Time-Dependent Queueing Networks and its Application to Blood Collection. *Operations Research for Health Care, accepted*
(Basis for Chapter 4.)

S.P.J. van Brummelen, K. van den Hurk, W.L. de Kort and N.M. van Dijk. Dynamic staff allocation at blood collection sites. *Submitted*
(Basis for Chapter 6.)

N.M. van Dijk, S.P.J. van Brummelen, R.J. Boucherie. Uniformization: basics, extensions and applications. *Performance Evaluation, accepted*
(Basis for Chapter 2.)

S.P.J. van Brummelen, W.L. de Kort and N.M. van Dijk. Combining appointments and walk-ins at Dutch blood collection sites. *In preparation*
(Basis for Chapter 7.)

J.H.J. van Sambeeck, S.P.J. van Brummelen and N.M. van Dijk. Blood type specific issuing policies to improve inventory management of red blood cells. *In preparation*
(Basis for Chapter 8.)

S.P.J. van Brummelen, N.M. van Dijk, W.L. de Kort and K. van den Hurk. Combining optimal shift scheduling and reallocation policies at Dutch blood collection sites. *In preparation*
(Basis for Chapter 9.)