MODELLING ACTIVITIES AT A NEUROLOGICAL REHABILITATION UNIT

by

Richard Max Wood

Thesis submitted to

Cardiff University

School of Mathematics

for the degree of

DOCTOR OF PHILOSOPHY

September 2011

Declaration

This work has not previously been accepted in substance for any degree and is not concurrently being submitted in candidature for any degree.

Signed:

Date:

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by giving explicit references. A bibliography is appended.

Signed:

Date:

Statement 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and inter-library loan, and for the title and summary to be made available to outside organisations.

Signed:

Date:

Acknowledgements

First and foremost I would like to thank my supervisors Jeff Griffiths and Janet Williams who have provided fantastic support and guidance throughout this project. Not only this, they have provided a friendly environment which has been a pleasure to work in.

My gratitude also to Jakko Brouwers, Angelo Lamberti and Jenny Thomas at Rookwood hospital who have given up much of their time and effort in the pursuit of this project. My thanks must also extend to the EPSRC through the LANCS initiative who have funded this work.

Finally, I would like to thank my parents and family who have been excellent not just during the PhD but throughout my university years. And last but not least I am obliged to my friends who, for almost three years, have had to put up with my rants about how rubbish visual basic is.

Summary

A queuing model is developed for the neurological rehabilitation unit at Rookwood Hospital in Cardiff. Arrivals at the queuing system are represented by patient referrals and service is represented by patient length of stay (typically five months). Since there are often delays to discharge, length of stay is partitioned into two parts: admission until date ready for discharge (modelled by Coxian phase-type distribution) and date ready for discharge until ultimate discharge (modelled by exponential distribution). The attributes of patients (such as age, gender, diagnosis etc) are taken into account since they affect these distributions. A computer program has been developed to solve this multi-server (21 bed) queuing system to produce steady-state probabilities and various performance measures.

However, early on in the project it became apparent that the intensity of treatment received by patients has an effect on the time, from admission, until they are ready for discharge. That is, the service rates of the Coxian distribution are dependent on the amount of therapy received over time. This directly relates to the amount of treatment allocated in the weekly timetables. For the physiotherapy department, these take about eight hours to produce each week by hand. In order to ask the valuable what-if questions that relate to treatment intensity, it is therefore necessary to produce an automated scheduling program that replicates the manual assignment of therapy. The quality of timetables produced using this program was, in fact, considerably better than its alternative and so replaced the by-hand approach. Other benefits are more clinical time (since less employee input is required) and a convenient output of data and performance measures that are required for audit purposes.

Once the model is constructed a number of relevant hypothetical scenarios are considered. Such as, what if delays to discharge are reduced by 50%? Also, through the scheduling program, the effect of changes to the composition of staff or therapy sessions can be evaluated, for example, what if the number of therapists is increased by one third? The effects of such measures are analysed by studying performance measures (such as throughput and occupancy) and the associated costs.

Conference Presentations

'LANCS Healthcare Modelling PhD symposium', (2009), *Cardiff, UK*.

'1st Student Conference in Operational Research', (2009), *Lancaster, UK*.

'51st Confernce of the Operational Research Society', (2009), *Warwick, UK*.

'LANCS Workshop on Real-world Scheduling in Healthcare', (2010), *Nottingham, UK*.

'2nd Student Conference in Operational Research', (2010), *Nottingham, UK*.

'24th European Conference on Operational Research', (2010), *Lisbon, Portugal*.

'14th Applied Stochastic Models and Data Analysis International Conference', (2011), *Rome, Italy*.

'37th Operational Research Applied to Health Services Conference', (2011), *Cardiff, UK*.

Contents

Chapter 1: Introduction

1.1 Background

The specialist neurological rehabilitation unit at Rookwood hospital is the application of the theoretical work considered in this study. Rookwood hospital is a major Regional Rehabilitation Unit located in Cardiff, the capital of Wales. In addition to the specialist Neurological Rehabilitation Centre (NRC) the hospital is also home to the Welsh Spinal Injury Centre, the Artificial Limb and Appliance Service and a geriatric day unit.

Rookwood hospital has a long history of rehabilitation medicine. In 1918 the site, formerly home to a British Colonel, was converted to a convalescent home for officers returning from the Great War. In subsequent years it was used to care for disabled ex-servicemen up until 1932 when it was transformed to a general hospital. The latter half of the 20th century saw the hospital develop as the premier location for specialist rehabilitation in Wales. It currently houses one of only twelve spinal rehabilitation units in the UK and one of only two neurological rehabilitation centres in Wales.

1.2 Neurological Rehabilitation

Neurological rehabilitation is the process that aids recovery following an Acquired Brain Injury (ABI). The Select Committee on Health in the UK (third report, 2000-2001 session) defines ABI as *'an injury to the brain that has occurred since birth'*. This includes Traumatic Brain Injury (TBI) which is the most common form of ABI (incidence is estimated at 250-300 per 100,000 in western developed countries – Campbell, 2000). In the UK, the National Institute for Clinical Excellence (NICE) claims that *'each year 1.4 million people attend hospitals in England and Wales after suffering a head injury, of whom 150,000 are admitted*

to hospital' (2003). In the same year it is reported in the National Clinical Guidelines for Rehabilitation following ABI (Turner-Stokes, 2003) that each year 25 people per 100,000 will sustain a moderate[1] to severe brain injury. Such patients are appropriate candidates for a specialist rehabilitation programme.

Figure 1.1 describes a typical pathway for a patient with moderate to severe ABI in Wales.

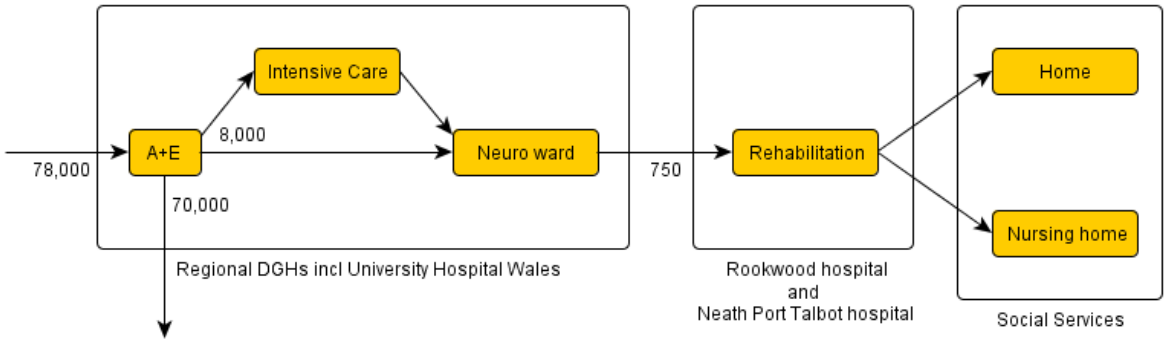


Figure 1.1 Typical patient pathway

The University Hospital Wales (UHW) is a large District General Hospital (DGH) located in Cardiff. It is the only hospital in Wales to contain a neurosurgery unit and is therefore the source of the vast majority of referrals for Rookwood NRC. The values represent annual demand on the various NHS facilities in Wales and are based[2] on the aforementioned studies. These are only approximate due to *'considerable variation in different parts of the country'* (Turner-Stokes, 2003).

Admission to the NRC at Rookwood hospital is by referral only. Patients are either directly referred by their hospital or are scouted by a clinical specialist from the NRC. Referrals that are deemed appropriate enter the queue; otherwise they are dismissed. This process is described in Figure 1.2. Note that due to the excessive demand for this service it is exceptional for a queue not to exist.

[1] See Teasell et al, 2003 for a definition of ABI severity
[2] Assuming the respective populations of England and Wales are 51m and 3m

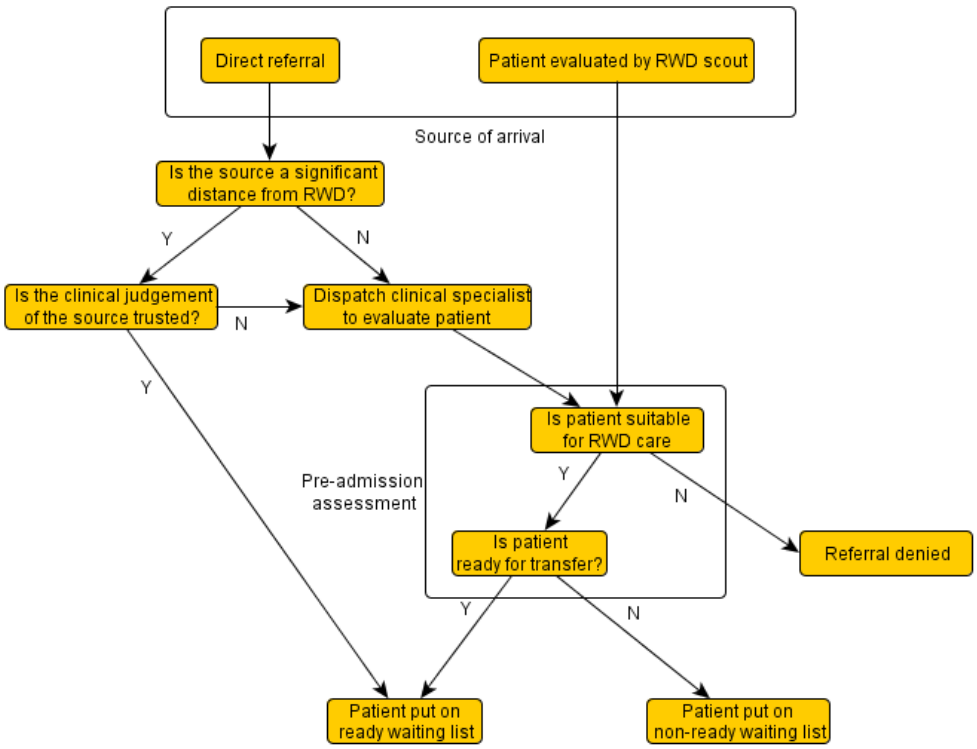


Figure 1.2 Referrals process

A *waiting list* is kept that documents the name of each referral and whether they are ready for transfer to the NRC or not. Each time an appropriate referral is received the size of the waiting list increases by one. The *ready waiting list* is defined as the 'queue' since admission is only possible when the patient is ready for transfer. A patient from the *non-ready waiting list* can join the queue when they become ready and likewise a patient from the queue can join the non-ready waiting list should they deteriorate. The size of the waiting list is reduced either by the withdrawal of a referral (due to transfer, discharge or death) or by an admission. An admission can only occur if there is sufficient capacity to meet the demands of the incoming patient. Since the queue is priority-based, the emphasis is placed on admitting the patient with the highest priority. In order to promote fairness priorities are determined by severity as well as time spent in queue. This ensures that less severe patients are not kept waiting for extreme lengths of time.

In the most basic of conceptions the capacity is dependent only on the availability of beds. Therefore, when a patient is discharged the bed is populated by the highest priority patient in the queue. However, this does not take into account the needs of the patient and the availability and proficiency of the staff. At Rookwood NRC such considerations are incorporated into the process of determining admissions; albeit in a qualitative manner. The

following example illustrates the complexity of this process. Assume that a bed is vacant but the availability and proficiency of the staff is insufficient to meet the needs of the highest priority patient in the queue. What do you do? If the supply-side (staff) issues are likely to persist then it may be best to admit the highest priority patient in the queue whose needs can be met. However, if another patient is close to discharge then it may be best to wait until this time so that the additional availability of staff could be used to adequately meet the needs of the highest priority patient. It is clear this process is far from an exact science. Figure 1.3 describes this process; following on from Figure 1.2.

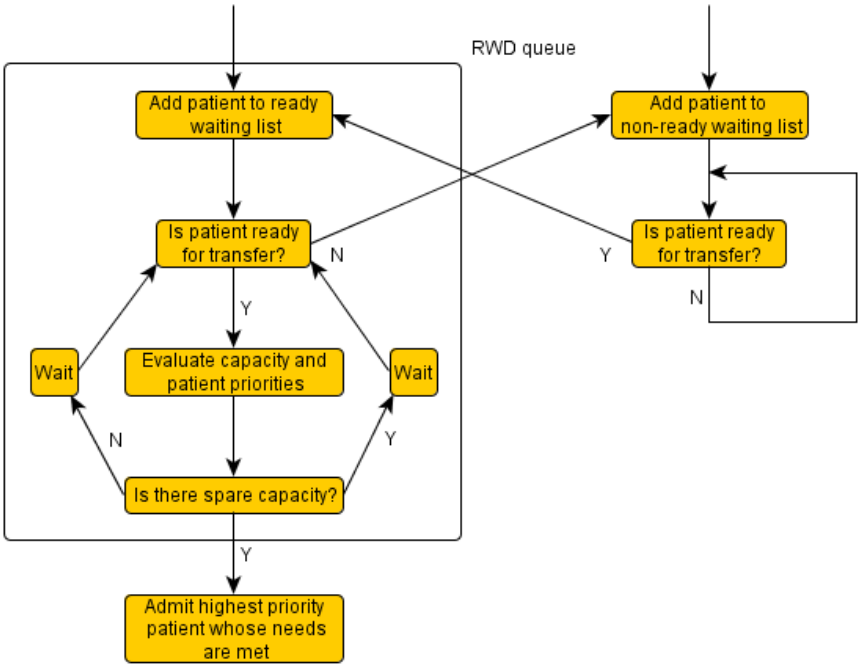


Figure 1.3 Admissions process

Once an appropriate referral has been cleared for admission the patient is transferred to the NRC at Rookwood hospital. This facility consists of two wards that, together, contain approximately 30 beds. Four of these beds are reserved for short-stay respite care and are used to provide relief for the carers of patients that have been discharged to the community. Since these beds are not used for rehabilitation they are, for the purpose of this project, not considered as part of the NRC. The availability of bedside oxygen prohibits the remainder of beds from being regarded as truly homogeneous. Since only eight of these have oxygen provision this can also complicate the admissions process. In recent years 21 beds have been used, on average, for non-respite care.

Within 24 hours of admission a comprehensive initial assessment of the patient is performed. In this, various performance measures, such as the FIM[3] (see Ch 2.2.1), are scored and initial goals are set. For the remainder of this week a basic level of care is provided to the patient. Treatment is provided to patients at the NRC by a multidisciplinary team involving physiotherapy, occupational therapy, speech and language therapy, psychology, dietetics, and orthopaedics (in order of decreasing size[4]) in addition to nursing and medical staff.

Following this, the patient enters a perpetual cycle of assessment and treatment until discharge. A ward round is undertaken each week to set short-term goals. Details are also recorded for the number and type of treatment sessions that are demanded, on behalf of each patient, to take place in the following week. These values represent the 'perfect world' level of treatment. Once determined, a treatment timetable for the forthcoming week can be constructed (see Chapters 3 and 4) for each patient. However, due to capacity considerations it is very unlikely that all demand for treatment sessions is met.

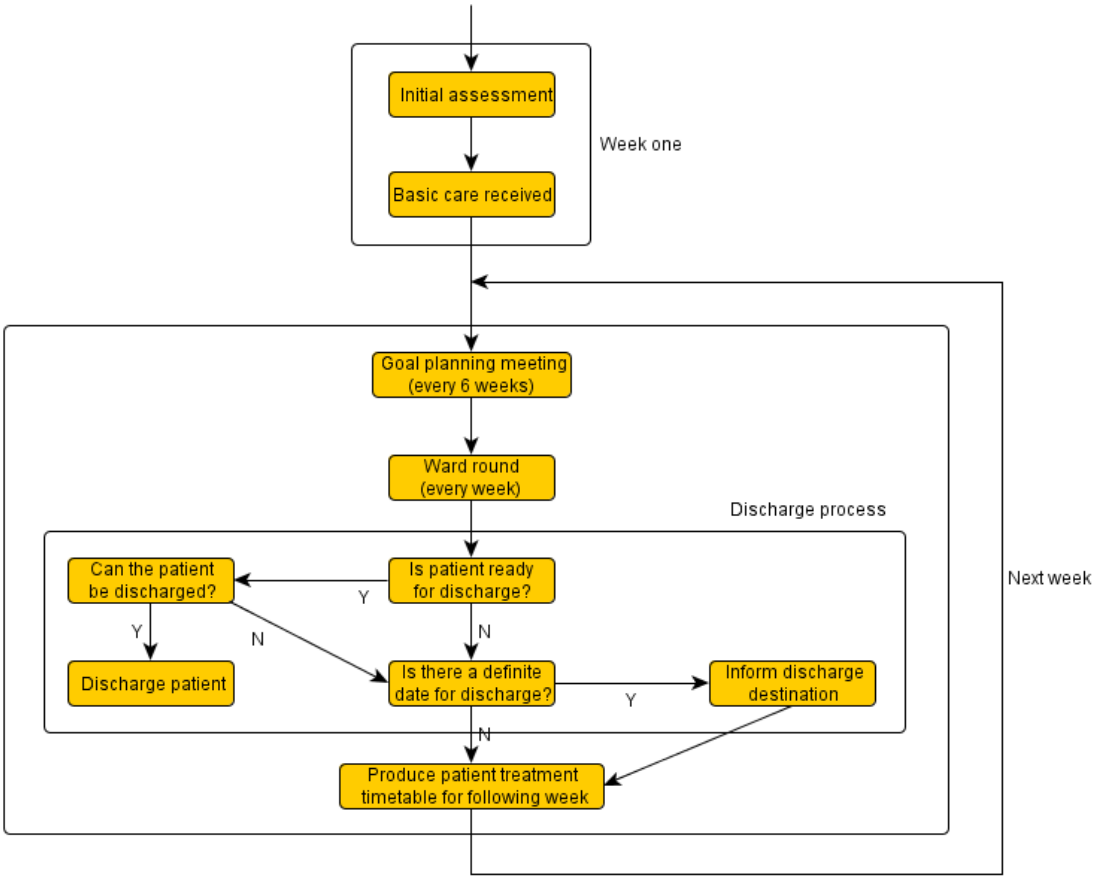


Figure 1.4 Treatment and discharge process

[3] Functional independence measure
[4] As measured by number of full time equivalents

An approximate discharge date is usually agreed upon at a goal planning meeting. These occur every six weeks and are used to set medium-term goals for the patient. They are also used to discuss discharge arrangements. Since it is difficult to accurately predict when a patient will be ready for discharge their condition must be continually assessed. When a patient is deemed to be ready the ability of the discharge destination to receive the patient must be evaluated. If the destination is not ready then the patient must remain at Rookwood hospital. This is commonly referred to as *bed-blocking* or *delayed discharge*.

The reasons for a delay in discharge are detailed by the Select Committee on Health (third report, 2001-2002 session) in Table 1.1.

Table 1.1 Reasons for delays in discharge

Reason for delay	% of delays
Awaiting completion of an assessment of future care needs and identifying an appropriate care setting	22.2
Awaiting social services funding for residential or home care	21.9
Awaiting care home placement	20.4
Awaiting further NHS care	11.5
Awaiting domiciliary package	6.7

Bed-blocking is a major problem at Rookwood NRC with some discharges being delayed for over a year. Since the average bed cost per day[5] is £480 this represents significant financial expense, not to mention the deterioration of patients kept waiting in the queue.

The average length of stay (LOS) is approximately five months with a cost of £72,000.

1.3 Mathematical Models

Operational Research (OR) techniques are often used to construct mathematical models of healthcare facilities. OR is a mathematical discipline defined by the Association of European Operational Research Societies as *'a scientific approach to the solution of problems in the management of complex systems'*. A mathematical model can be defined as a description of a system in mathematical language. There are many OR techniques that can be used to produce a mathematical model of a healthcare facility. These include continuous and discrete simulation and queuing theory.

[5] Figure obtained from Rookwood NRC

According to Nance, 1993 continuous simulation *'uses equational models, often of physical systems, which do not portray precise time and state relationships'*. It is suitable for *'systems in which the variables can change continuously'* (Ozgun & Barlas, 2009). System dynamics is a form of continuous simulation that is used to analyse a system of integrated states over time. These states are connected by various links that describe the dynamic interaction between components of the system. Differential equations are attached to these links to mathematically explicate such interactions. A *causal loop diagram* can also be used to graphically describe the system. Example applications of system dynamics include micro-economics (buying and selling of assets between firms) and ecology (birth and death of species).

It can also be used to model the flow of patients in a healthcare system. In this case the system could represent an NHS Trust and the states could represent individual hospitals. Feedback loops can be used to incorporate readmission or relapse. System dynamics is an effective modelling tool in this case because the system can be broken down into individual physical states. This is also true for a DGH which can be reduced to A+E, ICU, and a number of different wards. However, no such reduction exists for Rookwood NRC. Moreover, it would be inappropriate to model this facility through continuous simulation due to its low annual throughput. In order to accurately model admission and discharge (two major events at this facility) a determination of precise time and state relationships is required.

This can be achieved through Discrete Event Simulation (DES) which *'utilises a mathematical/logical model of a physical system that portrays state changes at precise points in simulated time'* (Nance, 1993). In the case of Rookwood NRC the principal discrete events are the arrival of a referral, an admission, and a discharge. If sample data is known for each of these then the inter-arrival distribution and service (LOS) distribution can be found. A computer package can then be used to simulate this system. Results for the probability of being in each state of the system and performance measures (e.g. mean waiting time) can then be output.

What has been described is a DES model of a queuing system. Queuing theory, on the other hand, is a mathematical branch of OR that can be used to derive analytic (i.e. exact, not simulated) results for a queuing system. For more complex systems, however, the problem can become mathematically intractable, and so simulation is used. Ozgun & Barlas, 2009

state that DES *'has been the major tool for arriving at conclusions about complicated queuing networks'* and that *'it is very rare to see a study that uses continuous simulation'*.

The intended approach is to deduce analytic solutions using queuing theory. Should this not be possible (i.e. if mathematically intractable) then DES could be used.

1.4 Queuing Theory

Queuing theory is a mathematical sub-discipline of Operational Research that is used in the analytical study of queues and queuing systems. Central to the subject is the concept of customers arriving at a service facility, waiting in line, being served, and leaving.

A very simple case is considered as an example. Assume that customers arrive at a single-server facility at random and with mean rate λ. The customer immediately enters service if the server is available and there is no queue. Otherwise, they join the back of the queue and wait their turn. Customers in the queue must wait for all those in front of them to begin service before they can proceed. Service times are exponential with mean rate μ. There is no limit on the number of customers in the system. This system is illustrated in Figure 1.5.

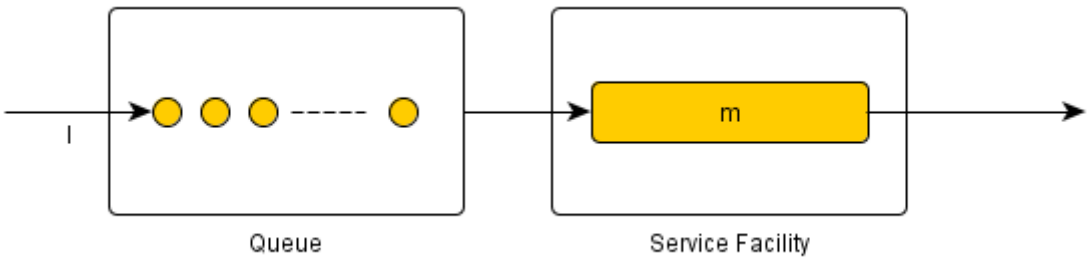


Figure 1.5 A simple queue

Kendall's notation provides a convenient classification of a queuing system in the form A | B | C | D | E where A is the inter-arrival distribution, B is the service distribution, C is the number of servers, D is the capacity of the system, and E is the queue discipline. Typically, *M* denotes Markovian inter-arrival/service times (i.e. exponentially distributed) and *FIFO* denotes the queue discipline first-in first-out. Therefore, the above example describes a $M|M|1|\infty|FIFO$ queuing system.

There are many mathematical methods that can be applied to solve such a system for both transient and stationary results. Essentially, transient results have a dependence on time and stationary, or steady-state, results assume the system has reached equilibrium. If the objective

is to investigate the behaviour of a queue of road traffic at rush hour then transient results are sought. On the other hand, if the aim is to investigate the behaviour of a continually operating machine in a factory then steady-state results are required.

Over the years queuing theory has been applied extensively to healthcare facilities (see Ch 2.1 for a review). The type of queuing model used is, of course, dependent on the healthcare facility under consideration. The number of servers could represent the number of GPs in a doctor's surgery or the number of beds in a hospital ward. If arrivals represent people and not referrals then it is likely that a restriction on the capacity of the system is required. For example, at a doctor's surgery there are a limited number of seats available in the waiting room. The discipline of the queue is also sensitive to the facility being modelled. It is typical for an A+E department to operate a priority-based queue so that those with the worst injuries are seen first. On the other hand, a day clinic may choose to operate a ticket-based system where arriving patients pick a ticket and are served in the order they arrive in (i.e. FIFO).

The inter-arrival and service distributions must also be considered. These are determined by fitting statistical distributions to sample data obtained from the facility (see Chapter 5). In such an exercise the objective is usually to find the distribution that produces the best approximation to the data. However, the mathematical tractability of constructing a queuing theoretic model is heavily dependent on the distributions that are used (see Ch 5.1). Therefore, it is necessary to attain a balance between the goodness of fit and the tractability of the associated queuing model.

Phase-type distributions are often used in queuing theory to this end. They are highly versatile and possess the necessary attributes that permit tractability. They are explained in greater detail in the following subchapter.

1.5 Phase-Type Distributions

A phase-type distribution is a stochastic process that describes a finite system of integrated Poisson processes occurring in phases. This concept was first introduced in Erlang, 1917 and generalised in Cox, 1955. A Poisson process is defined as follows. The process $N(t)$ is a Poisson process with rate λ if it is a counting process and:

1. $N(0) = 0$

2. The process has independent increments

3. $P\left(N\left(t+h\right)-N\left(t\right)=1\right)=\lambda h+o\left(h\right)$

4. $P\left(N\left(t+h\right)-N\left(t\right)\geq 2\right)=o\left(h\right)$

It follows that $P\left(N\left(h+t\right)-N\left(h\right)=n\right)=e^{-\lambda t}\left(\lambda t\right)^{n}\left(n!\right)^{-1}$, i.e. the Poisson distribution. The distribution of time between consecutive events is now sought. By letting T_i denote the time of the *i*-th event the following is obtained:

$$P\left(T_{n+1}>t\,|\,T_n=h\right)=P\left(N\left(h+t\right)-N\left(h\right)=0\,|\,T_n=h\right)=P\left(N\left(h+t\right)-N\left(h\right)=0\right)=e^{-\lambda t}$$

The time between consecutive events is exponentially distributed. Due to the memoryless property of this distribution the time until the next event in each phase is also exponentially distributed. The Poisson process is therefore a special case of a continuous-time Markov process (since the Markov property holds).

Moreover, a phase-type distribution can be represented by a finite continuous-time Markov process. This describes the time until absorption of a Continuous-Time Markov Chain (CTMC). In this description the phases are represented by transient states. The stochastic process begins in a transient state and ceases upon entry to the absorbing state. If there are *p* phases in the distribution then there are *p+1* states in the CTMC. The CTMC is defined on the state space $\left\langle 1,2,...,p,p+1\right\rangle$.

Continuous-time Markov processes can be characterised by their *infinitesimal generator matrix*, *Q*. This is a square matrix with dimension equal to the cardinality of the state space. The sum of values in each row must equal zero. Each non-diagonal element $Q_{i,j}$ represents the transition rate from state *i* to state *j*. In constructing the infinitesimal generator matrix it is necessary to consider the parameterisation of the CTMC. There are two options. The first requires a specification of the *rate parameter* for all transient states and any associated *transition probabilities*. The second requires a specification of the *transition rates* between all states. Figure 1.6 describes these options for an example CTMC.

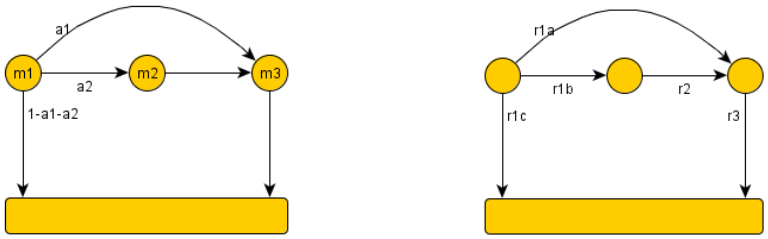


Figure 1.6 Alternate parameterisation for an example CTMC

The infinitesimal generator matrices for these are

$$Q=\begin{pmatrix} -m1 & a2\cdot m1 & a1\cdot m1 & (1-a1-a2)m1 \\ 0 & -m2 & m2 & 0 \\ 0 & 0 & -m3 & m3 \\ 0 & 0 & 0 & 0 \end{pmatrix}, Q=\begin{pmatrix} -(r1a+r1b+r1c) & r1b & r1a & r1c \\ 0 & -r2 & r2 & 0 \\ 0 & 0 & -r3 & r3 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Note that it is trivial to re-parameterise, i.e. can obtain rate parameters and transition probabilities in terms of transition rates or vice versa. In this example,

$$r1a=m1\cdot a1,\ r1b=m1\cdot a2,\ r1c=m1\cdot(1-a1-a2),\ r2=m2,\ r3=m3$$

Note also that the number of parameters that require estimation is independent of the choice of parameterisation. Let the formulation equivalent to the first and second option be referred to respectively as the *probability parameterisation* and the *rate parameterisation*.

This infinitesimal generator matrix can be partitioned as follows

$$Q=\begin{pmatrix} T & \tau \\ 0 & 0 \end{pmatrix} \tag{1.5.1}$$

Here, T is the $p\times p$ generator matrix for the transient states henceforth referred to as the *PH-generator matrix*. The $p\times 1$ vector τ represents the transition rates from the transient states to the absorbing state. Clearly $\tau_{i,1}=\sum_{j=1}^{p}T_{i,j}$. Therefore, T is the minimal representation of Q. What is not detailed by the PH-generator matrix, however, is the state in which the stochastic process begins. The *initial probability vector* α is used to provide this information. This is a $p+1$ dimensional stochastic row vector whose entries represent the

probability of starting the process in each of the states. The minimal representation of a phase-type distribution is therefore (α, T).

It can be shown (see Bladt, 2005 for proof) that the *pdf* [6] of a phase type distribution is

$$f(t) = P(X = t) = \alpha \exp(Tt)\tau \tag{1.5.2}$$

where $\exp(\cdot)$ is the matrix exponential. This method of derivation is henceforth referred to as the *matrix geometric approach*. It is particularly effective when performed by a computer program such as Matlab. The *cdf* [7] is therefore derived by integration

$$F(t) = P(X < t) = 1 - \alpha \exp(Tt)\mathrm{I} \tag{1.5.3}$$

since $\tau = T \cdot \mathrm{I}$ where I is a p dimensional column vector of ones. The *n*-th moments are given by the formula

$$E\left[X^n\right] = (-1)^n n! \alpha T^{-n} \mathrm{I} \tag{1.5.4}$$

Alternatively, the pdf can be obtained through the *Laplace transform approach*. The Laplace transform (LT) of the distribution of sojourn time[8] is derived in two steps. First, the convolutions associated with each possible transition pathway are summed. Then, each convolution is multiplied by the respective probability of transition though that pathway. The number of terms summed is equal to the number of possible pathways from start to absorption. The probabilities associated with each pathway must sum to one. For the example previously considered there are three possible pathways (assuming the stochastic process begins in the first state). The LT of the sojourn time distribution is therefore (using probability parameterisation)

$$LT\{f(t)\} = a1\frac{m1}{m1+s}\frac{m3}{m3+s} + a2\frac{m1}{m1+s}\frac{m2}{m2+s}\frac{m3}{m3+s} + (1-a1-a2)\frac{m1}{m1+s}$$

since $LT_S\left\{\lambda e^{-\lambda x}\right\} = \lambda(\lambda + s)^{-1}$. The pdf is derived by applying the inverse LT. This approach is possible since Cox, 1955 proved that any distribution with a rational LT can be represented in a phase-type structure.

[6] Probability density function
[7] Cumulative density function
[8] The time from the start of the stochastic process until absorption

1.6 Treatment Intensity

The objective of this project is to model activities at the NRC at Rookwood hospital. The main activities are the arrival of a referral and the admission and discharge of a patient. However, it was noticed[9] early on that LOS (discharge date minus admission date) is dependent, not only on the particular attributes of the patient that are readily available on arrival (age, gender, diagnosis etc), but also on the intensity of treatment received post-arrival. In fact, it is the active component of LOS that is affected by treatment intensity. This relates to the time, from admission, until the patient is ready for discharge. From this point until ultimate discharge they are said to be bed-blocking (Ch 1.2) and this period of time is referred to as the blocked component of LOS.

At Rookwood NRC treatment is received according to timetables that are produced weekly for each of the departments. These stipulate which patients receive what treatment, by whom, and for how long. An automated scheduling program is developed for one department (physiotherapy) in order to replicate the procedure in which treatment is allocated (by hand). In brief, the scheduling of treatment requires information on patient demand, details, priority and availability in addition to staff attributes and availability. Such measures are variable and the effect of changes to them on treatment intensity can be quickly determined using the scheduling program. The consequent effect on (active) LOS can then be deduced if the relationship between treatment intensity and (active) LOS is known.

Although the automated scheduling program was designed to replicate the manual timetabling procedure for the physiotherapy department it has since replaced the former by-hand approach. This is due to the many advantages that an automated program can bring to employee scheduling such as an improved solution, less manual input, and better data storage. Whilst these advantages typically provide sufficient motivation to produce an automated schedule this is not the case here. It is therefore worth re-affirming that the aim of the automated program is to replicate the week-to-week scheduling of physiotherapy treatment such that the effects of changes to the aforementioned patient and staff details on treatment intensity can be quickly evaluated. The automated program is therefore successful in satisfying this aim of replicating the week-to-week scheduling of physiotherapy treatment since it actually creates the week-to-week schedules of physiotherapy treatment.

[9] From the literature review and discussions with clinicians

1.7 Summary

In this project queuing theory and scheduling are used to model activities at the NRC at Rookwood hospital. The first phase is to produce an initial, representative model of the unit. This model can be described in the following four chronological stages.

1.7.1 Phase One: Initial construction

Historical data is used in the first stage to determine the characteristics of a small number of homogeneous patient groups. The objective is to deduce the range of values for appropriate branching variables (e.g. age, gender, diagnosis – Ch 6.2.6) that return a low variance of active LOS for each group.

The second stage uses the automated scheduling program (Chapters 3 and 4) to produce treatment timetables by fitting patient demand to staff supply. The aim is to use the program to produce average values of treatment intensity for each patient group that are approximately equal to their empirical counterparts. This is achieved by initialising the program with variable values that are equivalent to the current situation on the ground at Rookwood NRC. That is, the number of beds for each group, number and band level of staff, and number of hours worked each day should all be realistic.

Thirdly, these average values of treatment intensity are converted to average values of active LOS through a two-dimensional line graph for each patient group. These graphs are produced by using historical data to examine the relationship between treatment intensity and active LOS (Ch 6.4.2.3) for patients of each group. Assuming that these graphs and the average values of treatment intensity obtained through stage two are correct then the average values of active LOS should match their empirical counterparts (for each group).

In the final stage, the queuing theoretic model is constructed (Chapter 9). This consists of a number of disconnected homogeneous server queuing systems; one for each patient group (Ch 8.4.2.1). The distribution of referral arrivals for each queuing system is (justifiably – Ch 6.4.1) assumed Poisson and rates are derived from historical data (Ch 6.4.1). A Coxian phase-type and exponential distribution are used to model active and blocked LOS respectively for each system. These are fitted using derived formulae for the parameter estimates (Chapter 5), the Matlab program (Ch 6.5.1) and the average values of active LOS of stage three. Each queuing system can then be solved for steady-state probabilities and performance measures using a purpose built program encoded in Maple (Ch 8.3.1.6). Ultimately, steady-state

probabilities, performance measures and costs are output for the holistic queuing model (Ch 8.4.2.1).

1.7.2 Phase Two: Hypothetical scenarios

The second phase is to use the constructed model to consider hypothetical scenarios. Broadly speaking, these *what-if* questions relate to the *major policy decisions* (who to admit, what care to give, when to discharge) that clinicians have control over. For the purposes of the model such decisions are made not on an individual basis for particular patients but on a macro level. For example, consider the question: what if therapists work weekends? This affects the treatment received by all of the patients and relates to the second major policy decision (what care to give). It can be evaluated by making a change to the scheduling program at stage two. A consideration to the major policy decisions is shown throughout this thesis.

1.7.3 Motivation

The motivation for this project is simple. From the literature review (Chapter 2) no study has been found that applies queuing theory to a neurological rehabilitation setting. Such a study is needed since inpatient neurological rehabilitation is an expensive and highly sought after service. For the NRC at Rookwood hospital annual demand is about 375 (according to Ch 1.2). This significantly exceeds the annual throughput which, with 21 beds (on average) and a mean LOS of five months, is just fifty. The (average) bed cost per day stands at £480.

1.7.4 Outline

A review of relevant literature is provided in Chapter 2. Following this the current process of scheduling treatment is discussed in Chapter 3 with particular attention restricted to the physiotherapy department. Thereafter, the automated computer program for scheduling physiotherapy treatment is described in Chapter 4. This program forms the basis of stage two (Ch 1.7.1). Distribution fitting and parameter estimation are considered in Chapter 5. Phase-type distributions are central to this owing to the aforementioned advantages (Ch 1.5). Chapter 6 is divided into two parts with the first listing a number of models (of increasing complexity) alongside any associated pros and cons in addition to data requirements. The availability of data is considered in the second part and distributions are fitted to the LOS using the results of Chapter 5. A specially designed computer program that solves the multi-channel, fixed buffer queuing system with Markovian arrivals and Erlang service times for steady-state results is detailed in Chapter 7. This is extended to Coxian (to model active LOS)

plus exponential (to model blocked LOS) service times in Chapter 8. In Chapter 9 balking and reneging are introduced to control for the size of the queue as the number in system approaches capacity. The patient groups (stage one) are also determined and so too are the relationships between treatment intensity and active LOS (stage three) for each of these groups. Finally, the queuing model (stage four) is constructed (using the computer program of Chapter 8) and some hypothetical scenarios (phase two) are considered.

Chapter 2: Background and Literature Review

2.1 Introduction

Operational Research (OR) is a relatively new discipline of mathematics. It is thought to have been conceived in the UK in 1936 as part of an MOD study into enemy aircraft detection and interception. Subsequent operations were successful enough that in 1941 an Operational Research Section was established. The objective was to find the most effective utilisation of limited military resources by the use of quantitative techniques.

In the years after the Second World War the application of OR moved toward more domestic concerns. By the early 1950s over forty Operational Research sections had been established in Great Britain (Goodeve, 1953). These sections ranged in size and speciality. Many were based in the private sector whilst others could be found in government departments or research associations. Various applications of the discipline were studied; agriculture, civil aviation, the textile industry, property development, and healthcare. The paper entitled *'Operational Research in Medicine'* (Bailey, 1952) is perhaps the earliest publication that considers the application of OR in healthcare. Since then there has been a plethora of research within this field.

At the core of such research is the production of mathematical models that are, to some extent, representative of real-life behaviour. Richard Gibbs defines three types of healthcare resource allocation models in the book *'Operational Research applied to Health Services'* (Boldy, 1981). First considered are *macro-econometric* models. These contain aggregate variables such as consumption, supply and price of health services, and population attributes. They contain multiple linear equations and may be interpreted as a system dynamics type

model. *Behaviour simulation* models are considered next. These models are constructed by hypotheses that relate to the behaviour of physicians and patients. Finally, *system optimisation* models are considered. Here, resource allocations are determined based on a pre-defined objective function. There are many OR tools that can be employed to construct such models (see Ch 1.3). The most valuable of these with respect to modelling activities at Rookwood NRC is queuing theory (a justification is provided in Ch 1.3).

Queuing Theory (QT) is one of many sub-disciplines of OR. The first academic paper in this field was published by Agner Krarup Erlang in 1909 and was entitled *'The Theory of Probabilities and Telephone Conversations'*. The author continued his studies into the next decade and in 1917 published some of his most influential work. The paper *'Solution of some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges'* establishes formulae for loss and waiting time – these have since become prominent results in the field. Decades later in 1953, a standard for the characterisation of queues was introduced by David George Kendall. This is known as Kendall's notation and is described in Ch 1.4.

A year later, in 1954, a paper entitled *'Queuing for Medical Care'* was published by Norman Bailey. The author relates his study of an inpatient facility to Erlang's work on telephony by considering patients as telephone calls and hospital beds as telephone channels. The length of stay (LOS) is equivalent to the duration of the call. The author deduces the average waiting time (through Erlang's formula) and calculates the optimal number of beds required in the hospital. In subsequent years and decades, research interest in healthcare modelling through QT has developed and there now exist a multitude of studies.

QT has been applied to many healthcare related problems over the years. Cochran & Roche, 2007 use the discipline to estimate demand for inpatient beds at a level 1 trauma facility in the United States. The activities of an intensive care unit are modelled in Griffiths et al, 2006 in which the optimal bed allocation is determined and *what-if* type scenarios examined. Optimal bed allocation is also sought in de Bruin et al, 2007. This is found with respect to a maximum permitted number of refused admissions for emergency cardiac patients. The application of QT has also been extended to outpatient activities. Mayhew & Smith, 2008 use a queuing model to examine patient LOS in an accident and emergency department. Their aim is to evaluate the impact of the government's (2000) target to treat and discharge 98% of patients within four hours. Furthermore, a general medicine practice is modelled in Sobieraj, 2006 by an M | G | 1 queue. In addition to these rather general applications, QT has also been

applied to more specialised healthcare problems (for example, the use of a portable X-Ray machine is modelled in Abujudeh et al, 2005).

A more thorough review of QT applications in healthcare can be found in Fomundam & Hermann, 2007. The authors of this paper make an important distinction between model types. The first type defines models that '*predict how particular healthcare configurations affect delay in patient service and healthcare resource utilisation*' whereas the second type defines models that '*seek to determine the optimal allocation of resources necessary to attain the goals determined by healthcare providers and decision makers*'. Alternatively put, the former outputs performance measures based on healthcare configuration whilst the latter outputs healthcare configuration based on performance measures. These may be considered as equivalent to the behaviour simulation and system optimisation models previously defined by Richard Gibbs. As an example, a representative model of an accident and emergency (healthcare configuration) is constructed in Mayhew & Smith, 2008 and the service time information (performance measures) is output. This is then compared to the government's targets. De Bruin et al, 2007 exemplifies the second model type. The authors are given a maximum permitted number of refused admissions (performance measures) and seek to optimise the bed allocation (healthcare configuration) to avoid this being exceeded.

2.2 Neurological Rehabilitation

Neurological rehabilitation is defined in Wade, 1987 as *'the management of patients with neurological disability'*. However, this 'management' has been inconsistent over time. It is reported in Burke, 1995 that after the Second World War, Acquired Brain Injury (ABI) patients tended to be treated alongside stroke patients. However, this was problematic since these units were typically populated by geriatric patients. This led to an increasing number of ABI patients being admitted onto more general rehabilitation programmes. These were also unsuitable due to the significant cognitive and behavioural dysfunction associated with ABI. By the 1970s brain injured patients were being denied opportunities for rehabilitative treatment in favour of transfer to nursing homes or psychiatric facilities. The 1980s saw a marked change in this practice with the development of specialised neurological rehabilitation programmes across the United States and Europe. Rookwood NRC has operated such a programme since (about) 1990.

It is perhaps the recency of this conception that explains the absence of QT models within the literature. After a thorough search no articles relating QT to neurological rehabilitation could

be found. Despite this, there is evidence of extensive clinical research into the efficacy of these specialist programmes. These have been published widely within journals such as *Brain Injury*, *Disability and Rehabilitation*, *Head Trauma Rehabilitation*, and *Neurorehabilitation and Neural Repair*. The aim of this subchapter is to review such studies that are considered relevant to this project.

An important limitation must first be addressed. As mentioned, Rookwood NRC provides care to those who have Acquired Brain Injury (ABI). However, the majority of studies within the literature consider only Traumatic Brain Injury (TBI) (Teasell et al, 2007). This is problematic since there is evidence of disparity in functional recovery between TBI and non-TBI patients (Cullen et al, 2007). The authors report that '*TBI patients achieve greater functional improvements compared to non-TBI patients*'. Therefore, the limitations in the applicability of TBI studies to Rookwood NRC must be acknowledged.

2.2.1 Prognostic models for outcome and length of stay

Prognostic models are used to predict future values of outcome variables on the basis of current values of predictor variables. They can be used in neurological rehabilitation to predict patient performance measures (mortality, functional ability, LOS) on the basis of pre, or on, arrival attributes (age, gender, diagnosis). This is beneficial for the following reasons. Firstly, a smoother discharge can be facilitated for the patient since the discharge destination and date can be predicted far in advance. This decreases *bed-blocking* and so increases throughput. Secondly, support can be given to the *major policy decisions* (who to admit, what care to give, when to discharge) since a more accurate awareness of current and future capacity is available. This is because future demand for current and prospective (in queue) patients can be forecast.

Owing to these advantages there exist a vast number of prognostic models within the research community. In a recent review (Perel et al, 2006) 102 prognostic models for TBI were identified across 53 studies. A total of 89 predictor variables were used in the models – the most common being Glasgow Coma Scale[1] (50%), age (46%), and pupil reactivity (26%). Mortality (30%) and Glasgow Outcome Scale (28%) were the most widely used outcome variables. The authors conclude that '*although publications of prognostic models are very frequent their quality is relatively poor*'. Possible reasons are insufficient sample sizes (75% of models contained < 500 patients) and lack of external validity (only 11% of models were

[1] Teasdale & Jennett, 1974

validated externally). The lack of representation from low income countries (2% of models) is cited as a drawback to the review since such countries account for 90% of TBI cases. The authors point toward the (at the time, ongoing) MRC CRASH Trial to redress these concerns.

The Medical Research Council (MRC) CRASH Trial is a large-scale international study of TBI prognostic models (MRCCT collaborators, 2008). The study seeks to develop and validate prognostic models for death at 14 days and death or severe disability at six months after TBI. An unprecedented number of patients (10,008) were studied, with many from low-middle income countries (7,526). The authors construct a basic model using four predictor variables (age, GCS, pupil reactivity, presence of major extra-cranial injury) in addition to a more advanced model that requires CT scan results. The models are externally validated in a separate cohort of 8,509 TBI patients. In high income countries, such as the UK, age was found to be the most significant predictor of outcome.

However, outcome in this study (and many others considered in Perel et al, 2006) is restricted to mortality and not functional ability. This is of potential irrelevance to rehabilitation since *'ABI is not thought to affect life expectancy after the initial acute phase'* (Dept of Health, 1998). With respect to functional ability there are many assessment measures and scales that are available. The most common in TBI (Wilson et al, 1998) is the Glasgow Outcome Scale (Jennett & Bond, 1975). Two decades before, the Barthel Index (BI) was created (Mahoney & Barthel, 1965). The authors aim was to create a simple index of independence that was useful in scoring disability. However, this was considered '*too crude and simple*' (van der Putten et al, 1999) and as a consequence, the Functional Independence Measure (FIM) (Keith et al, 1987) was developed. Another measure that is widely used is the Disability Rating Scale (DRS) (Rappaport et al, 1982). There are many studies into the efficacy and accuracy of such measures. For example, Hall et al, 1996 investigate the ceiling effects of the FIM and DRS in TBI and van der Putten et al, 1999 compare the responsiveness of the BI and FIM in TBI.

Black et al, 2000 examine the effect of age, initial admission GCS, rehabilitation admission strength, standing balance and sitting balance on FIM scores at discharge. They find that these predictor variables account for 29% of the variance in the outcome measure. Age and sitting balance were found to be the most powerful predictors. Cognitive outcome (as measured by the Rancho Los Amigos Level of Cognitive Functioning) is studied in Finch et al, 1997. The authors conclude that the results from a simple examination of mental status

(involving tasks such as phrase repetition and eye movement) can account for 27% of the variance in this outcome measure. Many other studies point to early intervention in the rehabilitation of patients as a means of improving long-term functional ability. The articles Mackay et al, 1992 and Cope & Hall, 1982 support this.

In addition to functional ability on discharge, some authors extend their prognoses to consider LOS. Cowen et al, 1997 investigate predictor variables for outcome (measured by motor and cognitive FIM) and LOS in TBI. They conclude that functional ability and LOS can be predicted '*with a high degree of the variance explained by the data known at the time of rehabilitation admission*'. The predictor variables are found to be initial GCS score, CT findings, etiology, presence of fractures, age, length of acute hospitalisation and admission motor and cognitive FIM scores. A similar study is conducted in Feigenson et al, 1977 with respect to stroke rehabilitation. In this investigation, age was found to be unrelated to functional ability on discharge or LOS despite the inclusion of patients from a variety of age groups. Lehmkuhl et al, 1993 confine their research to factors that predict LOS and not outcome in their study of over 300 TBI patients. They find severity of injury, as measured by GCS, to be of statistical significance in predicting LOS.

A thorough review of predictor variables (split to pre-injury, injury, post-injury) for outcome and LOS of TBI patients can be found in Zasler, 1997.

2.2.2 The effect of length of stay on outcome

In addition to the multitude of pre/on arrival attributes that have been considered in the previous subchapter, there exist many post-arrival variables that also affect outcome (functional ability on discharge). Some of these are, to some extent, outside the control of the rehabilitation unit - such as illness or re-admission to acute care. Other variables are, however, under the direct control of clinicians. Patient LOS is one such variable. This is an important measure to consider since it relates to the major policy decisions (i.e. policy 3: when to discharge). It is intuitive to expect that longer LOS is associated with improved outcome.

A recent study in Sandhaug et al, 2010 aims to assess the effect of pre-injury and injury-related factors *'as predictors of early recovery*' for TBI patients in Norway. The authors deduce that '*a longer stay in the rehabilitation unit is associated with a better functional level at discharge'*. It must, however, be noted that only 55 patients are used within the study.

Spivack et al, 1992 investigate the outcome of (95) TBI patients – but in this study patients are split into either a short or long LOS group. The authors conclude that '*patients in the long LOS group consistently made more progress across all outcome variables than patients in the short LOS group'*. The effect of rehabilitation LOS on the FIM outcome measure is studied in Cowen et al, 1997. The conclusion is as expected with $p < 0.01$.

Note that there is considerable crossover with respect to the performance measures (LOS and outcome) and their predictor variables. The former subchapter begins by reviewing studies that predict outcome on the basis of pre/on arrival attributes and ends by extending these prognoses to LOS. The papers considered in this subchapter then support the expected result that LOS is indeed predictive of outcome. Therefore, the variables that predict LOS must also be included in the set of variables that are used to predict outcome.

2.2.3 The effect of treatment intensity on length of stay and outcome

Another post-arrival variable that affects outcome is treatment intensity. This is defined as the amount of treatment received over an interval of time. This is an important measure to consider since it relates to the major policy decisions (i.e. policy 2: what care to give). It is expected that treatment intensity will affect LOS (with outcome held constant) or outcome (with LOS held constant). This combined with the previously identified relationship between LOS and outcome infers a trilateral relationship between these three measures.

The need for a sufficiently intensive rehabilitation programme is reported in Chapter 1.5 of the National Clinical Guidelines for Rehabilitation following Acquired Brain Injury (Turner-Stokes, 2003). The authors advise that '*following acute ABI, patients should be transferred to a rehabilitation programme of appropriate intensity'*. However, in the same chapter, the authors cite a paper (which has since been published with complete results – Zhu et al, 2007) that asks *'What is the optimal amount of rehabilitation that patients should receive?'*, the answer being '*there is no established standard'*. This is indeed true. It is evident from the conflicting results of the many studies that there is a lack of agreement; not only with respect to optimal intensity but also with respect to the efficacy of varying treatment intensity in the first place. Due to the expected trilateral relationship between treatment intensity, LOS and outcome, these studies are henceforth reviewed.

Perhaps the earliest investigation into the effect of treatment intensity on outcome is that of Smith et al, 1981. The authors conduct an RCT[2] involving 133 '*appropriate'* stroke patients who have been discharged to the community from Northwick Park (a DGH in the UK). Patients were allocated at random to three groups; each involving a different level of treatment intensity. Results showed that improvement (measured at three and six months) was greatest in those receiving intensive treatment (twice the level of conventional treatment) and least in those who received no routine treatment. Improvement was intermediate in the group that received conventional treatment. The authors comment that not all patients are '*appropriate'* for the intensive treatment – over 300 patients *'too elderly or too frail'* were excluded from the study. Four years later a paper was published *'The Significance of Intensity of Rehabilitation of Stroke – A Controlled Trial'* (Sivenius et al, 1985). In the study, 95 stroke inpatients were divided into intensive and normal treatment groups. The amount of treatment received by each group varied over time since there was no fixed ratio (*cf.* Smith et al, 1981). At three months after stroke there was a statistically significant difference ($p < 0.05$) between the amounts received, but, this did not follow at six and twelve months. This trend was repeated when gains in ADL (Activities of Daily Living – an independence index also used in Smith et al, 1981) were studied.

Later that decade, a study (Blackerby, 1990) investigating the effect of treatment intensity on LOS for ABI patients was published. This differed somewhat when compared to its predecessors. The author considers the retrospective analysis of the LOS of patients before and after a global change to the rehabilitation programme. That is, the experiment and control groups were not treated simultaneously. Secondly, patient LOS is the dependent variable in this study and not outcome (note that the *'criteria for discharge from these programmes did not change'*). After the change in programme, LOS was found to have decreased by 31% – this represented a 1.5 month reduction on average. In addition, the variability in LOS was found to have decreased. This enables the more accurate prediction of discharge date ergo the facilitation of smoother discharge.

Some studies have investigated the effect of varying the intensity of specific treatments. In Slade et al, 2002 the intensity of physiotherapy and occupational therapy is increased by 67% for the experimental group in an RCT. Again, the outcome was held constant – *'No significant difference in discharge Barthel scores was found'* – and the effect on LOS was

[2] Randomised controlled trial

observed. The result was a significant ($p < 0.01$) reduction in LOS of 14 days. Five years later a similar study (Zhu et al, 2007) was conducted in which physiotherapy and occupational therapy were increased from two to four hours a day. Although there were no significant differences in the mean FIM scores at six and twelve months post-injury; this was not the case at three months ($p = 0.016$). This led the authors to conclude that *'Intensive rehabilitation in this study speeded up recovery rather than changed the final outcome'*.

The effects of changes to both treatment intensity and LOS are considered in Spivack et al, 1992. Patients are classified into high and low treatment intensity groups (for both their first month of rehabilitation and overall average daily amount) based on a median split. They are also classified into long and short LOS groups. 2 × 2 analyses of variance follow and are used to evaluate the effect of changes on cognitive and motor outcome. In a similar fashion to other studies, the authors control for pre/on arrival attributes by including these moderator variables as covariates in the analyses. Instead of deducing these statistically from their own cohort, they are obtained from the results of previous research. It is found that *'patients with longer LOS made more progress across all outcome variables than patients with shorter LOS'*. This is because patients with longer LOS were *'significantly more disabled'* initially and so needed more rehabilitation to *'reach the same levels at discharge as the short LOS patients'*. Treatment intensity was shown to have a significant effect on cognitive outcome measures (higher-level cognitive skills and Rancho level) but not motor outcome measures.

Research has also been undertaken in efforts to evaluate the effect of changes in intensity of individual therapy types on outcome and LOS. A multicentre survey involving 491 TBI patients is presented in Cifu et al, 2003 in which the effect of variations in speech, occupational, physical and psychological therapy is studied. In investigating the predictors of treatment intensity (as therapy levels were not fixed in the study) the authors find that *'a lower age and a shorter onset-admission interval were significant predictors of increased psychological therapy intensity'* and that *'FIM motor score at admission was a significant predictor of speech therapy'*. It was also found that treatment intensity (of any type) was not predictive of gains in the FIM cognitive score. Other results show that *'longer LOS in rehabilitation significantly predicted greater motor potential achieved'* and that *'speech and physical therapy were significant predictors of motor outcome at discharge'*. Many of the pre and on arrival attributes considered were found to be significant predictors of LOS, although, therapy intensity was not. A possible limitation to this survey is the *'limited variance in intensities among patients'*. Note that this survey is in fact a '*replicate'* of Heinemann et al,

1995. Whilst many similarities exist with respect to the study designs these congruities did not extend to the results. Contrarily Heinemann et al find that *'only intensity of psychology services seemed to have any relation to functional gain'*. The authors cite spontaneous recovery and floor effects of the FIM as possible explanations of this.

To conclude the review of this subchapter it can be firmly asserted that *'there is a tenuous relation between the intensity of therapy services provided after TBI and the functional outcome'* (Cifu et al, 2003). This can be seen by the noticeable disparities between the results of the studies considered above. However, on the whole *'more therapy is better'* (Cullen et al, 2007) and, according to the authors, despite the short-term benefits of high over low intensity rehabilitation *'these differences are gradually reduced over time'*. This is an interesting concept that is supported by the results of Sivenius et al, 1985 and Zhu et al, 2007.

2.2.4 Costing

Costs usually form an essential part of any study in healthcare. Beecham et al, 2009 explain a simple method of calculating patient costs at healthcare facilities – *'A unit cost for treatment and care is multiplied by the length of stay to obtain an average 'episode' cost for that placement'*. However, a more accurate estimation can be achieved if a distinction is made between fixed and variable costs. The fixed costs represent generic overheads whilst variable costs are patient-specific and relate to the particular treatment that is received. This is supported by Turner-Stokes, 2007 *'the cost of providing rehabilitation is largely determined by time and effort'*, that is, by LOS and treatment intensity.

There have been many papers published that have documented various costs associated with inpatient neurological rehabilitation. For example, in Beecham et al, 2009, the disposition of young adults with ABI is divided into four groups according to complexity. The groups range from short stay acute to long stay institutional with associated average costs of £240 (one-off) and £33,900 (per annum) respectively. Another example, in Whitlock et al, 1995, examines the *'Functional Outcome after Rehabilitation for Severe TBI'*. Severe TBI is defined in this study by a FIM score of 18 – the lowest value possible. The average daily cost of rehabilitation was '*roughly $1,000'* and the efficacy of the rehabilitation programme was confirmed by an average discharge score of 65.

However, *'it is not sufficient to demonstrate merely that rehabilitation interventions are effective or lead to improvements in quality of life. They must also be demonstrably cost-*

efficient' (Turner-Stokes, 2003). For a proposal to be cost-efficient, *'the initial investment in rehabilitation must be offset by the ongoing savings on care within a reasonably short timeframe'* (Turner-Stokes et al, 2006). The earliest review of TBI rehabilitation efficacy is published in Cope, 1995. In this, the author cites a study (Max et al, 1991) which estimates the total annual cost associated with brain injury in the US at $44bn. The authors of this study find that 54% of this is due to work loss and disability and assert that *'it is in this area that rehabilitation has the greatest potential to affect savings'*.

The validity of this assertion is strengthened by many subsequent studies, including Turner-Stokes et al, 2006. In this paper two methods are compared for evaluating the cost efficiency of specialist rehabilitation for ABI patients. Firstly, (297) patients are divided into three groups according to Northwick Park Dependency Score[3] on admission. For each of these groups the mean changes in NPDS, NPCNA[4] and FIM (from admission to discharge) are calculated. Following this the NPCNA is used to estimate the mean reduction in weekly cost of community care as a result of inpatient rehabilitation. The time taken to offset the cost of inpatient rehabilitation is then calculated by dividing the mean cost of inpatient rehabilitation by the mean reduction in weekly cost of community care for each group. Results show that the low dependency group has an offset time of 38 months whilst the medium and high dependency groups have times of 21 and 16 months respectively. Since ABI patients who have survived past acute care can expect normal life expectancies (Dept of Health, 1998) this can represent significant long-term savings.

The other method for evaluating cost-efficiency is related to the FIM measure. FIM efficiency is calculated by dividing the mean change in total FIM score by LOS. The values for the low, medium, and high dependency groups are 0.17, 0.25, and 0.16. Results for the low and medium groups *'are to be expected'* and follow a trend similar to that exhibited by the NPCNA. However, FIM efficiency of the high dependency group fails to sufficiently acknowledge the evident financial benefits associated with inpatient rehabilitation for this

[3] The NPDS (Turner-Stokes et al, 1998) is an *'ordinal scale that can be used to assess the impact of nursing time in a rehabilitation setting'*. It was designed as part of a *'larger project to develop a tool for predicting care needs in the community from hospital'*. Measures such as the FIM and BI are inappropriate to this end because although they *'have been shown to correlate with care needs'* they *'cannot be used to assess them directly, as they do not indicate the number of people required to help with a task, nor the time taken'* – this is the role of the NPDS

[4] The completion of the project was marked a year later with the publication of the Northwick Park Care Needs Assessment (Turner-Stokes et al, 1999). The NPCNA represents an algorithm that translates NPDS scores into an *'estimate of care needs in total care hours, or a timetable of care'* ahead of discharge to the community

group of patients. The authors cite the reputable *'floor effects'* of the FIM (van der Putten et al, 1999) as an explanation to this.

A year later, the study is extended (Turner-Stokes, 2007) to examine the cost-efficacy of providing ABI patients with longer rehabilitation programmes. This study aims to *'determine whether highly dependent patients who remain in rehabilitation beyond the average of 3-4 months continue to make meaningful change'*. This change must be justifiable with respect to cost-efficiency. Results are presented in a similar fashion as before. Although not implicitly referred to, it can be deduced from these results that after four months of rehabilitation the mean cost is £35,687 and the mean reduction in weekly community cost is equal to £833. Therefore, an offset time of ten months is estimated for a typical rehabilitation programme. However, since the mean LOS of these (51) patients is in fact six months, the effect of the additional time can be examined. The mean total cost is found to be £52,453 and the mean reduction in weekly community cost is equal to £950 – giving an offset time of 14 months. Note that the additional two months rehabilitation has reduced weekly community costs by £114 by reducing the amount of weekly care required from eight to five hours. The author concludes that *'the additional investment … was offset relatively quickly by long-term savings in the cost of care'*.

Other studies examine the cost-efficacy of providing ABI patients with higher intensities of treatment. In the previously reported study of Blackerby (1990) it is found that higher intensity treatment (from five to eight hours per day) reduces LOS by 48 and 53 days respectively for comatose and acute patients. Since the *'criteria for discharge from these programmes did not change'* this represents a saving in fixed costs of $16,950 and $18,504 respectively (fixed cost per day is $350). However, the author fails to acknowledge the significant effect of treatment intensity on variable costs. The following analyses are therefore undertaken using raw results from the study. The average variable cost per day is calculated as $435 since the average total cost per day is given as $785. If it is assumed that variable costs are linearly related to treatment intensity then the increase in treatment hours from five to eight hours per day lifts these costs from $435 to $696 per day. The average pre-change costs can be calculated as $119,940 (152.79 × $785) and $129,760 (165.3 × $785) respectively and post-change as $109,160 (104.36 × $1,046) and $117,601 (112.43 × $1,046) respectively. Therefore, the average savings are $10,780 for comatose patients and $12,159 for acute patients. Similar studies that also conclude favourably include Shiel et al, 2001 and Slade et al, 2002.

2.3 Queuing Theory Models (Geriatric)

Queuing Theory (QT) models applied to geriatric care are considered within this literature review since, in the absence of neurological rehabilitation models, they represent the closest alternative. This is due to the similarities between the transition of geriatric and ABI patients through healthcare facilities – typically this consists of acute/rehabilitative care followed by some form of continued care (e.g. community visits, nursing home). There exists a considerable amount of research within the field of geriatric care modelling with substantial contributions made by Queens University Belfast in collaboration with St George's Hospital in London. Some of these studies are systematically reviewed in order of complexity.

The specification of the inter-arrival and service distributions are fundamental requirements in the characterisation of a queuing system. With respect to the modelling of healthcare facilities, most studies either find or assume the negative exponential distribution (Ch 5.5.1) to be the most appropriate in modelling inter-arrival times. This assumption has been shown to hold water in the case of unscheduled arrivals at an inpatient facility (Young, 1965). More recently, Irvine et al, 1994 and Taylor et al, 1998 are among those that have verified this assumption for geriatric data. What is usually more challenging is the specification of the service distribution.

In an early study, (McClean & Millard, 1993), the service distribution for geriatric inpatient admissions is investigated. The study examined LOS data for 6,994 patients admitted to geriatric departments across a London borough over a 16 year period. The authors fit a two-term mixed-exponential distribution (Ch 5.5.2.1) to the data. This has previously given a *'good fit'* when applied to geriatric data in Millard, 1988 and Harrison & Millard, 1991. The distribution has been chosen to represent the two different types of patients – those requiring acute/rehabilitative care and those requiring long-stay care. Whilst this distribution produces a satisfactory fit, a better result is sought by applying other distributions to the LOS data. The motivation for this is the failure of the mixed-exponential distribution to represent a high probability density that is found at a low, but nonzero, LOS. This peak is as a result of early discharges or deaths. The author concludes by commenting that a *'log-normal and exponential mixture was seen to provide a better description'* but at the cost of *'the estimation of an additional parameter'*.

Later that decade, a Coxian phase-type distribution (Ch 5.5.4) is fitted to male geriatric service time (LOS) data in Faddy & McClean, 1999. The covariates age and year of

admission were incorporated into the distribution by allowing the dependence of model parameters on these variables through log-linear functions. The fitting procedure is maximum likelihood and computations are performed in Matlab. Again, the emphasis is firmly placed on finding a distribution that represents the early peak (describing frequent early discharges and deaths) as well as the heavy tail (describing exceptional but significant long stays). A four phase distribution is decided upon. Note that whilst the mixed-exponential distribution of McClean & Millard, 1993 had phases in parallel, this distribution has phases in series. This means that a greater LOS is, on balance, associated with a greater number of phases occupied prior to absorption (death, discharge, transfer). In addition to the quality of fit the authors also praise the *'interpretable structure'* of the distribution. For example, the phases could be interpreted as acute, rehabilitative, long stay and resident.

Some years later a novel approach to incorporate the effect of patient-related variables on LOS is published in McClean & Marshall, 2003. To this end, the authors use a conditional phase-type distribution; a Coxian phase-type distribution conditioned by a Bayesian belief network[5]. In this case, *'the phase-type distribution represents the continuous variable (LOS) and the BBN represents the network of interrelated variables'*. The aim is to produce accurate estimations of patient LOS based on appropriate causal variables. The model is defined as *'consisting of causal nodes belonging to the causal network and process nodes representing the phase type distribution'*. Figure 2.1 depicts the causal and process networks that are produced using the CoCo package (Badsberg, 1992).

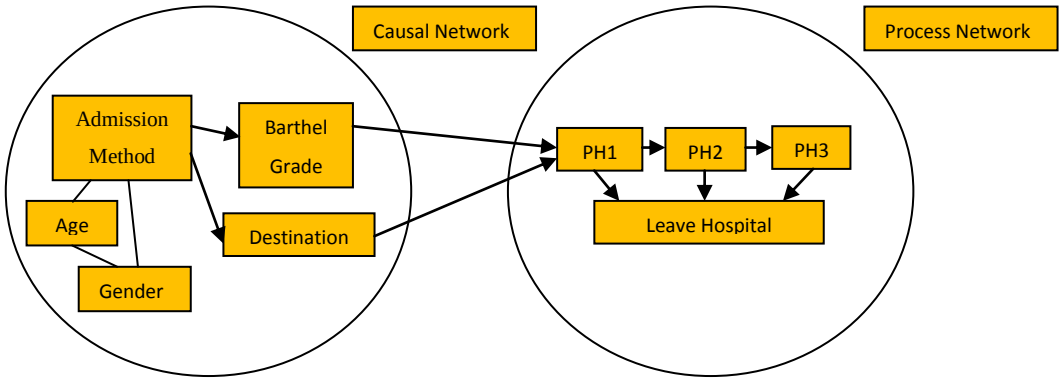


Figure 2.1 A conditional phase-type distribution for length of stay in a geriatric unit

[5] A BBN is a *'directed acyclic graph where each node represents a domain variable and each arc between nodes represents a probabilistic dependency, quantified using a conditional probability distribution for each node'* (Cheng & Greiner, 2001). They are used in Marshall et al, 2001 to predict discharge destination and LOS for geriatric patients based on a number of causal variables (age, gender, admission method etc)

An association has been found between age, gender, and admission method. Consequently, admission method has a direct influence on Barthel grade and destination which in turn have direct influence over LOS. Since there are three discrete options for destination and four for Barthel grade there exists twelve Coxian LOS distributions. These are calculated by maximum likelihood through the Nelder-Mead algorithm and are composed of either one or two phases. According to the authors, the benefits of such a model include *'providing users with a better understanding of the system'* and *'facilitating the inclusion of prior specialist knowledge about variable dependencies and causal relationships'*.

Other studies have included post-discharge care as a consideration within their model. This permits a more comprehensive representation of geriatric care as those admitted to geriatric wards often require some form of support on discharge. The definition of service time is thus extended from LOS in hospital to include the duration of time spent in the community (with care provided by social services). The mathematical composition of this service time is investigated in Faddy & McClean, 2005. The authors fit phase-type distributions of orders three and four to hospital and community LOS respectively. Age and year of admission are included as covariates within the model. The hospital phases are interpreted as initial assessment, diagnosis, rehabilitation and long-stay and care ended with transfer, discharge or death. It was found that *'75% of patients went through phases 1 and 2'* and that those who went through the first phase only were either discharged (younger patients with little wrong) or died (older patients with serious illness). If the disposition is discharge then the patient is discharged to the community and LOS modelled by the order three phase-type distribution. Care in the community ended with either re-admission to hospital or death. In conclusion, the authors comment that *'models based on phase-type distributions were appropriate for describing times spent in care'*.

A QT model for describing the movement of geriatric patients between hospital and the community is presented in Taylor et al, 1998. The authors consider a three phase service distribution (two phases for hospital and a single phase for community) with readmissions and Poisson arrivals. The expected number of geriatric patients in each of the phases over time is deduced in graphical form. The steady-state results can be derived from this by choosing a sufficiently large warm-up period and can be used to estimate the number of beds required. The model verification is then addressed. A cohort of patients admitted in December 1976 is considered. The observed numbers of these both in hospital and the community are recorded for various times after admission. These numbers are contrasted

with expected figures obtained from the model. These differences are *'not significant until 200 days'* according to chi-square goodness of fit tests. The assumption of Poisson arrivals is also tested and verified to be appropriate. Finally, the authors move to look at the effect of what-if scenarios on the number of beds required. These include *'a 10% reduction in the admission rate'* and *'a 10% increase in the rate of release into the community'*.

A similar investigation is published in McClean & Millard, 2007. This builds upon the earlier work of Faddy & McClean, 2005 by including the service time distribution (found in this study) within a QT model with Poisson arrivals. In addition, relative cost ratios are estimated for each of the phases based on a £150 daily cost for the acute care phase of hospital care. The effect of policy changes can be represented by adjustments to the transition rates between phases; the output being the impact on hospital and community costs. The authors describe the notion that *'keeping patients longer in hospital, and improving their fitness for discharge can reduce the transfer of patients to secondary care systems and may in the long-term both improve hospital performance and reduce costs'*.

The statement of this really does underline the congruity between geriatric and ABI care; not only in terms of the relationship between the hospital and the community but with regard to the advantages of providing greater functional ability on discharge (Turner-Stokes, 2007). Therefore, in the absence of neurological rehabilitation models, the mathematical formulations provided in these geriatric studies are of great interest when modelling the flow of ABI patients.

Chapter 3: Outline of Treatment Scheduling at Rookwood Neurological Rehabilitation Centre

3.1 Introduction

The Neurological Rehabilitation Centre (NRC) at Rookwood hospital provides specialist care to patients by means of a multidisciplinary team[1]. This team comprises six specialities; physiotherapy, occupational therapy, speech and language therapy, psychology, dietetics, and orthopaedics[2]. Therapists from these specialities form the *therapeutic departments* of the NRC whose efforts are directed toward improving cognitive and motor function. In addition, the nursing and medical departments are responsible for maintaining the physical health of the patient. Nursing staff operate around the clock and are available as and when required.

Treatment is provided by the therapeutic departments by means of pre-arranged sessions. These involve the congregation of patient and therapist(s) at a particular time and for a pre-determined duration. Such times are obtained from a timetable that is produced periodically and in advance of treatment. Timetabling is defined in Wren, 1996 as *'the allocation, subject to constraints, of given resources to objects being placed in space time, in such a way as to satisfy as nearly as possible a set of desirable objectives'*. Sessions may only be placed within typical working hours – approx 0800-1700 Monday to Friday. The scheduling process is undertaken entirely by hand.

[1] *'Multidisciplinary teams are made up of a group of professionals who work alongside one another to meet the needs of the patients'* (Embling, 1995)
[2] In approximate order of size measured by number of Full-Time Equivalents (FTEs)

The motivation for investigating the scheduling of treatment stems from the growing body of research that suggests a relationship between treatment intensity and length of stay (Ch 2.2.3). Since length of stay (LOS) is *'the primary significant predictor of rehabilitation charges'* (Cifu et al, 2003) it is important to study how this can be affected. It is also important since treatment intensity relates to policy two of the *major policy decisions* (what care to give) and LOS relates to policy three (when to discharge).

This chapter commences by providing an overview of the treatment scheduling process. Motivation for an automated scheduling program is then discussed and possible applications are presented. Attention is then restricted to the case of physiotherapy scheduling for a more in-depth commentary. An introduction to the physiotherapy scheduling procedure at Rookwood NRC is then provided in Chapter 3.4. Following this, a formulation of the problem in terms of soft and hard constraints is presented in Chapter 3.5. Various limitations are thereafter discussed in the conclusion to this chapter.

3.2 Scheduling Rehabilitation Treatment

3.2.1 Literature review

A literature search with respect to the scheduling of rehabilitation treatment yielded very few results. Furthermore, the studies that were found shared very little in common with the NRC at Rookwood hospital. Due to this lack of congruity only two papers are reviewed.

The scheduling of patient care activities is studied in Beggs et al, 1971. This is with respect to a 53 bed inpatient rehabilitation facility. The hospital contains a number of therapy disciplines – physical, occupational, and vocational. A computer program is designed that produces individualised care plans for new patients based on *'pre-admission diagnosis, impairments, and complications'*. This takes the form of a daily schedule containing the times at which various events and activities occur. This is then printed and copies are handed to each department as well as to the patient. The same schedule is then used day-after-day. Modifications can be made to the personalised schedule in response to changing circumstances.

Other less-related studies have also touched on the issue to some extent. Chien et al, 2008 investigate the scheduling of rehabilitation therapy in an outpatient setting by using a genetic algorithm. Since some of the treatment must be provided in a chronologic order they structure their problem as a hybrid shop scheduling problem – *'an open shop scheduling problem with*

partial precedence constraints'. Whilst this is an interesting study it has little commonality with the neurological unit at Rookwood hospital.

3.2.2 Treatment scheduling process at Rookwood NRC

Each week (*preparation week*) a schedule is produced at Rookwood NRC that specifies patient treatment for the forthcoming week (*treatment week*). This is the patient timetable overview (PTO) and provides details of all scheduled sessions for each patient across all departments. The PTO contains as many individual timetables as there are patients; with each timetable specifying a patient's treatment for that week. It is constructed on a department-by-department basis in order of increasing size[3]. This construction therefore represents a process of individual scheduling procedures that are now described.

At the beginning of the preparation week the PTO is reset before any pre-planned patient engagements (such as home visits) are added. It is then 'passed' on to the smallest department. This department then schedules their treatment around the periods of patient unavailability. When complete the PTO is passed on to the next department where the process is repeated. This is continued until treatment from all departments has been scheduled.

SAT	SUN	MON	TUE	WED	THU	FRI
Nursing staff produce a care plan		Smaller departments produce timetables		Occupational Therapy produces timetable	Physiotherapy produces timetable	PTO printed and distributed

Figure 3.1 Treatment scheduling process for the preparation week

This is to ensure that a fair and appropriate number of session allocations are achieved for each department. Since patient availability is reduced with each session allocation the search space of feasible allocation options is also reduced. This limits the extent to which subsequent departments can provide treatment and so reduces the quality of care respective of that speciality. This is problematic for any department that schedules after others; however, the problem is exasperated for smaller departments due to restricted flexibility. It is less of a problem for larger departments owing to a greater coverage of time (due to more FTEs).

[3] Measured by number of FTEs (full-time equivalents)

Therefore it can be noted that whilst the departmental scheduling procedures appear unconnected they are, in fact, related since the scheduling of treatment by one department affects the subsequent scheduling of treatment by another. This is largely because of the above-mentioned effect of session allocation on patient availability but is also because of departmental co-operation in the provision of sessions. This is when therapists from more than one department provide treatment to a patient and is an example of inter-disciplinary teamwork (see Mandy, 1996). Such sessions are typically scheduled and added to the PTO by a department whose turn it is to timetable. This is often as a result of informal discussions between the departments concerned.

The departmental scheduling procedures that are employed are at the discretion of the department. They typically concern the fitting of patient demand for treatment to the available supply of therapists. In the case of smaller departments this is done by roughly working out which patients need treatment and assigning as much as possible to available therapist times. However, a more formal approach is required for the larger departments. This is due to the added complexity associated with a larger search space of feasible session allocation options. Here, demand is typically determined in advance and fitted to the available supply of therapists by means of a simple algorithm. Some departments assign priority levels to patients in order to influence the likelihood that demanded sessions are received. By and large it is exceptional for demand to be less than supply.

3.3 Automated Scheduling Program

An automated program is proposed for inclusion within this project. Firstly, the background of such a program is presented. Secondly, the motivation is discussed. And thirdly, the capacity in which it can be applied is considered.

3.3.1 Literature review

An automated scheduling program is a computerised software package that is used in the production of timetables. They work by optimising a set of objectives whose values represent the quality of the timetable as defined by the user. This optimisation typically takes place with respect to a number of *hard constraints* whose satisfaction is essential. Whilst the actual scheduling is automated the user is still required to input certain details prior to the commencement of the program. These details provide a level of sensitivity with respect to the period of time that is being scheduled for.

Automated scheduling programs have been harnessing the power of computers for many decades. The first such example is in the 1960s where computer scheduling is applied in the field of education. Early in this decade a study is published that investigates the application of machines in the production of a high school master schedule that maps out *'who is teaching what, to whom, and at what time and place'* (Bush et al, 1961). Following this, a computer scheduling package introduced as GASP (Generalised Academic Simulation Program) is presented in Kiviat & Colker, 1964. This too is a program for producing high school master schedules. It is trialled at three different schools across the United States for the academic year 1963-1964. According to the principal of one such school *'the computer-built schedule has fewer conflicts than does the handmade schedule'* and has been produced with *'less over-all cost'*. These benefits are echoed from the other schools. However, disadvantages are also acknowledged. Another principle remarks that *'to use GASP well requires someone with the proper know-how, an expert'* and that *'the educator feels somehow uncomfortable in relinquishing a measure of control of the schedule to a computer or even a computer man'*.

In these early days of automated scheduling there have been related studies in the field of education that have considered university examination timetabling. A method for scheduling final examinations to yield a minimal number of conflicts is published in Broder, 1964. This involves assigning courses to blocks of exam periods and production by hand is described as *'a tedious and frustrating task'*. Therefore, a mathematical formulation of the problem is presented and solved through a computer program. The author then analyses the effects of a different number of blocks on the number of conflicts for two sample problems. In addition, it is mentioned that the objective can be extended to include the minimisation of consecutive exams.

One of the early studies involving the application of automated scheduling to healthcare is presented in Miller et al, 1976. The authors investigate the cyclical scheduling of nursing staff to treatment shifts (cycle length is typically five weeks). The program aims to satisfy both the administrator's need to *'have enough nurses on duty'* and the nurse's preference with regard to *'rotation patterns, weekends off, long working stretches'*. This is achieved firstly with the creation of an initial timetable. The work patterns of individual nurses are then varied whilst holding constant the patterns of others. Should these permutations result in a lower value of the cost function then the change is accepted; otherwise it is rejected. The terminating criterion is fulfilled when no change can be made that reduces the value of the cost function. This is known as *iterative improvement* (Ch 4.3.1).

Since these early studies there has been a plethora of further research within these fields. Burke & Petrovic, 2002, Burke et al, 1996 and Schaerf, 1999 review the developments made by automated scheduling in education whilst Petrovic & Berghe, 2008 review those made in nurse rostering. There have been many other applications of automated scheduling over the years; Elmaghraby & Park, 1974 study the scheduling of jobs on a number of identical machines, Yen & Birge, 2006 investigate the scheduling of airline crews, and Kendall, 2007 studies the scheduling of English football fixtures over the Christmas and New Year period.

Over the decades there have been many advances with regard to the automated scheduling process; existing optimisation algorithms have been developed, new ones have been found, computation times have decreased, GUIs are more user-friendly etc. Despite all of these, however, there are some things have remained unchanged – the benefits and drawbacks of an automated program are the same as they ever were back in 1961 when GASP was introduced. That is improved solution quality and reduced employee input at the expense of complexity to user and reluctance to relinquish control.

3.3.2 Motivation

The previously identified benefits associated with an automated scheduling program typically provide a sufficiently strong motivation to develop such a tool. However, this is not the case in this project. There is a substantial base of research to suggest a relationship between treatment intensity and LOS (Ch 2.2.3). With sufficient data from Rookwood NRC this relationship could be developed and used to condition a service time (LOS) distribution as part of a wider queuing theoretic model. This would enable the prognosis of LOS based on a typical intensity of treatment. It would also enable the extrapolation of the model to cover *what-if* type scenarios relating to changes in treatment intensity and their consequent effect on LOS, throughput, and other performance measures.

However, treatment intensity is not a measure that can be expressly regulated by the therapists. It is, in fact, dependent on variables relating to the case-mix of both therapist and patient through the scheduling procedure. These controllable measures (see Ch 3.4.3) can be thought of as the independent variables (such as staff availability and patient demand) and can easily be related to by therapists.

In order to ask what-if questions relating to treatment intensity it is necessary to produce a schedule with the test values of the controllable measures. For example, what if the number

of therapists was increased by 20%? Here, a couple of new therapists would be added to the roster and a schedule produced. Treatment intensity could then be deduced for this scenario. Because of the time it takes to produce a schedule and the number of potential scenarios that are of interest it is not a possibility to rely on the therapists to produce these schedules by hand. Hence an automated program that mimics the scheduling process is required.

If the automated program is used to replace the current by-hand approach then there are other advantages. Data on the amount of care received could easily be collected and stored. This can then be used to develop the relationships between treatment intensity and LOS. Also, data required for audit purposes could be output.

3.3.3 Application to Rookwood NRC

The automation of multilateral treatment scheduling for all departments is a possibility that is considered. This involves the simultaneous scheduling of sessions from all departments through a computer program. Whilst this ensures the possibility of finding an optimal solution (with respect to the PTO), the scale of restructure and lack of research potential mean this approach would be inappropriate.

Therefore, it is more appropriate to retain the sequential scheduling process of before whilst automating the individual procedures of each department. This is more amicable to the departments and requires less re-structuring. However, the automation of scheduling is not an appropriate course of action for every department. There is little to be gained from the introduction of such a program to the smaller departments. This is because an optimal or near-optimal solution can be found with relative ease by hand due to the small search space. It is of greater benefit to larger departments such as speech and language, occupational therapy, and physiotherapy. Scheduling by hand for these departments is both ineffective, as it is difficult to find the optimal solution, and inefficient, as the process is heavily time-consuming.

Nevertheless, there exist significant differences between how these larger departments schedule. These differences relate to operating procedures and scheduling procedures. For example, the speech and language department can provide bedside therapy whereas the occupational therapy department require the use of specific rooms. An example of the latter can be seen more generally in the scheduling procedure; a single employee is tasked with timetable production in the physiotherapy department whilst employees schedule their own

treatment in occupational therapy. These differences make it impractical to produce a generic automated program that can be used across these departments.

An automated program is therefore developed for the physiotherapy department only. This is because it is the largest department by a considerable measure (physiotherapy department has nine FTEs; next largest department is occupational therapy with only five FTEs, 2010). It is also because the current scheduling procedure of the physiotherapy department has an attractive foundation that an automated program can be built upon. This exists in the form of a clear framework of times at which patient sessions are permitted to occur as well as an algorithm governing assignment priorities. In addition, the largest study (Cifu et al, 2003) into the effect of treatment intensity on LOS ($n = 491$) found that the provision of physiotherapy was a significant predictor of motor outcome at discharge. Should this be the case then this grants considerable flexibility in influencing LOS or outcome by varying this measure. These qualities are useful in terms of providing coherent rules to which the program must adhere to and their existence ensures minimal disruption to the structure of the procedure for staff.

3.4 Scheduling Physiotherapy Treatment

3.4.1 Literature review

A literature search was undertaken with the aim of identifying studies in which physiotherapy treatment is scheduled. Unfortunately, very little was found. Two studies are briefly reviewed.

An outpatient physiotherapy clinic is considered in Bourque, 1980. The author states that the scheduling procedure *'influences patient punctuality and compliance, congestion of the premises, working atmosphere and quality of care'*. Since Rookwood NRC is an inpatient facility patient punctuality and congestion are measures that need not be taken into account. The appointment system described in this study *'relies on specific descriptions of therapies with explicit reference to the human and material resources involved'*. Since some therapy requires the consecutive use of different material resources over the course of a session the problem is formulated as a job shop scheduling problem. The difference is that *'once initiated a patient treatment must be performed without interruption'*. The aim is to homogenise the activities of therapists (i.e. distribute workloads evenly). This is achieved by evaluating the

objective function for each feasible appointment period and selecting one that produces a minimum. The objective function is defined as the sample variance of therapist activity.

Ogulata et al, 2008 also consider an outpatient physiotherapy clinic. Before their study the clinic operated in an inefficient manner due to the poor weekly scheduling procedure. Patients were allocated days in which their treatment would occur but were not given specific times or even time windows. This meant that patient waiting time was *'very long'*. In addition, the assignment of patients to physiotherapists was made *'randomly'* and *'fairness among them is not taken into account'*. The authors sought to *'maximise the number of patients treated'*, *'obtain a balanced distribution of patients among physiotherapists'*, and *'schedule patients to a particular period of time in a work-day'*. These are achieved sequentially by splitting the procedure into three stages. The first stage selects a list of patients for inclusion within the treatment week. The second stage assigns these patients to therapists (based on a fair allocation), and the third stage schedules the therapy to particular windows throughout a work-day. The authors conclude that *'a decision support system based on proposed mathematical models may further enhance the usefulness of these models'*.

3.4.2 Physiotherapy treatment sessions at Rookwood NRC

Table 3.1 provides details of patient treatment sessions that can be provided by employees from the Physiotherapy department. Additional commentary is provided for the types of session that require further explanation.

Table 3.1 Therapy session attributes

Session type:	Duration (minutes)	Number of physiotherapists	Number of patients	Departmental provider	Therapist preference	Specific day/time
Single	45 or 60	1	1	Physiotherapy	Primary/ secondary	–
Double	45 or 60	2	1	Physiotherapy	Primary/ secondary + 1 assistant	–
Triple	45 or 60	3	1	Physiotherapy	Primary/ secondary + 2 assistants	–
Group	45 or 60	1	2 – 4	Physiotherapy	Any	see text
Hydro	Undefined	Undefined	1	Physiotherapy	Any	see text
Stretch	30	1 or 2	1	Physiotherapy	Any	see text
Joint	Undefined	1 (typically)	1	Two depts	Primary, secondary	–
Patient meeting	45 or 60	1	1	Physiotherapy	Primary M, secondary M	see text
Goal planning meeting	Undefined	1	1	All	Primary M, secondary M	–
Other patient interactive slot	Undefined	Undefined	Undefined	Any	Any	–

Note an assistant is an untrained physiotherapist (band level less than five).

Group sessions and **stretches** can take place only in certain periods of time. A start and end time is specified for each availability period on each day of the week.

Hydrotherapy sessions take place in the purpose-built swimming pool of the hydrotherapy suite. Since the pool is used by the other wards of Rookwood hospital (spinal unit, ALAC[4]) there are specific access times available to the NRC. There is no formal requirement governing the ratio of physiotherapists to patients in the pool – this depends entirely on the ability of the staff and the condition of the patient(s). Neither is there a fixed amount of time

[4] Artificial limb and appliance centre

allocated for the individual patient sessions although they are predominantly of 30, 45, or 60 minute duration. An employee is usually required to be present at the poolside when any hydrotherapy sessions are taking place.

Patients are categorised as either requiring **stretches** for *maintenance* or *treatment preparation*. If for maintenance the stretches can be considered as stand-alone sessions whose assignment can be made independent of any other session type. If for treatment preparation it is desirable (but not essential) that any stretches be in advance of the treatment sessions that they are in preparation for – singles, doubles, triples and joint sessions[5]. The purpose of the stretch is to prepare the patient for their upcoming post-stretch session by reducing muscle tensions. This allows the patient to perform at a higher level in the session. If a stretch is not provided in advance of any post-stretch session then a basic stretch can be provided at the beginning of the session itself. This however is undesirable since it reduces the therapeutic time that is available for the single, double, triple or joint session.

Patient meetings, if scheduled, must take place in advance of any goal planning meetings (if scheduled). There is typically a maximum of one patient meeting each week and it is mandatory that either the primary or secondary therapist be in attendance.

Goal planning meetings take place every six weeks and are used to set targets on behalf of the patient for the duration of time running up to the next such meeting. They are attended by representatives of all departments[6]. Upon the exhibition of sufficient functional ability the nature of the meetings is changed from goal planning to discharge planning.

Other patient interactive slots are miscellaneous sessions that can involve any number of therapists from any number of departments. Their duration is also undefined.

3.4.3 Physiotherapy scheduling process at Rookwood NRC

The determination of the scheduled recipients of weekly physiotherapy treatment sessions (i.e. treatment intensity) is dependent on variables that relate to the case-mix of both therapist and patient. These variables are known as the *direct controls* and consist of patient demand, attributes, priority and availability in addition to therapist availability and attributes.

[5] Henceforth referred to as post-stretch sessions
[6] Must be primary or secondary therapist from physiotherapy department

Patient demand is deduced by means of a weekly ward- round. The number of requests for different types of physiotherapy sessions is thus recorded on behalf of each patient. This is carried out on a Wednesday afternoon and involves all therapists within the department.

Patient attributes consist of various properties that relate to the patients. Included within these is a specification of the patient's primary and secondary physiotherapist. The complexity of the patient and the therapist band level are factors that are taken into account when distributing primary and secondary roles. There may be a requirement for the primary or secondary therapist to be in attendance depending on the type of treatment session. There are other properties as well such as the preference to time of day, the number of employees required to stretch, and whether the patient attends lunch or not. However, these are typically known by most employees and are not formally documented.

Patient priority levels are determined in the timetabling phase of the process; just before any scheduling actually occurs. This six point scale describes the extent to which emphasis should be placed on patients receiving requested sessions. A high priority patient is more likely to receive sessions than a patient of lower priority *ceteris paribus*. This emphasis is in practice relative to the numbers of patients in each priority level.

Patient availability can be deduced from the Patient Timetable Overview. After the second largest department - occupational therapy - has scheduled their treatment sessions the PTO is 'passed' on to the physiotherapy department.

The **therapist attributes** relate to many definable aspects of the staff. Therapist band level is based on a nine point scale ranging from least qualified (assistant) to most qualified (superintendant). Technically, an employee with a band level of five or above is classed as a 'physiotherapist'; otherwise they are an 'assistant'. An important consideration is also the ability of the therapists to perform different types of treatment session.

Therapist availability is compiled over time within the staff diary. This contains pre-planned times at which therapists are unavailable for clinical or administrative duties within the treatment week. Such unavailability could be due to pre-arranged (fixed) sessions, goal planning meetings, other meetings, training, or annual leave. The staff diary exists in paper form and is not stored electronically.

SAT	SUN	MON	TUE	WED	THU	FRI
				PTO received	Physiotherapy timetable produced	PTO printed
				Ward round		
				Timetable preparation		

Figure 3.2 Physiotherapy scheduling process for preparation week

The chronological process of how physiotherapy treatment is scheduled is described in Figure 3.2. The PTO is typically received before 1500 hours on a Wednesday afternoon. The ward round is conducted at 1500 and is usually complete by 1615. Following this the *timetable preparation* is undertaken. This takes one employee about an hour and is officially described on the first page of Appendix 3.1. It is summarised forthwith.

First and foremost the type of week is determined. This falls into two categories: even and odd. This is since the shifts of employees are two-week periodic[7]. The basic details of each day of each week for each therapist are found in the unedited daily staff timetables (DSTs). These are stored as Excel Worksheets and feature the core structure of each day for each therapist (an example for an even-week Wednesday can be found in Appendix 3.2). They contain the start and end time of the day complete with various partitions throughout the day. These partitions are generally fixed and describe the durations of slots[8] that are permissible. In addition, any events, engagements or activities that occur regularly from week-to-week are included. These would include the ward round on the Wednesday but not the timetabling on the Thursday since there is no fixed employee for this task.

Secondly, the unedited DSTs are printed for the appropriate type of week (whether it be even or odd). Information on any events, engagements or activities that occur irregularly (i.e. are specific to that week) is then added from the staff diary to these printouts. This would include goal planning meetings but not regular commitments at other hospitals. The list of 'relevant information' on page one of Appendix 3.1 specifies all the events, engagements and activities that should have been entered at this stage.

[7] The employees are partitioned into a Red and Blue team. On a particular week one team is assigned to an *early* shift whilst the other is assigned to a *late* shift

[8] Note that sessions relate to the patient assignment whilst slots relate to the employee assignment

On the Thursday morning the edited paper copies of the DSTs are passed on to the employee tasked with scheduling for that week. This procedure typically takes one employee the whole day to complete and is regarded as a frustrating and arduous task. The official procedure governing the scheduling of physiotherapy can be found on the second and third page of Appendix 3.1. It is summarised in the following subchapter.

The conclusion of the process is marked by the printing and distribution of the PTO to patients at about 1500 on Friday.

3.4.4 Physiotherapy scheduling procedure at Rookwood NRC

Since the construction of the physiotherapy schedule consists of a number of steps (described in Ch 3.4.3) it may be described as a process of individual procedures. Perhaps the most integral procedure within this chain is the therapy scheduling itself. The official guidelines governing this allocation of treatment to patients is described on pages two and three of Appendix 3.1. It is briefly summarised forthwith.

Firstly, any multi-department sessions[9] that have been entered onto the PTO by another department are pencilled onto the DSTs[10]. Patients are then ascribed priority levels based on attributes regarding health, response to therapy, and timing of admission/discharge. The minimum score is zero whilst the maximum is five.

Following this, therapy allocations are made in order of session type. Hydrotherapy sessions are scheduled first, followed by group, stretches, and then singles, doubles, and triples. Each session type is scheduled in order of decreasing priority. That is, the highest priority patients are scheduled first followed by patients from a lower priority level. There are no guidelines governing the exact allocation procedure of sessions to patients. It is assumed that these allocations are made by the intuition and experience of the employee tasked with scheduling.

In terms of making an individual allocation the following practice has been observed. Firstly, the employee tries to assign the treatment to the most appropriate day. This is roughly determined by the current spread of sessions for that patient. The employee will try and ensure that treatment sessions are spaced as evenly as possible throughout the week; so will opt for the day with the least assignments. The PTO will then be studied to identify the times at which treatment is not possible due to other patient engagements. Then, the DSTs are used

[9] Treatment sessions that involve more than one department
[10] At this stage these contain all regular and irregular events, engagements and activities that are listed on page one of Appendix 3.1

to assess the availability of the therapists (starting with the most preferable). If there is no correspondence between availabilities of the therapist(s) and the patient then another therapist or day is selected.

There are many variables that affect the allocation of physiotherapy sessions. Generally, these are not formally detailed and it is employee intuition that accounts for their inclusion. For example, some patients respond better to treatment in the morning. This information is not recorded because such patients are known by the employee tasked with the timetabling. Appropriate allowances can then be made in the scheduling procedure of the sessions for such patients. This is important since *'not taking patients' preferences into account tends to affect compliance and punctuality'* (Bourque, 1980).

This is a problem for an automated scheduling program since the effects of such variables cannot be incorporated unless there is a clear specification of their values. The following subchapter identifies these variables and their effects on the quality of the timetable.

3.5 Problem Formulation

The basic components of a typical optimisation problem are as follows:

- **Decision Variables** – these affect the value of the objective function and constraints
- **Objective Function** – whose value is to be minimised
- **Constraints** – violations of which are to be avoided

There are many different types of constraint within an optimisation problem. These often exist in a hierarchy of importance in which the satisfaction of some is valued more highly than the satisfaction of others. The most important constraints are known as hard constraints whose satisfaction is essential for the solution to be valid or feasible. Less important constraints are referred to as soft constraints. Here, it is the addition of these that form the objective function.

The determination of these constraints can be partly attributed to the scheduling guidelines (see Appendix 3.1); however, these lack a sufficient description of the assignment procedure. An invaluable practice has therefore been to understand the various intuitive motivations of therapists when scheduling treatment by hand. This understanding matured over many meetings with therapists who were involved in the timetabling procedure. The rationale for

various treatment allocations was deduced by questioning their thought processes when scheduling and by posing various what-if questions to them.

The constraints are compiled using standard notation from set theory and Boolean algebra. A complete specification of the syntax used and its meaning can be found in Appendix 3.3. Note that variables operated on by a tilde are dummy variables whose only use is in aiding the explanation of constraints within this subchapter.

3.5.1 Hard constraints

These exist as discrete restrictions within the scheduling problem that are *'rigidly enforced'* (Burke & Petrovic, 2002). A specification of the hard constraints is provided in Appendix 3.4.

3.5.2 Soft constraints

The satisfaction of the soft constraints is *'desirable but not absolutely essential'* (Burke & Petrovic, 2002). They are typically referred to as the *objectives* of the scheduling problem. They differ to hard constraints insofar as they do not represent discrete violations. Instead, they take on a continuous value that reflects the extent to which that objective has been satisfied. The summation of these (usually weighted) values defines the *objective function*, the aim of which is to minimise. Note that care should be taken in the interpretation of objective functions of this type since the soft constraints can differ in units of measurement.

$$Obj = W_1 \cdot SCON1 + W_2 \cdot SCON2 + ... + W_n \cdot SCONn \qquad (3.5.2.1)$$

The value of this function can therefore be used to measure the quality of the solution. However, its ability to do so is dependent on the adequacy of the soft constraints to capture the effects of decision variables that affect the calibre of the schedule. Appendix 3.1 formally describes two such variables – session type and priority. These are used to govern the allocation precedence of treatment sessions. Since the assignment of stretches follow group sessions it can be surmised that unmet needs of group sessions are less desirable than unmet needs of stretches. Therefore a greater weight needs to be placed on group session unmet needs. In addition, priority levels are used to control assignments within each session type allocation. It is stated in Appendix 3.1 that *'the highest scoring patients get timetabled first'* and that *'unmet needs should only affect the low priority patients'*. It is the specification of these two variables and their effects on solution quality that enable the construction of the first soft constraint.

SCON 1: Minimise unmet needs

$$SCON1 = \sum_{\forall i} \dot{w}_{cap_i} \cdot \begin{pmatrix} w^1_{a1} \cdot \left(ds^1_i\right)^2 + w^1_{a2} \cdot \left(dd^1_i\right)^2 + w^1_{a3} \cdot \left(dt^1_i\right)^2 + w^1_{a4} \cdot \left(dg^1_i\right)^2 + w^1_{a5} \cdot \left(dh^1_i\right)^2 + \ldots \\ + w^1_{a6} \cdot \left(dst^1_i\right)^2 + w^1_{a7} \cdot \left(dpm^1_i\right)^2 + w^1_{a8} \cdot \left(dgm^1_i\right)^2 + w^1_{a9} \cdot \left(djs^1_i\right)^2 + w^1_{a10} \cdot \left(dop^1_i\right)^2 \end{pmatrix} \quad (3.5.2.2)$$

Unmet needs (*type I*) are defined as treatment sessions that have been demanded but not scheduled[11]. The number of unmet needs is equivalent to the remaining demand for a particular type of treatment session (i.e. initial demand minus number scheduled). This is represented by $d*^1_i$ in the above where $*$ denotes the type of treatment session. These values are squared since it is undesirable for unmet needs to contribute similar costs should the remaining demand be different. For example, two patients of similar attribute both demand three stretches but are currently only scheduled one and two respectively. It is possible for one of the patients to receive another session. It is obvious that this patient should be the one with the greatest number of unmet needs – and this would be the choice of a therapist producing the schedule by hand. However, if the value of *SCON1* depended linearly on remaining demand then no distinction would be made between the patients. The addition of the power function discourages this since the objective value has a quadratic dependence on the number of unmet needs.

SCON 2: Optimise the spread of sessions throughout the week

$$SCON2 = \sum_{\forall i} \dot{w}_{cap_i} \cdot \sum_{k=1}^{3} w^2_k \cdot \sum_{x=1}^{n^k_i} \left| ma_{i,k,x} - mo_{i,k,x} \right|^{w^2_\alpha} \quad (3.5.2.3)$$

An uneven spread of treatment sessions can reduce the efficacy of any therapy received. This is because patients benefit from a balanced and proportionate treatment schedule both mentally and physically. The mental advantages are borne from having a consistent regime of treatment. Physically, the functional ability of the patient may deteriorate if they do not receive a sufficient amount of therapy over time. For example, a patient who receives all of their treatment on a Monday could be mentally overwhelmed by the amount of therapy but could also deteriorate due to not being seen on the other days.

An undesirable spread of treatment sessions can be deterred by reducing the time between the scheduled (actual) times, ma, and corresponding optimal times, mo. This is done by adding

[11] *Type II* are defined as treatment sessions that have been scheduled but not received

an amount to the objective function that is dependent on the time difference, $ma - mo$. This dependence must be nonlinear so as to add proportionally more cost when this time difference is significantly large. The associated power factor is given by w_α^2.

The patient treatment session types are partitioned into the following three groups:

1. Group sessions
2. Stretches (if for maintenance[12])
3. Singles, doubles, triples, hydro, joint sessions, other patient interactive sessions, patient meetings, goal planning meetings

This is because it is necessary for group sessions and stretches to be spread independently of any other session types. For example, assume there are three group sessions and two stretches. The optimal positions (for a five day week) are

M	**T**	**W**	**T**	**F**
G	St	G	St	G

and not, for example,

M	**T**	**W**	**T**	**F**
G	G	G	St	St

The remainder of the sessions are contained in the third group in order to improve the likelihood that employee hours are spread evenly among the patients for each day.

In the mathematical specification of *SCON2*, the session type group is given by $k \in \{1,2,3\}$ whilst the number of sessions of that group for that patient is given by n_i^k. The weights $w_k^2 > 0$ are used to vary the emphasis placed on sessions from different groups obtaining their optimal spreads. Since the guidelines of Appendix 3.1 state that group sessions are scheduled before stretches, this could provide a rationale for $w_1^2 > w_2^2$.

The scheduled and actual times are defined on the scale depicted below. Since the second week (7-14) describes the treatment week that is being scheduled for then $7 \le ma_{i,k,x} \le 12$. As the integer values on this scale are used to denote 0000 hours the $ma_{i,k,x}$ are generalised to the mid-points of these days, i.e. $7.5 \le ma_{i,k,x} \le 11.5; ma_{i,k,x} - 0.5 \in \mathbb{Z}$.

[12] Since stretches for treatment preparation must be scheduled with consideration to the assignment of singles, doubles, triples, joint sessions (see Ch 3.4.2)

MON	TUE	WED	THU	FRI	SAT	SUN	MON	TUE	WED	THU	FRI	SAT	SUN	
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

The optimal times follow similar restrictions but in this case integer values are also permissible, i.e. $7.5 \le mo_{i,k,x} \le 11.5;\ 2 \cdot mo_{i,k,x} \in \mathbb{Z}$. This is because it is possible for an optimal time to straddle two days; thus showing no preference to either. Although it is only possible to schedule sessions in the forthcoming week (7-12), awareness must be shown to the sessions received in the previous week (0-5). This is so that the optimal times can reflect the aim to evenly space treatment; not just from the beginning of the forthcoming week, but from the time and day that treatment was last received ($T_{i,k}$) for that session group.

Therefore, the raw values of the optimal times are calculated by the formula

$$mo_{i,k,x} = \frac{14 - T_{i,k}}{n_i^k + 1} \cdot x + T_{i,k} \qquad x = 1,...,n_i^k \qquad (3.5.2.4)$$

However, this might lead to an instance in which $mo_{i,k,1} < 7$ or $mo_{i,k,n_i^k} > 12$. If treatment can be scheduled on weekends then this is no problem. However, if, like at present, treatment only occurs on weekdays then such an instance is obviously impermissible and so some form of action is required to enforce validity. Note that since $T_{i,k} \le 5$ then $mo_{i,k,n_i^k} \le 12$ if $mo_{i,k,1} \ge 7$. Therefore, if $mo_{i,k,1} < 7$ then the optimal times must be compressed from the left to ensure that $mo_{i,k,1} \ge 7$ (and so $mo_{i,k,n_i^k} \le 12$) using

$$mo_{i,k,x} = \frac{7}{n_i^k} \cdot (x - 1) + 7 \qquad x = 1,...,n_i^k \qquad (3.5.2.5)$$

If this results in $mo_{i,k,n_i^k} > 12$ then the optimal times must be compressed from both the left and right;

$$mo_{i,k,x} = \frac{5}{n_i^k} \cdot (x - 1) + 7 \qquad x = 1,...,n_i^k \qquad (3.5.2.6)$$

SCON 3: Maximise number of sessions at preferred times

$$SCON3 = \sum_{\forall i} \dot{w}_{cap_i} \cdot \sum_{\forall d} \sum_{\forall s} \tilde{\delta}^{3}_{i,d,s} \tag{3.5.2.7}$$

For some patients the efficacy of treatment is dependent on the time of day. Preferential times are generalised to morning ($apr_i = 1$) and afternoon ($apr_i = 2$).

The objective is to improve the likelihood that physiotherapy treatment sessions are assigned to these preferred times. This aim is formulated by adding a value to the objective function for instances when this is not the case. This value, as before, is controlled for by the priority level of the patient.

In the case that a patient has no preference, the dummy variable, $\tilde{\delta}^{3}_{i,d,s}$, attains a value of zero. This is also the case should the treatment occur at a preferred time. A value of one is only returned if the treatment occurs at a time that is not of preference to the patient. The criterion for determining whether a session is in the morning or afternoon is described below.

$$\tilde{\delta}^{3}_{i,d,s} = \begin{cases} 0 \quad if \quad \begin{cases} apr_i = 0 \\ apr_i = 1 \quad \wedge \quad pps_{i,d,s} < le_d \\ apr_i = 2 \quad \wedge \quad ppe_{i,d,s} > ls_d \end{cases} \\ 1 \quad if \quad \begin{cases} apr_i = 1 \quad \wedge \quad pps_{i,d,s} \geq le_d \\ apr_i = 2 \quad \wedge \quad ppe_{i,d,s} \leq ls_d \end{cases} \end{cases} \tag{3.5.2.8}$$

SCON 4: Maximise number of sessions with primary or secondary therapists

$$SCON4 = \sum_{\forall i} \dot{w}_{cap_i} \cdot \sum_{\forall d} \sum_{\forall s} \tilde{\delta}^{4}_{i,d,s} \cdot \tilde{v}^{4}_{i,d,s} \tag{3.5.2.9}$$

Each patient has a primary (app_i) and a secondary (asp_i) therapist. It is preferential (sometimes essential) for these therapists to be involved in the provision of certain physiotherapy sessions[13].

For such sessions a cost (w_1^4) is therefore added to the objective function when the primary or secondary therapist is not present. A lesser cost (w_2^4) is added for sessions in which the primary is absent but the secondary is present.

[13] Singles, doubles, triples, joint sessions, patient meetings, goal planning meetings

$$\tilde{\delta}^4_{i,d,s} = \begin{cases} 0 & if \quad ppst_{i,d,s} = ccih \vee ccig \vee ccist \vee cciop \\ 1 & if \quad ppst_{i,d,s} = ccis \vee ccid \vee ccit \vee ccijs \vee ccipm \vee ccigm \end{cases} \tag{3.5.2.10}$$

$$\tilde{v}^4_{i,d,s} = \begin{cases} w^4_1 & if \quad app_i \notin ppea_{i,d,s} \quad \wedge \quad asp_i \in ppea_{i,d,s} \\ w^4_2 & if \quad app_i \notin ppea_{i,d,s} \quad \wedge \quad asp_i \notin ppea_{i,d,s} \end{cases} \qquad w^4_2 \gg w^4_1 \tag{3.5.2.11}$$

SCON 5: Minimise prevalence of multi-disciplinary cumulative sessions

$$SCON5 = \sum_{\forall i} amdc_i \cdot \dot{w}_{cap_i} \cdot \sum_{\forall d} \sum_{s=1}^{nps_{i,d}-1} \exp\left(\frac{-\left| ps_{i,d,s+1} - pe_{i,d,s} \right|}{w^5_\alpha} \right) \tag{3.5.2.12}$$

Some patients do not respond well to back-to-back treatment sessions. This is typically because of agitation or a lack of endurance. It can result in patient refusal to participate or an unsatisfactory outcome of the therapy session.

For the affected patients ($amdc_i = 1$) a value is added to the objective function that is dependent on the time between consecutive sessions within a day (from all departments). It is intuitive that this value should be very large when the time between sessions is very small but should reduce as time increases. After a certain amount of time the value should be negligible. To this end, an exponential distribution was proposed to therapists. This was agreed upon with a parameter value $w^5_\alpha = 12$.

SCON 6: Optimise placement of patient meetings and goal planning meetings

$$SCON6 = \sum_{\forall i} \dot{w}_{cap_i} \cdot \tilde{\delta}^6_i \tag{3.5.2.13}$$

If a goal planning meeting is scheduled then it is a hard constraint (*HCON16*) that any patient meeting occurs before this session. This is because, in this case, the objective of the patient meeting is to discuss the upcoming goal planning meeting with the patient. However, it is undesirable for both of these to occur on the same day.

The dummy variable, $\tilde{\delta}^6_i$, is therefore used to determine whether this is the case. If so, a cost is added to the objective function (dependent on priority level).

$$\tilde{\delta}^6_i = \begin{cases} 1 & if \quad \exists! d : \exists s_p, s_g : \quad ppst_{i,d,s_p} = ccipm \quad \wedge \quad ppst_{i,d,s_g} = ccigm \\ 0 & otherwise \end{cases} \tag{3.5.2.14}$$

3.5.3 Moderate constraints

In the previous subchapters a number of hard and soft constraints have been introduced. Hard constraints represent discrete violations whose satisfaction is mandatory whilst, in this particular application, soft constraints represent a continuous value whose satisfaction is desirable but not essential. A new type of constraint is forthwith introduced whose definition and comprehension is in difference to those previously discussed.

Moderate constraints can be thought of either as hard constraints whose satisfaction is nonessential or soft constraints with a discrete number of violations. The satisfaction of these can be regarded as being more important than soft constraints but less important than hard constraints. Since their satisfaction is nonessential the validity of the solution does not depend on the number of violations. It is thought that by representing these constraints as discrete violations a more concise picture of their incidence can be formed. This will allow for a more targeted approach in reducing the number of violations (Ch 4.5.2). The priority level of the patient is not a consideration within the specification of these constraints.

Note that the third and fourth moderate constraints are only applicable for patients who receive stretches for treatment preparation. This is since the assignment of treatment is more complex for these patients and so requires additional constraints. This complexity is due to the absence of independence between the assignment of stretches and post-stretch sessions[14] (see Ch 3.4.2). *MCON3* relates to the volumes of stretches and post-stretch sessions that are assigned to the schedule whilst *MCON4* considers the optimal positions of these assignments within the week.

MCON 1: Demand and supply

$$MCON1 = \sum_{\forall i} \sum_{k=1}^{3} \tilde{\delta}_{i,k} \qquad (3.5.3.1)$$

This constraint relates to the number of sessions that are scheduled and demanded from specific session groups. The aim is to ensure that at least one session is scheduled from each session group if demand exists. The groups of sessions that this is applicable for is as follows:

1. Groups
2. Stretches

[14] Singles, doubles, triples, joint sessions

3. Singles, doubles, triples

Therefore, a violation is incurred for each instance in which no assignments are made when demand is nonzero. This is formulated as follows:

$$\tilde{\delta}_{i,1} = \begin{cases} 1 & if \quad dg_i > 0 \quad \wedge \quad dg_i^2 = 0 \\ 0 & otherwise \end{cases}$$
$$\tilde{\delta}_{i,2} = \begin{cases} 1 & if \quad dst_i > 0 \quad \wedge \quad dst_i^2 = 0 \\ 0 & otherwise \end{cases} \qquad (3.5.3.2)$$
$$\tilde{\delta}_{i,3} = \begin{cases} 1 & if \quad ds_i + dd_i + dt_i > 0 \quad \wedge \quad ds_i^2 + dd_i^2 + dt_i^2 = 0 \\ 0 & otherwise \end{cases}$$

MCON 2: Session spread

$$MCON2 = \sum_{\forall i} \sum_{k=1}^{3} \max\left(\sum_{\forall d} \tilde{\delta}_{i,k,d}^{m2\min}, \sum_{\forall d} \tilde{\delta}_{i,k,d}^{m2\max} \right) \qquad (3.5.3.3)$$

The importance of having an even spread of sessions has been introduced in the specification of *SCON2* (see Ch 3.5.2). The quality of the spread in this constraint is, however, determined by a different means.

Constraint violations occur when there is a difference of more than one between the numbers of sessions scheduled on any two days. This is with respect to the following session groups:

1. Groups
2. Stretches
3. Singles, doubles, triples, joint sessions, hydro

The first two session groups are equivalent to those of *SCON2*. However, the third contains fewer patient sessions. This is because it is easier and more important to obtain an even spread with respect to these. It is easier because there are fewer fixed sessions such as goal planning/patient meetings and other patient interactive slots. It is more important because these sessions relate to patient treatment *c.f. SCON2*[15]. Since these are intensive sessions a patient may become exhausted if too many are carried out over a period of time. Conversely, if not enough are carried out then the condition of the patient can deteriorate. For these

[15] *SCON2* considers an even spread with respect to employee time and not patient treatment time

reasons the satisfaction of this constraint takes precedence over *SCON2* – hence achieving status as a moderate constraint.

The number of violations is determined as follows. Firstly, the minimal ($\tilde{v}_{i,k}^{m2\min}$) and maximal ($\tilde{v}_{i,k}^{m2\max}$) permissible values for the daily number of sessions are deduced. These are defined as the floor and ceiling functions of the average number of sessions for each day of the week[16].

$$\tilde{v}_{i,k}^{m2\min} = \left\lfloor \frac{\sum_{\forall d} \tilde{n}_{i,d}^{k}}{\bar{d}} \right\rfloor \quad , \quad \tilde{v}_{i,k}^{m2\max} = \left\lceil \frac{\sum_{\forall d} \tilde{n}_{i,d}^{k}}{\bar{d}} \right\rceil \qquad \bar{d} = \begin{cases} 5 & \text{if weekends excluded} \\ 7 & \text{if weekends included} \end{cases} \tag{3.5.3.4}$$

Secondly, a score of either 0 or 1 is given to each day of the week depending on whether the number of sessions scheduled is contained within the acceptable range.

$$\begin{aligned} \tilde{\delta}_{i,k,d}^{m2\min} &= \begin{cases} 0 & if \quad \tilde{n}_{i,d}^{k} \geq \tilde{v}_{i,k}^{m2\min} \\ 1 & if \quad \tilde{n}_{i,d}^{k} < \tilde{v}_{i,k}^{m2\min} \end{cases} \\ \tilde{\delta}_{i,k,d}^{m2\max} &= \begin{cases} 0 & if \quad \tilde{n}_{i,d}^{k} \leq \tilde{v}_{i,k}^{m2\max} \\ 1 & if \quad \tilde{n}_{i,d}^{k} > \tilde{v}_{i,k}^{m2\max} \end{cases} \end{aligned} \tag{3.5.3.5}$$

The number of violations is then calculated as the supremum of the summations of these values over each day of the week.

MCON 3: Treatment preparation: Session volumes

$$MCON3 = \sum_{\forall i} asm_i \cdot \tilde{\delta}_i^{m3} \tag{3.5.3.6}$$

It is strongly desirable that the volumes of stretches and singles, doubles, triples and joint sessions scheduled are appropriately related to the volumes demanded. This is because the intended purpose of these sessions can be determined by the number of which are demanded. For an effective consideration of this issue the satisfaction of *MCON4* is assumed. That is, those stretches and post-stretch sessions that are scheduled are optimally placed within the week. This means that there should be as few days as possible which contain a stretch without a post-stretch session or vice versa. It also means that the stretch is in advance of any post-stretch session. Three cases are considered.

[16] Where $\tilde{n}_{i,d}^{k}$ is used to represent the number of scheduled sessions for session group *k* for patient *i* on day *d*

First considered is the case in which the demand for stretches is greater than the demand for singles, doubles, triples, and joint sessions. If the demand for post-stretch sessions is less than the number of treatment days in a week then it can be assumed that some of these stretches are intended for maintenance. However, at this stage these maintenance sessions are expendable. The emphasis, in this first case, is on ensuring that each post-stretch session is preceded by a stretch. For this to be achieved the number of stretches assigned must be greater than or equal to the number of post-stretch sessions assigned (assuming an even spread of sessions, i.e. *MCON2* satisfied). It is automatically achieved if the number of stretches scheduled is greater than or equal to the number of treatment days in a week (currently five).

$$if \quad dst_i > ds_i + dd_i + dt_i + djs_i \quad then \quad \tilde{\delta}_i^{m3} = \begin{cases} 1 & dst_i^2 < ds_i^2 + dd_i^2 + dt_i^2 + djs_i^2 \quad \wedge \quad dst_i^2 < 5 \\ 0 & otherwise \end{cases}$$

The second case describes the situation in which the demand for stretches and post-stretch sessions are equal. In this case there can be no possible intention to provide maintenance sessions. Following a similar rationale as above an equivalent determination of the dummy variable is arrived at;

$$if \quad dst_i = ds_i + dd_i + dt_i + djs_i \quad then \quad \tilde{\delta}_i^{m3} = \begin{cases} 1 & dst_i^2 < ds_i^2 + dd_i^2 + dt_i^2 + djs_i^2 \quad \wedge \quad dst_i^2 < 5 \\ 0 & otherwise \end{cases}$$

The third case considers the situation in which the number of stretches demanded is fewer than the number of post-stretch sessions demanded. Whilst it is still beneficial for a post-stretch session to follow a stretch it is, in this case, by no means mandatory. This implies that either the patient is not too dependent on stretches or that it is acceptable to provide a basic stretch as part of their post-stretch session. It also implies that it is wasteful for such a patient to be assigned a stretch on a day in which there are no post-stretch sessions. If the demand for the stretches is less than the number of treatment days in a week then it can be assumed that some of these post-stretch sessions are intended for assignment in the absence of a stretch. However, at this stage the assignment of these stand-alone post-stretch sessions is unimportant. The issue here is ensuring that there are no days in which there are stretches but no post-stretch sessions. For this to be achieved the number of post-stretch sessions assigned must be greater than or equal to the number of stretches assigned (assuming an even spread of sessions, i.e. *MCON2* satisfied). It is automatically achieved if the number of post-stretch

sessions scheduled is greater than or equal to the number of treatment days in a week (currently five).

$$if \quad dst_i < ds_i + dd_i + dt_i + djs_i \quad then \quad \tilde{\delta}_i^{m3} = \begin{cases} 1 & dst_i^2 > ds_i^2 + dd_i^2 + dt_i^2 + djs_i^2 \\ & \wedge \quad ds_i^2 + dd_i^2 + dt_i^2 + djs_i^2 < 5 \\ 0 & otherwise \end{cases}$$

MCON 4: Treatment preparation: Optimal placements

For stretches and post-stretch sessions to be optimally placed (regardless of the scheduled volumes) there should be as few days as possible in which a stretch occurs without a post-stretch session or vice versa (*MCON4.1*). In addition, the stretch should be in advance of any post-stretch sessions (*MCON4.2*).

MCON 4.1: Same day principle

$$MCON\,4.1 = \sum_{\forall i} \frac{\left\| \tilde{n}_i^{dst} - \tilde{n}_i^{dsdtjs} \right| - \tilde{n}_i^{dso} \right|}{2} \qquad (3.5.3.7)$$

The number of violations should reflect the number of sub-optimal days[17] that are scheduled in relation to the minimal number of sub-optimal days that are possible with the scheduled volumes. For example, assume there are three stretches and two post-stretch sessions scheduled. Although in the first example there is a sub-optimal day (Wednesday) this is acceptable since there is not another day that contains just a post-stretch session without a stretch. This is in contrast with the second example which would sustain a violation.

M	**T**	**W**	**T**	**F**
St		St	St	
P-st			P-st	

M	**T**	**W**	**T**	**F**
St		St	St	
	P-st		P-st	

The constraint is structured this way to eliminate any double-counting affiliated with an undesirable volume allocation of stretches and post-stretch sessions (*MCON3*).

For each patient the number of days which contain a stretch ($\tilde{n}_i^{dst}$) is calculated. Then the number of days which contain either a single, double, triple, or joint session ($\tilde{n}_i^{dsdtjs}$) is determined in addition to the number of sub-optimal days ($\tilde{n}_i^{dso}$).

[17] Days that contain a stretch with no post-stretch session or vice versa

$$\tilde{n}_i^{dst} = \sum_{\forall d} \tilde{\delta}_{i,d}^{m41st} \quad , \quad \tilde{n}_i^{dsdtjs} = \sum_{\forall d} \tilde{\delta}_{i,d}^{m41sdtjs} \quad , \quad \tilde{n}_i^{dso} = \sum_{\forall d} \tilde{\delta}_{i,d}^{m41so} \tag{3.5.3.8}$$

The dummy variables are defined as:

$$\begin{aligned}
\tilde{\delta}_{i,d}^{m41st} &= \begin{cases} 1 & if \quad \exists s : ppst_{i,d,s} = ccist \\ 0 & otherwise \end{cases} \\
\tilde{\delta}_{i,d}^{m41sdtjs} &= \begin{cases} 1 & if \quad \exists s : ppst_{i,d,s} = ccis \vee ccid \vee ccit \vee ccijs \\ 0 & otherwise \end{cases} \\
\tilde{\delta}_{i,d}^{m41so} &= \begin{cases} 1 & if \quad \begin{cases} \left(\exists s : ppst_{i,d,s} = ccist \quad \wedge \quad \nexists s : ppst_{i,d,s} = ccis \vee ccid \vee ccit \right) \\ \vee \\ \left(\nexists s : ppst_{i,d,s} = ccist \quad \wedge \quad \exists s : ppst_{i,d,s} = ccis \vee ccid \vee ccit \right) \end{cases} \\ 0 & otherwise \end{cases}
\end{aligned} \tag{3.5.3.9}$$

MCON 4.2: Order of precedence

$$MCON\,4.2 = \sum_{\forall i} \sum_{\forall d} \tilde{\delta}_{i,d}^{m42} \cdot \left(\tilde{n}_{i,d}^{m42a} + \tilde{n}_{i,d}^{m42b} \right) \tag{3.5.3.10}$$

The number of violations of this constraint is comprised of two parts: the number of post-stretch sessions with no preceding stretch ($\tilde{n}_{i,d}^{m42a}$) and the number of stretches without a proceeding post-stretch session ($\tilde{n}_{i,d}^{m42b}$).

The parts are formulated as follows:

$$\begin{aligned}
\tilde{n}_{i,d}^{m42a} &= \sum_{\forall s} \tilde{v}_{i,d,s}^{m42a} \quad , \quad \tilde{n}_{i,d}^{m42b} = \sum_{\forall s} \tilde{v}_{i,d,s}^{m42b} \\
\tilde{v}_{i,d,s}^{m42a} &= \begin{cases} 1 & if \quad ppst_{i,d,s} = ccis \vee ccid \vee ccit \vee ccijs \quad \wedge \quad \nexists s_a \in [0,s) : ppst_{i,d,s_a} = ccist \\ 0 & otherwise \end{cases} \\
\tilde{v}_{i,d,s}^{m42b} &= \begin{cases} 1 & if \quad ppst_{i,d,s} = ccist \quad \wedge \quad \nexists s_a \in (s, npps_{i,d,s}] : ppst_{i,d,s_a} = ccis \vee ccid \vee ccit \vee ccijs \\ 0 & otherwise \end{cases}
\end{aligned} \tag{3.5.3.11}$$

A violation is only permissible if there are nonzero allocations of both stretches and post-stretch sessions on that day, i.e.

$$\tilde{\delta}_{i,d}^{m42} = \begin{cases} 1 & \tilde{n}_{i,d}^{m42a} > 0 \quad \wedge \quad \tilde{n}_{i,d}^{m42b} > 0 \\ 0 & otherwise \end{cases} \tag{3.5.3.12}$$

3.6 Conclusion

An overview of the treatment scheduling process at Rookwood NRC has been presented within this chapter. Attention has then been restricted to the process of scheduling physiotherapy treatment. The assignment procedure for physiotherapy treatment sessions is then described in the fourth subchapter. Finally a mathematical formulation of this scheduling procedure is produced. Some limitations are now addressed.

First is the issue of obtaining a globally optimal schedule (with respect to the assignment of treatment sessions from all of the therapeutic departments). Since the proposed process does not involve the simultaneous assignment of all treatment sessions a globally optimal schedule is not possible.

Second is the problem that whilst one department has an automated scheduling program, the others do not. There are no negative implications of this on producing relationships between treatment intensity and length of stay (see Ch 3.3.2) for these departments – provided that data on received treatment is collected. Neither is there a problem in evaluating the effects of changes to treatment intensity on LOS. However, since treatment intensity is linked to the direct controls (e.g. patient demand, priority – see Ch 3.4.3) through the department schedule it is not possible to evaluate the effect of changes to these on LOS. This is a problem since treatment intensity is not directly controllable by the therapists and so its use as an independent variable is limited.

Third is a complication associated with multi-department sessions. The problem exists when the demand for the session has been specified but not a day or time. Whilst this is not a problem for those sessions that are fixed in advance, such as goal planning meetings, it is a problem for joint sessions[18]. The problem is borne from the informal way in which these are scheduled. By and large, if a department wishes to request a joint session with another department they must wait until it is their turn to timetable. At which point they can only request joint sessions with departments that are yet to schedule since the other departments would have finalised their allocations. This problem is therefore an obvious concern for those departments scheduling towards the end of the cycle.

Whilst it is not feasible to resolve the problems associated with the first two limitations it is possible to resolve those of the third. It is proposed that joint sessions and other multi-dept

[18] The majority of multi-department sessions are joint sessions

sessions (where appropriate[19]) are scheduled first-thing on a Monday morning. This could take place as a brief meeting involving representatives from participating departments. The demand could be discussed at the outset of such a meeting followed by the allocation of therapists to patients. However, this is not without drawback. Since the predictability of future patient demand reduces over time it can be ineffective to schedule treatment too far in advance. In this case, joint sessions would be scheduled 7-12 days in advance of the actual treatment. In this time the requirement for such a session could have surpassed, so too could the ability of the patient to receive it.

If an automated scheduling program is used to replace the current by-hand approach in the physiotherapy department then the process of scheduling treatment may change somewhat. Some aspects, however, will remain the same. The PTO will still be passed from occupational therapy to physiotherapy and there will still be a requirement for a ward round to determine patient demand. In addition, there will still be a need for timetable preparation but this would now be in respect to readying the automated program. Should this occur at the same time and for a similar duration then the maximal run-time for the program is approximately one and a half days from Wednesday (late afternoon) to Friday (morning). This would give the staff an opportunity to check the quality of the timetable before distribution at 1500 hours. It would also allow time for any possible alterations that need to be made.

<table>
<tr><th>SAT</th><th>SUN</th><th>MON</th><th>TUE</th><th>WED</th><th>THU</th><th>FRI</th></tr>
<tr><td rowspan="4"></td><td rowspan="4"></td><td rowspan="4"></td><td rowspan="4"></td><td>PTO received</td><td rowspan="4">...running…</td><td rowspan="3">Physiotherapy timetable produced and checked</td></tr>
<tr><td>Ward round</td></tr>
<tr><td>Timetable preparation</td></tr>
<tr><td>Start physiotherapy scheduling</td><td>PTO printed</td></tr>
</table>

Figure 3.3 Possible physiotherapy scheduling process for preparation week with the employment of an automated scheduling program

[19] Some are fixed in advanced (e.g. home visit involving physiotherapy and occupational therapy)

It has been mentioned previously (Ch 3.3.1) that the two main advantages of an automated program are improved solution quality and a reduction in employee time. The physiotherapy department therefore has much to gain from an automated scheduling program for the following reasons. Firstly, the complications associated with the constraints of Chapter 3.5 make it very hard for an employee to produce a solution of optimal (or even near-optimal) quality. And secondly, the therapist tasked with scheduling takes approximately eight hours each week to produce the timetable – this time could be used to provide approximately as many as eight treatment sessions (Appendix 3.2).

Whilst these advancements are obviously advantageous they do not epitomise our motivation for investigating treatment scheduling. It is the relationship between the direct controls, treatment intensity and LOS that rationalise this interest in timetabling (Ch 3.3.2). The automated scheduling program that has been developed for the physiotherapy department is presented in the next chapter. It is hoped that this program acts as an achievement that other departments can aspire to.

Chapter 4: The Automated Scheduling Program

4.1 Introduction

This chapter is concerned with the development of an automated scheduling program for the assignment of physiotherapy treatment at the Neurological Rehabilitation Centre (NRC) at Rookwood hospital. The mathematical formulation of the problem is laid out in Chapter 3.5 and is used as a foundation for this work. This contains a specification of hard and soft constraints whose respective satisfaction is mandatory and desirable. Using other information such as treatment session details (Ch 3.4.2) and knowledge of the scheduling procedure (Ch 3.4.4) an automated program is developed.

This program must be built from scratch since it represents a very particular scheduling problem. Such an approach is necessary because *'the unique characteristics of different organisations mean that specific mathematical models and algorithms must be developed for personnel scheduling solutions in different areas of application'* (Ogulata et al, 2008). The development of an automated program to represent this scheduling problem is not a straightforward task. Before it is constructed it is necessary to make a number of important observations and decisions. Firstly, a classification of the scheduling problem is considered. To this end, a number of common and relevant scheduling problems are classified in accordance to three criteria. Secondly, a number of heuristics and meta-heuristics are introduced. These are heuristic-based algorithmic frameworks that can be used to approximately solve the scheduling problem.

Following this, some information on the graphical user interface (GUI) is presented. This includes details of the layout and some navigational options. Also included is a brief

walkthrough of how to use the program to construct a timetable. Chapter 4.5 contains an explanation of how the scheduling procedure actually works. The commentary of the GUI is then revisited as a number of post-production options are explained.

4.2 Classification of the Scheduling Problem

There are many different types of scheduling problems that can be found within the literature. For the benefit of this chapter a *scheduling problem* is defined as a well-known and accepted combinatorial optimisation problem that is associated with the assignment of some kind of resource in a particular domain of application. For example, the examination timetabling problem is a scheduling problem in which university examinations are assigned to particular periods of time. The resource is the examination rooms and the domain of application is education.

After reviewing many papers and books it has become apparent that scheduling problems can be classified by three criteria. The first concerns the domain of application. The majority of scheduling problems can be categorised into the following fields: education, manufacturing, personnel, computing and sports. In the domain of education, Schaerf, 1999 splits the scheduling problems to school timetabling[1], course timetabling[2], and examination timetabling. The single processor scheduling problem and multi-processor scheduling problem are two examples that may be found within the computing domain.

The second criterion is the type of scheduling problem. The scheduling problems that are typically studied are derived from a more broad and general type of scheduling problem. Routing, staff assignment and shop scheduling are all rather general scheduling problems that contain many variants. For example, shop scheduling[3] contains the open-shop, flow-shop and job shop scheduling problems (JSSP) whilst routing problems include various vehicle routing problems (such as multi-trip and dynamic VRPs) in addition to the travelling salesman problem (TSP).

The third distinction is the type of scheduling activity that is performed. Up to this point the terms timetabling and scheduling have been used interchangeably with no difference intended in their interpretation. A distinction is made between the activities of scheduling, timetabling, sequencing and rostering in Wren, 1996. The author describes scheduling as *'the allocation,*

[1] The weekly scheduling of school classes
[2] The weekly scheduling of university course lectures
[3] The allocation of *N* jobs to *M* machines. Can involve precedence constraints that govern the order in which jobs must be processed by machines. A machine can only process one job at a time

subject to constraints, of resources to objects being placed in space-time, in such a way as to minimise the total cost of some set of the resources used' e.g. VRP and JSSP. Timetabling is defined similarly although the aim here is to *'satisfy as nearly as possible a set of desirable objectives'*. Examples include the class and examination timetabling problems. Sequencing is *'the construction, subject to constraints, of an order in which activities are to be carried out'* e.g. the flow-shop scheduling problem and the travelling salesman problem. Finally, rostering is defined as *'the placing, subject to constraints, of resources into slots in a pattern'*.

Of these definitions timetabling is the activity that best describes the assignment of physiotherapy treatment at Rookwood NRC. This is because the aim is not to minimise the cost of resources used but to satisfy a set of objectives (Ch 3.5). Rostering is excluded as an option since this activity relates to the assignment of shifts. These definitions enable a scheduling problem to be classified on the basis of the scheduling activity rather than just the type of problem or the application domain. This is beneficial since it could be assumed that just because some scheduling problems belong to the same type of problem they are fundamentally similar. For example, the job-shop and flow-shop scheduling problems both belong to the shop scheduling problems but whilst the job-shop problem relies on scheduling the flow shop problem relies on sequencing.

Since the scheduling problem at Rookwood NRC is very specialised (Ch 3.5) it is difficult to classify it as a particular type of scheduling problem. The review paper *'Staff scheduling and rostering: A review of applications, methods and models'* (Ernst et al, 2004) is used to help with this classification. From the 'Demand Modelling' section (Ch 2.1) it can be surmised that the specification of demand at Rookwood NRC can be defined as task based demand. This is applicable since *'tasks are usually defined in terms of a starting time and duration … and the skills required to perform the task'*. Using the definitions of assignments in Chapter 2.5 it can be inferred that the scheduling problem at Rookwood hospital is very closely related to a task Assignment problem. A formalisation of the problem is presented in Chapter 4.1 of Shen et al, 2003. Depending on the objective of the problem Lo, 1988 states that *'the problem of finding an optimal assignment of tasks to processors is found to be NP-hard'*. The scheduling problem at Rookwood hospital may also be comparable to an open-shop scheduling problem. These are similar to job-shop scheduling problems but have no precedence constraints. It has been shown that this problem is NP-complete (Gonzalez & Sahni, 1976) when the number of machines (staff) is greater than or equal to three.

These inconsistencies are typical in classifying a scheduling problem because the classifications can be non-unique. That is, a particular scheduling problem can belong to more than one of the classes of the above criteria. For example, in the domain of education, it is stated in Schearf, 1999 that *'classification is not strict, in the sense that there are some specific problems that can fall between two classes'* (of problem type). This is echoed in Wren, 1996 in which the author states that *'some problems may fit more than one of the above definitions'* (of scheduling activity). In addition, there are many authors that report equivalence (of some kind) between the groups of a particular classification criterion: '*computer and manufacturing processes'* (Blazewicz et al, 2007); *'shop world and multi-processor computer world'* (Bugnon et al, 1995); *'the TSP and a simple job sequencing problem'* (Wren, 1996).

The point here is that the classification of scheduling problems is by no means an exact science. There are many inconsistencies and discrepancies between the classifications that are made in the literature, and those that are made here. Although it is not essential to provide a classification of a particular scheduling problem it is, however, a useful exercise. This is because when a problem is classified there are many resources that can be exploited to provide information about similar problems that have been studied.

4.3 Exact and Approximate Methods

The scheduling of physiotherapy treatment at Rookwood NRC can be presented as a multi-objective combinatorial optimisation problem. The aim is to determine the solution that yields the lowest cost. The cost of any solution is attainable and is dependent on the extent to which the objectives are satisfied. There exist two types of method that can be used to solve this problem; exact and approximate methods.

Exact methods *'are guaranteed to find the optimal solution'* (Dumitrescu & Stutzle, 2003). An example is an exhaustive search that enumerates every possible solution. However, as the size of the instance increases the running time becomes *'forbiddingly large, even for instances of fairly small size'* (Woeginger, 2003). To increase the efficiency of these methods *'pruning rules'* are developed to discard parts of the search space that cannot contain the optimal solution (Stutzle, 1998). These give rise to familiar methods such as branch-and-bound, dynamic programming and linear and integer programming.

Approximate methods do not guarantee optimality but they are able to find *'good quality solutions in a limited time'* (Puchinger & Raidl, 2005).They may be categorised into two groups; heuristics and metaheuristics. A heuristic algorithm is defined in Reeves, 1993 as a *'technique which seeks good (i.e. near optimal) solutions at a reasonable computational cost without being able to guarantee either feasibility or optimality, or even in many cases to state how close to optimality a particular feasible solution is'*. These methods are typically developed for a particular problem and so *'cannot be used to solve another one'* (Ehrgott & Gandibleux, 2000). A metaheuristic, on the other hand, is defined in (Osman & Laporte, 1996) as *'an iterative generation process which guides a subordinate heuristic by combining intelligently different concepts for exploring and exploiting search space'*. It has been suggested that local search is the *'most successful class'* of approximate methods (Dumitrescu & Stutzle, 2003).

From this point forward the discussion regarding the solution to the scheduling problem at Rookwood NRC is restricted to approximate methods only. This is because it is assumed that the scheduling problem is NP-hard and so cannot be solved in polynomial time (see Ch 7.1 of Burke & Kendall, 2005). In addition, the local search algorithms of approximate methods are *'intuitively understandable, flexible, generally easier to implement than exact algorithms, and in practice have shown to be very valuable when trying to solve large instances'* (Stutzle, 1998).

4.3.1 Heuristics

Approximate methods may be categorised as *constructive* (Ch 4.5.1) or *local search*. Local search methods form a general class of heuristics (Osman & Laporte, 1996) that have become *'widely accepted'* for solving combinatorial optimisation problems (Aarts & Lenstra, 2003). *'The principle of local search is to refine a given initial solution point in the solution space by searching through the neighbourhood of the solution point'* (Wu et al, 1997). In the field of task assignment this often involves either the movement of a task from one employee to another or the exchange of tasks between two employees. The most basic form of local search involves the acceptance of a new solution only if it produces a lower cost function value. This is known as *iterative improvement* (see Chapter 7 of Burke & Kendall, 2005). The termination criterion is achieved when the neighbourhood of the current solution contains only solutions of higher cost function value. The problem is that a final solution will be locally optimal, but not necessarily globally optimal.

4.3.2 Metaheuristics

These *'general search schemes'* (Stutzle, 1998) are described in Ehrgott & Gandibleux, 2000 as '*powerful techniques applicable generally to a large number of problems'*. They are developed to avoid any local optima whilst retaining the simplicity and generality of a local search approach (Osman & Laporte, 1996). A number of metaheuristics have been inspired by natural phenomena such as ant colony optimisation (ACO), evolutionary algorithms (EAs incl. genetic and memetic algorithms), simulated annealing (SA), and neural networks. Others that do not share this link with nature include tabu search (TS), GRASP, iterated local search (ILS), and variable neighbourhood search. These metaheuristics can also be categorised by whether they are population-based (e.g. EAs, ACO) or single-point search (TS, SA, GRASP, ILS) methods. The difference is that in single-point search only one solution is manipulated at each iteration.

4.4 Graphical User Interface I

The automated program for scheduling physiotherapy treatment at Rookwood NRC has been written in Visual Basic 6.0 (for MS Excel). This is for reasons that relate to practicality and convenience. Many of the employees within the physiotherapy department have very limited experience working with computers and MS Excel is used to perform most computer-based tasks that are undertaken. For a more comprehensive guide to the scheduling program the user documentation is available from the author.

4.4.1 The patient timetable overview

The scheduling program requires details of patient availability that can be computationally input from the patient timetable overview, or PTO. However, the PTO (an MS Excel file) was incompatible with the schedule and so it was necessary to produce an amended version.

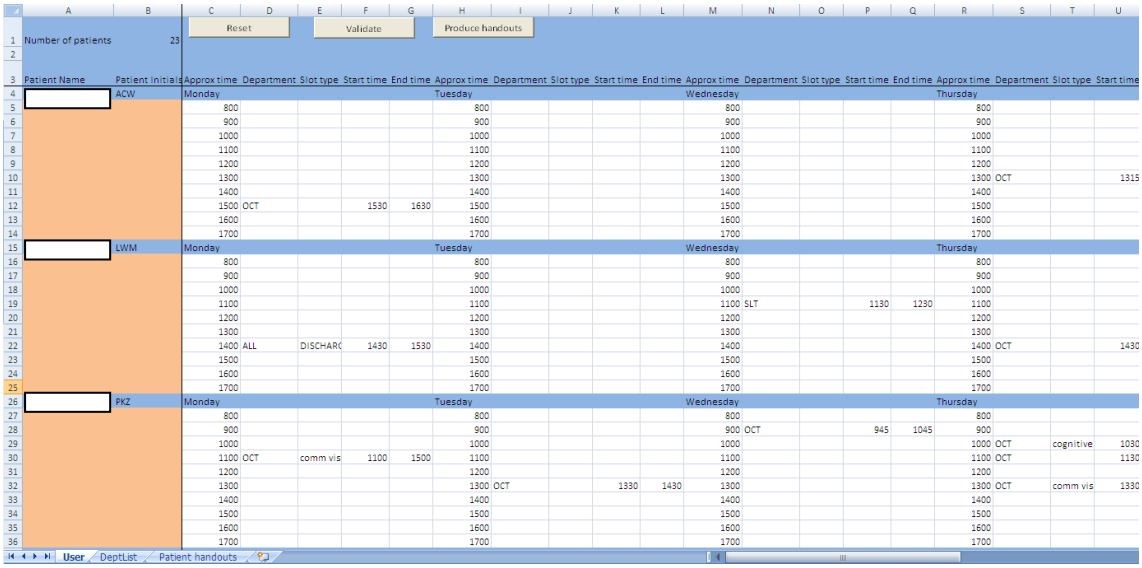


Figure 4.1 The PTO after input from departments other than physiotherapy

At the beginning of the *preparation week*[4] the PTO is reset. This is done by entering the number of patients for the forthcoming week and clicking 'Reset'. This builds an empty framework for each of the patients – removing any treatment details. The patient names and any planned treatment sessions may then be entered in by any of the other departments. For each treatment session it is necessary to input the initials of the department(s) providing the treatment in addition to the start and end time of the session. The 'session type' field is optional.

It has been noticed that accuracy and consistency is a significant problem in the existing PTO. A major problem was the inability to input an end time for a session. Also, there was no separation between the fields of 'department' and 'session type'. The new version redresses these concerns. It also adds a validation feature that can be used to identify any immediate problems with the user input. By clicking 'Validate' the program checks for any violations of the following restrictions:

- A department must be entered if a session contains a start and/or end time
- The department initials of a session must conform with those on the list of department initials
- A start time must be entered if a session contains an end time and/or department

[4] The week before the week that is being scheduled for

- An end time must be entered if a session contains a start time and/or department
- The start time of a session must be before the end time of a session
- Session duration must be of 5 minute multiple
- The start time of a later session must be after the end time of an earlier session

Once the PTO has been completed by all other departments and validated (Figure 4.1) it is ready for use by the scheduling program of the physiotherapy department.

4.4.2 The scheduling program

The scheduling program consists of three stages. In chronological order these are 'Construction', 'Implementation', and 'Analysis'. Only one stage can be active at any time and movement through the stages is one-way.

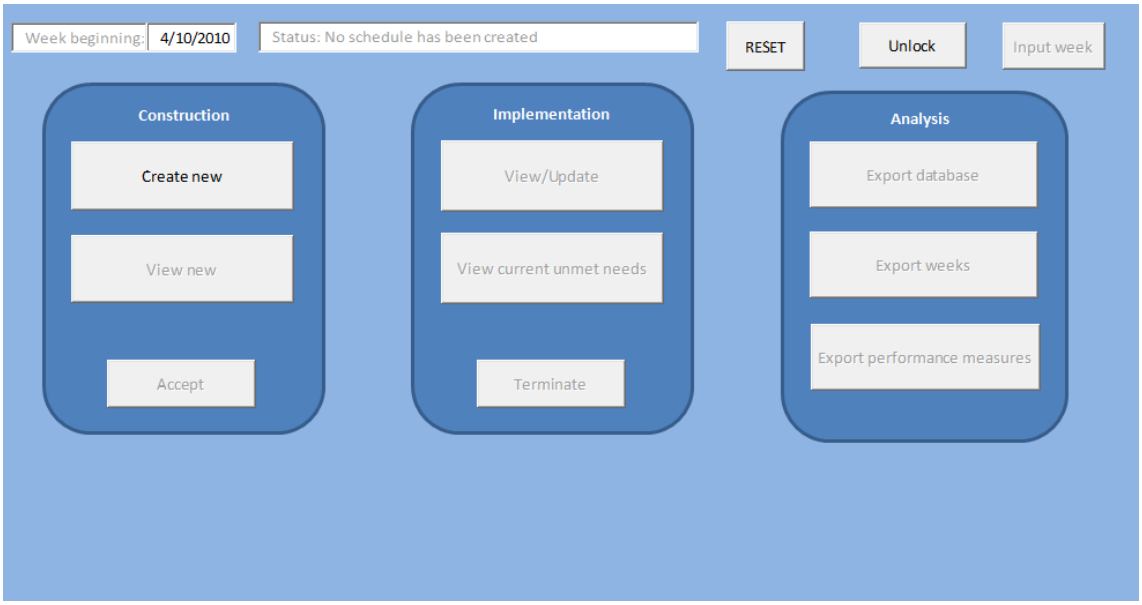


Figure 4.2 Main menu of the scheduling program

Once the scheduling program is opened it must be reset (see Figure 4.2). The status will then read 'No schedule has been created' and a date can be added for the forthcoming week. All the buttons within the three stages should be disabled with the exception of 'Create new'.

Upon clicking this button a number of inputs are required from the user. These relate to the *direct controls* that have been discussed in the previous chapter. Before the user can enter (or modify) any of these inputs the relevant information on patient availability is automatically loaded from the PTO. This allows the user to update any information on the PTO without

having to restart the schedule. There is also an ‘Options’ button that takes the user to another screen which contains many details that can be modified.

Figure 4.3 ‘Create new’ menu

The first two inputs (Figure 4.3) require a specification of patient details and demand. Each button takes the user to a screen where the respective information can be entered. Data associated with patients who are not new or have not since been discharged are carried through from the previous week. Upon completion of these fields the user is returned to the screen displayed in Figure 4.3.

The third input is the type of week. As mentioned previously the physiotherapy department operates an even/odd week rotation. The framework DSTs[5] (that contain partitions, lunch, admin and regular events) for each type of week are stored in the program[6]. A facility for a ‘custom week’ is also available.

Once a type of week is chosen, the program creates a copy of the typical DSTs from that week. The user may now click 'Input multi-dept sessions' or 'Input staff availabilities/slots'. By selecting the former, the program cycles through all sessions that have been entered onto the PTO that involve the physiotherapy department (e.g. goal planning meetings). For each of

[5] Daily staff timetables (see Ch 3.4.3)
[6] They may be modified by clicking the ‘Options’ button of Figure 4.3

these the user is asked for the initials of the physiotherapy employee(s) that are attending the session. The program automatically enters these sessions onto the copies of the DSTs.

The user may also make other modifications to the (copied) DSTs, such as altering the partitions of employee slots or restricting availability for an employee who is ill or at a meeting. It is also possible to manually add treatment sessions. Since it has been determined that hydrotherapy and other patient interactive sessions are of a very specific nature and are not appropriate for computer assignment, they may be manually entered at this stage.

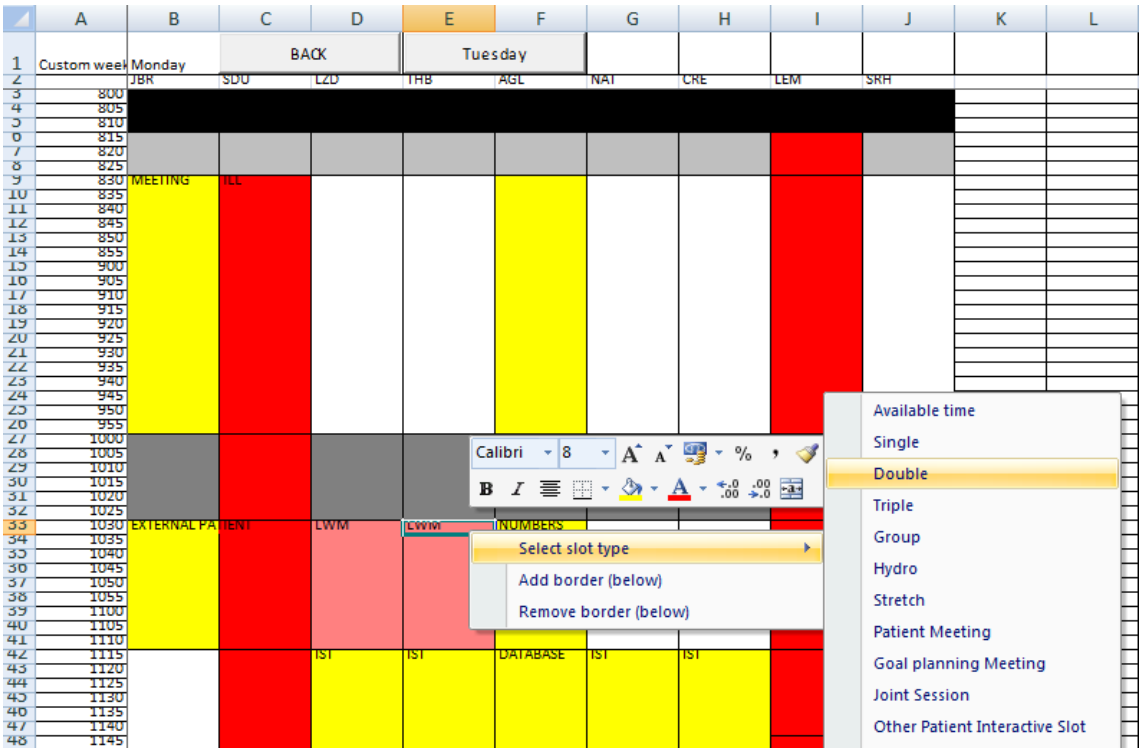


Figure 4.4 'Input staff availabilities/slots' menu (Monday)

The example depicted in Figure 4.4 shows the DST (copy) for a Monday morning. The DST at this time is based on the framework DST for that day but contains modifications that are specific to that week. This example showcases how a double may be manually added for patient 'LWM' from 1030 to 1115 hours. Note how the colour of the session denotes its type. A consistency check for all sessions on that day is automatically undertaken whenever the user attempts to navigate from this screen.

Upon completion of these five inputs it is necessary to specify the time and date that the automated program can run until. Since Excel has a single-thread engine it is not possible to introduce a feature that can stop the program at any given time.

Once satisfied the user may click 'GO!'. First of all, the program checks for any hard constraint violations. If some are found, they are pointed out to the user who is then directed to the 'Create New' screen in order to rectify them. Once no violations are found, all that remains is to enter the 'session attributes' for any (applicable) sessions that have been manually entered. The possible options for each session are as follows:

1. Internally fixed[7]; externally fixed[8]; therapist fixed[9]
2. Internally flexible; externally fixed; therapist fixed
3. Internally flexible; externally fixed; therapist flexible
4. Internally flexible; externally flexible; therapist flexible

4.5 Construction

The schedule is constructed in three stages based on the hierarchical satisfaction of hard, moderate and soft constraints (Ch 3.5). The first stage produces a valid[10] and possibly feasible[11] solution. The second stage optimises this solution with respect to the moderate constraints; ensuring that no hard constraints are broken as a result. Finally, the third stage attempts to minimise the cost function whilst making sure that no (more) hard or moderate constraints are violated.

Note that the automated program is only responsible for the assignment of patient meetings, patient admin, group sessions, stretches, and single, double, and triple sessions. That being said, the quality of the solution is dependent on the assignment of all sessions – not just those that are scheduled by machine.

Note also that the explanations provided in this subchapter are heavily abbreviated. For a more complete understanding the reader is referred to the code.

4.5.1Stage One: Initial construction

The objective at this stage is to effectively and efficiently produce an initial solution that contains no violation of any hard constraint. Following this the local search algorithms of stages two and three can be employed.

[7] The session cannot be moved to any other place within the week
[8] The session cannot be removed from the DST/PTO
[9] The session cannot be provided by any other employee(s)
[10] Satisfies all hard constraints; not all events placed
[11] Satisfies all hard constraints; all events placed

The development of an initial solution is made possible through specialised *constructive algorithms*[12]. Each type of patient session that is assigned by computer has a specialised constructive algorithm that produces an initial allocation of sessions. A generic algorithm is not possible due to the characteristic differences between the assignment procedures of different session types.

The constructive algorithms that are used employ a greedy heuristic at each iteration[13] to select the 'best' assignment for the current solution. First, appropriate elements (patients or patient sessions) are added to a *candidate list*. These elements are then ranked according to the quality of the solution their assignment would achieve. At each iteration, the 'best' element is selected for assignment. Greedy heuristics are *'often'* used in the construction of an initial solution (Ehrgott & Gandibleux, 2000). Whilst this approach *'rarely leads to an optimal solution'* (Bondy & Murty, 2008) it does *'typically construct reasonably good starting points for a local search algorithm'* (Stutzle, 1998).

Many procedures and heuristics are used within these constructive algorithms to produce an initial solution that is not just valid and/or feasible but is also of high quality. This reduces the workload of subsequent local search algorithms. It could also enhance the quality of the final solution because local search algorithms can struggle in highly constrained situations. Note that whilst the constructive algorithms are *stochastic*[14] they are not *randomised*[15].

Patient meetings

The first sessions that are scheduled by the automated program are patient meetings (see Ch 3.4.2). The assignment of these sessions is given precedence over others because the demand is typically very low (zero or one). The rationale for this follows from the sequential scheduling process (see Ch 3.3.2) in which smaller departments schedule before larger ones. The constructive algorithm for the assignment of patient meetings is described as follows.

[12] These *'generate solutions from scratch by adding to an initially empty solution components in some order until a solution is complete'* (Stutzle, 1998)
[13] The process that concerns the assignment of one session
[14] Random numbers ensure a different solution is generated on each run
[15] Elements are not selected at random from a restricted candidate list of 'good' solutions (see Resende & Ribeiro, 2003)

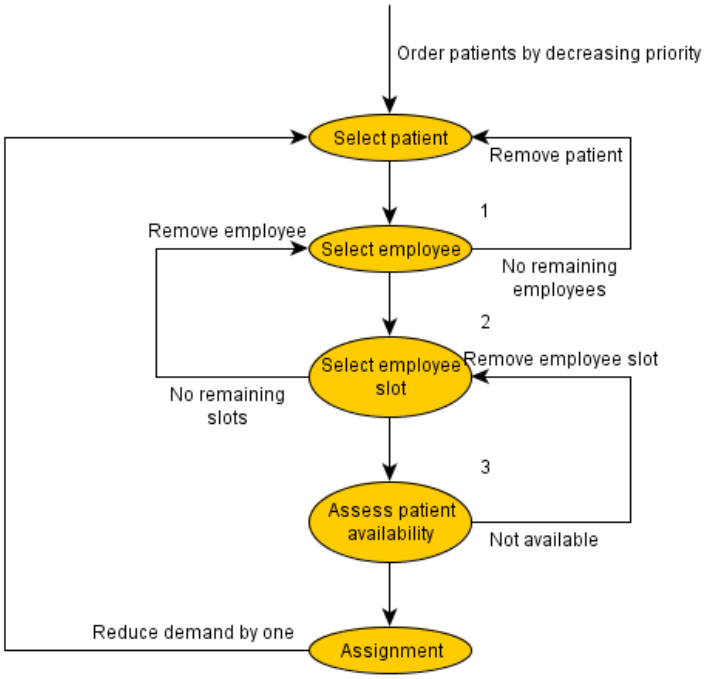


Figure 4.5 Assignment of patient meetings

1. The primary and secondary therapists are first and second choice respectively. If both are found to be unsuitable[16] or unavailable, then any other suitable employee may be selected.

2. A suitable employee slot is selected at random. An employee slot is suitable if: it is empty (i.e. contains no patients); is of 45/60 minute duration; does not clash with lunch-time if the patient attends lunch ($aal_i = 1$) and if a goal planning meeting is scheduled then the patient meeting must end before this begins. If a slot is of 45/60 minute multiples then it is appropriately partitioned into 45/60 minute slots.

3. Slots that are not contained within group or stretch availability periods are given precedence over others to deter assignments at these times. This ensures that search space is available for allocations of group sessions and stretches. Attempts are also made to select slots that are at a preferred time of day (am/pm) for appropriate patients ($apr_i > 0$).

Groups

Group sessions are assigned next. These are allocated in advance of patient admin, single, double and triple sessions because these sessions can occur at the specific times that group

[16] If not proficient in the provision of patient meetings (i.e $tpm_j = 0$ in Appendix 3.3)

sessions can occur (Ch 3.4.2) this is not the case vice versa. Therefore the overall heuristic governing the initial construction could be likened to a *saturation degree* heuristic. These select the *'event which has the smallest number of valid periods available for scheduling'* (Burke & Petrovic, 2002). Assume that selection is made at random unless otherwise stipulated.

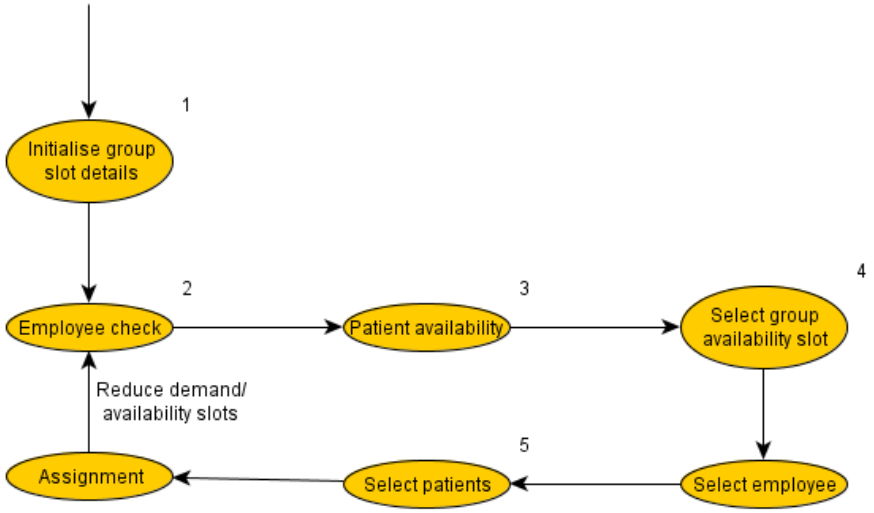


Figure 4.6 Assignment of group sessions

1. Determine whether any group slots have been manually input.
2. Assess the proficiency of staff ($tg_j = 1$) and their availability at each group slot availability period.
3. Determine the number of patients that are available at each group slot availability period.
4. Select a slot that contains the greatest number of available patients on a day that contains the fewest number of group slot assignments.
5. By level of priority.

The approach taken in this (and following) constructive algorithm uses *adaptive greedy heuristics*. They are adaptive because the assignment procedure at each iteration *'takes into account previous decisions in the construction'* (Osman & Laporte, 1996). This tends to produce a higher quality assignment.

Stretches

There is no practical reason why group sessions are assigned before stretches since group and stretch availability periods typically occupy different times. They must, however, be assigned in advance of patient admin and single, double, and triple sessions for reasons that have been previously acknowledged. Assume that selection is made at random unless otherwise stipulated.

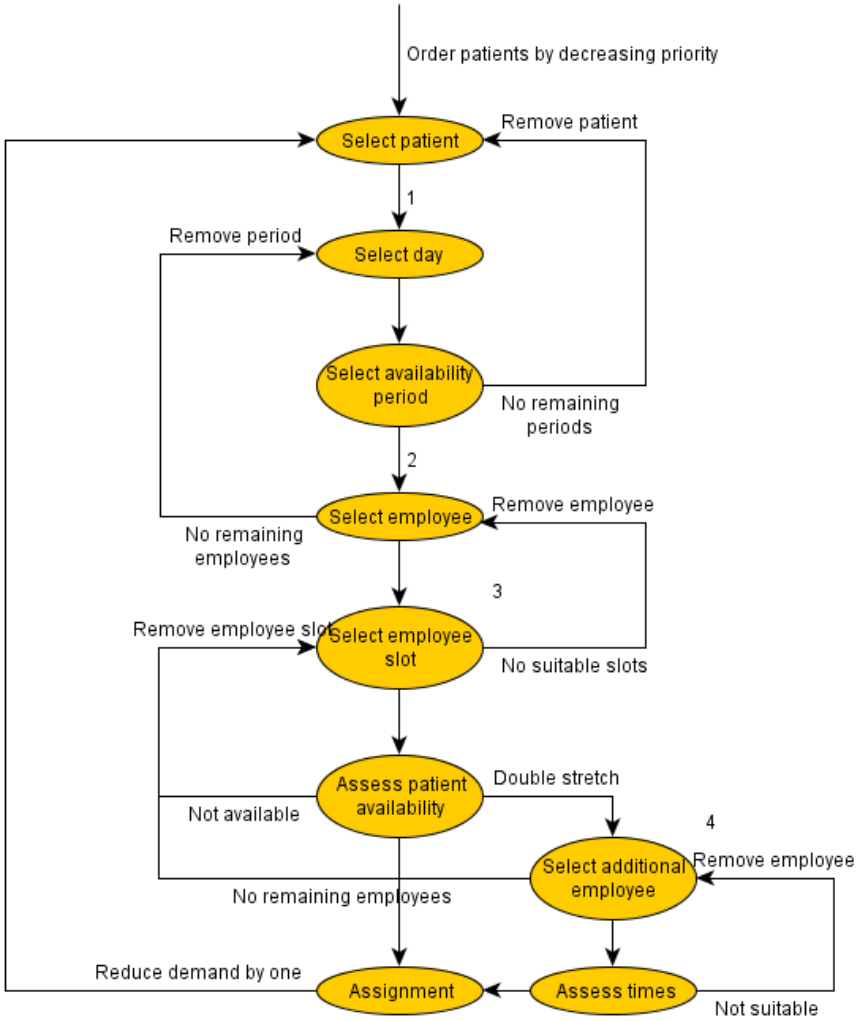


Figure 4.7 Assignment of stretches

1. An initial list of all days is constructed. First, any days that do not contain any stretch availability periods (see Ch 3.4.2) are removed. Second, the minimum number of stretches assigned for that patient on each of the remaining days is calculated. Any days which contain more than this minimum are removed. This enhances the likelihood of an equal spread of stretches. Furthermore, if stretches are for treatment

preparation (Ch 3.4.2) then any of the listed days are removed if they contain fewer than the maximal number of joint sessions (as calculated for the listed days). This consideration is made only for joint sessions because single, double and triple sessions have not yet been assigned.

2. An initial list of all employees is constructed. Any employees that cannot 'lead' stretches ($tst_j < 2$) are removed. Any employees whose slots do not hold the same start and end time as the availability period are removed.

3. If the (empty) employee slot is of multiples of 30 minutes but not 60 minutes then every $(2 \cdot i - 1) - th$ slot is selected before every $(2 \cdot i) - th$ slot $\forall i = 1..(\text{slot length} / 60)$. This is to enhance the likelihood that other session(s) of 60 minute length can be assigned to the remaining time if it remains empty. For example, if a stretch is assigned from 830-900 or from 930-1000 in the even week Wednesday of Appendix 3.2 then this still allows time for a 60 minute session to be assigned.

4. Employees that only assist ($tst_j = 1$) are selected before those that lead ($tst_j = 2$). This is because it is unnecessary for a stretch to have two lead employees assigned.

Patient admin

The constructive algorithm for patient admin sessions is similar to that of patient meetings. However, there is no dependence on goal planning meetings so these sessions can take place at any time in the week. Also, since group sessions and stretches have, at this point, been assigned there is no reason to deter assignments within these availability periods.

Singles, doubles, and triples

These sessions may be assigned using one constructive algorithm. This is because the assignment procedures are very similar; the only practical difference being the number of employees that it is necessary to assign.

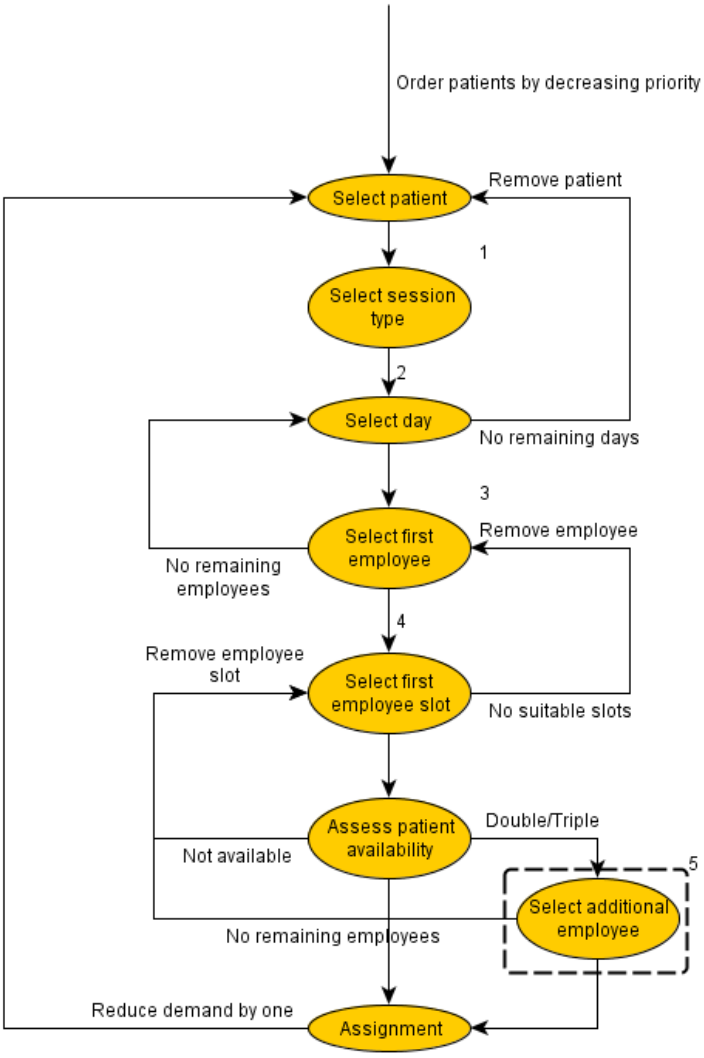


Figure 4.8 Assignment of singles, doubles, triples

1. Session types are selected on the basis of decreasing complexity (i.e. triples then doubles then singles). This represents a *largest degree first* type heuristic that is used because *'events with a large number of conflicts are more difficult to schedule and so should be tackled first'* (Burke & Petrovic, 2002).

2. An initial list of all days is constructed. The aim is to select a day that contains the least number of single, double, triple, joint, hydro and other patient sessions assigned for that patient. Therefore, any days that contain more than the minimum number assigned are removed. Furthermore, if stretches are required for treatment preparation then any of the listed days are removed if they contain fewer than the maximal number of stretches (as calculated for the listed days).

3. Providing the primary and secondary therapists can lead ($tl_j = 1$) and are above the minimum band level of the patient ($tb_j \geq amb_i$) they are the first and second choice respectively. If both are unsuitable or unavailable then any other employee fulfilling these criteria is eligible for selection. In order to ensure that high-band therapists are not wasted the employees with the lowest band levels are selected first.

4. An initial list of all employee slots is constructed. Any slots that are non-empty are removed. So too are any slots that are neither a multiple of 45 nor 60 minutes. Applicable slots are then appropriately partitioned into 45 or 60 minute slots. Any slots are removed that clash with lunch-time should the patient attend lunch.

5. If the session is a double or a triple then additional employee(s) are required. Such employees must be able to assist ($ta_j = 1$). Since a lead therapist has already been assigned it is preferable for the remaining employees to be assistants (i.e. $tb_j < 5$). These are selected before higher-band employees and their availability is assessed. If all employees are found to be unavailable then a different first employee slot is sought.

4.5.2 Stage Two: Moderate constraint optimisation

The objective at this stage is to reduce the number of moderate constraint violations (Ch 3.5.3) whilst respecting hierarchical precedence. That is, a change to the solution that reduces the number of moderate constraint violations can only be accepted if there remain no hard constraint violations. A hierarchy also exists within the moderate constraints. The satisfaction of higher-order moderate constraints is more important than the satisfaction of those that are of lower-order. The order is principally determined by the number that precedes the decimal point e.g. *MCON1* is of higher-order than *MCON2*. In addition, there exists a hierarchy within each of the moderate constraints. This is determined by the number that follows the decimal point e.g. *MCON3.2* is of higher-order than *MCON3.3*. As an example, if a new solution is found that contains fewer violations of *MCON2.2* then this can only be accepted if there are any no (more) violations of *MCON2.1*, *MCON1*, or any *HCON* (hard constraint).

The optimisation that is sought at this stage is reliant on specialised *local search algorithms* (see Ch 4.3.1) to effectively and efficiently reduce the number of moderate constraint violations. It is necessary to develop these algorithms singly for individual constraints

because of the differences that relate to their definitions. That being said, each of the algorithms is broadly governed by a similar form of local search. Firstly, a patient is selected that has one or more violations of the *MCON* in question. Secondly, the algorithm intelligently identifies any 'problem' sessions that are causing the violations. The neighbourhood of the solution is then appropriately inspected to determine whether neighbouring solutions deliver an improvement. If such a neighbour is found then the current solution is replaced and the next violation is addressed. This iterative process is illustrated below. When all violations have been studied the next moderate constraint is considered.

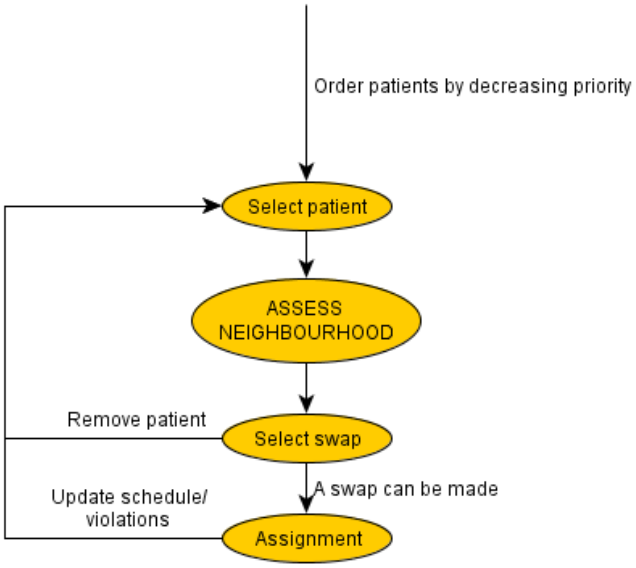


Figure 4.9 General reassignment process for moderate constraint optimisation

It can take much time to assess the quality of a neighbourhood. Therefore, to promote efficiency, only a fraction of neighbouring solutions are assessed. The neighbourhood is trimmed to contain only the solutions that result in fewer violations of the moderate constraint in question. This is done on a violation-by-violation basis. Firstly the session(s) that incur a particular violation are identified and then either a MOVE or SWAP is performed on these to generate a new solution. A MOVE is a one-way reassignment of a 'problem session' to an empty employee slot(s) whilst a SWAP involves the exchange of the problem session with another patient session. MOVEs and SWAPs can be *internal* or *external*. If internal then all of the sessions involved in the MOVE or SWAP remain assigned to the PTO and DSTs but their positions within the week are adjusted. Otherwise, the 'problem' session is removed from the week and either replaced with an unmet need (an unscheduled patient session for which demand exists) or left empty.

The *search process* that governs the acceptance of neighbouring solutions must also be addressed. Two popular options are *first improvement* and *best improvement*. First improvement selects the first neighbouring solution found that is of better quality than the original. Best improvement uses an exhaustive search to examine all neighbouring solutions and selects the one that is of the highest quality. It is intuitive that the former is more efficient whilst the latter is more effective. Best improvement (or *steepest descent*) is used here because the principal drawback of this (inefficiency) is negligible as the neighbourhoods are trimmed. In addition, it is found in Hansen & Mladenovic, 2006 that best improvement gives *'better and faster results'* (to the TSP) when the initial solution is generated using greedy constructive heuristics (*c.f.* Ch 4.5.1).

A swap is rejected on the basis of the hierarchical precedence that is introduced at the outset of this subchapter. It is likely, however, that at least one swap remains that reduces the number of violations of the *MCON* in question whilst incurring no higher-order constraint violations. In such an event the lower-order constraint violations are inspected in turn. At each step any swaps are rejected that contain less violation reductions than the best possible swap with respect to the *MCON* in question. This process is terminated when only one swap is left or when the lowest-order *MCON* has been considered. In addition, there is a hierarchy within individual moderate constraints that is based on patient priority. If the number of *MCON* violations remains the same between two swaps then the individual patient *MCON* violations must be inspected to assess whether a low-priority patient has lost a violation at the expense of a high-priority patient gaining one. If a swap is not found that produces a better quality solution then the patient is removed from the list.

Although not referred to at the time, the procedures for calculating the number of *MCON* violations (Ch 3.5.3) have been carefully developed to represent a minimal *distance to feasibility*. This term is typically used to quantify the extent to which the solution is infeasible. In this case it is used to specify the minimal number of SWAPs or MOVEs necessary to achieve *MCON* satisfaction. For example, if *MCON2.1* was equal to two for a particular patient then at the very least two sessions would need to be reassigned in order to reduce this to zero. An understanding of this 'distance' is useful since it provides a lower bound on the complexity that is associated with efforts to satisfy the moderate constraints.

MCON1

It has been found that in the vast majority of cases the violations of *MCON1* (see Ch 3.5.3 for details) can be reduced to zero by modifying the constructive algorithms of stage one. Most of these algorithms work by firstly selecting a patient (by order of priority) and then attempting to make assignments until all demand is satisfied. The modification is to use a two-phase approach. On the first phase patients are selected (as before) but this time only one assignment is attempted. Once all patients have been considered phase two is initiated. This attempts to assign the remaining demand for the patients. By making this simple adjustment the need to develop specialised algorithms for the satisfaction of this *MCON* is redundant.

MCON2

The reassignment of problem session(s) in this moderate constraint is based on internal MOVEs and SWAPs. To avoid repetition in the explanation of the (similar) specialised algorithms for each of the individual moderate constraints only one is described in detail.

The objective of the specialised algorithm for *MCON2.1* is to produce an even spread of group sessions for all patients. The algorithm for **'Assess Neighbourhood'** (in Figure 4.9) is described in Figure 4.10. Note that selection is random if not otherwise specified. Note also that any reassignment is referred to as a swap. This is appropriate since a MOVE is equivalent to a SWAP with an empty employee slot(s).

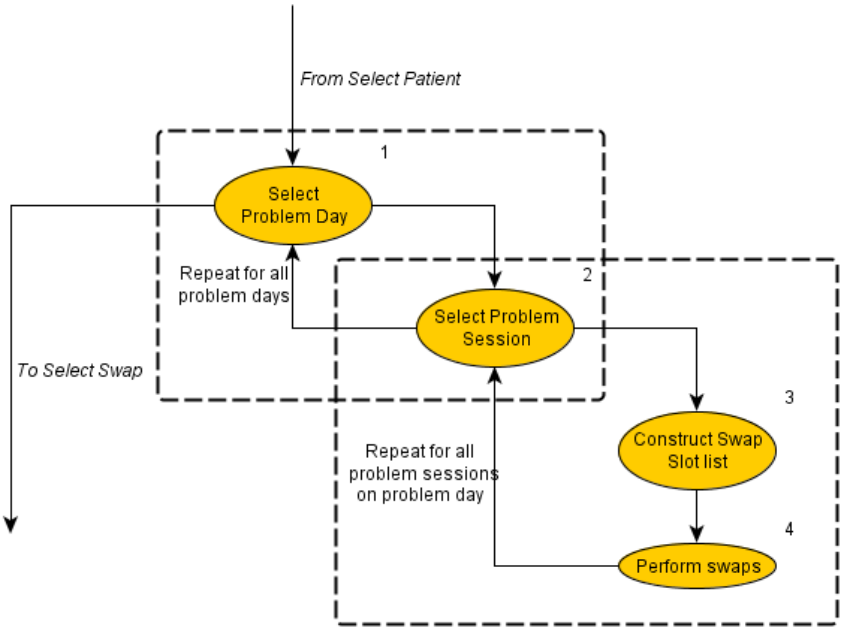


Figure 4.10 Reassignment selection process for MCON2

1. A problem day has more than one group session in difference with the day that contains the least number of group session assignments for the patient.

2. A problem session is any group session on the problem day that is not internally fixed. A problem slot refers to the employee slot(s) that contain the problem session. A 'problem employee' is an employee that provides the problem session.

3. A 'swap slot' is any 'suitable' employee slot on any 'swap day'. A swap day is a day that contains the least number of group session assignments for the patient. A suitable employee slot is contained within a group availability period; has a group session assigned; is of appropriate duration (45/60 minutes); and is not internally or therapist fixed. In addition, if the problem session is 'therapist fixed' then the problem employee must match the 'swap employee'.

4. The problem patient is swapped with each of the 'swap patients' from each of the swap slots in turn. If there is space within a swap slot then a one-way swap is also made. This process requires the construction and use of temporary details for each of the swaps. Only the details of the problem and swap entities are replicated in order to promote efficiency. Each swap receives a 'swap number' and relevant details (*HCON* and *MCON* violations and *SCON* value) are stored against this.

The specialised algorithm of *MCON2.2* differs from this since swap slots are restricted to stretch availability periods and problem sessions can require two problem employees. The specialised algorithm of *MCON2.3* relaxes the requirement of swap slots to be contained in availability periods, but does introduce the possibility of a third problem employee. This further increases the complexity since the neighbourhood grows exponentially with the addition of a problem employee to a problem session. If there are *n* possible swap slots for problem employee 1 then there are $n-1$ swap slots for problem employee 2. The number of trial solutions is therefore approximately equal to n^2 for two employees and n^3 for three employees (since *n* is typically large).

MCON3

The objective of the specialised algorithms for *MCON3* is to obtain consistency with respect to the requested and assigned volumes of stretches and single, double, triple and joint sessions for patients who require stretches for treatment preparation. The reassignment of problem session(s) in this moderate constraint is based on external MOVEs and SWAPs.

Again, to avoid repetition in the explanation of the (similar) specialised algorithms for each of the individual moderate constraints only one is described in detail.

The number of violations of *MCON3.1* represents the shortfall in the number of assigned stretches. This constraint can be satisfied by removing single, double, triple, or joint sessions and/or assigning stretches for appropriate patients. Note that a stretch can only be added if the patient has a nonzero number of unmet needs in stretches.

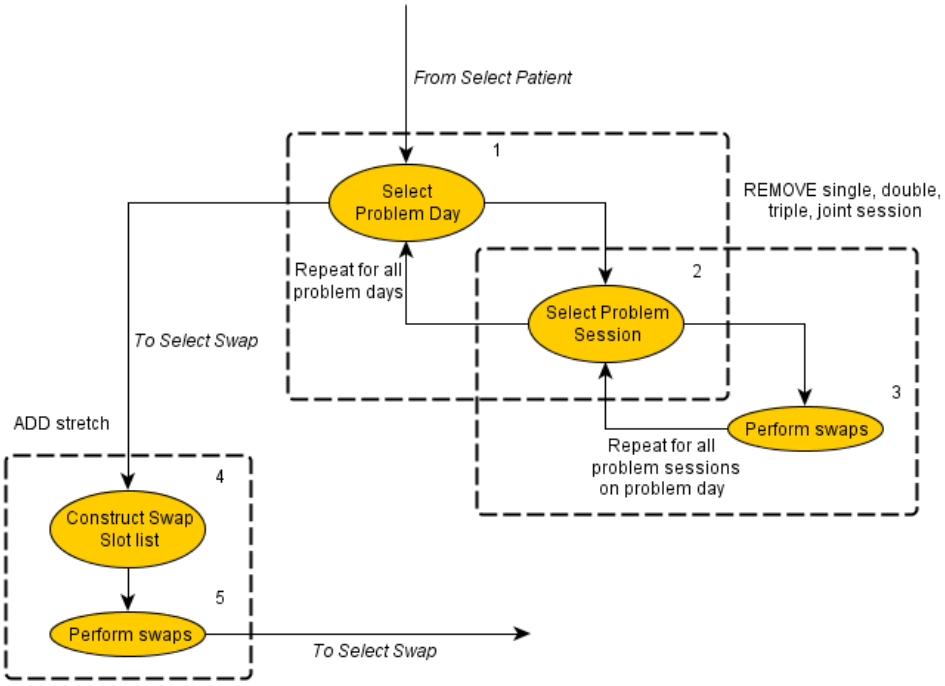


Figure 4.11 Reassignment selection process forMCON3

1. A problem day is defined as any day that contains a single, double, triple, or joint session.

2. A problem session is defined as any single, double, triple or joint session on the problem day that is not externally fixed.

3. An external swap is performed between the problem slot(s) of the problem session and any appropriate unmet need(s). For example if there are two problem slots then any double sessions that are unmet will be externally swapped in turn. So too will any two single sessions (from different patients). In addition the one-way movement of the problem session to an unmet need is also considered.

4. Any slot that is contained in stretch availability periods; is of 30 minute duration; and is not externally or therapist fixed.

5. An external swap is performed between the unmet stretch of the problem patient and any of the swap slots in turn. If the problem patient requires two employees to provide a stretch then two swap slots must be selected.

MCON4

Both algorithms for *MCON4.1* and *MCON4.2* employ internal swaps. Constraint violations are only possible for patients who require stretches for treatment preparation.

MCON4.1

The number of violations of *MCON4.1* reflects the number of sub-optimal days that are scheduled in relation to the minimal number of sub-optimal days that are possible with the scheduled volumes. A sub-optimal day contains either (i) a stretch without a post-stretch session[17] or (ii) vice versa.

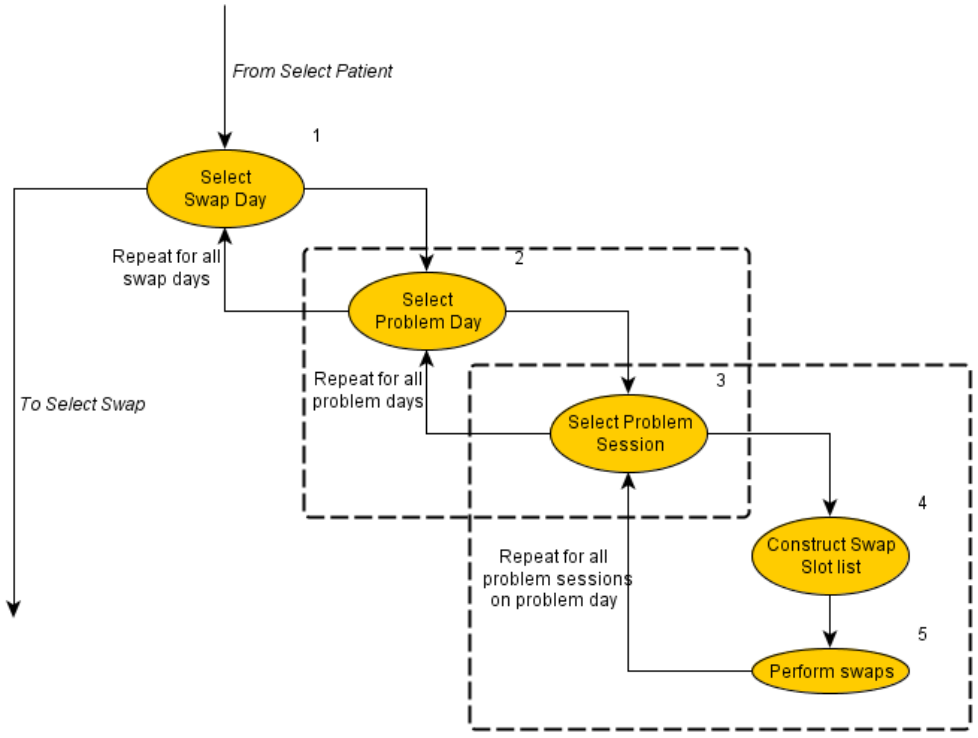


Figure 4.12 Reassignment selection process for MCON4.1

1. A swap day is defined as any sub-optimal day.

[17] Single, double, triple, or joint session

2. If the swap day is a type (i) sub-optimal day then a problem day is any day that contains either two or more post-stretch sessions or one post-stretch session with no stretches. If type (ii) then a problem day is any day that contains either two or more stretches or one stretch with no post-stretch sessions.

3. A problem session is any of the complimentary sessions[18] on the problem day.

4. Any slot that is on the swap day; is equal in duration to the problem session; and is not internally fixed. If the problem session is therapist fixed then the problem employee must match the swap employee.

5. An internal swap is performed between the problem slot(s) and each of the swap slot(s) in turn. The number of problem and swap slots that are swapped depends on the number of employees assigned to the problem session.

This systematic approach has been specifically designed so that all possible options are considered. For example, it is possible for a patient session to be used as both a problem and a swap session within one iteration. This evaluation of all appropriate neighbouring solutions means that the one producing the highest quality solution can be returned.

MCON4.2

The number of violations represents the number of instances in which it is possible for a stretch to be in advance of a post-stretch session, but is not.

[18] If type (i) then any post-stretch session; if type (ii) then any stretch

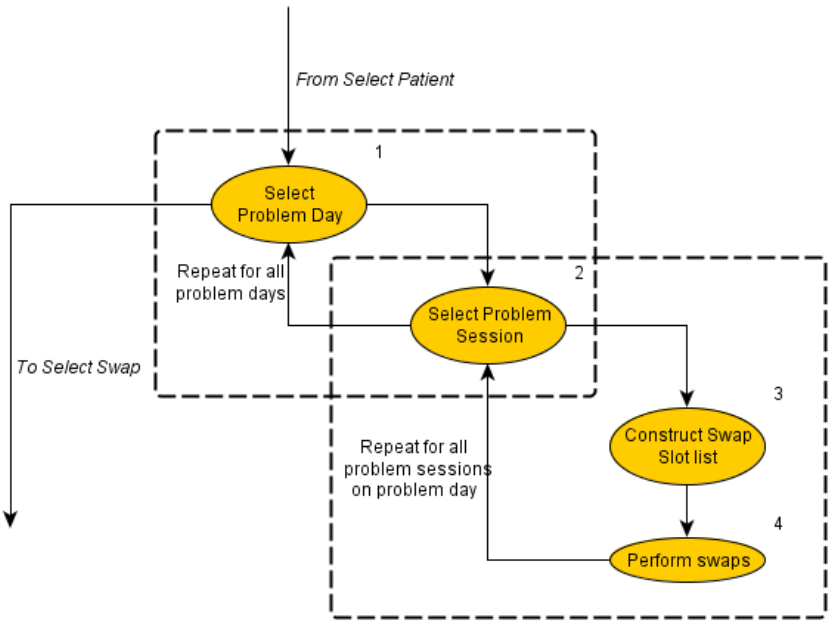


Figure 4.13 Reassignment selection process for MCON4.2

1. A problem day is defined as any day that contains either a post-stretch session with no preceding stretch or a stretch with no proceeding post-stretch session. The day must contain a nonzero assignment of both stretches and post-stretch sessions.

2. If, on the problem day, there are post-stretch session(s) that lack a preceding stretch then the problem sessions are defined as these post-stretch sessions in addition to any stretches that occur on that day for the patient in question. It is these sessions that can be (appropriately) swapped with others that can reduce this particular violation. Conversely, if there are stretches that lack a proceeding post-stretch session then the problem sessions are defined as these stretches in addition to any post-stretch sessions that occur on that day for the patient in question.

3. If the problem session is a stretch then the list contains any slot that is on the problem day; is of 30 minute duration; and is not internally fixed. In addition, the slot must end before the start of the post-stretch session that is immediately before the problem session. If the problem session is a post-stretch session then the eligibility criteria differs insofar as the slot must be of 45/60 minute duration and must start after the end of the stretch session that is immediately after the problem session. If the problem session is therapist fixed then the problem employee must match the swap employee.

4. An internal swap is performed between the problem slot(s) of the problem session and each of the swap slot(s) in turn. The number of problem and swap slots that are swapped depends on the number of employees assigned to the problem session.

4.5.3 Stage Three: Soft constraint optimisation

The objective at this stage is to reduce the value of the *cost function*[19] whilst respecting hierarchical precedence. This is done by using a generic local search algorithm to inspect appropriate neighbourhoods of the solution. If neighbouring solutions are found to contain (more) moderate or hard constraint violations then they are rejected. If they are found to contain fewer moderate constraint violations they are accepted with respect to the hierarchy that is addressed in the introduction to Ch 4.5.2. The possibility of this exists because the specialised local search algorithms for the *MCON*s can only consider solutions that are one neighbourhood away.

The neighbourhood is iteratively explored as follows. In summary, an employee slot is selected at random. This can be empty (no patients assigned) or non-empty. If it is empty then a one-way external swap is performed to attempt to assign some unmet need. If it is non-empty then both internal and external swaps are performed.

[19] The weighted addition of the soft constraint values (see Ch 3.5.2)

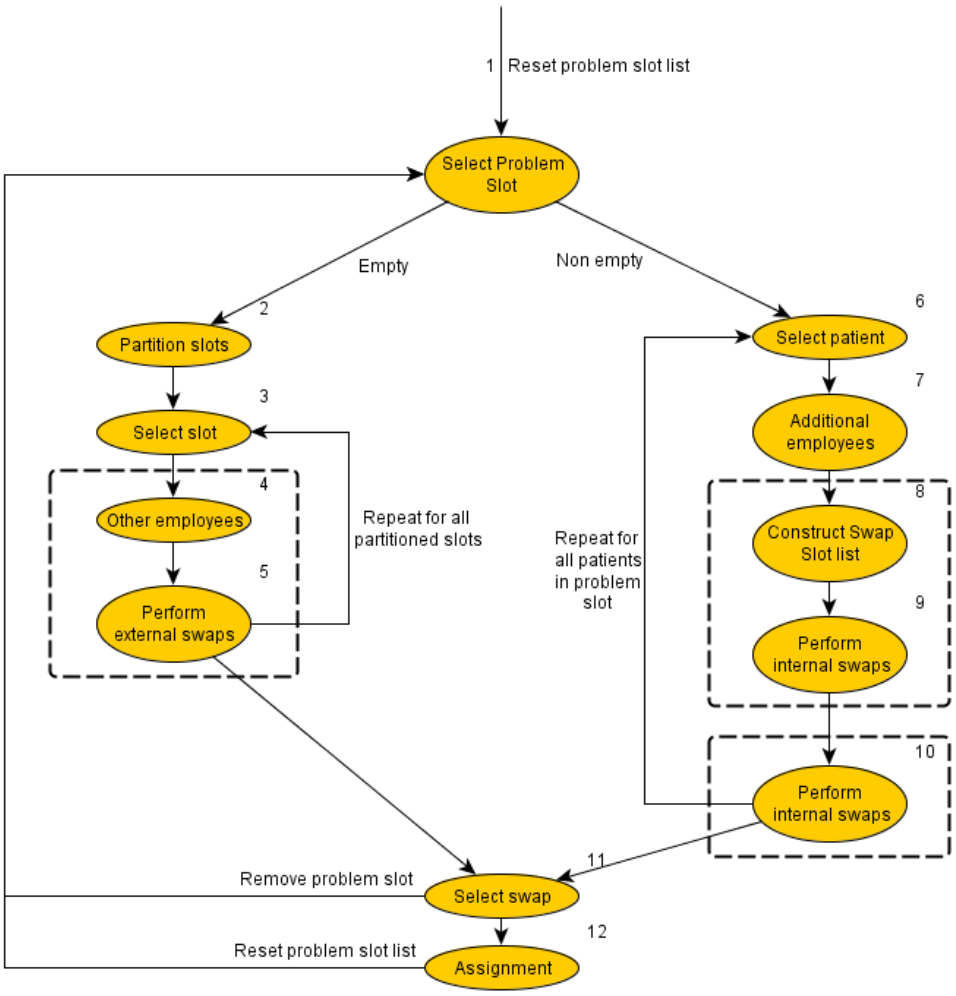


Figure 4.14 Generic reassignment process for soft constraint optimisation

1. This list contains any employee slots that are not internally, externally and therapist fixed and are not 'complicated'[20]. Suitable slots must either be of 30, 45 or 60 minute duration or must take place within a stretch availability period.

2. If the problem slot is contained within a stretch availability period then all empty employee slots that are contained within this stretch availability period on that day are partitioned. This occurs, where appropriate, at 30 minute increments from the start of the availability period to the end. It is necessary to do this so that otherwise unsuitable slots (i.e. those that exceed 30 minutes) can be made available for the assignment of single and double stretches. Such slots are not permanently partitioned this way because it must also be possible for other sessions (of 45/60 minute duration) to be assigned within these periods.

[20] Slots/sessions that involve more than one patient and more than one employee

3. More than one slot is possible only when the problem slot is contained in a stretch availability period and at least two slots have been created as a result of the partition. The selected slot is thus known as the problem slot.

4. Other employee slots are assessed to determine whether any additional employees have availability at the same time and for the same duration as the problem employee. If the problem slot is of 30 minute duration then a maximum of one additional employee can be selected. This is because the maximum number of employees attributed to a stretch session is two. This will therefore permit the possibility that a double stretch unmet need can be assigned. If of 45/60 minute duration then a maximum of two additional employees can be selected; making possible the assignment of a triple session unmet need or a single and double session unmet need etc. Any appropriate slots are also referred to as problem slots.

5. An external swap is performed between the problem slot(s) and each (appropriate combination) of the unmet needs.

6. Each patient assigned to the problem slot is, in turn, selected as the problem patient.

7. The problem session corresponding to the problem slot is examined to determine whether there are any more employees assigned and thus if there are any more problem slots.

8. This list contains any slot that has duration greater than or equal to that of the problem slot and is not internally fixed. If the problem slot is a stretch then any slots that are not contained in any of the stretch availability periods are excluded. If the problem session is therapist fixed then the problem employee must match the swap employee.

9. An internal swap (of patients) is performed between the problem slot(s) of the problem session and each of the swap slot(s) in turn. If a swap slot is empty then a one-way swap is made.

10. An external swap is performed between the problem slot(s) of the problem session and each (appropriate combination) of the unmet needs.

11. A swap is selected based on the hierarchical precedence that has been previously discussed. However, if there remain a nonzero number of swaps after removing those that lead to (more) moderate or hard constraint violations then the following apply. If

there are fewer moderate constraint violations under any of the remaining swaps than the current solution then the neighbour that produces the lowest cost function value is selected. If there are no such reductions in moderate constraint violations then one of the following three options apply[21].

a. Accept a neighbour if the cost function value is less than the original (i.e. iterative improvement – see Ch 4.3.1)

b. Accept a neighbour if the cost function value is less than or equal to the original

c. Accept a neighbour if the cost function value is less than or equal to the original. Accept also if a random number, $(0,1]$, is less than $\exp\left(\frac{\text{current cost - swap cost}}{T}\right)$. This approach is known as *Simulated Annealing*. This naturally inspired memoryless metaheuristic *'starts with a high temperature (T) to perform a coarse search of the solution space. The temperature is then reduced to focus on a specific region'* (Osman & Laporte, 1996). There are a vast number of studies into how the temperature can be reduced. For this project a simple option is taken that involves the multiplication of the temperature by a *reduction factor* (< 1). For experimental purposes this multiplication can be carried out after a successful iteration; after an unsuccessful iteration; or after any iteration

12. If a neighbouring solution is accepted then the list of problem slots is reset (see 1.). Otherwise the problem slot(s) are removed from the list. This is so that in the absence of reassignments they are not selected again. The search terminates when no problem slots remain on the list. This approach shares commonality with Tabu Search. The objective of this memory-based metaheuristic is to *'prevent the reversal, or sometimes repetition, of certain moves – by rendering certain attributes of these moves forbidden'* (Glover, 1990). Upon acceptance of a neighbour any partitions between empty slots contained within a stretch availability period are removed.

[21] This depends on what the user has input in the options menu of Figure 4.3

4.6 Graphical User Interface II

Upon completion of the schedule the user will be directed to the following screen:

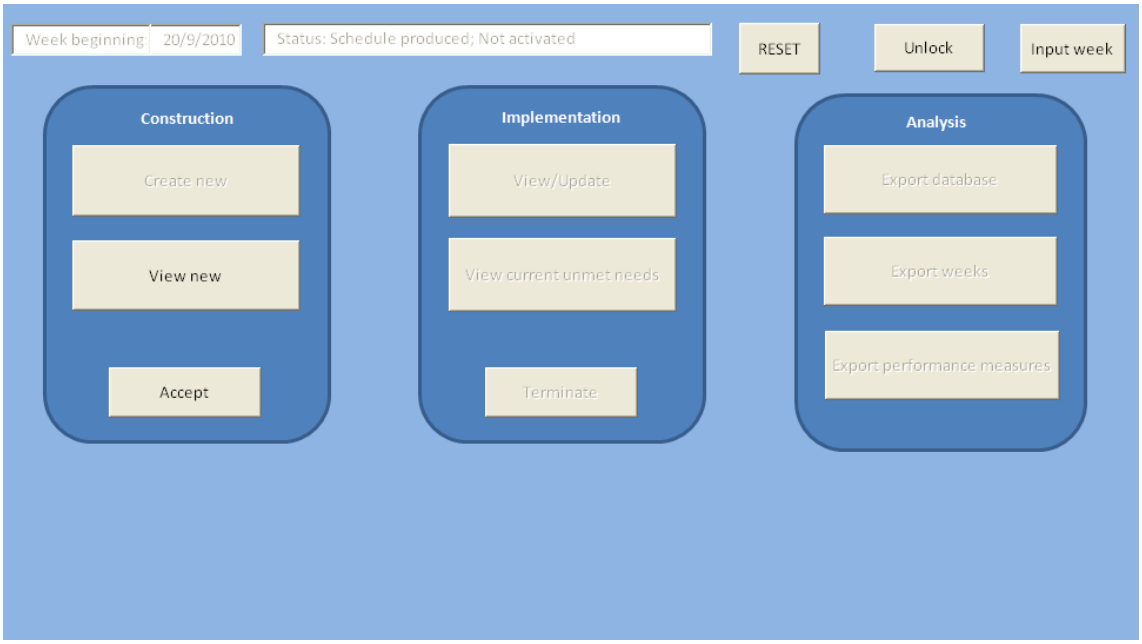


Figure 4.15 Main menu

Before pressing 'Accept' the user should inspect the schedule that has been produced. The following screen is displayed once 'View new' is pressed:

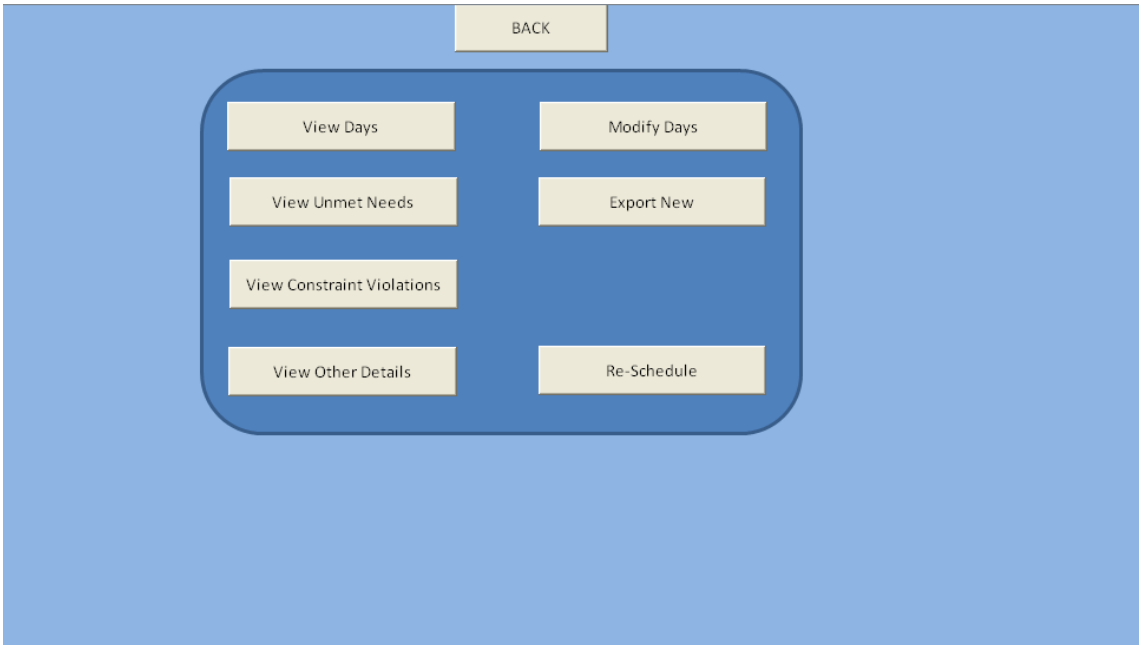


Figure 4.16 'View new' menu

- 'View Days' - The generated DSTs may be viewed (see Appendix 4.1 for an example)
- 'View Other Details' – Enables the user to view various attributes of the generated schedule, such as
    - percentage of (appropriate[22]) sessions provided by primary and secondary therapists
    - percentage of sessions assigned to preferred time of day
    - amount of employee time attributed to clinical, admin, other work
- 'Modify Days' – Enables the user to modify the generated schedule. Basic validity and consistency checks are automatically executed upon a modification. Schaerf, 1999 states that *'most of the systems allow the user at least to modify the final output'*
- 'Re-schedule' – This returns the user to the screen depicted in Figure 4.3. Note that upon return all buttons are enabled with the exception of week type and 'Input multi-dept sessions'. If 'Input staff availabilities/slots' is pressed then the user is directed to the DSTs equivalent to those pre-automation.

Once the user is content with a schedule the 'Export New' button can be pressed. This automatically copies all details that are represented on the LHS of Figure 4.16 to a macro-free Excel workbook. Coded patient and employee names are replaced by full names. Following this the 'Accept' button of Figure 4.15 can be pressed. The PTO is then overwritten to contain all physiotherapy sessions. The 'Produce Patient Handouts' button (on the PTO) can then be pressed. This converts the information stored on the 'User' worksheet to a format that is easier to read for the patients. An example is given in Appendix 4.2.

The 'Implementation' phase is then activated. This disables all buttons from the 'Construction' phase. If sessions go as planned then it is not necessary for any action to be taken. However, if there is a *shock* such as employee unavailability or patient illness or refusal, then an employee must make this altercation to the DSTs in the 'View/Update' section. Note that the specification of unmet needs is related to these DSTs and so is kept up-to-date. Once the treatment week is complete the 'Terminate' button can be pressed.

This activates the 'Analysis' phase which enables various measures to be output. Details on therapies received and requested could then be exported directly into a specifically designed database (see Ch 6.3.2). The DSTs of the treatment week and a selection of performance

[22] Only includes sessions that a primary or secondary is necessary, i.e. excludes groups and stretches

measures can also be exported to macro-free workbooks. From here the two types of unmet needs (Ch 3.5.2.1) can be studied.

Note that once a schedule has been produced but not yet activated the 'Input Week' button of Figure 4.15 is enabled. This allows a schedule produced by hand to be analysed. Hard and moderate constraint violations in addition to cost function value can be output once all the slots have been entered onto the DSTs. This allows for a quantitative comparison between any solutions produced by hand and machine.

4.7 Results

For three weeks in late 2010 treatment schedules have been produced by hand and by the computer program. These are now compared.

Table 4.1 Comparison of timetable production by hand and by program (12 hrs run time)

Week start (2010):	20 Sep		4 Oct		1 Nov	
Method:	hand	program	hand	program	hand	program
# MCON1 violations	2	0	0	0	0	0
# MCON2 violations	6	1	5	0	0	0
# MCON3 violations	0	0	0	0	0	0
# MCON4 violations	4	0	0	0	4	0
SCON value	280,435	147,385	207,720	125,350	318,050	151,765
Avg # clinical minutes demanded	279	279	290	290	258	258
Avg # clinical minutes scheduled	236	264	270	262	195	192
Avg # slots with neither primary or secondary therapist	48%	28%	37%	16%	37%	18%
Average total minutes available for work	976	976	906	906	891	891
Average total minutes worked	753	787	785	742	756	767

It can be seen that the computer program obtains far fewer violations of the moderate constraints. In two cases these are reduced to zero. Based on the large number of violations by hand it can be presumed that the first week is a 'tough' week in which demand is much greater than supply. The soft constraint value (or cost function) for the computer produced schedule is, on average, 46% lower than the solution obtained by hand. It is obviously not possible to always satisfy demand and this is supported by the figures for the average number of clinical (not admin) minutes demanded and scheduled. It has been observed that in order to

avoid *MCON* violations the computer program generally allocates fewer clinical minutes than by hand. This is evident in weeks two and three. It is not so in the first week and so it is postulated that the computer program is superior in dealing with 'tough' weeks in this respect. By hand twice as many (appropriate) sessions are scheduled with neither the patient's primary nor secondary therapist in attendance when compared to the computer produced schedule. Here, only one fifth of (appropriate) sessions are scheduled with neither a primary nor secondary. Finally, it must be noted that, in general, the average number of clinical minutes scheduled is proportional to the average number of minutes worked (by therapists). However, this is not always the case (see week 3) since clinical sessions can be of different duration (45/60 minutes) and do not include patient admin.

Figure 4.17 and Figure 4.18 show how the moderate constraint violations and soft constraint value decrease over time. Simulated annealing (11.c of Ch 4.5.3) is used as the soft constraint optimisation algorithm.

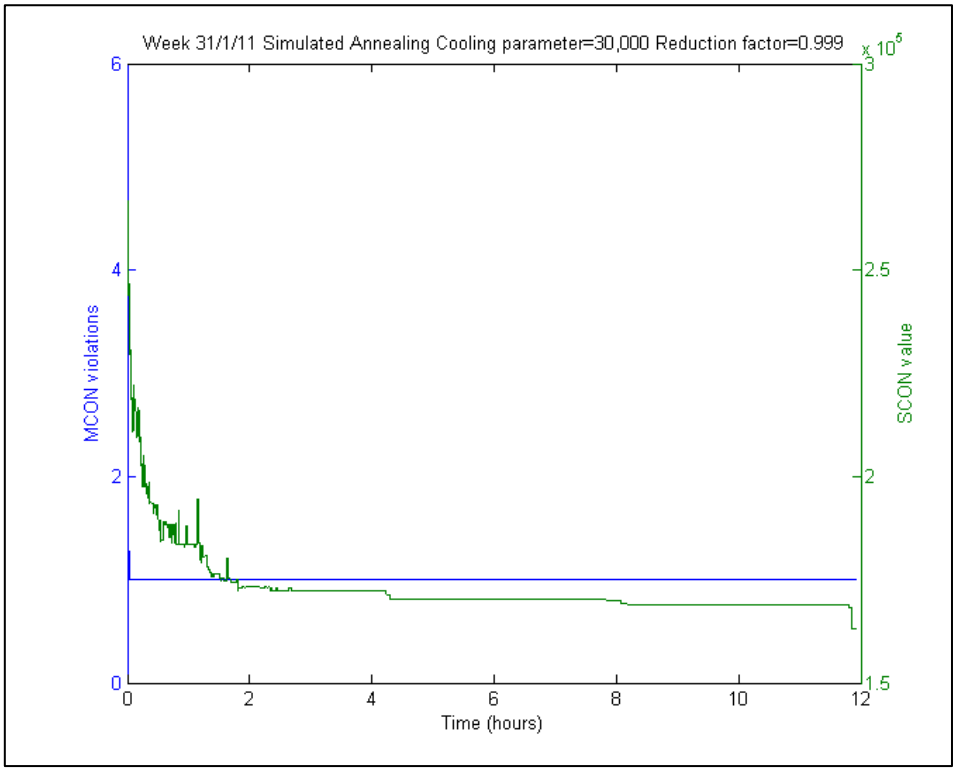


Figure 4.17 Moderate and soft constraint reduction over time

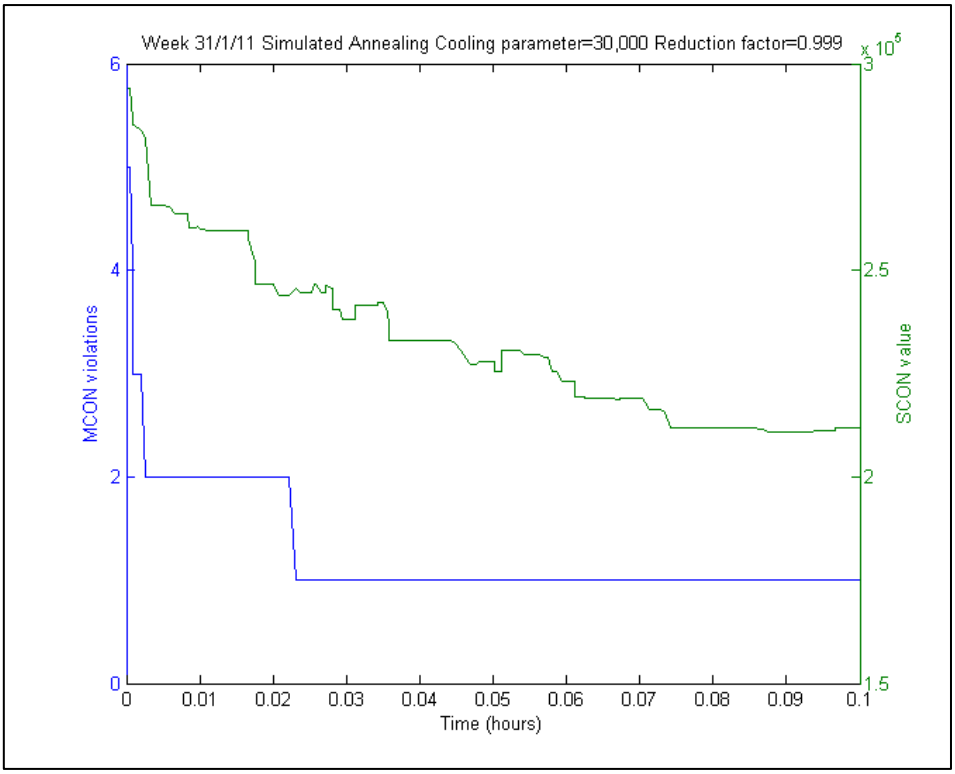


Figure 4.18 Moderate and soft constraint reduction over time

In the twelve hours the program was given to run it took two seconds for stage one (initial construction), 20 seconds for stage two (moderate constraint optimisation) and the remainder of time for stage three (soft constraint optimisation). In stage two eight iterations were performed which reduced the number of moderate constraint violations from five to three and the cost function from 294,190 to 285,190. In stage three 23,253 iterations were performed which further reduced the number of moderate constraint violations to one and the cost function to 163,280.

A stage three iteration therefore takes on average 1.86 seconds to compute. The precise time is dependent on the number of therapists assigned to the problem slot. If one therapist is assigned then the number of swap slots is about 100 and so 100 unique solutions are evaluated within the iteration. If two therapists are assigned then approximately 10,000 unique solutions are evaluated, i.e. for each of the 100 swap slots of the first therapist there are 100 swap slots for the second. Therefore, approximately one million unique solutions are evaluated within an iteration if three therapists are assigned to the problem slot. In this example, there were 5,054 iterations involving one therapist, 17,026 involving two and 1,173 involving three. It can thus be deduced that approximately 1.3 billion solutions were evaluated as part of the soft constraint optimisation.

It can clearly be seen from Figure 4.17 that the notion of diminishing returns is prevalent here. In practice, the availability of time must be balanced with solution quality in determining a reasonable timescale to allow the program to run for. In this example, 92% of the (overall) reduction in cost function is achieved within the first two hours. Is the remaining 8% worth the further ten hour wait?

It is also obvious from Figure 4.17 that, in general, the cost function decreases as time elapses. However, this is not strictly decreasing since simulated annealing does not prohibit the acceptance of swaps that lead to a worse objective function, i.e. lower soft constraint value. Since the reduction factor decreases the cooling parameter (Ch 4.5.3 11.c) at each iteration the acceptance of such swaps is less likely as time passes. This is evident from Figure 4.17 where there are many quite large spikes in the initial two hours but nothing noticeable afterwards. This is because the algorithm has found an area of the search space that it seeks to develop. The existence of this volatility in the initial stages is also a reason for allowing the program a sufficiently large amount of time to run – otherwise the program could terminate whilst on a ‘spike’.

This particular problem is not an issue for steepest descent (see 11.a of Ch 4.5.3) which only accepts swaps that improve the cost function or ‘steepest descent (+)’ (see 11.b) which accepts swaps that lead to an improved or equivalent cost function. These algorithms are compared, alongside simulated annealing, for the week used before (albeit with different random numbers – note initial cost function value),

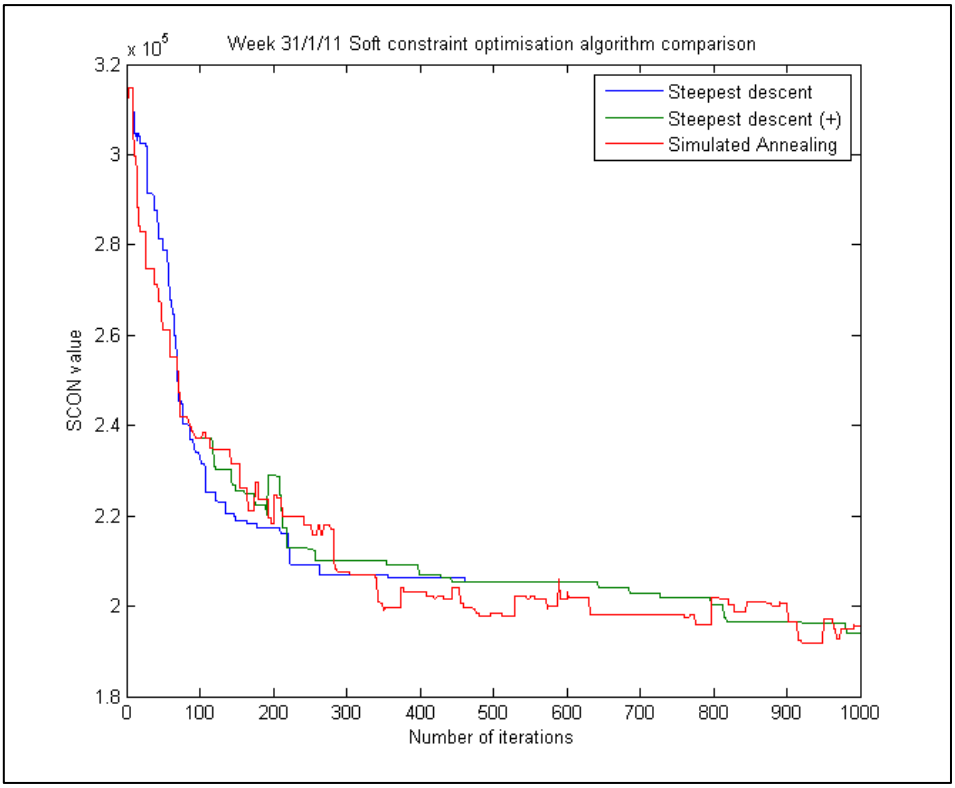


Figure 4.19 Comparison of soft constraint optimisation algorithms

The volatility of simulated annealing in this implementation is evident from Figure 4.19. If the program is stopped after 1,000 iterations (roughly half an hour) then 'steepest descent (+)' would have produced the best solution. This is despite simulated annealing producing a better solution around the 920th iteration mark. Note that the steepest descent algorithm terminates at the 589th iteration when a swap cannot be made that improves the cost function. By accepting swaps that neither decrease nor increase the cost function the second algorithm continues to improve upon this solution by further exploring the search space. Ultimately, when left to run for twelve hours, simulated annealing outperforms its rivals by a considerable measure.

4.8 Conclusion

A number of classification criteria for categorizing the scheduling problem of the physiotherapy department at Rookwood NRC are presented at the outset to this chapter. The use of exact and approximate methods for solving such a problem is then evaluated. Based on the decision to use approximate methods, the three stages of construction and optimisation are detailed. A brief step-by-step guide to the scheduling program is then presented. A number of important points and issues are forthwith discussed.

First the method of solution is considered. A three-stage local search based approximate approach is used to solve the hierarchical multi-objective combinatorial optimisation problem. In stage one an adaptive stochastic greedy constructive heuristic is used. In stage two specialised local search heuristics are used with a steepest descent based selection criteria. In stage three a generic local search heuristic is used alongside simulated annealing and elements of tabu Search.

However, this method is problematic since it is only the first-order neighbouring solutions that are checked. It is possible for a vastly improved solution to lie in the neighbourhood of a worse-performing neighbour to the original. Should steepest descent be used exclusively then this would not be considered. Simulated annealing is a method that seeks to combat this problem. However, since this problem is highly constrained (with respect to *MCON* and *HCON* violations) then its ability to do so is limited.

Other single-point metaheuristics (see Ch 4.3.2) that could conveniently be included to combat this problem are ILS and GRASP. ILS works by modifying a solution to provide a different one that cannot be found in the neighbourhood of the original. GRASP performs multiple runs simultaneously to construct different solutions. Population based metaheuristics (such as GAs) are inappropriate since the problem has been formulated otherwise. Other more unorthodox methods have been studied and found to be interesting but ultimately inappropriate, such as Puchinger & Raidl, 2005. In this article the authors investigate various combinations of exact and approximate methods in the solution of combinatorial optimisation problems. Results are favourable.

A dynamic cost function could also be used to improve the quality of the final solution. This would involve real-time adjustments to weight values depending on the progress of the heuristic. For example, if moves and swaps are being persistently rejected because they incur unacceptable increases in a particular soft constraint then the weight attached to this constraint could be reduced over time until fluidity is regained. Also, if *MCON* violations were hampering the acceptance of neighbouring solutions then these could be transferred to continuous value soft constraints whose initial weights are very high but are reduced over time.

The rationale for the construction of stage two *MCON* optimisation algorithms may be questionable. This is because the generic algorithm of stage three has the potential to achieve similar results – albeit in a longer time. However, since the scheduling program must respond

efficiently and effectively to, what are sometimes, very limited time constraints, the effort expended in the construction of stage two algorithms is justifiable.

A further consideration is that of stochastic network design – the ability of a schedule to respond to shocks. A shock is an unexpected change in the environment that, to some extent, prohibits the use of the original schedule. Schedules based on this concept are designed to be *robust* and *flexible* in the event of such a shock occurring. Essentially, these terms relate to the ability of a schedule to adapt to the new environment. At Rookwood NRC a shock represents either excess demand for treatment sessions or excess supply of treatment slots. The former occurs when a patient unexpectedly requires more treatment or when an employee becomes unavailable. The latter occurs when a patient becomes unavailable, perhaps due to illness. Excess supply is more favourable as it is easier to satisfy (by finding a suitable unmet need or using slot for admin).

The automated schedule underwent trials (debugging and fine-tuning weights[23]) from the summer of 2010 through to the end of the year. It was finally adopted by the physiotherapy department in January 2011. Up to the time of writing it has been used consistently on a weekly basis and has enjoyed praise from its users. Since its introduction it has required negligible maintenance and has not crashed once. This concludes the chapters that concern the specification and development of an automated scheduling program. The next chapter marks the move from scheduling to queuing theory by discussing methods for fitting statistical distributions to empirical data.

[23] See Appendix 4.3 for approximate values of weights

Chapter 5: Distribution Fitting and Parameter Estimation

5.1 Introduction

This chapter is concerned with the fitting of a probability distribution to patient length of stay (service time) at the Neurological Rehabilitation Centre (NRC) at Rookwood hospital. Patient length of stay (LOS) is defined as the total number of days spent in Rookwood NRC for a particular episode of care. This may also be referred to as the *sojourn time*. The process of fitting a distribution involves the selection of a probability distribution and the subsequent estimation of associated parameters. It is essential to fit a parametric model to the data because of the ability to utilise the many results often associated with well known distributions (Laplace transforms, moment generating functions etc). If the data being modelled represents service or inter-arrival times then useful performance measures (such as mean length of queue, mean wait in queue etc) may be derived depending on the distribution used. However, if a good fit between the raw data and the approximating distribution is not attained then any associated results may be unreliable. A good fit can be achieved by choosing a suitable distribution and selecting optimal parameter estimates.

Selecting a suitable distribution requires not only an idea of the probability density function shape but also an insight into the potential advantages and drawbacks of each one. Advantages and drawbacks may relate to the mathematical complexity of the distribution in analysing the queue. The optimality of parameter estimates is dependent on the methodology that is used as well as the competency in deriving the results through the chosen process. Methods of parameter estimation include least squares, method of moments and maximum

likelihood. Increasing the flexibility may be possible for some but not all distributions and relates to the manipulation of the design of the distribution. This may be seen in the case of the shifted exponential distribution by the introduction of a new parameter or by the addition of a 'phase' in a phase-type distribution (Ch 1.5).

The statistical distributions considered are split into two groups based on the underlying foundation of the distribution. Firstly, the following non phase-type distributions are considered:

1. Log-normal
2. Gamma
3. Weibull
4. Log-logistic

Secondly, the following phase-type distributions are considered:

1. Exponential
2. Hypo-exponential (*aka* generalised Erlang)
3. Hyper-exponential (*aka* sum of exponentials)
4. Coxian phase-type
5. Acyclic phase-type
6. General phase-type

Such distributions have been selected because of their potential ability to model LOS for patients at Rookwood NRC. Good fitting distributions are likely to be those that represent the behaviour of a typical patient's LOS, i.e. positively skewed with a long-sided tail (Marazzi et al, 1998). The log-normal, gamma and Weibull have been used historically for such reasons (Hanson, 1973, Lee & Hahn, 1970, Steadman et al, 2001, Harper, 2002, Marazzi et al, 1998). The log-logistic distribution also exhibits such properties and can be similar to the log-normal distribution (Dey & Kundu, 2010) except for having a heavier tail (Ghalebsaz-Jeddi et al, 2009). Faddy et al, 2009 compare the gamma and log-normal distributions with a phase-type distribution, concluding that the phase-type gives the best fit due to its versatility.

The application of phase-type distributions to patient LOS is, however, not new and has been widely covered in the research community. McClean & Millard, 1993 and Griffiths et al, 2006 fit the hyper-exponential; Faddy, 1995 fits the generalised Erlang; and Xie et al, 2005 fit the Coxian to name but a few. A significant advantage of phase-type distributions is their mathematical tractability in deriving performance measures in stochastic models. This is because they are composed of a system of inter-related Poisson processes that exhibit the memoryless property (Stewart, 2009). They are also highly flexible and can *'approximate arbitrarily closely any distribution with support on the positive reals'* (Bladt, 2005). An advantage pertinent to the modelling of healthcare facilities is the ability to view the phases as having physical meaning. This is of particular interest when describing patient transition through a number of facilities (A+E, ICU, …) or through a number of stages of care within a facility (short-term, medium-term,…). Kolker, 2008 acknowledges the limitation of a non-phase-type distribution when modelling an emergency department consisting of many steps from registration to discharge.

A method for parameter estimation is now required so that the distributions can be tuned to fit the data as best as their design permits. There are many ways of achieving this. The least squares approach minimises the sum of squares of the residuals of the observed values from the approximating distribution. There are two categories of the least squares methodology – linear and non-linear. Whereas the former possesses a closed form solution, the latter does not and may only be solved through a process of iterative refinement. The method of moments is an alternative approach that matches sample moments with unobservable population moments. The sample moments are obtained, in absolute value, from the data whereas the population moments are determined from the theoretical distribution. However, the results from this method are often considered inferior to those of the method of maximum likelihood. This approach aims to determine the parameter values that maximise the likelihood of the sample data. It has been shown to include many asymptotic properties that are beneficial when using a sufficiently large amount of data. All three methods are described in the next subchapter.

Upon the fitting of these distributions to sample data a criterion must be developed to select the most suitable. There are many goodness of fit tests that can be applied to determine whether a distribution approximates data sufficiently well. These include numerical assessments such as Pearson's chi-square, Kolmogorov-Smirnov, Cramer-von-Mises and Anderson-Darling tests in addition to graphical assessments such as probability plotting

(D'Agostino & Stephens, 1986). However, if the aim is to determine the most suitable distribution then it is necessary to compare the fits of the distributions. A distinction is made between the comparison of nested distributions, i.e. distributions from the same family, and non-nested distributions. The former permits the use of Neyman-Pearson hypothesis testing (Atkinson, 1970). This gives rise to the likelihood-ratio test which can be used to test whether the distribution with more parameters fits significantly better than its alternative. However, when the distributions are non-nested then this test cannot be used. Faddy et al, 2009 compare the log-normal, gamma and phase-type distributions by a comparison of the calculated Bayesian Information Criterion values and generalised Pearson statistics of each distribution. Marazzi et al, 1998 compare the log-normal, gamma and Weibull by robust Cox tests (Cox, 1962, Victoria-Feser, 1997) in addition to a self-developed assessment whilst Dey & Kundu, 2010 compare the log-logistic and log-normal distributions by log-likelihood values, K-S distance and chi-square values. Such findings further extend the motive for the separation of distributions into phase and non phase-type within this chapter.

This chapter commences by introducing three common parameter estimation methods. Ch 5.3 then addresses some numerical analytic approaches that are often required by the parameter estimation methods of Ch 5.2. Next, parameter estimation is considered for non phase-type distributions. And finally phase-type distributions are considered in Ch 5.5.

5.2 Parameter Estimation Methods

Three methods of parameter estimation are introduced and described. These methods have been used extensively within the literature for distribution fitting (Marshall & Zenga, 2010).

5.2.1 Least squares

The objective of the least squares method is to minimise the sum of squares of the residuals of the observed values from the model values. The residuals may relate to the perpendicular offsets of the data from the curve if there are uncertainties in both the *x* and *y* measurements of the data. However, most often, it is the vertical offsets that are chosen, providing a fitting function that estimates values of *y* given *x*, with the implicit assumption that there is negligible error in the *x* values.

The above objective may be equivalently stated as minimising

$$S = \sum_{i=1}^{n} r_i^2 \qquad (5.2.1.1)$$

where $r_i = y_i - f(x_i;\beta)$ is the *i*-th residual and β is a vector of *m* parameters.

Hence, the optimal parameter estimate is given by the value of β that minimises (5.2.1.1), i.e.

$$\hat{\beta}_{LS} = \arg\min \|r_i\|^2 \qquad (5.2.1.2)$$

This may be found by partial differentiation of (5.2.1.1) with respect to the *m* parameters, thus finding the gradient equations, and setting equal to zero,

$$\frac{\partial S}{\partial \beta_j} = 2\sum_{i=1}^{n} r_i \frac{\partial r_i}{\partial \beta_j} = 0 \qquad j = 1, 2, ..., m \qquad (5.2.1.3)$$

Since $r_i = y_i - f(x_i;\beta)$ it is trivial to show that (5.2.1.3) reduces to

$$\frac{\partial S}{\partial \beta_j} = 2\sum_{i=1}^{n} r_i \frac{\partial f(x_i;\beta)}{\partial \beta_j} = 0 \qquad j = 1, 2, ..., m \qquad (5.2.1.4)$$

For the case where the approximating function, $f(x_i;\beta)$, contains only a linear combination of the parameters, this is called a linear least squares problem. In such cases a closed form solution is possible and so calculation is typically straightforward. However, when the approximating function is non-linear then a closed form solution does not exist and so calculation may require time-consuming iterative refinement techniques.

It is possible to 'linearise' such functions by means of a linear transformation thus enabling the calculation of parameter estimates without the need for numerical analysis. However, this process can produce unreliable results because the error values on the *y* measurements are also transformed. Hence, since the variances of the residuals are unequal it cannot be inferred from the Gauss-Markov theorem that the parameter estimate(s) are unbiased (as this requires homoscedasticity). Fraile & Garcia-Ortega, 2005 identify this issue with application to the exponential distribution, recognising that the linear transformation is in essence placing varying weights on the error terms that results in heteroscedasticity.

5.2.2 Method of moments

The *j*-th sample moment may be defined as

$$m_j = \frac{1}{n}\sum_{i=1}^{n} x_i^j \qquad (5.2.2.1)$$

The *j*-th population moment may be defined as

$$M_j = E\left[X^j\right] \tag{5.2.2.2}$$

This expectation can be calculated as follows:

$$E\left[X^j\right] = \int_{-\infty}^{\infty} x^j f(x)dx \tag{5.2.2.3}$$

Or alternatively through the moment generating function, $M_X(t)$, of the distribution

$$E\left[X^j\right] = M_X^{(j)}(0) = \frac{d^j M_X}{dt^j}(0) \tag{5.2.2.4}$$

Parameter estimates are then deduced from setting equal respective sample and population moments. Gross & Juttijudata, 1997 conclude that equating the first two moments in a standard two parameter distribution is not usually sufficient. The authors recommend that it may be necessary to capture at least five moments in such cases (albeit with lower-order moments receiving more importance). There are many examples in the literature where the method of moments has been applied to phase-type distributions (see Johnson & Taaffe, 1989, Telek & Heindl 2002, Bobbio et al, 2005).

5.2.3 Maximum likelihood

Pawitan, 2001 states that:

'Assuming a statistical model parameterised by a fixed and unknown θ, the likelihood $L(\theta)$ is the probability of the observed data x considered as a function of θ'

The objective is to find $\hat{\theta}$ that maximises the likelihood, i.e.

$$\hat{\theta} = \arg\max L(\theta \mid x_1, ..., x_n) \tag{5.2.3.1}$$

where

$$L(\theta \mid x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{5.2.3.2}$$

It is often easier to work in terms of the log-likelihood $l(\theta \mid x_1, ..., x_n) = \ln L(\theta \mid x_1, ..., x_n)$ in which case the aim is to find $\hat{\theta}$ that maximises this function, i.e.

$$\hat{\theta} = \arg\max l\left(\theta \mid x_1, ..., x_n\right) \tag{5.2.3.3}$$

The validity of this approach is ensured by the logarithm being a strictly increasing function.

Maximum likelihood estimation includes some attractive asymptotic properties. Consistency ensures that as the sample size tends to infinity the estimator tends to its true value (Severini, 2000). The estimators are also the most efficient (Cousineau et al, 2004), meaning they attain the Cramer-Rao lower bound (for parameter variance) as the sample size tends to infinity. An estimator is said to attain asymptotic normality if the distribution of the estimator tends to the Gaussian distribution as the sample size tends to infinity. Ferguson, 1996 reports the maximum likelihood estimator to have this property provided the second derivative of the *pdf* exists and is continuous.

It is common to require the use of iterative techniques in order to approximate estimator values and these are therefore discussed in the following subchapter.

5.3 Numerical Analysis Methods

The objective of the optimisation methods presented in this subchapter is to minimise the value of a function by manipulation of its parameters.

5.3.1 Newton-Raphson method

For a differentiable function $f(x)$ the following formula may be used in repetition until an approximation for x is determined that minimises the value of $f(x)$:

$$x_{n+1} = x_n - \frac{f\left(x_n\right)}{f'\left(x_n\right)} \tag{5.3.1.1}$$

In the case of parameter estimation when an estimate is not readily available analytically the above process may be used (where the parameter is represented by x)

On application to maximum likelihood, with $\underline{\theta}$ as a vector of unknown parameters and $g(\underline{\theta})$ as a vector of first derivatives of the log-likelihood function respective of the parameters, the following may be obtained as a result of the Taylor expansion of $g\left(\underline{\theta}_0\right)$:

$$\underline{\theta}_{n+1} = \underline{\theta}_n - \frac{g\left(\underline{\theta}_n\right)}{H\left(\underline{\theta}_n\right)} \tag{5.3.1.2}$$

where H is defined as the Hessian matrix of second derivatives of the log-likelihood function (Garthwaite et al, 2002).

5.3.2 Fisher's method of scoring

This method can be seen as a modification of the Newton-Raphson method (Osborne, 1992) in which the Hessian in (5.3.1.2) is replaced by its expectation. Thus

$$\underline{\theta}_{n+1} = \underline{\theta}_n - \frac{g(\underline{\theta}_n)}{E\left[H(\underline{\theta}_n)\right]} = \underline{\theta}_n + \frac{g(\underline{\theta}_n)}{I\left[\theta_n\right]} \tag{5.3.2.1}$$

where $I\left[\theta_n\right]$ is the Fisher's information matrix.

One advantage is the absence of need to compute second derivatives (Garthwaite et al, 2002) because of the relation

$$E\left[\frac{\partial^2 l(\underline{\theta};x)}{\partial\theta_i\partial\theta_j}\right] = -E\left[\left(\frac{\partial l}{\partial\theta_i}\right)\left(\frac{\partial l}{\partial\theta_j}\right)\right] \tag{5.3.2.2}$$

5.3.3 Nelder-Mead 'simplex' algorithm

The Nelder-Mead algorithm (1965) is an optimisation method that belongs to the general class of *direct search methods*, that is, it does not use gradient information within the search (Wright et al, 1998). Instead, function values are taken from a set of sample points and such information is used to continue sampling. The Nelder-Mead algorithm is part of a large subclass of direct search methods that operate a simplex of approximations to an optimal point (Kelley, 1999). Moreover, this *n*-dimensional simplex contains *n+1* vertices that are ordered according to decreasing function value, i.e. $f(x_1) \le f(x_2) \le .... \le f(x_{n+1})$. The algorithm works by attempting to replace the worst performing vertex by one that gives a lower function value. The terminating criterion is met when there is sufficiently small difference between the best and least performing vertices, i.e. $f(x_{n+1}) - f(x_1) < \varepsilon$.

Note that depending on the complexity of the approximating function these methods can be extremely sensitive to the initial value(s) used at the first iteration. An unsuitable choice of initial value(s) may result in infeasibility or convergence to a solution that locally optimal but not so globally.

5.4 Non Phase-Type Distributions

The distributions considered within this subchapter are all continuous univariate probability distributions that are supported on a semi-infinite interval $[0,\infty)$ for a non-negative random variable. Parameter estimation for these distributions is based on the method of moments and maximum likelihood (described in subchapters 5.2.2 and 5.2.3 respectively).

5.4.1 Log-normal distribution

The log-normal distribution is a probability distribution in which the logarithm of the random variable is normally distributed. This is expressed as follows:

$$F(x;\mu,\sigma) = N(\ln x) \tag{5.4.1.1}$$

By differentiation (Aitchison & Brown, 1957):

$$f(x;\mu,\sigma) = \frac{1}{\sigma x\sqrt{2\pi}} \exp\left(\frac{-(\ln x-\mu)^2}{2\sigma^2}\right) \tag{5.4.1.2}$$

This distribution has previously been fitted to service times. Brown et al, 2005 fit a log-normal distribution to service time data from a telephone call centre, stating a very close but not exact fit. In this paper the authors test the goodness of fit visually via a histogram of the logarithm of service times (checking for Gaussian shape) and a log-normal quantile-quantile plot of service times. A Kolmogorov-Smirnov test fails owing to the authors' reluctance to infer an exact fit. Lawrence, 1984 uses a log-normal distribution to approximate the length of strikes in the UK in the 1960s using a refined moment approach to estimate the parameters. Applications to patient LOS have also been made by Hanson, 1973 and Lee & Hahn, 1970 with the authors of both papers citing satisfactory fits. A major drawback to applying the log-normal distribution to service time is that it does not possess a closed form Laplace transform (Wagner & Geyer, 1995). This is a problem as it substantially restricts the ability to analyse queuing systems (Fischer et al, 2001). To generalise, this is a problem for all heavy-tailed (typically sub-exponential) distributions including the Weibull and Pareto distributions. A solution to this is to approximate such distributions with a phase-type distribution (Greiner et al, 1999) or to approximate the Laplace transform of the distribution (as described in Shortle et al, 2003 by the transform approximation method).

5.4.1.1 Maximum likelihood

By application of the natural logarithm to the likelihood of (5.4.1.2) the log-likelihood function is obtained;

$$l(x_i;\mu,\sigma) = \ln\prod_{i=1}^{n}\frac{1}{x_i} - \frac{1}{2}n\ln 2\pi\sigma^2 - \sum_{i=1}^{n}\frac{(\ln x_i - \mu)^2}{2\sigma^2} \qquad (5.4.1.1.1)$$

From partial derivatives of (5.4.1.1.1) with respect to the parameters the following estimators are deduced:

$$\hat{\mu} = \frac{\sum_{i=1}^{n}\ln x_i}{n} \qquad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(\ln x_i - \hat{\mu})^2}{n} \qquad (5.4.1.1.2)$$

5.4.1.2 Method of moments

The first and second population moments are

$$\exp(\mu + \frac{1}{2}\sigma^2) \qquad \exp(2\mu + 2\sigma^2) \qquad (5.4.1.2.1)$$

By equating these with the first and second sample moments (see Ch 5.2.2);

$$\hat{\mu} = 2\ln\bar{x} - \frac{1}{2}\ln\frac{\sum_{i=1}^{n}x_i^2}{n} \qquad \hat{\sigma}^2 = \ln\frac{\sum_{i=1}^{n}x_i^2}{n} - 2\ln\bar{x} \qquad (5.4.1.2.2)$$

5.4.2 Gamma distribution

The gamma distribution is a continuous probability distribution with shape parameter $k > 0$ and scale parameter $\theta > 0$ whose probability density function is defined as

$$f(x;k,\theta) = \frac{x^{k-1}e^{-x/\theta}}{\theta^k\Gamma(k)} \qquad (5.4.2.1)$$

This distribution is a generalisation of the Erlang distribution where *k* is restricted to positive integer values. An exponential distribution with rate parameter λ is given if *k=1* and $\theta = 1/\lambda$. The tail of the gamma distribution is shorter than that of the log-normal distribution looked at previously (Marazzi et al, 1998). The gamma distribution is used by Faddy et al, 2009 as part of a goodness of fit comparison of continuous distributions fitted to LOS data. It is compared with the log-normal and a phase-type distribution and was found to perform worse than its competitors when fitted to the data. However, a crucial advantage of the gamma distribution is that a closed form of the Laplace transform exists.

5.4.2.1 Maximum likelihood

The log-likelihood function of (5.4.2.1) may be simplified to

$$l = (k-1)\sum_{i=1}^{n} \ln x_i - nk \ln \theta - n \ln \Gamma(k) - \frac{\sum_{i=1}^{n} x_i}{\theta} \tag{5.4.2.1.1}$$

The partial derivative of (5.4.2.1.1) with respect to the parameter θ set equal to zero yields

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{nk} \tag{5.4.2.1.2}$$

Whilst the partial derivative of (5.4.2.1.1) with respect to *k* can be written

$$\frac{\partial l}{\partial k} = \ln k - \Psi(k) + \frac{1}{n}\sum_{i=1}^{n} \ln x_i - \ln \bar{x} \tag{5.4.2.1.3}$$

where Ψ is the Digamma function defined $\Psi(k) = \Gamma'(k)/\Gamma(k)$.

The Newton-Raphson method (an iterative refinement technique) is used to solve (5.4.2.1.3). The second partial derivative of (5.4.2.1.1) with respect to *k* is

$$\frac{\partial^2 l}{\partial k^2} = \frac{1}{k} - \Psi'(k) \tag{5.4.2.1.4}$$

Therefore, using (5.3.1.1) the iterative process is defined

$$k_{i+1} = k_i - \frac{\ln k_i - \Psi(k_i) + \frac{1}{n}\sum_{i=1}^{n} \ln x_i - \ln \bar{x}}{\frac{1}{k_i} - \Psi'(k_i)} \tag{5.4.2.1.5}$$

An initial value of *k* may be taken from a method of moments approximation or calculated as follows.

By setting (5.4.2.1.3) equal to zero and approximating the Digamma function as

$$\Psi(k) \approx \ln k - \frac{1}{2k} - \frac{1}{k(12k+2)} \tag{5.4.2.1.6}$$

the following is obtained:

$$\frac{1}{2k}+\frac{1}{k(12k+2)}\approx \ln \bar{x}-\frac{1}{n}\sum_{i=1}^{n}\ln x_i \qquad (5.4.2.1.7)$$

When the RHS of (5.4.2.1.7) is set to *p* and the quadratic equation is used a formula for the initial value of *k* is deduced;

$$k_0=\frac{3-p\pm\sqrt{p^2+18p+9}}{12p} \qquad p=\ln \bar{x}-\frac{1}{n}\sum_{i=1}^{n}\ln x_i \qquad (5.4.2.1.8)$$

Minka, 2002 provides an alternative approach that (according to the author) can converge in about four iterations. The algorithm is obtained via a generalised Newton method that the author explains in Minka, 2000. The estimation of the shape parameter is also considered in Miller & Bhat, 1997 in which the authors apply the distribution to model inter-arrival times of a renewal process.

5.4.2.2 Method of moments

The first and second population moments (as derived from the MGF: $(1-\theta t)^k$) are

$$\theta k \qquad \theta^2 k(k+1) \qquad (5.4.2.2.1)$$

On equating these with sample moments, the following estimators are obtained:

$$\hat{k}=\frac{\bar{x}^2}{\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2}=\frac{\bar{x}^2}{s^2}=\frac{1}{CV_s^2}$$
$$\hat{\theta}=\frac{\bar{x}}{\hat{k}}=\frac{1}{\bar{x}}\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2\right)=\frac{s^2}{\bar{x}} \qquad (5.4.2.2.2)$$

where *s* denotes the sample standard deviation and CV_s denotes the sample coefficient of variation.

5.4.3 Weibull distribution

The Weibull distribution is a two-parameter continuous probability function with shape parameter $k>0$ and scale parameter $\lambda>0$ and probability density function

$$f(x;\lambda,k)=\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-\left(\frac{x}{\lambda}\right)^k} \qquad (5.4.3.1)$$

The Weibull distribution has been used extensively in reliability modelling. In its two-parameter form (as shown in 5.4.3.1) it has been used to model failure times of machines etc (see Prabhakar et al, 2004 and references therein). It has also had application to patient LOS; such as in Perez-Hoyos et al, 2000, Steadman et al, 2001 and Harper, 2002. On application to queuing theory the Weibull distribution presents the same problem as described for the log-normal distribution in relation to the nature of its Laplace transform (Fischer et al, 2001).

5.4.3.1 Maximum likelihood

The log-likelihood of the pdf given above may be written

$$l = n\ln\frac{k}{\lambda} + (k-1)\sum_{i=1}^{n}\ln\frac{x}{\lambda} - \sum_{i=1}^{n}\left(\frac{x}{\lambda}\right)^{k} \qquad (5.4.3.1.1)$$

The partial derivative of this with respect to λ set equal to zero yields the following estimator for the scale parameter:

$$\hat{\lambda} = \left\{\frac{1}{n}\sum_{i=1}^{n}x_i^k\right\}^{\frac{1}{k}} \qquad (5.4.3.1.2)$$

The partial derivative of (5.4.3.1.1) with respect to *k* may be simplified to

$$\frac{\partial l}{\partial k} = \frac{n}{k} + \sum_{i=1}^{n}\ln x_i - \frac{n}{\sum_{i=1}^{n}x_i^k}\sum_{i=1}^{n}x_i^k\ln x_i \qquad (5.4.3.1.3)$$

by replacing the λ's with the value of $\hat{\lambda}$ from (5.4.3.1.2).

When (5.4.3.1.3) is set equal to zero, no closed form solution for *k* exists so iterative techniques may be applied – the Newton-Raphson method being the obvious choice. This approach has been suggested by Al-Fawzan, 2000, Gove, 2003 and Cohen, 1965. A trial and error approach involving linear interpolation is also put forward by Cohen, 1965. Balakrishnan & Kateri, 2008 introduce a graphical method to solve (5.4.3.1.3).

Further differentiation of (5.4.3.1.3) gives the second derivative necessary for the Newton-Raphson method. Gove, 2003 suggests using the Method of Moments estimators as initial values. On finding the value of $\hat{k}$ the value of $\hat{\lambda}$ may be found through substitution into (5.4.3.1.2).

5.4.3.2 Method of moments

The first and second population moments of the Weibull distribution are

$$\lambda\Gamma\left(1+\frac{1}{k}\right) \qquad \lambda^2\Gamma\left(1+\frac{2}{k}\right) \tag{5.4.3.2.1}$$

On equating the first sample and population moments;

$$\hat{\lambda} = \frac{\bar{x}}{\Gamma\left(1+1/k\right)} \tag{5.4.3.2.2}$$

By substitution of (5.4.3.2.2) for λ in the second moment equation;

$$n\Gamma\left(1+1/k\right)\sum_{i=1}^{n} x_i^2 - \Gamma\left(1+2/k\right)\left(\sum_{i=1}^{n} x_i\right)^2 = 0 \tag{5.4.3.2.3}$$

the solution of which may be found through an iterative process.

A related approach employed by Al-Fawzan, 2000 and Gove, 2003 is to equate the sample and population values of the coefficient of variation in order to find an estimator for *k*. This is in equivalence to the traditional approach of equating individual moments because the sample coefficient of variation is a function of the first two sample moments, i.e. $CV_s = \sqrt{\frac{m_2}{m_1^2} - 1}$.

Therefore an estimator for *k* is obtained when the following equation is solved:

$$CV_s = CV_p = \frac{\sigma}{\mu} = \frac{\sqrt{E\left[X^2\right] - E\left[X\right]^2}}{E\left[X\right]} = \frac{\sqrt{\Gamma\left(1+2/k\right) - \left\{\Gamma\left(1+1/k\right)\right\}^2}}{\Gamma\left(1+1/k\right)} \tag{5.4.3.2.4}$$

5.4.4 Log-logistic distribution

The log-logistic distribution is a probability distribution that is derived by the application of the logarithmic transformation to the logistic distribution. This transformation is in equivalence to that required to obtain the log-normal distribution from the normal distribution (see Ch 5.4.1). Therefore the probability density function is as follows:

$$f(x;\beta,\alpha)=\frac{\frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}}{\left(1+\left(\frac{x}{\alpha}\right)^{\beta}\right)^{2}} \qquad \alpha,\beta>0 \qquad (5.4.4.1)$$

The log-logistic distribution is typically applied to the fields of hydrology, reliability and economics (see Ashkar & Mahdi, 2006, Srivastava & Shukla, 2008, Wang, 2008 and references therein). Attempts have also been made in fitting the distribution to service time. Kolker, 2008 used a program called stat:fit to fit various distributions to inpatient LOS in an emergency department concluding that the log-logistic performs best. It is possible to estimate parameters for this distribution by a number of methods including maximum likelihood, method of moments, and (ordinary and generalised) least squares since the pdf, cdf and inverse cdf may all be specified explicitly (Ahmad et al, 1988). However, the literature details many other methods that are employed. Ashkar & Mahdi, 2006 propose a generalised moments approach whilst Shoukri et al, 1988 take advantage of the existence of the closed form inverse cdf in equating probability weighted moments. A drawback of this distribution, shared by both the Weibull and log-normal distributions, is the absence of a closed form Laplace transform.

5.4.4.1 Maximum likelihood

The log-likelihood of the pdf given in (5.4.4.1) is

$$l=\ln L=n\ln\frac{\beta}{\alpha}+(\beta-1)\sum_{i=1}^{n}\ln\frac{x_i}{\alpha}-2\sum_{i=1}^{n}\ln\left(1+\left(\frac{x_i}{\alpha}\right)^{\beta}\right) \qquad (5.4.4.1.1)$$

By partial differentiation with respect to the parameters and equating to zero;

$$\begin{aligned}&\sum_{i=1}^{n}\frac{(x_i/\alpha)^{\beta}}{1+(x_i/\alpha)^{\beta}}-\frac{n}{2}=0\\&\sum_{i=1}^{n}\frac{\ln(x_i/\alpha)(x_i/\alpha)^{\beta}}{1+(x_i/\alpha)^{\beta}}-\frac{1}{2}\sum_{i=1}^{n}\ln\frac{x_i}{\alpha}-\frac{n}{2\beta}=0\end{aligned} \qquad (5.4.4.1.2)$$

These are solved by a successive iterative method in which initial values of the parameters, α_0,β_0, are inserted into the first equation with β fixed and solved to give a new value of α. Then, α_1,β_0 are inserted into the second equation with α fixed and solved to give a new

value of β. These two steps represent one iteration and the process is continued until sufficient convergence to the maximum likelihood estimators $\hat{\alpha}, \hat{\beta}$ is attained.

The equations may be solved at each step by the Newton-Raphson method, i.e.

$$\alpha_{m+1} = \alpha_m + \frac{\sum_{i=1}^{n} \frac{(x_i/\alpha)^\beta}{1+(x_i/\alpha)^\beta} - \frac{n}{2}}{\sum_{i=1}^{n} \frac{\beta(x_i/\alpha)^\beta}{\alpha\left(1+(x_i/\alpha)^\beta\right)^2}}$$

$$\beta_{m+1} = \beta_m - \frac{\sum_{i=1}^{n} \frac{\ln(x_i/\alpha)(x_i/\alpha)^\beta}{1+(x_i/\alpha)^\beta} - \frac{1}{2}\sum_{i=1}^{n} \ln\frac{x_i}{\alpha} - \frac{n}{2\beta}}{\sum_{i=1}^{n} \frac{\{\ln(x_i/\alpha)\}^2 (x_i/\alpha)^\beta}{1+(x_i/\alpha)^\beta} + \frac{n}{2\beta^2}} \qquad (5.4.4.1.3)$$

5.4.4.2 Method of moments

The first and second population moments of the log-logistic distribution are

$$\frac{\alpha b}{\sin b} \qquad \frac{2\alpha^2 b}{\sin 2b} \qquad (5.4.4.2.1)$$

where $b = \pi/\beta$. Equating the first population and sample moments gives rise to the estimator

$$\hat{\alpha} = \frac{\bar{x}\sin b}{b} \qquad (5.4.4.2.2)$$

Equating the second population and sample moments and substituting in (5.4.4.2.2) yields

$$2\bar{x}^2 \sin^2 b - \frac{\sum x_i^2}{n} b \sin 2b = 0 \qquad (5.4.4.2.3)$$

which can be solved by the Newton-Raphson method.

However, a more simple (Shoukri et al, 1988) and effective (Singh & Guo, 1995) approach is to use the class of probability weighted moments to derive estimators for which

$$\hat{\alpha} = \frac{\hat{W}_0^2 \sin b}{\pi\left(2\hat{W}_1 - \hat{W}_0\right)} \qquad \hat{\beta} = \frac{\hat{W}_0}{2\hat{W}_1 - \hat{W}_0} \qquad (5.4.4.2.4)$$

where $W_r = \frac{1}{n}\sum_{i=1}^{n} x_i \prod_{j=1}^{r} \frac{i-j}{n-j}$.

5.5 Phase-Type Distributions

A phase-type distribution models the distribution of transient time in a finite state continuous-time Markov chain (CTMC). The distributions considered within this subchapter are all continuous univariate probability distributions that are supported on a semi-infinite interval $[0,\infty)$ for a non-negative random variable, *X*. Their minimal representation is (α, T) where α is the *initial probability vector* and T is the *PH-generator matrix*. The pdf can be expressed

$$f\left(x;\alpha,T\right) = \alpha e^{Tx} \cdot (-T1) \tag{5.5.1}$$

where 1 is an order *p* column vector of ones (see Ch 1.5 for an introduction to phase-type distributions). Parameter estimation for these distributions is based on the method of moments (Ch 5.2.2) and maximum likelihood (Ch 5.2.3).

5.5.1 Exponential distribution

The exponential distribution describes the time between events in a Poisson process (see Ch 1.5). An advantage of this distribution is the memoryless property. This enables the tractable derivation of performance measures in queuing theoretic models. For a distribution to be memoryless means that for a random variable *X* the conditional probability $P\left(X > s+t \mid X > s\right)$ is equal to the probability $P\left(X > t\right)$ for all $t, s \geq 0$. The exponential distribution is the simplest class of phase-type distributions and has a coefficient of variation equal to one. Its pdf is

$$f\left(x;\lambda\right) = \lambda e^{-\lambda x} \tag{5.5.1.2}$$

This distribution forms the building blocks of any phase-type distribution. Each phase within a phase-type distribution is exponentially distributed and it is the number and inter-relation of these that uniquely describe *T*, the PH-generator matrix.

5.5.1.1 Maximum likelihood

The log-likelihood of (5.5.1.2) is

$$l = \ln L = n \ln \lambda - \lambda \sum x_i \tag{5.5.1.1.1}$$

The partial differential of this set equal to zero yields

$$\hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{x}} \qquad (5.5.1.1.2)$$

5.5.1.2 Method of moments

Equating the first sample and population moments returns an estimator equal to (5.5.1.1.2).

5.5.2 Hyper-exponential distribution

The hyper-exponential distribution is a phase-type distribution whose coefficient of variation is greater than one (see Stewart, 2009 for proof). Represented as a CTMC the states may be ordered in a *parallel* arrangement as follows:

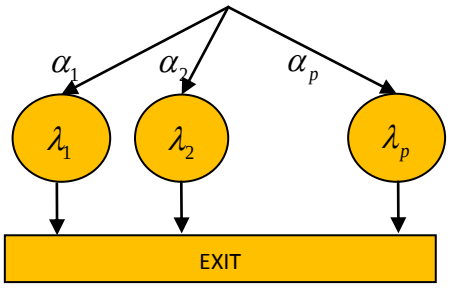


Figure 5.1 CTMC representation of a hyper-exponential distribution

Thus the initial probability vector and the PH-generator matrix (both order *p*) are given by

$$\alpha = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_p \end{pmatrix} \qquad T = \begin{bmatrix} -\lambda_1 & 0 & 0 & 0 \\ 0 & -\lambda_2 & 0 & \\ 0 & 0 & \ddots & \vdots \\ 0 & & \cdots & -\lambda_p \end{bmatrix} \qquad (5.5.2.1)$$

Because T is a diagonal matrix the matrix exponential e^{Tx} may be written

$$\begin{bmatrix} e^{-\lambda_1 x} & 0 & 0 & 0 \\ 0 & e^{-\lambda_2 x} & 0 & \\ 0 & 0 & \ddots & \vdots \\ 0 & & \cdots & e^{-\lambda_p x} \end{bmatrix} \qquad (5.5.2.2)$$

Therefore, by (5.5.1), the pdf of sojourn time is

$$f(x) = \sum_{i=1}^{p} \alpha_i \lambda_i e^{-\lambda_i x} \qquad (5.5.2.3)$$

The Laplace transform of (5.5.2.3) is found to be

$$L\{f(x)\} = \alpha_1 \frac{\lambda_1}{\lambda_1 + s} + \alpha_2 \frac{\lambda_2}{\lambda_2 + s} + ...\alpha_p \frac{\lambda_p}{\lambda_p + s} \qquad (5.5.2.4)$$

by using the Shift theorem to derive $L\{f_i(x)\} = \int_0^\infty \lambda_i e^{-\lambda_i x} e^{-sx} dx = \frac{\lambda_i}{\lambda_i + s}$ with $f_i(x) = \lambda_i e^{-\lambda_i x}$.

The number of parameters that needs to be estimated is $(2p-1)$ since $\sum \alpha_i = 1$ (thus one of the α_i's is determined by the values of the others).

5.5.2.1 Two-term hyper-exponential distribution

A hyper-exponential distribution with two phases is considered. Such a distribution has been applied to LOS for geriatric patients in McClean & Millard, 1993 with the two phases being used to represent acute/rehabilitative care and long-stay. This requires the estimation of three parameters. However, by fixing α the number of moment equations required to solve is reduced to two (α is set to 0.5 in Whitt, 1984). An alternative method (Johnson & Taaffe, 1991) is to assume balanced means, i.e. $\alpha\lambda_1^{-1} = (1-\alpha)\lambda_2^{-1}$. This too requires the estimation of two, and not three, parameters.

It is possible to select the λ_i such that the mean service time of the distribution, $\sum \alpha_i / \lambda_i$, is equal to a specified value, say $1/\lambda$. This is of particular benefit when comparing the characteristics of more than one distribution whilst enforcing homogeneity of mean service time. It is possible to modify the pdf of the hyper-exponential distribution in order to incorporate this. This is done by taking the pdf as $f(x) = \sum p\alpha_i^2 \lambda e^{-p\alpha_i \lambda x}$.

5.5.2.1.1 Maximum likelihood

The log-likelihood of (5.5.2.3) with $p = 2$ and $\alpha = \alpha_1 = 1 - \alpha_2$ can be written

$$l = \sum_{i=1}^{n} \ln\left[\alpha\lambda_1 e^{-\lambda_1 x_i} + (1-\alpha)\lambda_2 e^{-\lambda_2 x_i}\right] \qquad (5.5.2.1.1.1)$$

which must be solved iteratively since explicit solutions do not exist. The Nelder-Mead algorithm is the chosen approach for two reasons. Firstly, it is a direct search method and so does not require calculation of derivatives (this can become time-consuming when many parameters are considered). Secondly, in early iterations the convergence to some (but not all) local optima is avoided. A program that employs the Nelder-Mead algorithm to deduce maximum likelihood estimators has been encoded in Matlab – see Ch 6.5.1 for details.

5.5.2.1.2 Method of moments

To determine the first and second population moments the relationship $L_s\{f(x)\} = E\left[e^{-sX}\right]$ is exploited by replacing *s* by *-t* to give the moment generating function (MGF) of the sojourn time, *X*. This gives $M_X(t) = \alpha\frac{\lambda_1}{\lambda_1 - t} + (1-\alpha)\frac{\lambda_2}{\lambda_2 - t}$. Differentiating this once and setting *t* equal to zero yields the first population moment and further differentiation yields the second and third. Equating with the sample moments m_1, m_2, m_3 gives

$$\begin{aligned} m_1 &= \frac{\alpha}{\lambda_1} + \frac{1-\alpha}{\lambda_2} \\ m_2 &= \frac{2\alpha}{\lambda_1^2} + \frac{2(1-\alpha)}{\lambda_2^2} \\ m_3 &= \frac{6\alpha}{\lambda_1^3} + \frac{6(1-\alpha)}{\lambda_2^3} \end{aligned} \qquad (5.5.2.1.2.1)$$

The first of these equations yield

$$\hat{\alpha} = \frac{m_1 - \lambda_2^{-1}}{\lambda_1^{-1} - \lambda_2^{-1}} \qquad (5.5.2.1.2.2)$$

Substitution of this into the second equation of (5.5.2.1.2.1) returns

$$\hat{\lambda}_1^{-1} = \frac{0.5m_2 - m_1\lambda_2^{-1}}{m_1 - \lambda_2^{-1}} \qquad (5.5.2.1.2.3)$$

On substituting (5.5.2.1.2.2) into the third equation of (5.5.2.1.2.1);

$$\left(m_1 - \lambda_2^{-1}\right)\left(\lambda_1^{-2} + \lambda_1^{-1}\lambda_2^{-1} + \lambda_2^{-2}\right) = \frac{1}{6}m_3 - \lambda_2^{-3} \qquad (5.5.2.1.2.4)$$

and by the subsequent substitution of (5.5.2.1.2.3) into (5.5.2.1.2.4);

$$\left(12m_1^2 - 6m_2\right)\lambda_2^{-2} + \left(2m_3 - 6m_1m_2\right)\lambda_2^{-1} + \left(3m_2^2 - 2m_1m_3\right) = 0 \qquad (5.5.2.1.2.5)$$

This may be solved by the quadratic equation, i.e.

$$\hat{\lambda}_2^{-1} = \frac{B \pm \sqrt{A}}{6D} \qquad (5.5.2.1.2.6)$$

where the auxiliary variables are defined as

$$\begin{aligned} A &= B^2 - 6CD \\ B &= 3m_1m_2 - m_3 \\ C &= 3m_2^2 - 2m_1m_3 \\ D &= 2m_1^2 - m_2 \end{aligned} \qquad (5.5.2.1.2.7)$$

By substitution of (5.5.2.1.2.6) into (5.5.2.1.2.3) it can be shown that

$$\hat{\lambda}_1^{-1} = \hat{\lambda}_2^{-1} \mp 2\frac{\sqrt{A}}{6D} = \frac{B \mp \sqrt{A}}{6D} \qquad (5.5.2.1.2.8)$$

depending on the value taken as λ_2^{-1} in (5.5.2.1.2.6).

Therefore the estimators for the three parameters are

$$\begin{aligned} \hat{\lambda}_{1,2}^{-1} &= \frac{B \pm \sqrt{A}}{6D} \\ \hat{\alpha} &= \frac{m_1 - \hat{\lambda}_2^{-1}}{\hat{\lambda}_1^{-1} - \hat{\lambda}_2^{-1}} \end{aligned} \qquad (5.5.2.1.2.9)$$

A similar method of estimation is used in Rider, 1961. A general check to ensure moment consistency has been developed in Karlin & Studden, 1966 for use with a non-negative random variable. The necessary and sufficient conditions for the first three sample moments are $m_1 \geq 0; m_2 - m_1^2 \geq 0; m_1m_3 - m_2^2 \geq 0$. On extending this to the hyper-exponential distribution these conditions remain necessary in determining moment consistency but are not sufficient (Whitt, 1982). The author then proposes that the values $m_1, m_2, m_3, \ldots$ are the moments of the hyper-exponential distribution if and only if $m_1, m_2/2!, m_3/3!, \ldots$ satisfy the general moment consistency check (proof omitted). Therefore, the necessary and sufficient conditions for the second-order hyper-exponential distribution are

$$m_1 \geq 0$$
$$m_2 m_1^{-2} - 2 \geq 0 \qquad (5.5.2.1.2.10)$$
$$2m_3 m_1 - 3m_2^2 \geq 0$$

If a constraint is not satisfied then the sample moment values may be adjusted (Telek & Henidl, 2002). If, for example, the first two constraints are satisfied and the third is not then the value of the third sample moment may be adjusted so that its value is set to the closest possible value necessary to obtain consistency.

5.5.3 Hypo-exponential distribution

The Hypo-exponential distribution is a phase-type distribution whose coefficient of variation is less than one. Whilst the hyper-exponential distribution has phases in parallel the hypo-exponential distribution has phases in series. The CTMC for the order *p* hypo-exponential distribution is given in Figure 5.2.

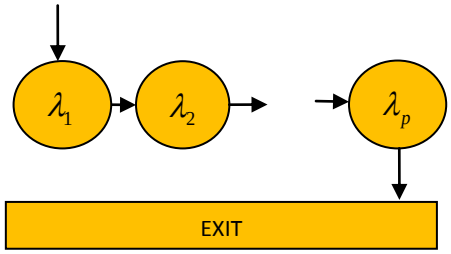


Figure 5.2 CTMC representation of a hypo-exponential distribution

Therefore the initial probability vector and the PH-generator matrix (both order *p*) are

$$\alpha = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix} \qquad T = \begin{bmatrix} -\lambda_1 & \lambda_1 & 0 & 0 \\ 0 & -\lambda_2 & \lambda_2 & \\ 0 & 0 & \ddots & \vdots \\ 0 & & \cdots & -\lambda_p \end{bmatrix} \qquad (5.5.3.1)$$

It can be shown that the squared coefficient of variation is $CV^2 = \dfrac{\sum \lambda_i^2}{\left(\sum \lambda_i\right)^2}$ whose upper bound is equal to one. The Cauchy-Schwarz inequality may be used in determining the value of the lower bound. By substituting $a_i = \lambda_i$ and $b_i = 1 \ \ (i = 1..p)$ in $\left(\sum a_i b_i\right)^2 \leq \left(\sum a_i\right)\left(\sum b_i\right)$ the following limit for the lower bound is obtained:

$$CV^2 = \frac{\sum \lambda_i^2}{\left(\sum \lambda_i\right)^2} \geq \frac{1}{p} \Rightarrow CV \geq \frac{1}{\sqrt{p}} \qquad (5.5.3.2)$$

The upper limit of the coefficient of variation of a 'generalised' hypo-exponential distribution (where initial entry to any of the transient states is permitted) can be shown to be unbounded (Telek & Heindl, 2002).

5.5.3.1 Two-term hypo-exponential distribution

The sojourn time can be written $x = x_1 + x_2$ where x_1 and x_2 represent the amount of time spent in the first and second phases with probabilities $f_1(x_1)$ and $f_2(x_2)$ respectively. Therefore the joint probability is expressed as $f_1(x_1) f_2(x - x_1)$. The pdf of sojourn time is deduced by integrating this joint probability over all values of x_1;

$$f(x) = \int_0^x f_1(x_1) f_2(x - x_1) dx_1 \qquad (5.5.3.1.1)$$

Therefore

$$f(x) = \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2}\left(e^{-\lambda_2 x} - e^{-\lambda_1 x}\right) \qquad (5.5.3.1.2)$$

This result may be confirmed by the *Laplace transform approach* or by the evaluation of (5.5.1) as part of the *matrix geometric approach*.

5.5.3.1.1 Maximum likelihood

The log-likelihood of (5.5.2.1.2) is

$$l = \ln L = n \ln\left(\frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2}\right) + \sum_{i=1}^{n} \ln\left(e^{-\lambda_2 x_i} - e^{-\lambda_1 x_i}\right) \qquad (5.5.3.1.1.1)$$

which may be solved by the Nelder-Mead algorithm.

5.5.3.1.2 Method of moments

As previously stated the Laplace transform of the exponential distribution with parameter λ_i is

$$L\{f_i(x)\} = \int_0^\infty \lambda_i e^{-\lambda_i x} e^{-sx} dx = \frac{\lambda_i}{\lambda_i + s} \qquad (5.5.3.1.2.1)$$

If two of these distributions are considered in series then the Convolution theorem may be applied to deduce

$$L\{f(x)\} = L\{f_1(x)\} \cdot L\{f_2(x)\} = \left(\frac{\lambda_1}{\lambda_1 + s}\right) \cdot \left(\frac{\lambda_2}{\lambda_2 + s}\right) \qquad (5.5.3.1.2.2)$$

By differentiation of the MGF (found by replacing *s* by *-t* in (5.5.3.1.2.2)) the first two theoretical moments are found to be

$$\lambda_1^{-1} + \lambda_2^{-1} \qquad 2\left[\lambda_1^{-2} + \lambda_1^{-1}\lambda_2^{-1} + \lambda_2^{-2}\right] \qquad (5.5.3.1.2.3)$$

On equating the first sample and population moments;

$$\hat{\lambda}_1^{-1} = m_1 - \lambda_2^{-1} \qquad (5.5.3.1.2.4)$$

and equating the second sample and population moments and substituting in (5.5.3.1.2.4) for λ_1^{-1} returns

$$\lambda_2^{-2} + \left(-m_1\right)\lambda_2^{-1} + \left(m_1^2 - 0.5m_2\right) = 0 \qquad (5.5.3.1.2.5)$$

This may be solved by the quadratic formula to give

$$\hat{\lambda}_{1,2}^{-1} = \frac{m_1}{2} \pm \frac{1}{2}\sqrt{2m_2 - 3m_1^2} \qquad (5.5.3.1.2.6)$$

making use of (5.5.3.1.2.4). This is confirmed by Whitt, 1982. The feasible values of the two sample moments are as follows. The first sample moment must be non-negative. It has already been proved that $p^{-0.5} \le CV \le 1$ and with $p = 2$ it is simple to derive that $1.5m_1^2 \le m_2 \le 2m_1^2$.

5.5.3.2 Erlang distribution

The Erlang distribution is obtained from the hypo-exponential when all the λ_i's representing transition rates between states are of equal value. It may also be seen as a special case of the gamma distribution when the shape parameter is confined to the set of positive integers. The coefficient of variation of this distribution is fixed at the lower bound of that of the hypo-exponential distribution, i.e. $CV = 1/\sqrt{p}$. This is because of the absence in variation of transition rates (homogeneity) associated with the Erlang distribution. Aldous & Shepp, 1987

prove a significant result, namely the title of their paper, *'The least variable phase-type distribution is Erlang'*. Therefore the lowest possible value of the coefficient of variation for an order *p* phase-type distribution is $1/\sqrt{p}$.

Generalising the result obtained in (5.5.3.1.2.2) to a series of *p* phases with equal transition rates gives

$$L\{f(x)\} = \left(\frac{\lambda}{\lambda + s}\right)^p \tag{5.5.3.2.1}$$

On application of the Shift theorem to (5.5.3.2.1);

$$f(x) = \frac{\lambda^p x^{p-1} e^{-\lambda x}}{(p-1)!} \tag{5.5.3.2.2}$$

A hyper-Erlang distribution may be defined as the sum of Erlang distributions (see Johnson & Taaffe, 1989 and Wang et al, 2008 for details).

5.5.3.2.1 Maximum likelihood

The log-likelihood of (5.5.3.2.2) is

$$l = \ln L = np \ln \lambda + (p-1)\sum \ln x_i - \lambda \sum x_i - n \ln(p-1)! \tag{5.5.3.2.1.1}$$

The partial differential of this (w.r.t λ) set equal to zero yields

$$\hat{\lambda} = \frac{np}{\sum x_i} = \frac{p}{\bar{x}} \tag{5.5.3.2.1.2}$$

A closed-form solution for the shape parameter, *p*, is intractable. There are two options regarding the estimation of this parameter. The first method simply involves assessing the value of (5.5.3.2.1.1) (substituting in (5.5.3.2.1.2) for λ) for a variety of integer values of *k*. When a value k_i has been found such that $L(k_{i-1}) < L(k_i)$ and $L(k_{i+1}) < L(k_i)$ then $\hat{k} = k_i$. This approach is valid because the Erlang distribution is unimodal, thus implying its log-likelihood is also unimodal. The estimator should be found with relative ease – as according to Miller & Bhat, 1997 *'in queuing contexts, usually only small values of k are relevant'*. An alternative is suggested by Miller, 1999 based on the workings of Dahiya, 1981. This aims to

find p such that $L(p-1|x) = L(p|x)$. It follows that $\hat{p}$ is defined as the greatest integer less than or equal to p, i.e. $\hat{p} = \lceil p \rceil$. It is therefore necessary to solve

$$\frac{\left(\frac{p-1}{\bar{x}}\right)^{n(p-1)} \prod x^{p-2} e^{-n(p-1)}}{\Gamma(p-1)^n} = \frac{\left(\frac{p}{\bar{x}}\right)^{np} \prod x^{p-1} e^{-np}}{\Gamma(p)^n} \qquad (5.5.3.2.1.3)$$

Note the denominator of the Erlang pdf has been replaced by that of a gamma pdf. This is because $p \notin \mathbb{Z}^+$ at this stage. By manipulation of (5.5.3.2.1.3) the equation to be solved for p can be simplified to

$$\left(1-\frac{1}{p}\right)^p = \frac{\left(\prod x_i\right)^{\frac{1}{n}}}{e\bar{x}} \qquad (5.5.3.2.1.4)$$

by making use of the result $\Gamma(z+1) = z\Gamma(z)$.

5.5.3.2.2 Method of moments

The population moments of the Erlang distribution can be deduced from the moment generating function that can be obtained in accordance with the approach employed in subsection 5.5.2.1.2. Hence, by replacing s by $-t$ in (5.5.3.2.1) it follows that $M_X(t) = (1-t/\lambda)^{-p}$. By differentiation the first and second moments are

$$\frac{p}{\lambda} \qquad \frac{p}{\lambda^2}(1+p) \qquad (5.5.3.2.2.1)$$

Equating these with the respective sample moments gives

$$\hat{\lambda} = \frac{p}{m_1} \qquad \hat{p} = \frac{m_1^2}{m_2 - m_1^2} = \frac{1}{CV_s^2} \qquad (5.5.3.2.2.2)$$

5.5.4 Coxian phase-type distribution

The Coxian distribution is a phase-type distribution whose coefficient of variation occupies the set of positive real values $1/\sqrt{p} < CV < \infty$. It may be seen as a generalisation of the hypo-exponential distribution in which transition from any transient state to the absorbing state is permitted. For this phase-type distribution a choice must be made with regard to parameterisation (see Ch 1.5 for details on parameterisation options).

Firstly the *probability parameterisation* is considered. The CTMC representation of the Coxian distribution with this parameterisation is illustrated in Figure 5.3. Here the β_i represent the probability of transition from state *i* to *i+1* for $i = 1, 2, ..., p-1$.

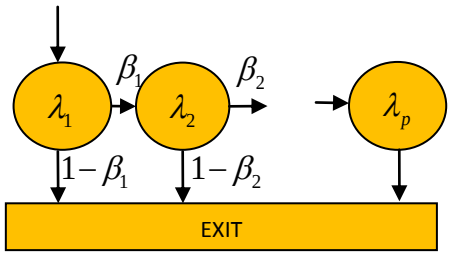


Figure 5.3 CTMC representation of a Coxian distribution with probability parameterisation

Thus the initial probability vector, (bi-diagonal) PH-generator matrix and *exit vector* are

$$\alpha = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}, T = \begin{bmatrix} -\lambda_1 & \beta_1\lambda_1 & 0 & 0 \\ 0 & -\lambda_2 & \beta_2\lambda_2 & \\ 0 & 0 & \ddots & \vdots \\ 0 & & \cdots & -\lambda_p \end{bmatrix}, -T.1 = \begin{bmatrix} (1-\beta_1)\lambda_1 \\ (1-\beta_2)\lambda_2 \\ \vdots \\ \lambda_p \end{bmatrix} \qquad (5.5.4.1)$$

where $\lambda_i (i = 1..p)$ represents the rate parameter of phase *i*. The transition rates between states may be summarised as

$$\begin{aligned} Rate(s_i \to s_{i+1}) &= \nu_i = \beta_i\lambda_i & i &= 1..p-1 \\ Rate(s_i \to s_{p+1}) &= \mu_i = (1-\beta_i)\lambda_i & i &= 1..p-1 \\ Rate(s_i \to s_{p+1}) &= \mu_i = \lambda_i & i &= p \end{aligned}$$

where s_i represents the *i*-th state.

Secondly the *rate parameterisation* is considered. The CTMC representation of the Coxian distribution with this parameterisation is illustrated in Figure 5.4. Here the υ_i represent the rate of transition from state *i* to *i+1* for $i = 1, 2, .., p-1$ whilst the μ_i represent the rate of transition from state *i* to the absorbing state.

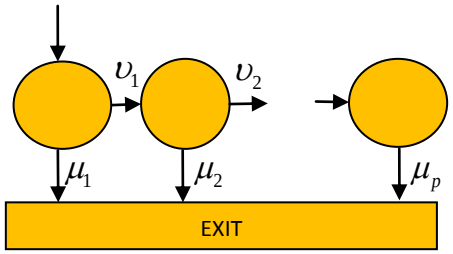


Figure 5.4 CTMC representation of a Coxian distribution with rate parameterisation

Thus the following is obtained:

$$\alpha = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}, T = \begin{bmatrix} -(\nu_1 + \mu_1) & \lambda_1 & 0 & 0 \\ 0 & -(\nu_2 + \mu_2) & \lambda_2 & \\ 0 & 0 & \ddots & \vdots \\ 0 & & \cdots & -\mu_p \end{bmatrix}, -T.1 = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \qquad (5.5.4.2)$$

The probability of transition between states may be summarised as

$$P\left(s_i \to s_{i+1}\right) = \beta_i = \frac{\nu_i}{\nu_i + \mu_i} \qquad i = 1..p-1$$

$$P\left(s_i \to s_{p+1}\right) = 1 - \beta_i = \frac{\mu_i}{\nu_i + \mu_i} \qquad i = 1..p-1$$

$$P\left(s_i \to s_{p+1}\right) = 1 \qquad i = p$$

Note that a Coxian distribution requires the estimation of $\left(2p-1\right)$ parameters.

The Coxian distribution has been used in many fields of mathematics since its introduction (Cox, 1955) in 1955 by David Cox. One such area is healthcare for which it has become increasing popular to model survival times by the distribution (Marshall & Zenga, 2009). Faddy, 1998 identifies the adequacy of this distribution in reducing the trade-off between wide applicability and efficient parameterisation and claims that *'little is lost by restricting attention to this Coxian sub-class of phase-type distributions'* (The author reasons this by stating a result produced in O'Cinneide, 1989 that *'any acyclic phase-type distribution* (see Ch 5.5.6) *has an equivalent representation in Coxian form'*). However, this claim is refuted by example in Fackrell, 2009 in which the author finds that a 24-phase Coxian distribution is unable to outperform a six-term general phase-type distribution when fitted to LOS data. This

example of over-parameterisation, a problem often associated with the general phase-type distribution, has also been observed when fitting Coxian distributions in Faddy, 1994.

A generalisation of the Coxian distribution is to relax the condition that the process must begin in the first phase.

5.5.4.1 Two-term Coxian distribution

This distribution has been used historically to model LOS. For example in McClean & Millard, 2007 geriatric LOS is modelled by this distribution with acute care and long stay represented by the two phases.

The probability parameterisation and Laplace transform approach are used to derive the pdf due to the low order of the distribution. In order to assist in the formulation of the Laplace transform of sojourn time a graph of the convex combination of all possible pathways is drawn (Figure 5.5). Note that the considered distribution has thus been transformed into a mixture of generalised Erlang distributions (Commault & Mocanu, 2003).

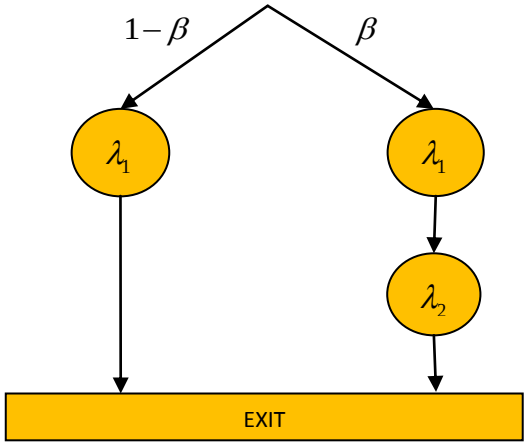


Figure 5.5 Convex combination of possible pathways in a two-term Coxian distribution

This representation of all possible pathways in terms of a mixture of generalised Erlang distributions enables a more convenient approach to the formulation of terms within the Laplace transform of sojourn time. Therefore

$$L\{f(x)\} = (1-\beta)\frac{\lambda_1}{\lambda_1 + s} + \beta\frac{\lambda_1}{\lambda_1 + s}\frac{\lambda_2}{\lambda_2 + s} \qquad (5.5.4.1.1)$$

and by application of the Shift theorem;

$$f(x) = (1-\beta)\lambda_1 e^{-\lambda_1 x} + \beta \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1}\left(e^{-\lambda_1 x} - e^{-\lambda_2 x}\right) \qquad (5.5.4.1.2)$$

This result may be confirmed by the calculation of (5.5.1) with α and T replaced by their respective values in (5.5.4.1) with $p = 2$. In cases where *T* is not in diagonal form the calculation of the matrix exponential, e^{Tx}, may be time-consuming. Mathematical packages such as Matlab and Maple can be used to produce quick and reliable results.

In extending to higher-order cases $(p > 2)$ the derivation of the pdf becomes increasingly difficult via the Laplace transform approach when compared to the matrix geometric approach. This is, in part, down to the increased complexity of setting up $L\{f(x)\}$, but principally it is because of the extent to which desktop computers can be employed in calculating $f(x)$ through the matrix geometric approach.

Equivalent results are obtained using rate parameterisation.

5.5.4.1.1 Maximum likelihood

The log-likelihood of (5.5.4.1.2) is

$$l = \sum_{i=1}^{n} \ln\left((1-\beta)\lambda_1 e^{-\lambda_1 x_i} + \beta \frac{\lambda_1 \lambda_2}{\lambda_2 - \lambda_1}\left(e^{-\lambda_1 x_i} - e^{-\lambda_2 x_i}\right) \right) \qquad (5.5.4.1.1.1)$$

which can be solved by the Nelder-Mead algorithm.

5.5.4.1.2 Method of moments

The Laplace transform of (5.5.4.1.1) may be simplified to

$$L\{f(x)\} = \frac{(\lambda_2 - (\beta - 1)s)\lambda_1}{(\lambda_1 + s)(\lambda_2 + s)} \qquad (5.5.4.1.2.1)$$

By employing the method (see 5.5.2.1.2 for details) that is used regularly throughout this subchapter the following estimators are deduced:

$$\hat{\lambda}_1 = \frac{B \pm \sqrt{A}}{C}$$
$$\hat{\lambda}_2 = \frac{2\left(m_1\hat{\lambda}_1 - 1\right)}{m_2\hat{\lambda}_1 - 2m_1} \qquad (5.5.4.1.2.2)$$
$$\hat{\beta} = \left(m_1 - \hat{\lambda}_1^{-1}\right)\hat{\lambda}_2$$

There always exists two real solutions for λ_1 (Augustin & Buscher, 1982), however, it is possible that one solution may be negative. This possibility is eliminated by taking $B + \sqrt{A}$ as the numerator of λ_1.

The following notation introduced in Altiok, 1985 is employed with the aim of attaining a more convenient form for the estimators:

$$X = \lambda_1 + \lambda_2 \qquad Y = \lambda_1\lambda_2 \qquad (5.5.4.1.2.3)$$

By manipulating these two equations;

$$\hat{\lambda}_{1,2} = \frac{X \pm \sqrt{X^2 - 4Y}}{2} \qquad (5.5.4.1.2.4)$$

and by calculation of X and Y with parameter values from (5.5.4.1.2.2),

$$X = \frac{2\left(3m_1m_2 - m_3\right)}{3m_2^2 - 2m_1m_3} = \frac{2B}{C} \qquad Y = \frac{6\left(2m_1^2 - m_2\right)}{3m_2^2 - 2m_1m_3} = \frac{6D}{C} \qquad (5.5.4.1.2.5)$$

The necessary and sufficient conditions required to achieve moment consistency are as follows. The first sample moment must be non-negative. The second sample moment is bounded only from below – it is inferred from $0.5 \le CV_s^2 \le \infty$ that $m_2 \ge 0.5m_1^2$. With regard to the bounds of the third sample moment a distinction must be made between the case when $CV_s^2 > 1$ and $0.5 \le CV_s^2 \le 1$. Numerous authors (van der Heijden, 1998, Papadopoulos, 1998) have articulated that for case $CV_s^2 > 1$ the third moment bound is equivalent to that of the third equation of (5.5.2.1.2.10), i.e. the third moment bound of the two-term hyper-exponential distribution obtained in Whitt, 1982. For case $0.5 \le CV_s^2 \le 1$ the method introduced in van der Heijden, 1998 is adopted. Firstly the (non-negativity) restrictions on the rate parameters are converted to restrictions on the auxiliary variables

$\left(X \geq 0, Y > 0, X^2 \geq 4Y\right)$. Then, these restrictions are subsequently converted to restrictions on the third sample moment (using $0 \leq \beta \leq 1$) to obtain

$$\frac{3m_1^3\left(CV_s^2+1\right)^2\left(1+\sqrt{2-2CV_s^2}\right)}{3-CV_s^2+2\sqrt{2-2CV_s^2}} \leq m_3 \leq 6m_1^3 CV_s^2 .$$

If the alternate (rate) parameterisation is used then the following estimators are obtained through the matrix geometric approach. Here μ_1 and μ_2 represent the transition rates from the first and second nodes to the absorbing state whilst ν_1 represents the transition rate from the first to the second phase.

$$\begin{aligned}\hat{\mu}_1 &= \frac{2\left(B-3m_1D\right)}{C} \\ \hat{\mu}_2 &= \frac{B \pm \sqrt{A}}{C} \\ \hat{\nu}_1 &= \frac{\hat{\mu}_2^2 D}{m_2\hat{\mu}_2 - 2m_1}\end{aligned} \qquad (5.5.4.1.2.6)$$

It can be shown through algebraic manipulation $\left(\nu_1 = \beta\lambda_1; \mu_1 = (1-\beta)\lambda_1; \mu_2 = \lambda_2\right)$ that the results of (5.5.4.1.2.2) and (5.5.4.1.2.6) are equivalent.

5.5.5 Acyclic phase-type distribution

The acyclic phase-type distribution may be regarded as a continuous-time Markov process in which no state is visited more than once. This is analogous to having an upper-triangular representation of the PH generator matrix. Therefore, in probability parameterisation

$$\begin{aligned}\alpha &= \begin{pmatrix}1 & 0 & \cdots & 0\end{pmatrix} \\ T &= \begin{bmatrix}-\lambda_1 & \beta_{1,2}\lambda_1 & \cdots & \beta_{1,p}\lambda_1 \\ 0 & -\lambda_2 & & \beta_{2,p}\lambda_2 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & -\lambda_p\end{bmatrix}, -T.1 = \begin{bmatrix}(1-\beta_{1,2}-\ldots-\beta_{1,p})\lambda_1 \\ (1-\beta_{2,3}-\ldots-\beta_{2,p})\lambda_2 \\ \vdots \\ \lambda_p\end{bmatrix}\end{aligned} \qquad (5.5.5.1)$$

or in rate parameterisation

$$\alpha = \begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}$$

$$T = \begin{bmatrix} -(\mu_1 + \nu_{1,2} + ... + \nu_{1,p}) & \nu_{1,2} & \cdots & \nu_{1,p} \\ 0 & -(\mu_1 + \nu_{2,3} + ... + \nu_{2,p}) & & \nu_{2,p} \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & -\mu_p \end{bmatrix}, -T.1 = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} \quad (5.4.5.1)$$

where $\nu_{i,j} (i = 1..p-1, j = i+1..p)$ represents the transition rate from state *i* to *j* and $\mu_i (i = 1..p)$ is as before (see Ch 5.5.4). The probabilities of transition are summarised as

$$P\left(s_i \to s_j\right) = \beta_i = \frac{\nu_{i,j}}{\mu_1 + \nu_{i,j} + ... + \nu_{i,p}} \qquad i = 1..p-1, j = i+1..p$$

$$P\left(s_i \to s_{p+1}\right) = 1 - \beta_i = \frac{\mu_i}{\mu_1 + \nu_{i,j} + ... + \nu_{i,p}} \qquad i = 1..p-1, j = i+1..p$$

$$P\left(s_i \to s_{p+1}\right) = 1 \qquad i = p$$

where s_k represents the *k*-th state.

A theorem published in Cumani, 1982 states *'A triangular PH-representation is equivalent to an ordered Cox representation of at most the same order'*. This result means that the same distribution may be obtained but with the estimation of fewer parameters. It is also supported by Dehon & Latouche, 1982 and O Cinneide, 1987. There are a number of methods that may be employed to facilitate this reduction – see Pulungan & Hermanns, 2008 and references therein for details. However, it is not the case that analysis of the acyclic distribution may be omitted from this review since the complete data models will not be the same in such a distribution as in a Coxian (Asmussen et al, 1996). This can result in the production of different distributions for the acylic and Coxian distributions depending on the parameter estimation algorithms employed. It also depends on the prevalence of local minima/maxima and saddle points of the likelihood function if the maximum likelihood method of estimation is used.

This is an important distribution to consider because it not only increases the flexibility of the distribution (at the expense of greater parameter estimation) but can also be used to exemplify a physical representation of the phases as compartments. This physical representation has been adopted when modelling LOS for geriatric patients. Xie et al, 2005 use a four phase distribution to model LOS for patients in residential and nursing home care for which two of

the phases are used to represent short and long stay in each institution. This may be viewed as a Coxian distribution with the addition of a nonzero transition probability from the first node (of residential care) to the first node of nursing home care.

5.5.5.1 Three-term acyclic phase-type distribution

A two-term distribution will not be considered as this would be equivalent to the consideration of a two-term Coxian distribution. In order to capture the characteristics of this distribution a three-term model is therefore examined.

$$\alpha = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}, T = \begin{bmatrix} -\lambda_1 & \beta_1\lambda_1 & \beta_2\lambda_1 \\ 0 & -\lambda_2 & \beta_3\lambda_2 \\ 0 & 0 & -\lambda_3 \end{bmatrix}, v = \begin{bmatrix} (1-\beta_1-\beta_2)\lambda_1 \\ (1-\beta_3)\lambda_2 \\ \lambda_3 \end{bmatrix} \tag{5.5.5.1.1}$$

The probability parameterisation has been selected in retrospect after the consideration of both parameterisations yielded the rate parameterisation to have a more complex representation of the moments and the pdf. The pdf is therefore as follows:

$$\begin{aligned} f\left(x;\lambda_1,\lambda_2,\lambda_3,\beta_1,\beta_2,\beta_3\right) &= \lambda_1\left(1-\beta_1-\beta_2\right)e^{-\lambda_1 x} + \\ &+ \frac{\lambda_1\left(\beta_1\lambda_2\lambda_3+\beta_2\lambda_2\lambda_3+\beta_1\beta_3\lambda_1\lambda_2-\beta_1\lambda_1\lambda_2-\beta_2\lambda_1\lambda_3\right)e^{-\lambda_1 x}}{\left(\lambda_1-\lambda_2\right)\left(\lambda_1-\lambda_3\right)} + \\ &+ \frac{\beta_1\lambda_1\lambda_2\left(\lambda_2-\lambda_3-\lambda_2\beta_3\right)e^{-\lambda_2 x}}{\left(\lambda_1-\lambda_2\right)\left(\lambda_2-\lambda_3\right)} + \frac{\lambda_1\lambda_3\left(\beta_2\lambda_2+\beta_1\beta_3\lambda_2-\beta_2\lambda_3\right)e^{-\lambda_3 x}}{\left(\lambda_1-\lambda_3\right)\left(\lambda_2-\lambda_3\right)} \end{aligned} \tag{5.5.5.1.2}$$

5.5.5.1.1 Maximum likelihood

By application of the logarithm to the likelihood function of (5.5.5.1.2) the log-likelihood function is obtained as

$$\begin{aligned} l = \sum_{i=1}^{n} \ln\Big[&\lambda_1\left(1-\beta_1-\beta_2\right)e^{-\lambda_1 x_i} + \\ &+ \frac{\lambda_1\left(\beta_1\lambda_2\lambda_3+\beta_2\lambda_2\lambda_3+\beta_1\beta_3\lambda_1\lambda_2-\beta_1\lambda_1\lambda_2-\beta_2\lambda_1\lambda_3\right)e^{-\lambda_1 x_i}}{\left(\lambda_1-\lambda_2\right)\left(\lambda_1-\lambda_3\right)} + \\ &+ \frac{\beta_1\lambda_1\lambda_2\left(\lambda_2-\lambda_3-\lambda_2\beta_3\right)e^{-\lambda_2 x_i}}{\left(\lambda_1-\lambda_2\right)\left(\lambda_2-\lambda_3\right)} + \frac{\lambda_1\lambda_3\left(\beta_2\lambda_2+\beta_1\beta_3\lambda_2-\beta_2\lambda_3\right)e^{-\lambda_3 x_i}}{\left(\lambda_1-\lambda_3\right)\left(\lambda_2-\lambda_3\right)} \Big] \end{aligned} \tag{5.5.5.1.1.1}$$

This may, in conventional fashion, be solved by the Nelder-Mead algorithm.

5.5.5.1.2 Method of moments

An alternative method of equating moments is developed due to the complexity associated with the traditional analytical approach that has been used thus far (see Bobbio et al, 2005 and Horvath & Telek, 2007 for the intricacies associated with an analytical approach).

The method that is used seeks to minimise the square of the deviation of sample moments from theoretical moments. This method is a heuristic in the sense that it produces approximations and not exact results (see Ch 4.3). Higher value weights may be placed on lower-order moments to emphasise the greater importance in solving these (Gross & Juttijudata, 1997). Johnson & Taaffe, 1990 employ this approach with three moments. The authors develop the following nonlinear programming problem:

$$\min: w_1\left(CV_s - CV_p\right)^2 + w_2\left(\gamma_s - \gamma_p\right)^2 \qquad (5.5.5.1.2.1)$$

where $\gamma_{s,p}$ are the sample and population coefficients of skewness defined as

$$\gamma_p = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{E\left[X^3\right] - 3\mu\sigma^2 - \mu^3}{\sigma^3} \text{ and } \gamma_s = \frac{m_3 - 3\bar{x}s^2 - \bar{x}^3}{s^3}.$$

The authors conclude by suggesting that *'NLP methods have the potential to become a powerful tool for selecting approximating PH distributions'*.

The approach that is used forthwith is as follows:

$$\min: w_1\left(m_1 - M_1\right)^2 + w_2\left(m_2 - M_2\right)^2 + w_3\left(m_3 - M_3\right)^2 + ... \qquad (5.5.5.1.2.2)$$

in which theoretical (population) moments are calculated by the formula $M_n = E\left[X^n\right] = (-1)^n n!\alpha T^{-n}1$. The first three moments are found to be

$$\begin{aligned}
M_1 &= \frac{1}{\lambda_1} + \frac{\beta_1}{\lambda_2} + \frac{\beta_1\beta_3 + \beta_2}{\lambda_3} \\
M_2 &= 2\left[\frac{1}{\lambda_1^2} + \frac{\beta_1}{\lambda_1\lambda_2} + \frac{\beta_1}{\lambda_2^2} + \frac{\beta_1\beta_3 + \beta_2}{\lambda_1\lambda_3} + \frac{\beta_1\beta_3}{\lambda_2\lambda_3} + \frac{\beta_1\beta_3 + \beta_2}{\lambda_3^2}\right] \\
M_3 &= 6\left[\frac{1}{\lambda_1^3} + \frac{\beta_1}{\lambda_1^2\lambda_2} + \frac{\beta_1}{\lambda_1\lambda_2^2} + \frac{\beta_1}{\lambda_2^3} + \frac{\beta_1\beta_3 + \beta_2}{\lambda_1^2\lambda_3} + \right. \\
&\left. + \frac{\beta_1\beta_3}{\lambda_1\lambda_2\lambda_3} + \frac{\beta_1\beta_3}{\lambda_2^2\lambda_3} + \frac{\beta_1\beta_3 + \beta_2}{\lambda_1\lambda_3^2} + \frac{\beta_1\beta_3}{\lambda_2\lambda_3^2} + \frac{\beta_1\beta_3 + \beta_2}{\lambda_3^3}\right]
\end{aligned} \qquad (5.5.5.1.2.3)$$

The optimisation can be performed via the Nelder-Mead simplex algorithm.

5.5.6 General phase-type distribution

The general phase-type distribution has the highest flexibility (Chakravarthy & Alfa, 1997) of all of the phase-type distributions covered thus far. This is because it imposes no restriction on state transitions and entry into initial state. This distribution is therefore a generalisation of all phase-type distributions covered within this chapter. By the same token, and noting Aldous & Shepp, 1987, the coefficient of variation is within the interval $1/\sqrt{p} \le CV < \infty$. In rate parameterisation it is described

$$\alpha = \begin{pmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_p \end{pmatrix},$$

$$T = \begin{bmatrix} -(\mu_1 + \nu_{1,2} + \dots + \nu_{1,p}) & \nu_{1,2} & \cdots & \nu_{1,p} \\ \nu_{2,1} & -(\mu_2 + \nu_{2,1} + \dots + \nu_{2,p}) & & \nu_{2,p} \\ \vdots & & \ddots & \vdots \\ \nu_{p,1} & \nu_{p,2} & \cdots & -(\mu_p + \nu_{p,1} + \dots + \nu_{p,p}) \end{bmatrix} \qquad (5.5.6.1)$$

The number of parameters required for estimation is therefore $p^2 + p - 1$ (i.e. $p-1$ from α and p^2 from T). Table 5.1 documents the number of parameters requiring estimation for both the Coxian and the general phase-type distribution.

Table 5.1 Comparison of number of parameters requiring estimation

# phases:	1	2	3	4	5	6	7	8	9	10
Coxian	1	3	5	7	9	11	13	15	17	19
General	1	5	11	19	29	41	55	71	89	109

It is because of this that the general phase-type distribution has not seen the scale of application and interest that the Coxian distribution has had within the research community. Some authors (Garg et al, 2009, Fackrell, 2009) refer to the general phase-type distribution as being over-parameterised; a problem defined by the fitting of excessive parameters to a model. This can result in a very good approximation for the sample data but a less than adequate fit for other samples from within the population (Svolba, 2006). However, as exemplified in Fackrell, 2009 (and mentioned in Ch 5.5.4) the author finds a 24-phase Coxian distribution to perform worse than a six-phase general distribution. Note here that the Coxian distribution requires the estimation of 47 parameters whilst the general distribution requires the estimation of six fewer.

The general phase-type distribution is of particular interest when a physical representation of a problem is sought. Griffiths et al, 2006 fit such a compartmental model to the HIV/AIDS epidemic in which the compartments (phases) represent different stages of the disease. Fackrell, 2009 also applies the general phase-type distribution to the field of healthcare by modelling six departments of a hospital (Theatre, ICU, Ward 1, etc). The author describes the cyclic transitions between the departments and remarks on how the sojourn time in each department can be modelled by distributions other than the single phase exponential distribution (e.g. Erlang, Coxian etc).

Two variations of the general phase-type distribution are forthwith considered: the feedback Erlang and the feedback Coxian. These distributions were introduced in O'Cinniede, 1997 and are equivalent to the respective Erlang and Coxian distributions of the same order albeit with a nonzero probability of transition from the final transient state to the first state. The respective number of parameters required for estimation is 2 and 2*p*. These distributions are considered firstly because of their relative simplicity when compared to distributions with a denser PH-generator matrix, and secondly, because the feedback concept could be useful in modelling readmissions or relapse within a healthcare facility.

5.5.6.1 Two-term Erlang distribution with feedback

First considered is a two-term feedback Erlang distribution which is the simplest case of a general phase-type distribution that exhibits cyclic behaviour. Its pdf cannot be found through the traditional Laplace transform approach because a convex combination of all pathways (see Ch 5.5.4.1) cannot be readily formulated. Therefore the pdf in probability parameterisation is found through the matrix geometric approach;

$$f\left(x;\lambda,\beta\right)=\frac{\lambda\left(1-\beta\right)\left(e^{-\lambda\left(1-\sqrt{\beta}\right)x}-e^{-\lambda\left(1+\sqrt{\beta}\right)x}\right)}{2\sqrt{\beta}} \qquad (5.5.6.1.1)$$

where β is the feedback probability. The Laplace transform is calculated to be

$$L(s;\lambda,\beta)=\frac{\left(1-\beta\right)\lambda^2}{\left(s+\lambda\right)^2-\beta\lambda^2} \qquad (5.5.6.1.2)$$

which is in agreement with (equation 25 of) Commault & Mocanu, 2003.

5.5.6.1.1 Maximum likelihood

The log-likelihood function of (5.5.6.1.1) to be maximised is

$$l = n\ln\left(\frac{(1-\beta)\lambda}{2\sqrt{\beta}}\right) + \sum_{i=1}^{n}\ln\left(e^{-\lambda\left(1-\sqrt{\beta}\right)x} - e^{-\lambda\left(1+\sqrt{\beta}\right)x}\right) \qquad (5.5.6.1.1.1)$$

5.5.6.1.2 Method of moments

The first and second theoretical moments, as derived through (5.5.6.1.2), are $\frac{2}{\lambda(\beta-1)}$ and $\frac{2(3+\beta)}{\lambda^2(\beta-1)}$. On equating with sample moments the estimators of β and λ are found to be $3C_s^2$ and m_1/D.

5.5.6.2 Two-term Coxian distribution with feedback

The two-term feedback Coxian distribution is now considered. For equivalent reasons to those issued in Ch 5.5.6.1 the pdf in probability parameterisation is deduced to be

$$f(x;\lambda_1,\lambda_2,\beta_1,\beta_2) = \frac{1}{2F}\left[\left(F+\lambda_1-\lambda_2\right)e^{-\frac{1}{2}(G+F)x} + \left(F-\lambda_1+\lambda_2\right)e^{-\frac{1}{2}(G-F)x}\right]\left(1-\beta_1\right)\lambda_1 + \frac{\beta_1\lambda_1}{F}\left[e^{-\frac{1}{2}(G-F)x} - e^{-\frac{1}{2}(G+F)x}\right]\left(1-\beta_2\right)\lambda_2 \qquad (5.5.6.2.1)$$

where $F = \sqrt{\left(\lambda_1-\lambda_2\right)^2 + 4\beta_1\beta_2\lambda_1\lambda_2}$, $G = \lambda_1+\lambda_2$, $\beta_1 = \mathrm{P}\left(S_1 \to S_2\right)$ and $\beta_2 = P\left(S_2 \to S_1\right)$. The Laplace transform of (5.5.6.2.1) is

$$L = \frac{\left(s+\lambda_2-\beta_1 s-\beta_1\beta_2\lambda_2\right)\lambda_1}{s^2+\left(\lambda_1+\lambda_2\right)s+\lambda_1\lambda_2-\beta_1\beta_2\lambda_1\lambda_2} \qquad (5.5.6.2.2)$$

5.5.6.2.1 Maximum likelihood

The log-likelihood function of (5.5.6.2.1) to be maximised is

$$l = \sum_{i=1}^{n}\ln\left\{\frac{1}{2F}\left[\left(F+\lambda_1-\lambda_2\right)e^{-\frac{1}{2}(G+F)x} + \left(F-\lambda_1+\lambda_2\right)e^{-\frac{1}{2}(G-F)x}\right]\cdot\left(1-\beta_1\right)\lambda_1 + \frac{\beta_1\lambda_1}{F}\left[e^{-\frac{1}{2}(G-F)x} - e^{-\frac{1}{2}(G+F)x}\right]\left(1-\beta_2\right)\lambda_2\right\} \qquad (5.5.6.2.1.1)$$

5.5.6.2.2 Method of moments

The method employed within subchapter 5.5.5.1.2 (given by the objective (5.5.5.1.2.2)) is used here. This is due to the complexity associated with obtaining analytic results by equating moments in the usual fashion.

The first three theoretical (population) moments are as follows

$$M_1 = -\frac{\beta_1\lambda_1 + \lambda_2}{\lambda_1\lambda_2\left(\beta_1\beta_2 - 1\right)}$$
$$M_2 = \frac{2\left(\beta_1\lambda_1^2 + \lambda_2^2 + \beta_1\lambda_1\lambda_2 + \beta_1\beta_2\lambda_1\lambda_2\right)}{\left\{\lambda_1\lambda_2\left(\beta_1\beta_2 - 1\right)\right\}^2} \qquad (5.5.6.2.2.1)$$
$$M_3 = -\frac{6\left(\beta_1\lambda_1^3 + \left(\beta_1 + \beta_1\beta_2 + \beta_1^2\beta_2\right)\lambda_1^2\lambda_2 + \beta_1\lambda_1\lambda_2^2 + \lambda_2^3 + 2\beta_1\beta_2\lambda_1\lambda_2^2\right)}{\left\{\lambda_1\lambda_2\left(\beta_2 - 1\right)\right\}^3}$$

5.6 Conclusion

A review of distribution estimation with application to healthcare studies has been presented. The aim of this review is to facilitate a greater understanding of probability distributions that are of potential interest when modelling arrival and length of stay data for healthcare facilities. A survey of the relevant literature is undertaken in the first subchapter. Three popular and commonly used methods of parameter estimation are introduced in Ch 5.2 and the following subchapter establishes the iterative techniques that are sometimes required to solve them. The estimation of parameters for four non phase-type distributions is addressed in Ch 5.4 and various types of phase-type distributions are investigated in Ch 5.5. An evaluation of limitations is now presented.

Perhaps the most significant limitation of this review is the number of distributions considered. By examining a greater number of suitable distributions (e.g. Burr, logistic, Rayleigh) one could hold greater confidence that the best approximating distribution of those applied is in fact the best approximating distribution possible. This concept also has application to the flexibility of the phase-type distributions. By examining higher-order phase-type distributions (or those with a greater number of nonzero transition probabilities) it is expected that better approximations could be achieved. However, this would be at the expense of increased mathematical complexity in obtaining analytic results and may risk over-parameterisation.

Another limitation of this review is the number of parameter estimation methods that are considered. Although three were originally discussed only two are formally applied. This lack of diversity could result in sub-optimal approximations of the data. The validity and accuracy of the methods used may also be questionable. The maximum likelihood method requires the independence of sample values; however, this could be affected by patient readmission. The accuracy of the method of moments is dependent on the number of moments matched. Within this review two or three moments are typically matched. However, there is a difference of opinion with regard to a sufficient number. Osogami & Harchol-Bolter, 2003 suggest that data is well-represented if three moments are equated whilst Gross & Juttijudata, 1997 advise matching at least five.

A further limitation relates to the accuracy of the results obtained when iterative techniques are employed. It has been previously acknowledged that the optimality and feasibility of final solutions found via such techniques can be very sensitive to initial values. The problem is therefore to determine initial values that allow the globally optimal solution to be found. If possible, analytic results from the alternative estimation method can be used as initial values. For example, if the method of moments returns explicit estimators, then these may be used within an iterative process for maximum likelihood. However, if neither method can produce analytic results then there is a problem. This can be dealt with when using the method of maximum likelihood by plotting a graph of the likelihood surface. If there is only one parameter to be estimated then a two-dimensional graph may be plotted and the initial value may be selected as a point near to the global optimum. A similar approach may be used should two parameters require estimation. When faced with three or more free parameters then two parameters may be plotted *ceteris paribus*. This is of significant complication when high-order phase-type distributions are considered.

The final matter addressed is that of purpose. Many authors have developed simple and user-friendly programs capable of deducing parameter estimates quickly and easily for a wide variety of phase-type distributions, for example EMPht (Olsson, 1998) and PHFit (Horvath & Telek, 2002). However, a substantial base of theory would not have been covered if these programs had been used. It was felt by the author that this theory is a valuable foundation for subsequent consideration in relation to queuing theory.

Chapter 6: Model Development and Data Evaluation

6.1 Introduction

The aim of this project is to model the activities of the Neurological Rehabilitation Centre (NRC) at Rookwood hospital. The two phases are *initial construction* (Ch 1.7.1) and *hypothetical scenarios* (Ch 1.7.2). The data used in the initial construction of the model can also be used to produce valuable statistics that can support the *major policy decisions* (who to admit, what care to give, when to discharge). Such data can be used to create a prognostic tool for estimating various patient-related attributes such as length of stay (LOS).

The accuracy of the model is determined by its ability to represent reality. This is dependent on two factors – the quality of the design and the quality of the data. The quality of the design is determined by the complexity of the model in relation to the complexity of the system. That is, a complicated system requires an equally complex model to capture all the underlying trends and variables within it. A sufficient understanding of the system is essential to create a model that is of appropriate complexity. A number of models of varying complexity are considered in Ch 6.2.

For a model to be pertinent to a particular environment a level of sensitivity to the characteristics of that environment is required. This is done by using raw data obtained from the environment to estimate model parameters. The quality of this data is dependent on the amount collected and the reliability of the collection process. If raw data is unavailable then model parameters can be estimated by experienced clinicians. These are issues discussed in Ch 6.3.

A brief statistical analysis of the data is presented in Ch 6.4. Finally, a number of distributions are fitted to patient LOS by the method of moments and maximum likelihood.

6.2 Model Development and Data Requirements

The data requirements for models of varying complexity are discussed forthwith. The first model represents the most simplistic and complexity is thereafter increased. Note that in some cases this is progressive in the sense that former models are used as a foundation that is built upon. The subchapter concludes with the specification of the most representative model that is considered. The objective is to establish options with regard to model design – it is not to decide upon a particular model.

Each model has potential benefit as well as drawback. In general, simpler models require less data (as fewer parameters require estimation) but are limited in their representation of reality (since important aspects of the data are not considered as variables). More complex models have the ability to incorporate such trends but carry a risk of being over-fitted. They also permit a wider range of hypothetical *what-if* type scenarios as a result of containing more variables as parameters.

6.2.1 Queuing theory model

This model is assembled from basic queuing theory principles. The suitability of queuing theory for use in modelling patient transition through a healthcare facility is discussed in Chapter 1.3. Let λ be the mean arrival rate of patients at the queue and μ be the mean service rate. The number of service channels is equivalent to the number of beds.

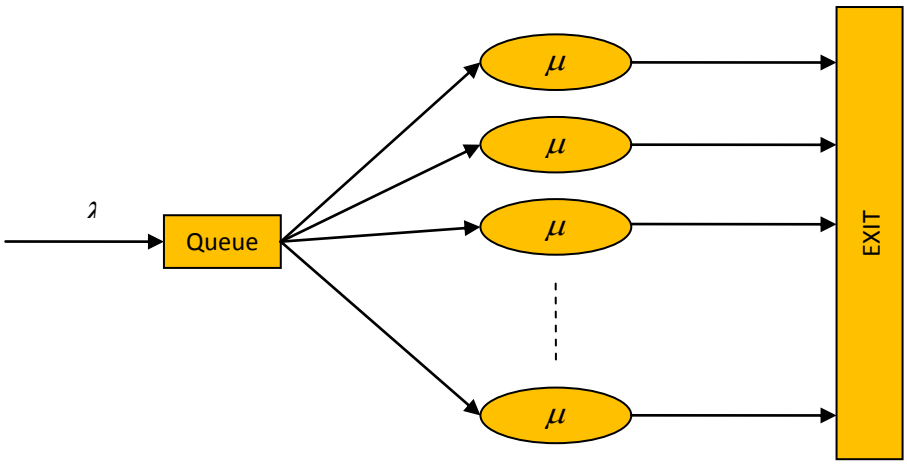


Figure 6.1 Basic multi-server queuing system

Data is required for the following variables:

- **Date referral received**
- **Date of admission**
- **Date of discharge**

Upon the determination of the (inter) arrival and service distributions (see Chapter 5) performance measures such as mean wait in queue and mean service time can be studied. The ability to deduce such measures is dependent on the fitted distributions (see Ch 5.4.1). The model may now be extrapolated to study what-if type scenarios. However, due to the simplicity of this model, only a limited number of scenarios may be examined – a change in the number of service channels (beds) and modification to the arrival or service distribution. Whilst the first scenario details a reasonable policy option (hospitals often review the number of beds) it is much more difficult to find a practical connection with the other two. For example, the lowest mean wait in queue could be achieved by changing the service distribution to a two-term hypo-exponential but how would this change be replicated on the ground? It is clear that a more advanced model is required.

6.2.2 Separation of LOS to active and blocked components

A simple and effective advancement of the model is to partition total LOS (service time) into *active LOS* (ALOS) and *blocked LOS* (BLOS). ALOS is defined as the length of time (days) for which a patient is subject to rehabilitative treatment. The time at which the patient becomes ready for discharge defines the end of this rehabilitative care. Thereafter, their category is downgraded from *therapeutic* to *nursing* and treatment levels are reduced from *rehabilitative* to *maintenance*. Such patients are said to be *bed-blocking* (see Ch 1.2). BLOS is defined as the length of time (days) for which a patient is subject to maintenance treatment. If the patient is promptly discharged then they incur a zero BLOS, however, this is exceptional. Maintenance treatment is regarded as necessary (as in its absence patients would deteriorate) but wasteful (since resources are diverted from those engaged in rehabilitative care). This model is depicted in Figure 6.2 where μ_1 represents the active service rate and μ_2 represents the blocked service rate.

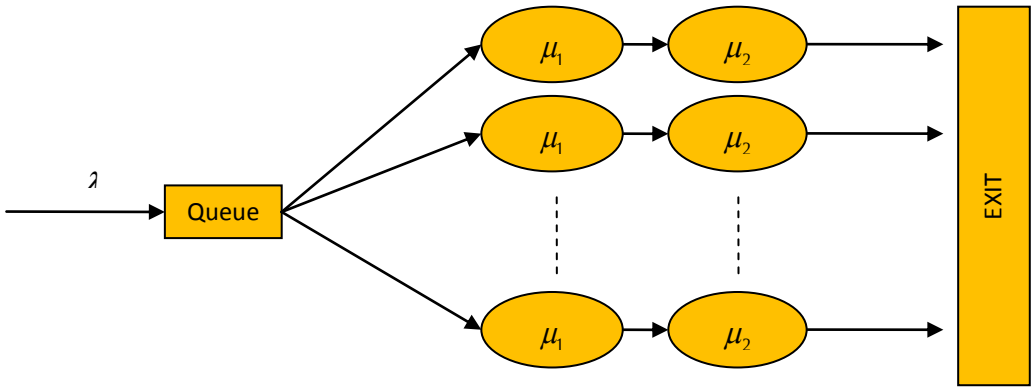


Figure 6.2 Queuing system with length of stay partitioned into active and blocked components

The variable that data are required for is

- **Date ready for discharge**

This enables the derivation of distributions for ALOS and BLOS and thus provides greater versatility in model design. It is practically speaking more meaningful to consider changes to ALOS and BLOS distributions than changes to LOS as a whole. This is because realistic policy decisions target either the ALOS or the BLOS independently of the other. For example, if through a more proactive arrangement with social services the mean BLOS was found to have decreased, then μ_2 may be increased without affecting μ_1 or its distribution.

6.2.3 Separation of active LOS to departmental active LOS

Patients may only be categorised as ready for discharge when all departments within the hospital (physiotherapy, occupational therapy etc) are in agreement that the patient is ready respective of their specialities. Hence date ready for discharge is in fact an upper bound of all such departmental readiness dates. If *dept active LOS* defines the length of time (days) for which a patient is subject to rehabilitative treatment for within a department then ALOS is an upper bound of these dept ALOSs.

The variable that data is required for is

- **Date ready for discharge (dept)**

From this, distributions of dept ALOS can be produced for each department. The distributions may then be subject to adjustments in accordance to what-if scenarios and the effect on performance measures can be analysed. This is of particular interest because of the

potentially avoidable cost associated with patient maintenance. An extreme case of this is forthwith exemplified.

Assume that the ALOS of one department dominates that of all others. This is illustrated in Figure 6.3 where yellow and grey shading are used to represent rehabilitative and maintenance departmental treatment respectively.

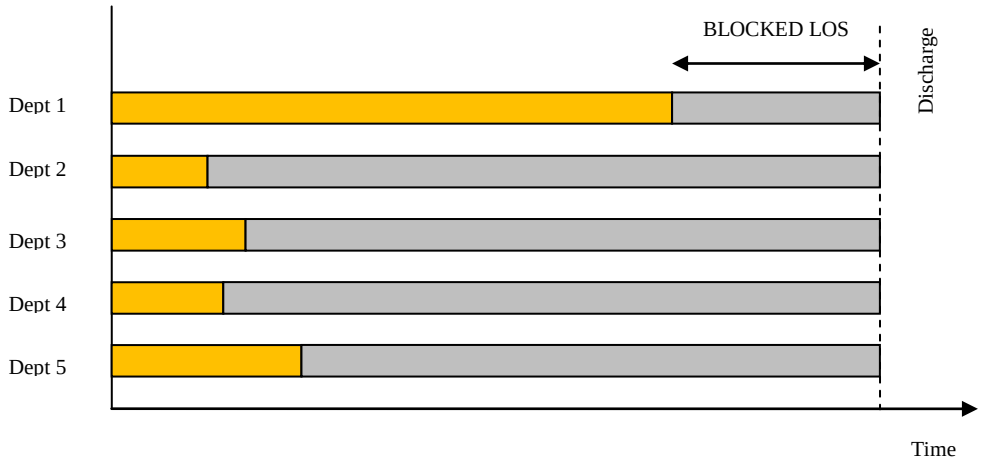


Figure 6.3 Active and blocked length of stay for each department

Despite early readiness from most departments, maintenance treatment must be provided to the patient for the remainder of time until discharge. Although this level of treatment is not as costly as rehabilitative treatment it still represents a significant cost to the departments. There are two ways of reducing this.

The first is to decrease BLOS. However, this is problematic since it is not directly controlled by Rookwood hospital – the principle problem being the punctuality of the discharge destination in accepting the patient. The second option is to improve inter-department co-ordination so that there is greater alignment between the dates of departmental readiness. Such a policy translates to the example as follows. A solution is to either reduce the ALOS of Dept 1 or raise the ALOS of the others. Reducing Dept 1 ALOS has additional benefit in increasing throughput due to the reduction in overall ALOS *ceteris paribus*.

A potential tool for controlling the ALOS of a department is the intensity of treatment.

6.2.4 Treatment intensity

Treatment intensity is defined as the rate at which treatment is provided. The intensity of treatment is the largest controllable post-arrival factor that affects ALOS (see Ch 2.2.3 for a

literature review of this concept). To summarise, it is widely accepted that greater treatment intensity reduces patient LOS (Blackerby, 1989, Spivack et al, 1992) by advancing functional gains (Sivenius et al, 1985). Because LOS is the *'primary significant predictor of rehabilitation charges'* (Cifu et al, 2003) then this is certainly a valuable tool.

This relationship between ALOS and treatment intensity holds at a departmental level for some, but not all, types of therapy (Heinemann et al, 1995, Cifu et al, 2003). Treatment is split into four therapy types (speech, occupational, physical, and psychological) in Cifu et al, 2003. The authors then study the effects of varying intensities of these therapies on cognitive and motor outcome. They find that whilst the intensity of speech and physical therapy were predictive of motor outcome, there were no therapy types that were predictive of cognitive outcome.

To incorporate the effect of departmental treatment intensity within the model, data for the following variable is required **for each department**:

- **Number of hours of clinical treatment received (until date ready for discharge)**

This concept is integrated within the model as follows. Assume that the distribution of ALOS for a particular department is given by Figure 6.4.

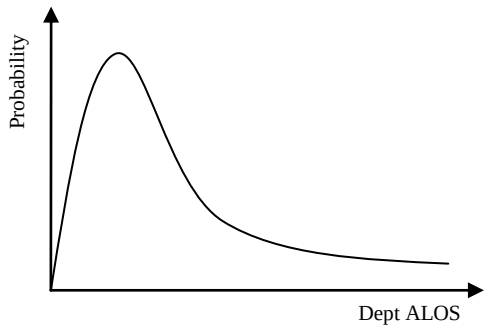


Figure 6.4 Probability density of departmental active length of stay

This can be used to deduce the mean and variance of ALOS. With the introduction of a third variable, treatment intensity, it becomes necessary to move to a three-dimensional graph. For illustrational purposes, however, it is easier to look at two two-dimensional graphs. The first (Figure 6.5) considers the relationship between service (treatment) intensity and service time (ALOS) by means of a *service curve*.

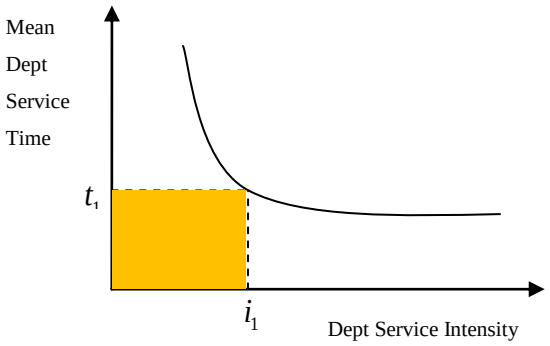


Figure 6.5 Relationship between departmental service intensity and service time

The dotted lines of Figure 6.5 illustrate an example in which the mean service time, t_1, can be deduced from a given service intensity, i_1. The area occupied by the dotted lines ($i_1 \cdot t_1$) is defined as the *expected service bulk* (in yellow). This represents the total amount of service that the patient is expected to receive from the department. Note that if dept treatment intensity is not predictive of dept ALOS then the service curve would approach horizontality. Figure 6.6 details the variance of the expected service time.

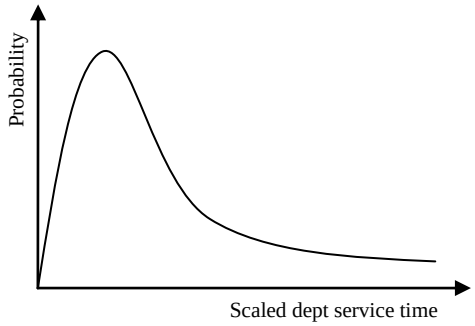


Figure 6.6 Probability density of scaled departmental active length of stay

Note that the shape of this distribution is equal to that of Figure 6.4. The dept ALOS axis values are re-scaled so that the mean of this distribution is equal to the mean service time obtained in Figure 6.5. Thus, a duplicate representation of Figure 6.4 is achieved when the mean service time obtained in Figure 6.5 is equal to the mean service time of Figure 6.4. A similar shape distribution is obtained for the variance of service bulk because of its linear dependence on service time (i.e. $bulk = i \cdot t$).

This extension of the model has the following benefits and drawbacks. The inclusion of treatment intensity as a variable within the model enables its effect on ALOS to be

appreciated. This is a valuable consideration since the intensity of departmental treatment can differ significantly among patients.

Such advances in the model design also serve to cover a more realistic range of what-if type scenarios. In the previous model (Ch 6.2.3) scenarios relate to changes in the distribution of dept ALOS. However, these changes are not easily replicated in real-life. In this model, technical modifications can still be made (i.e. to the service curve and distribution of scaled dept service time) but simple changes can also be made to treatment intensity. Such changes involve an increase or decrease to the dept service intensity of Figure 6.5. This is much more practical and relatable for therapists. However, it must be noted that a substantial amount of data are required in order to develop the relationships between departmental ALOS and treatment intensity.

Unfortunately, treatment intensity cannot be expressly regulated by therapists (see Ch 3.3.2). Treatment intensity is, in fact, an output of the scheduling procedures used by each of the departments. By uniting the automated scheduling program and the queuing model (Chapter 9), the effects of the *direct controls* (see Ch 3.4.3) of the therapists can be evaluated for the system. For departments that do not operate an automated scheduling program it is necessary to collect data on requested treatment so that an approximate ratio between treatment demanded and received can be determined.

6.2.5 Destination attributes

Patients are only considered ready for discharge from a department when they have exhibited a sufficient ability in that speciality to meet the demands of their prospective discharge destination. For example, if a patient is to be discharged to a multi-storey house that is on top of a steep hill then a high level of motor ability is necessary. It is likely that such a patient would require a greater amount of physiotherapy treatment in order to achieve this. Outcome is therefore defined as functional ability on discharge.

It is intuitive to hypothesise some form of trilateral relationship between outcome and treatment intensity and LOS. Outcome is a consideration in many articles that are mentioned in Ch 6.2.4. As three related variables are considered it is logical to hold one of these constant whilst another is varied. The effect of this is measured by the consequent change in the remaining variable. Blackerby, 1989 holds constant patient outcome and varies treatment intensity – the effect on LOS is then studied. Here the independent variables are treatment

intensity and outcome and the dependent variable is LOS. Other authors (Sivenius, 1985) choose outcome as the dependent variable and so hold constant LOS. Spivack et al, 1991 also studies the effect on patient outcome but varies both LOS and treatment intensity.

An approach to incorporate outcome within the model is proposed. First, a number of *destination levels* are constructed for each department. These represent echelons of required functional ability on discharge (outcome). Their simplest interpretation can be viewed as a diagonal shift in the service curve that represents a change in expected service bulk. The number of echelons is determined by the significance of the *destination attributes* (discharge destination and postcode) in dictating departmental outcome requirements. For example, consider the occupational therapy (OT) department. Assume that required patient outcome significantly differs between the following destinations: nursing/residential home, home with family support, home without family support.

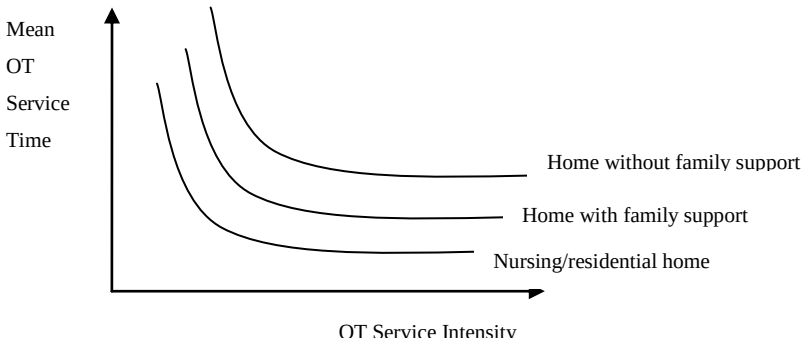


Figure 6.7 Relationship between occupational therapy service intensity and mean service time for a number of example destination levels

According to Figure 6.7 the required OT outcome for discharge to a nursing or residential home is less than that required for discharge to the patient's home with family support. This is intuitive since a patient discharged to a nursing or residential home is likely to be more reliant on assistance with everyday tasks such as self-care and cooking. The graph also reflects the further ability required should the patient be discharged home without family support.

The following variables may be defined as destination attributes;

- **Discharge destination**
    - **Complexity of home – *if destination is own home***

- **Family support**
- **Discharge postcode**

The variables family support and complexity of home could be Boolean or could be supported on some numeric scale. The value of these variables can affect the target destination level the patient is prescribed. For example, a patient cannot be assigned a level categorised as own home with family support if they do not have their own home or family support. The target destination level is also influenced by the attributes of the patient such as age, gender, condition etc (more on this in Ch 6.2.6). For example, it might be unreasonable to select home as a discharge destination for elderly or highly complex patients.

The inclusion of destination attributes within the model has the following advantages. Firstly, it improves accuracy. There can be significant differences in the discharge criteria for certain destinations and if this is not included then homogeneity among patients is incorrectly assumed. Secondly, the target destination level may be used as a variable when assessing the impact of what-if type scenarios. For example, the effects of increasing discharge to own home (where appropriate) could be studied. This is of particular interest when considering wider cost implications. Should patients be prescribed higher-order destination levels then the greater degree of outcome achieved would reduce the cost borne by social services post-discharge. These savings could be of significant magnitude for younger patients.

6.2.6 Treatment control factors

There are many factors that influence the care (treatment intensity, ALOS, outcome) that patients receive at Rookwood NRC. These factors can be used as part of a prognostic model to predict, say, LOS (see Ch 2.2.1) and as part of a queuing theoretic model.

Consider first a prognostic model. The aim here is to predict the value of a measure for a subject given a number of attributes pertinent to that subject. In Rookwood NRC the measure is ALOS and the attributes are the characteristics that relate individually to that patient. Since the most important time to make such prognoses is before[1] or at the time of[2] admission these characteristics must be captured pre or on-arrival. A review of literature relevant to the specification of such characteristics can be found in Ch 2.2.1. Many details of possible characteristics that could be used within the model are contained within this review. One of

[1] To make decisions about who to admit
[2] To inform social services (or other discharge destination) about expected discharge date

these studies (Perel et al, 2006) contains a survey of prognostic models for TBI patients. This is the largest[3] systematic review of such models to date. The authors find the following attributes to be the most common characteristics analysed as predictor variables:

- **GCS**
- **Age**
- **Pupil reactivity**

The condition of requiring factors pre or on arrival is relaxed when considering a queuing model. This is because it is unimportant whether the variable is a prospective predictor of LOS (e.g. condition on arrival) or a retrospective predictor of LOS (e.g. condition on discharge) – the only aim is to reduce the variance of LOS. Three approaches to include the effect of such factors within the model are considered. The first uses covariates to control for the effects of certain variables. In Faddy & McClean, 1999 the authors use age and year of admission as covariates in their study of a geriatric facility. The second approach is to use a Bayesian belief network to condition a distribution, such as in McClean & Marshall, 2003, with a three-term phase-type distribution. They find a relationship between admission method, age, gender, destination and Barthel grade. The third approach is to determine, by some means (e.g. CART[4] analysis), a number of homogenous patient groups for which distributions can be fitted to respective LOS data.

The advantages of including the effects of patient attributes are as follows. First and foremost, the queuing model is more accurate since the heterogeneity in LOS of patients is taken into account. In addition, what-if type scenarios relating to particular patient characteristics can be explored. For example, the effect of admitting more complex patients could be studied. For the prognostic model, the reduction in variance of predicted ALOS for different types of patient serves to reduce BLOS since a more accurate predicted date of discharge can be given to the discharge destination (which increases the likelihood they are ready for the patient). Finally, if data for these variables is collected pre-arrival the predictions can be used to assist admission decisions. For example, if high band therapists are over-stretched then the disastrous effect of admitting a complex patient could be shown.

[3] 66 models
[4] Classification and regression tree

These advantages come at the price of additional model complexity. The amount of data could also limit the effective identification of predictor variables.

6.2.7 Summary

In summary, a number of possible components have been considered as add-ons to the basic queuing theory model of Ch 6.2.1. Some of these components are free-standing and other build on the foundations set by other components. Several choices with regard to the selection of these components are depicted in Figure 6.8 where the number contained in each circle corresponds to the component contained in the subchapter extension of Ch 6.2. That is (1) corresponds to the queuing model of Ch 6.2.1.

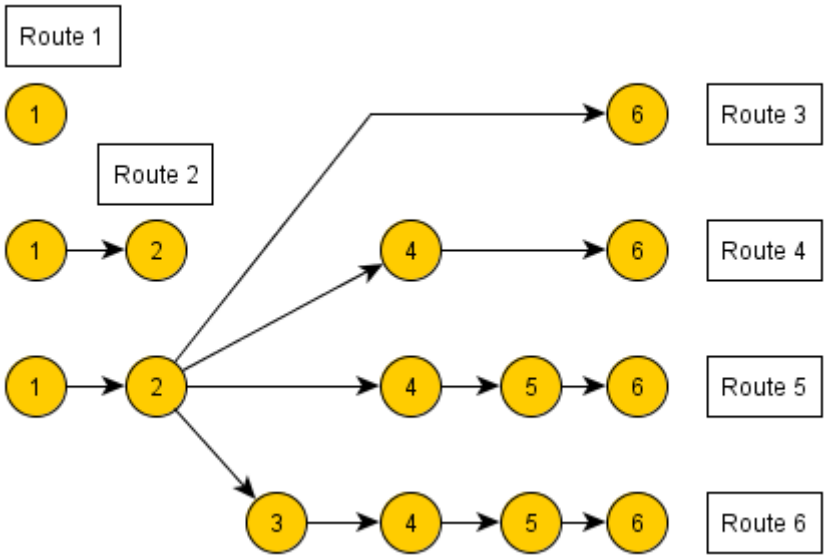


Figure 6.8 Possible types of model

It has been stated that models of greater complexity have the ability to better represent activities at Rookwood NRC. This infers more reliable prognoses of treatment, LOS and outcome through a reduction in the variance of such predictions. A complex design also has benefit with respect to what-if type scenarios. Not only can a broader number of these be considered but the variables used to control such studies are shown to be more relatable. However, these advantages come at the expense of a more complicated model in addition to the requirement for a significant amount of data. Therefore all of these factors must be taken into account when considering model design.

Finally, the capacity of Rookwood hospital is determined by the availability of therapists and the demand of patients. The facility is said to be operating at full capacity if the demand is

equal to the availability, that is, if there is no excess demand or no excess availability. It is important that decisions are made with respect to a consideration of capacity since resources are limited (and quite often demand exceeds supply).

6.3 Data Collection

In order to construct any of the models introduced in the previous subchapter a separation between the analysis of arrival and service data is made. The Arrivals dataset (Ch 6.3.1) contains the variables that are pertinent to the arrival of referrals and the behaviour of the queue whilst the Service dataset (Ch 6.3.2) contains variables that are relevant to the admission and discharge of patients.

Table 6.1 Availability of data

Variable	Source (raw dataset)	From	To
Date referral ready for transfer to RWD	Referrals dataset	January 2010	April 2011
Date referral removed			
Referral outcome			
CRN number	Physio Notes dataset	January 2003	May 2011
Admission number			
Date of birth			
Admission date			
Discharge date			
Gender	NHS Trust data request QH556	January 2003	January 2011
Home postcode			
Primary diagnosis description/code			
Referring practice			
Admission source description			
Admission method description			
Discharge destination description			
Date ready for discharge	Therapies database	January 2007 (*incomplete*)	January 2011 (*incomplete*)
Weekly total for physiotherapy received			

It is necessary to strike a balance between the quality and quantity of data. The Physio Notes dataset does, in fact, stretch back to discharges since 1990. Whilst this would result in a larger amount of data the relevance of this data would be questionable (since it is likely that clinical practices have changed over such a large time). Such problems can be averted by considering data on a smaller timescale. However, with an average LOS of roughly five months and only about 21 beds the throughput of the unit is insufficient to capture enough observations in a limited amount of time. It has been decided that eight years (2003-2011) would give a sufficient quantity of data whilst ensuring that practices had not changed too much (the current superintendent physiotherapist has been employed since 2003).

6.3.1 Arrivals dataset

This dataset, stored as an MS Excel file, is constructed from just one source (raw dataset).

Referrals dataset

Every Thursday of each week since January 2010 the Referrals dataset has captured a snapshot of the waiting list for Rookwood NRC. This list contains the names of patients referred and whether they are ready for transfer or not. For this project, an arrival is defined as the submission of a referral that is ready or the instance at which a non-ready referral becomes ready. This is since it is only the ready referrals that are considered for admission. As mentioned in Chapter 1.2 the size of the waiting list can be reduced by either an admission or the removal of a referral. A referral is removed if the patient is transferred, discharged or deceased. For the purposes of this project, a ready referral can also be removed from the ready waiting list, or queue, if the patient deteriorates and becomes not-ready. Note that patients who never become ready are not of interest.

Data is extracted as follows. The weekly waiting lists are inspected chronologically and the names corresponding to each ready referral are recorded. For each of these a date of arrival is found in addition to the date at which the referral is removed from the ready waiting list and the associated outcome. Observations are thereafter sorted by date referral received.

6.3.2 Service dataset

This is too stored as an MS Excel file and is comprised from the other (three) sources.

Physio Notes dataset

This dataset forms the foundation of the service dataset. It contains the variables given in Table 6.1 for each patient episode where columns represent variables and rows observations.

Columns representing data from other variables are removed. Data is sorted by discharge date.

NHS Trust dataset

Following a data request in early 2011 the QH556 dataset has been received as an MS Excel worksheet. This dataset contains variables (mentioned in Table 6.1) that can be used as predictors of LOS/outcome. Such details are extracted for as many of the patients on the Physio Notes dataset as possible by matching CRN number, surname and date of birth. Since not all patients on the Physio Notes dataset appear on the NHS Trust dataset the patients that data is available for is a subset of the total number under consideration. The service dataset now contains a list of patients from the Physio Notes dataset, many of which have additional fields obtained through the NHS Trust dataset.

Therapies dataset

A database capable of collecting data for all variables outlined in Ch 6.2 (including date ready for discharge and treatment intensity) has been developed in MS Access. Databases offer *'improved reporting, operational efficiency, interdepartmental communication, data accuracy, and capability for future research'* (Vreeman et al, 2006). The motivation for developing a database is centred on the deficiency in variables that existing data has been collected for; particularly those relating to treatment intensity. The aim is to be able to collect prospective data for all such variables whilst being able to store historic data for the existing variables.

Firstly, the relationships between the variables were determined. This is necessary to ensure consistency between the electronic tables which store the data. Appendix 6.1 contains a screenshot of the 'relationships' in MS Access. Secondly, the GUI was developed. This was handled by one of the technical assistants at Rookwood hospital. And thirdly, the database was debugged to ensure stability.

As well as collecting data it was hoped that the database would be used as part of the scheduling process (Figure 6.9). When the PTO is reset (at the beginning of the week), a list of names, initials and CRN numbers for current patients would be copied from the database to the PTO. Whenever availability data is input to the schedule from the PTO these details would also be transferred. Finally, when the 'Export database' button is pressed (see Figure

4.2), the summary information pertinent to physiotherapy treatment received and requested would be transferred from the schedule to the database.

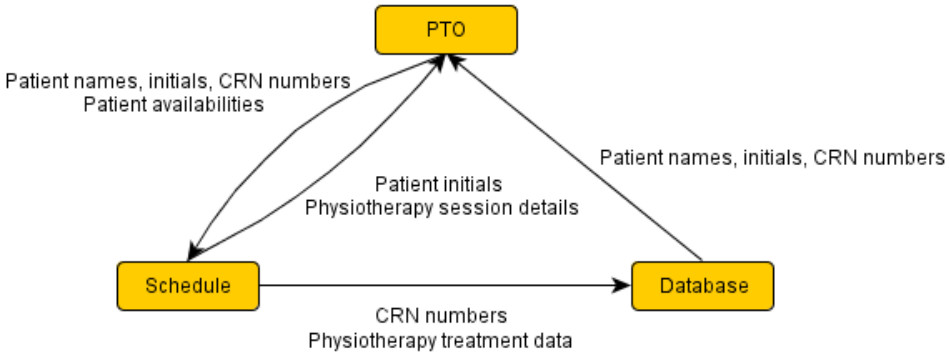


Figure 6.9 Transfer of data between the PTO, schedule, and database

However, despite our best efforts the database was not used to this end. In fact, the overall take-up of the database was poor. This is attributed to a lack of accountability of those tasked with inputting data in addition to a resentment to change. The data that was available for date ready for discharge and treatment intensity is added to the service dataset for the relevant patient episodes.

6.4 Preliminary Analysis of Data

The statistics and graphs contained in this subchapter are automatically produced[5] from the arrivals dataset and the service dataset which are stored as Excel files. This allows for fast and simple analysis (including updated parameter estimates) as and when new data is input.

6.4.1 Arrivals dataset

In the sixteen months from January 2010 until April 2011 there have been **183 ready referrals** (see Ch 6.3.1) in the queue for Rookwood NRC. Figure 6.10 depicts the arrival and removal date of each of these. Note that fifteen observations are left-censored, twelve are right-censored, and no observation is both left and right censored.

[5] Through the activation of some purpose-written VBA code

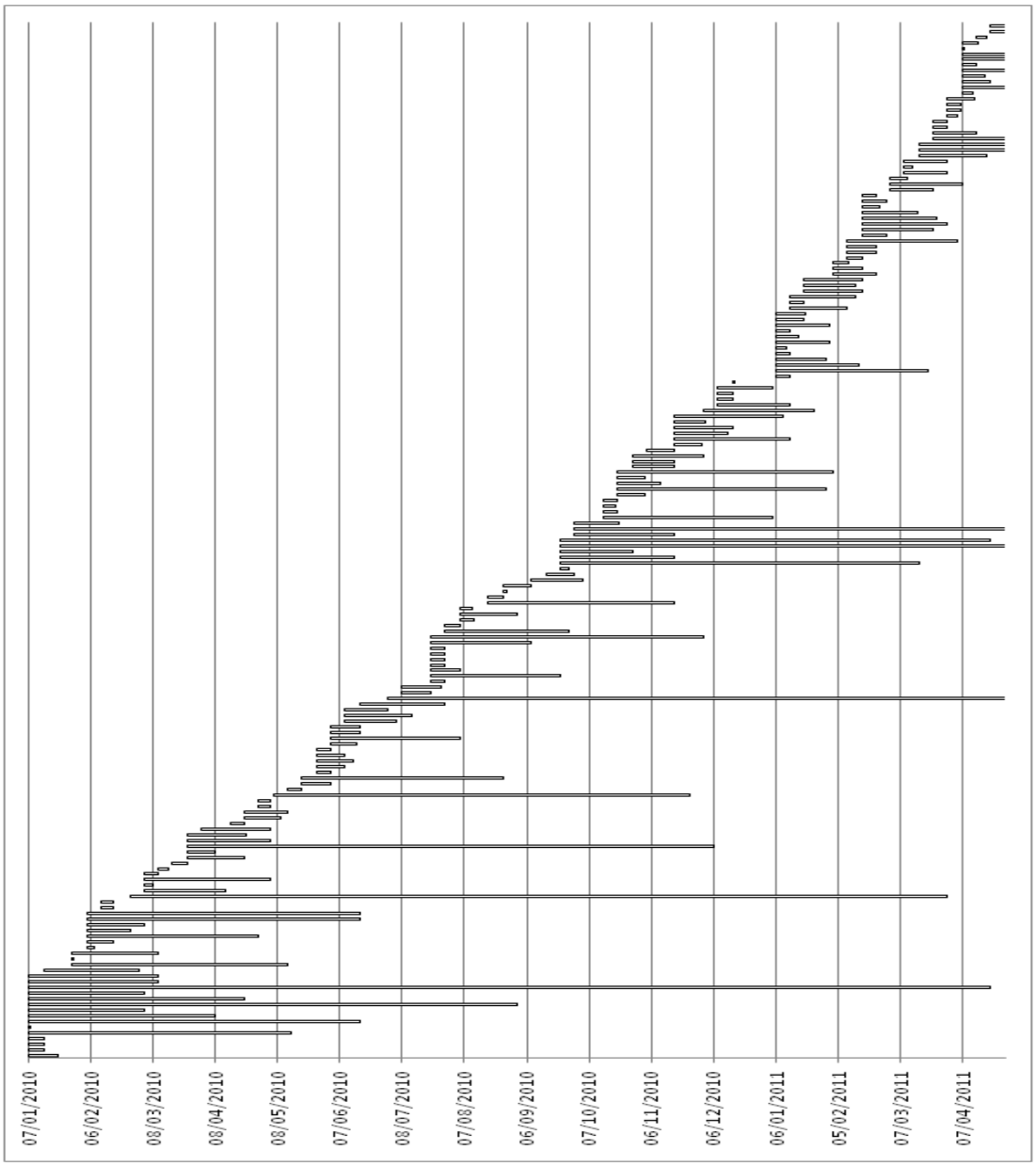


Figure 6.10 The arrival and removal of ready referrals over time

The mean arrival rate is deduced by the total number of arrivals in a given interval of time divided by the length of that interval. The first week is excluded since it is not known how many of the fifteen referrals are new arrivals for that week. With 168 arrivals in the remaining 462 days there are **0.36364 arrivals per day** or roughly 2.5 arrivals per week.

Quite often in the study of queuing theory, arrivals are modelled by a Poisson process, i.e. exponentially distributed inter-arrival times (see Ch 1.5). Such a decision is typically motivated by mathematical tractability and appropriateness – since many arrival processes are inherently random. The appropriateness of the exponential distribution to model inter-arrival times for Rookwood NRC is forthwith investigated. The specific time of arrival of

each referral is not recorded in the referrals dataset. All that can be deduced from this is the week in which referrals are added to the ready waiting list. The appropriate unit of time to consider is therefore the week. The concept of bulk arrivals is eminent here since more than one arrival can occur at a particular instant of time. However, it must be noted that in reality the arrival of ready referrals occur continuously in time.

First, the 167 inter-arrival times between each of the 168 referrals are found. The minimum number of weeks between arrivals is zero whilst the maximum is three. Since the time between arrivals is discrete and not continuous the geometric distribution is used to test for randomness[6]. The *pdf* of this distribution is given by $p(1-p)^k$ on support $0 < p \leq 1, k \in \mathbb{Z} \geq 0$ and with mean $p^{-1} - 1$. The empirical mean inter-arrival time is approximately 0.4 and so p is equal to five sevenths. Figure 6.11 is thus obtained.

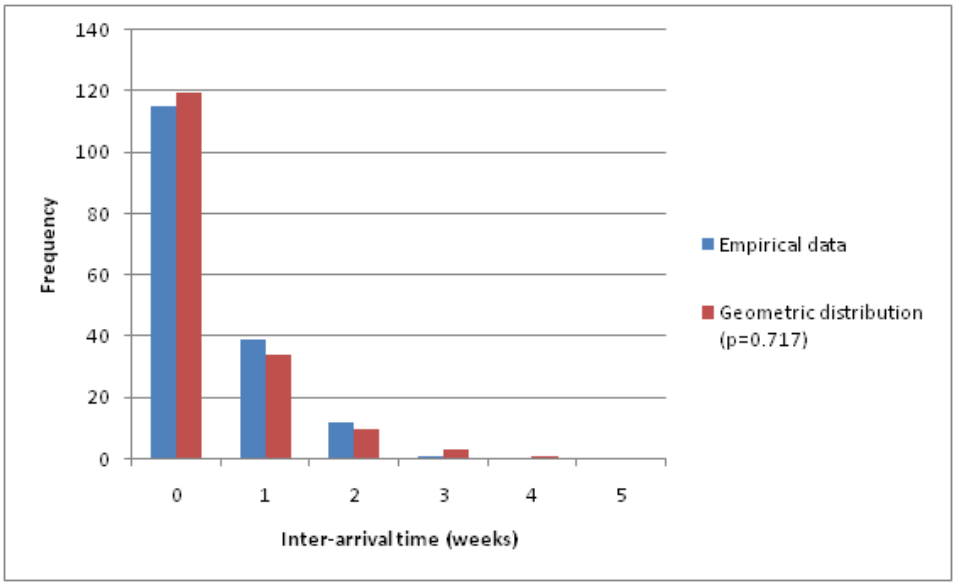


Figure 6.11 Frequency plot for empirical data and geometric distribution

The empirical data shows sufficient[7] congruence with the geometric distribution to conclude that arrivals occur at the queue (*ready waiting list*) at random and thus in line with a Poisson process.

Since a date is stipulated for the removal of ready referrals from the queue the appropriate unit to consider in evaluating departures is the day. Table 6.2 provides some basic statistics

[6] This is because the geometric distribution is the discrete equivalent of the exponential distribution. It too attains the memoryless property
[7] Visually; not statistically tested

for the **length of time spent in the queue**. Note that here the (twelve) right-censored observations are excluded since the reason for their removal is unknown.

Table 6.2 Analysis of length of time spent in queue

Outcome:	RWD		Other		Total
Censored:	left-cens	non-cens	left-cens	non-cens	Both
n	3	40	12	116	171
mean	110	27.4	94.67	34.52	38.4
s.d.	97.68	41.04	122.58	52	61.65
min	1	0	7	0	0
max	238	257	469	399	469

Approximately[8] **one quarter** of ready referrals result in **transfer to Rookwood NRC**. Excluding censored data such referrals spend, on average, one week fewer in the queue than those that are removed for other reasons. A program in Visual Basic has been written to determine the number of referrals in the queue on each of the 470 days. The midnight census is said to take place at the end of each day, i.e. new referrals are included but those removed are not. Figures 6.12 displays a line graph of the number in queue over time and Figure 6.13 contains a histogram of the number in queue.

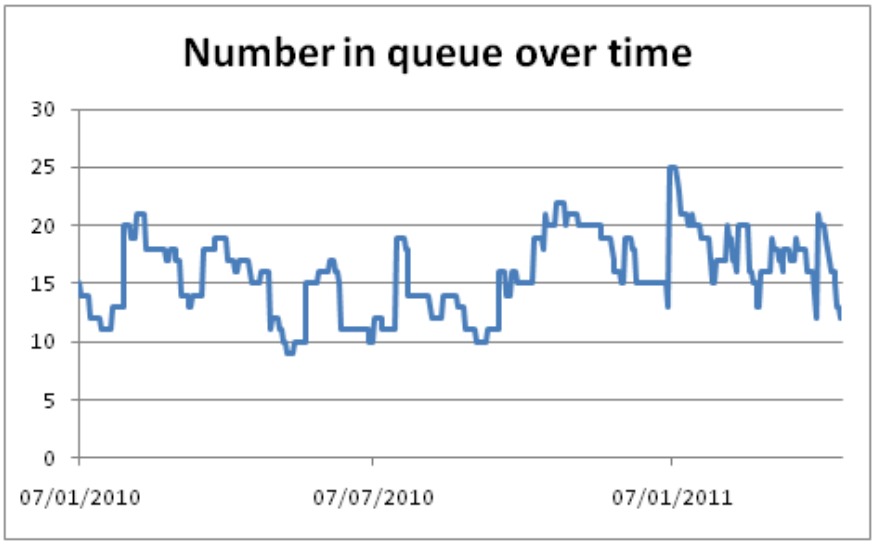


Figure 6.12 Number in queue from January 2010 to April 2011

[8] 25.64% non-censored data only; 25.15% non-censored and left-censored data

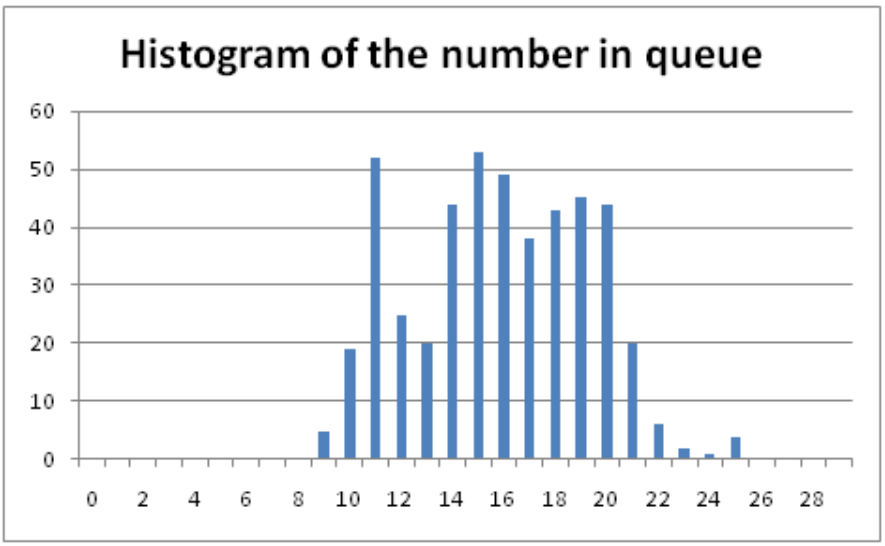


Figure 6.13 Histogram of the number in queue from January 2010 to April 2011

The **average number of ready referrals in the queue** is **15.83** with a standard deviation of 3.43. This gives a **coefficient of variation** of **0.217**.

It is presumed that the queue size has an effect on whether referrals are removed for reasons other than transfer to Rookwood NRC. That is, if there are many referrals in the queue then the likelihood of *reneging*[9] is larger than when there are fewer. To investigate this postulation the number in system (at the beginning of each day) and the reason for removal (transfer to Rookwood or other) is considered for each removal. The probability of reneging is therefore deduced.

[9] Reneging occurs when queuing customers leave before entering service

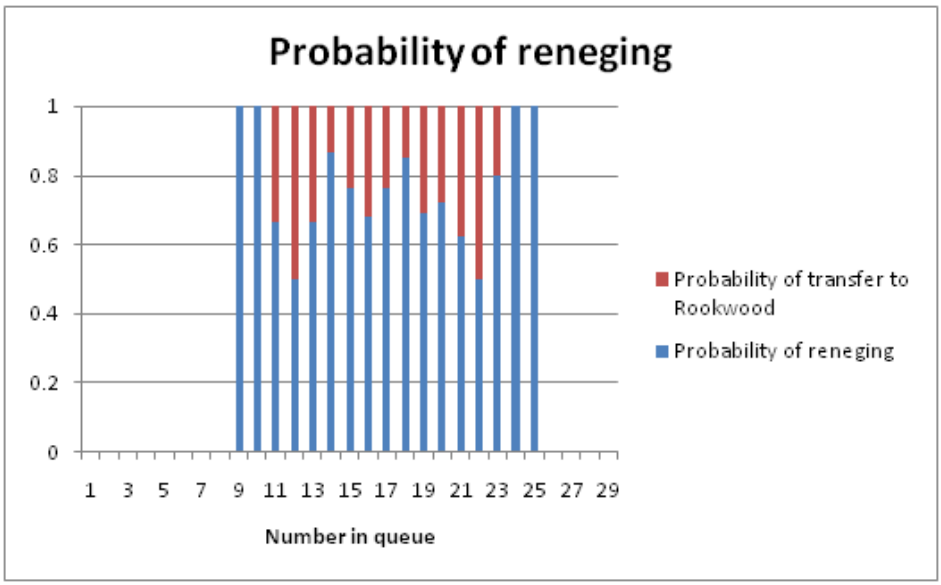


Figure 6.14 Probability of reneging based on the number in queue

Figure 6.14 provides no evidence to support the earlier postulation: there appears to be no relationship between reneging and queue size. The relationship between the probability of reneging and the length of time spent in queue is now investigated.

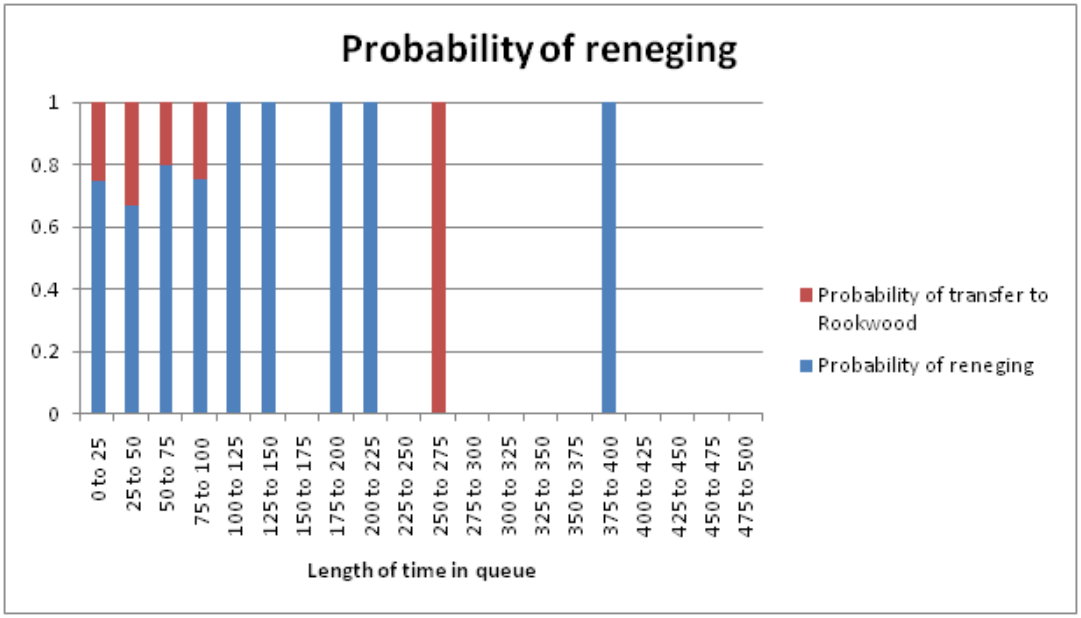


Figure 6.15 Probability of reneging based on length of time spent waiting

However, as can be seen from Figure 6.15, there is no conclusive trend.

It must finally be noted that whilst the data indicates the continuous presence of a queue over the sixteen months, it cannot be concluded that all available beds were occupied during this

period. It could well be that the availability of therapists or other resources was in fact the limiting factor of the occupancy at Rookwood NRC (see Ch 1.2).

6.4.2 Service dataset

The VBA program that derives statistics and graphs from the service dataset allows the user to specify an upper and lower bound on LOS for episodes that are used in the analyses. The lower limit is set to exclude the episodes of patients who were discharged on the same day as their admission. Those with LOS of 1,000 days or more are defined as residents and are also excluded from the analyses.

6.4.2.1 Summary statistics

There have been 384 patient episodes from the beginning of 2003 until May 2011. Twenty-two of these do not have NHS Trust data (i.e. gender, diagnosis etc) and a further four have LOS greater than or equal to 1,000 days. No patient episode has a LOS of zero days.

The **number of patient episodes** used for this analysis is therefore **358**.

Table 6.3 details the basic statistics for **patient episode LOS** (measured in days).

Table 6.3 Basic statistics

	mean	*s.d.*	m_1	m_2	m_3	m_4	m_5
LOS	149.3	152	149.3	45408.4	20986875	1.25×10^{10}	8.63×10^{12}

Note m_k is the *k*-th non-central sample moment. The **coefficient of variation** is **1.018**. The shortest LOS is one whilst the longest is 933 days. The skewness is 2.1 and the excess kurtosis is 5.5. Figure 6.16 displays the annual average LOS for those discharged in the year specified.

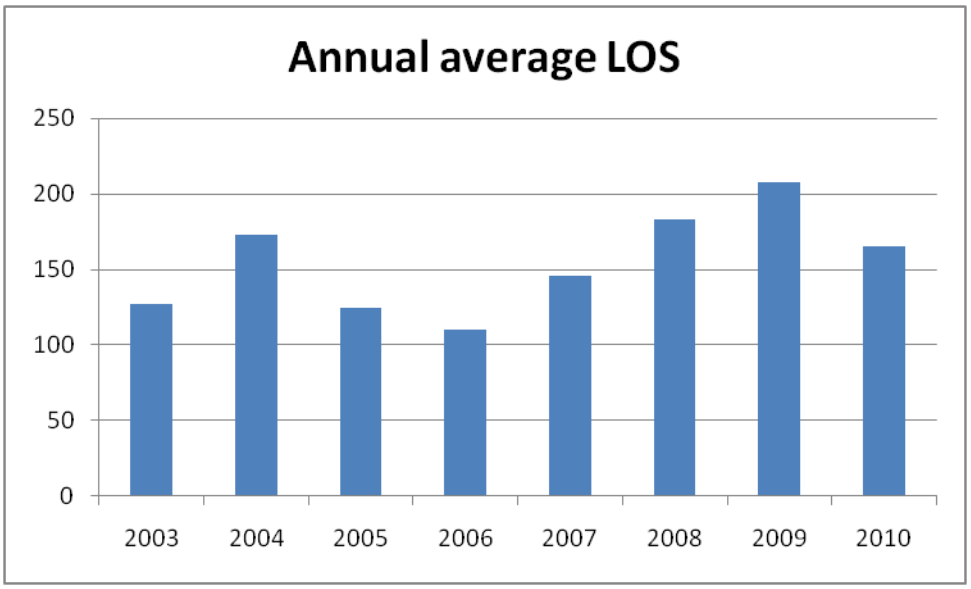


Figure 6.16 Average length of stay for each year from 2003 to 2010

The mean number of episodes per patient is 1.03. The maximum number of episodes for a patient is three. Basic statistics are produced for stand-alone episodes (i.e. one patient, one episode) and those that are part of a number of episodes (i.e. one patient, more than one episode) in Table 6.4.

Table 6.4 Basic statistics for stand-alone and multiple episodes

Episode type:	Stand-alone	Multiple
n	338	20
mean LOS	151.8	107.8
s.d.	154.1	103.4
min	1	10
max	933	455

Table 6.5 Basic statistics for multiple episodes for first, second, and third admission

Episode #:	1	2	3
n	9	9	2
mean LOS	103.2	132	19
s.d.	67.7	129.7	1
min	38	11	18
max	308	455	20

Nine patients have multiple episodes. Therefore, there are **347 patients** in the dataset.

The **mean Age on Admission** is **44.4 years** with a standard deviation of 15.8. The patient episodes are sorted into appropriate bins. For each bin there is a specification of the number of patient episodes, the mean age on admission and the mean LOS.

Table 6.6 Basic statistics for episodes categorised by age on admission

	0 – 9	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79	80 – 89	90 – 99	100 – 109
n	0	18	58	65	79	70	50	13	4	1	0
mean age	.	17.6	24.3	35.1	45	53.9	63.5	37.5	84	92	.
mean LOS	.	185.7	106.8	145	158.5	156.1	176.1	136.5	110.3	17	.
s.d.	.	157.1	112.9	154.7	140.8	171.6	173.1	122.2	58.5	.	.
min	.	6	1	7	4	4	2	7	17	17	.
max	.	611	542	933	780	918	757	455	178	17	.

The following procedure is employed (in VBA) for the remainder of variables that are studied. First, the unique values for the variable under consideration are determined. For each of these a tally is produced for the number of patient episodes that attain this value. Mean values of LOS are also calculated.

Table 6.7 Basic statistics for episodes categorised by gender

	Male	Female
n	218	140
mean LOS	144.9	156.2
s.d.	146.1	160.6
min	1	2
max	933	918

Table 6.8 Basic statistics for episodes categorised by local health board

	Cardiff & Vale	Aneurin Bevan	Cwm Taf	Powys	Hywel Dda	Abertawe Bro Morgannwg	Betsi Cadwaladr
n	187	91	48	11	7	6	5
mean LOS	155.8	145.7	134.2	112.3	193.7	1355.3	190
s.d.	162.4	134.1	161.9	92.3	167	64.9	113.7
min	1	6	4	25	10	35	36
max	933	735	780	329	472	211	353

Table 6.9 Basic statistics for episodes categorised by primary diagnosis (top 7; $n \geq 10$)

	ABI	MS	TBI	Anoxia	GBS	Other	MS & other neuropathies
n	187	44	42	26	22	14	10
mean LOS	172.8	95.7	132.8	193.1	99	104.6	83.5
s.d.	170.3	100.9	126.8	155.1	88.3	91	131.2
min	4	2	6	17	9	4	17
max	933	488	542	561	415	347	472

Table 6.10 Basic statistics for episodes categorised by admission source

	Cardiff & Vale hosp	NHS hosp	Usual residence	Temp	Unknown
n	229	78	49	1	1
mean LOS	146.3	200.1	73.6	352	387
s.d.	148.1	172.4	85.1	.	.
min	4	6	1	352	387
max	933	780	349	352	387

Table 6.11 Basic statistics for episodes categorised by admission method

	Transfer	Waiting list	A+E	Emergency (GP)	Emergency (OP clinic)	Other immediate
n	305	43	4	2	2	1
mean LOS	160.4	78	116	39	63.5	387
s.d.	156.7	97.9	68.6	1	27.5	.
min	4	1	16	38	36	387
max	933	352	184	40	91	387

Table 6.12 Basic statistics for episodes categorised by discharge destination (top 5; n>5)

	Usual residence	Internal transfer (gen)	Other Trust (gen)	Non-NHS nurs. home	Temporary residence
n	212	61	50	15	6
mean LOS	100.5	253.8	208.2	206.9	114.7
s.d.	109.1	215.4	136.6	111.5	135
min	1	16	7	25	28
max	735	933	561	408	413

This data is of great interest since it shows that, despite large variances, there is a clear difference in mean LOS between patient episodes of certain groups. For example, a patient in their twenties can expect to spend less time in RWD than any other patient below the age of ninety on admission (Table 6.6). Also, those admitted by transfer can expect to have a LOS over twice that of those admitted by waiting list (Table 6.11). Conversely, gender appears to have little effect on patient episode LOS (Table 6.7).

To determine the statistical significance of such differences a number of hypothesis tests can be performed. A t-test can be used to assess whether there is a difference in mean LOS between two groups whilst an ANOVA test can be used to assess the difference between multiple groups. As an example, the data in Table 6.4 is used to test whether there is a difference in mean LOS between stand-alone episodes and those that are part of multiple episodes.

Therefore, $n_1 = 338, \bar{x}_1 = 151.8, s_1 = 154.1, n_2 = 20, \bar{x}_2 = 107.8, s_2 = 103.4$. The test statistic for Welch's t-test is therefore $t = \left(\bar{x}_1 - \bar{x}_2\right) \Big/ \sqrt{s_1^2 n_1^{-1} + s_2^2 n_2^{-1}} = 1.789$. The degrees of freedom is calculated by the formula $v = \left(s_1^2 n_1^{-1} + s_2^2 n_2^{-1}\right)^2 \Big/ \left(s_1^4 n_1^{-2}\left(n_1 - 1\right)^{-1} + s_2^4 n_2^{-2}\left(n_2 - 1\right)^{-1}\right) \approx 24$. For the two-tailed test ($H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 \neq \mu_2$) the critical region (95%) is $|C| > 2.064$. Since the test statistic is not within this region the null hypothesis is not rejected – there is insufficient evidence to suggest a difference in means. The *p*-value is 0.086 (since $z_{24,1.789} = 0.043$). This means that there is only a 8.6% chance that the means are the same.

6.4.2.2 Occupancy

All 384 patient episodes from 1st January 2003 until May 2011 are used to obtain daily figures for bed occupancy (in addition to the current patients as of May 2011).

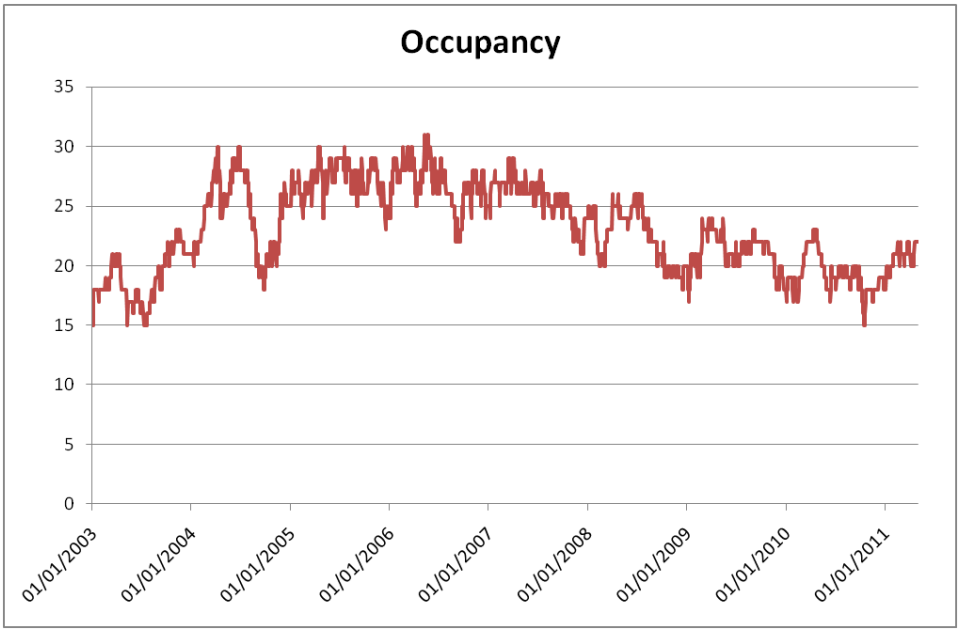


Figure 6.17 Bed occupancy from January 2003 to May 2011

The minimum number of beds in use is fifteen whilst the maximum is 31. The **mean number of occupied beds is 23.16**. Observe the below-average occupancy of 2003 and, to a lesser extent, 2005. This is as a result of ward closures due to refurbishment. During these times the NRC operated the same workforce across only one of the two wards. Interestingly enough, the annual average LOS was less at these times than others (see graph in previous subchapter – note 2006 is also low, probably due to a time lag as figures represent LOS of those discharged in that year). This supports the notion that treatment intensity does indeed have an effect on LOS.

From mid 2005 to mid 2007 it was commonplace for over 25 beds to be in use at any time. However, since then, due to the restructuring of health services in Wales, the NRC at Rookwood hospital has had to cater for a more complex type of patient. With no increase in personnel, this has meant that the bed occupancy has reduced. In recent years there have been, on average, about 21 patients in the unit at any time.

6.4.2.3 Treatment intensity

Date ready for discharge and treatment intensity are two very important variables that enable the model to progress further than the basic version characterised by Route 1 of Ch 6.2.7. If data is available for date ready for discharge then the LOS can be partitioned to its active and blocked component (Ch 6.2.2). If, in addition, data is available for treatment intensity then the effects of this influential variable can be appreciated (Ch 6.2.4).

Unfortunately, such data has not been collected over the years at Rookwood NRC. What is available, however, is the physiotherapy date ready for discharge and the physiotherapy treatment intensity. This being said both sets of data are only available for 89 of the 358 patient episodes considered in the analysis. This is because such data has not been routinely collected over the years. The data that is available has been obtained through either the database (see Ch 6.3.2) or clinical estimation (so not only the quantity but the quality is questionable). Figure 6.18 displays a scatter plot of physiotherapy treatment intensity versus physiotherapy active LOS (where physiotherapy active LOS is the difference between date of admission and physiotherapy date ready for discharge).

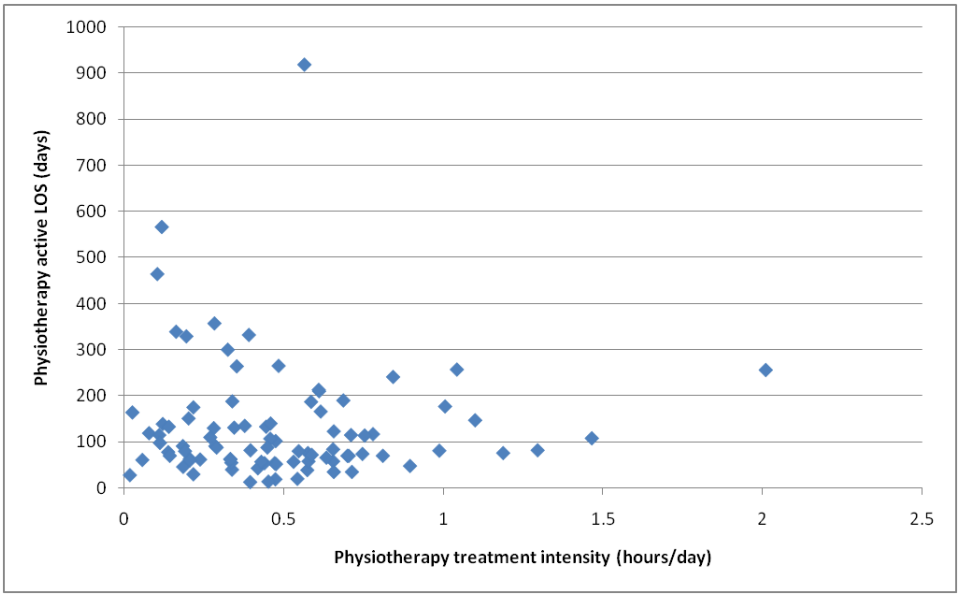


Figure 6.18 Relationship between physiotherapy treatment intensity and active LOS

As postulated in Ch 6.2.4 the underlying shape is similar to that of Figure 6.5.

If the relationship between active LOS and treatment intensity is to be included within the model then one of two assumptions must be made. First, it could be assumed that any change to physiotherapy treatment intensity is replicated by the other departments. For example, if

physiotherapy treatment intensity increased by 20% then it would be assumed that other departments replicate this change for the effect on active LOS to be representative. Second, it could be assumed that the physiotherapy active LOS is approximately equal to the (overall) active LOS. In which case, the effect of varying physiotherapy intensity would be the same for (overall) active LOS as it would for that of the physiotherapy department. Clinicians at Rookwood NRC have some affinity to this notion as it is typically the patient's motor functioning that dictates discharge readiness.

6.5 The Distribution of Length of Stay

Appropriate statistical distributions (see Chapter 5) are fitted to the LOS of the 358 patient episodes detailed in Ch 6.4.2.1.

A graph is produced for each approximating distribution considered. Empirical data is represented by small fixed-width columns whose height is calculated by the number of observations contained in the interval divided by the product of interval length and total number of observations. The approximating distribution is represented by a line graph. Log-likelihood values are deduced for each approximation. However, this measure should be treated with caution since it can promote over-fitting. The Akaike Information Criterion (Akaike, 1974) is therefore considered. This statistic consists of two parts: the first rewards goodness of fit whilst the second penalises the use of an excessive number of parameters. If L is the Likelihood function and k is the number of parameters then $AIC = -2\ln(L) + 2k$. The Bayesian Information Criterion (Schwarz, 1978) is similar but contains a larger penalty. If n is the number of observations then $BIC = -2\ln(L) + k\ln(n)$. Unlike log-likelihood the lowest value is sought. The chi-square test statistic is also deduced. These four statistics can be used to compare the appropriateness of approximating distributions.

For the graphical representation of the empirical data the lower and upper bounds are zero and 500 and the interval width is ten. Optimal parameter values for the calculation of the chi-square test statistic are now sought. Sturges, 1926 predicts the number of classes, k, by the result $k = \lceil \log_2(n) + 1 \rceil$ where n is the sample size (358), ergo, $k = 10$. Scott, 1979 predicts the class width, h, by the result $h = 3.49\sigma n^{-1/3}$ where σ is the standard deviation (152), thus, $h = 74.7$. To avoid sparsely populated classes the (already truncated) data can be curtailed above at about 500. Over 95% of the data is used since there are only thirteen patient episodes with $500 \le LOS \le 1{,}000$. Therefore $k \approx 7$ according to Scott's result. However, by

considering the corresponding histograms it was observed that these values of *k* were insufficient in adequately representing the features of the data. A more effective and intuitive approach is to survey the histograms produced by various parameter values that are considered appropriate. A lower and upper bound of zero and 510 and a class width of 30 are found to sufficiently represent the features of the sample data. The number of classes is therefore 17.

6.5.1 The Matlab fitting program

A purpose-built program has been developed in Matlab to deduce parameter estimates for approximating distributions.

First, the LOS data (obtained through the VB program) is copied to a single column within a '.txt' file. The program is then initialised by providing the upper bound and interval length for the graphical representation of the empirical data and the upper bound and number of classes for the calculation of the chi-square test statistic. A description of the approximating distribution must be provided. If the distribution is non phase-type then the pdf is input. Otherwise, it is more convenient to input the *initial probability vector* and *PH-generator matrix* (see Ch 1.5).

It is then necessary to determine parameter estimates for the approximating distribution. Maximum likelihood (ML) estimators are determined through the Matlab optimiser *fminsearch*. This uses a derivative-free simplex method (similar to that of Ch 5.3.3) to determine the variable values that minimise the value of a function. In this case, the function is the negative log-likelihood of the distribution under consideration. To eliminate the possibility that inappropriate parameter values are returned the optimiser *fminsearchbnd*[10] is used. This allows upper and lower bounds to be placed on the parameters. The minimal value of the function under these constraints is then sought. However, good[11] initial values of the parameters are required by the optimiser. These are obtained from the method of moments (MOM) parameter estimates.

There is no feature within the program to computationally derive the MOM estimators. They can be calculated through the results of Chapter 5. The densities corresponding to these two sets of estimators are plotted on the graph that contains the empirical data. Results for the four goodness of fit measures are also output.

[10] Developed by John D'Errico
[11] See note at the end of Ch 5.3

6.5.2 Non phase-type

Parameter estimation by ML and MOM is possible for each of the distributions considered here. The distributions are fitted by the purpose-built Matlab program for non phase-type distributions. The MOM estimators are derived from Ch 5.4.

6.5.2.1 Log-normal

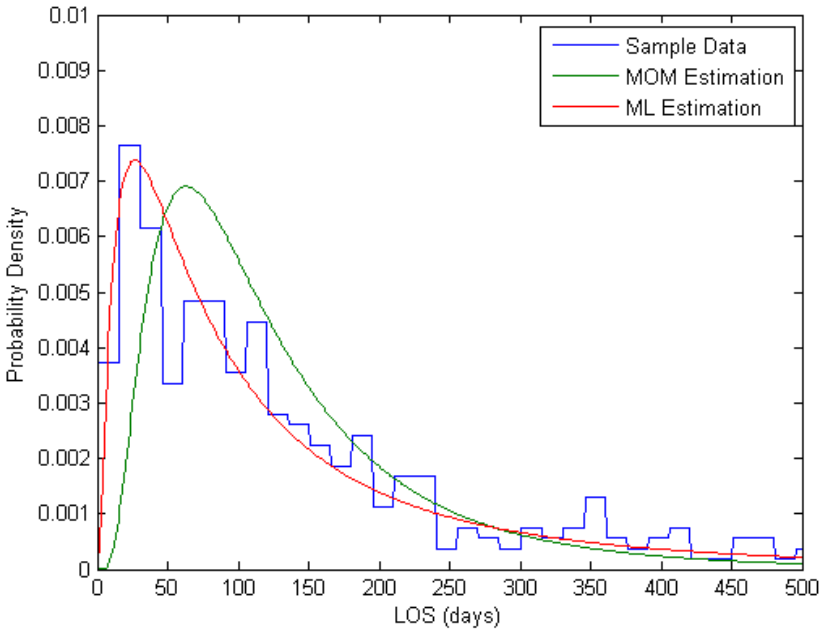


Figure 6.19 Log-normal approximation to length of stay

6.5.2.2 Gamma

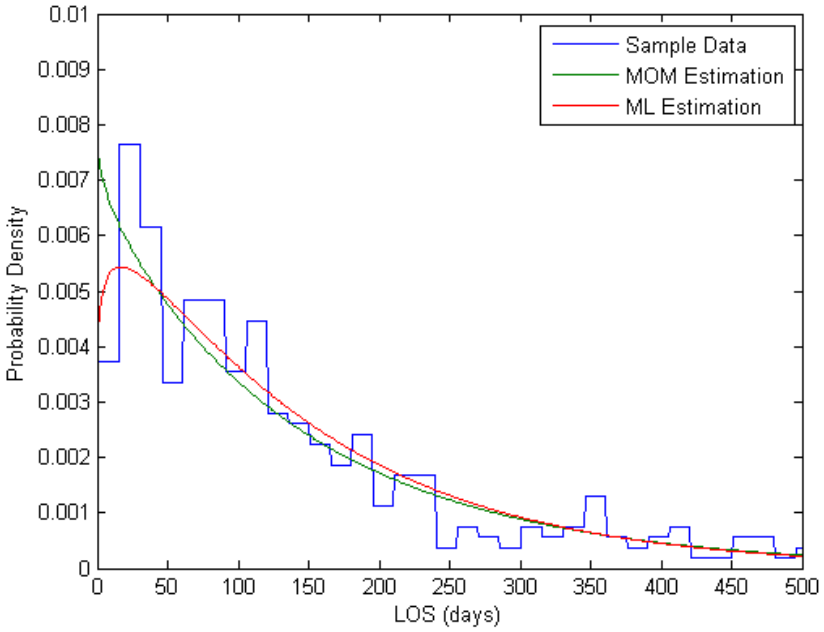


Figure 6.20 Gamma approximation to length of stay

6.5.2.3 Weibull

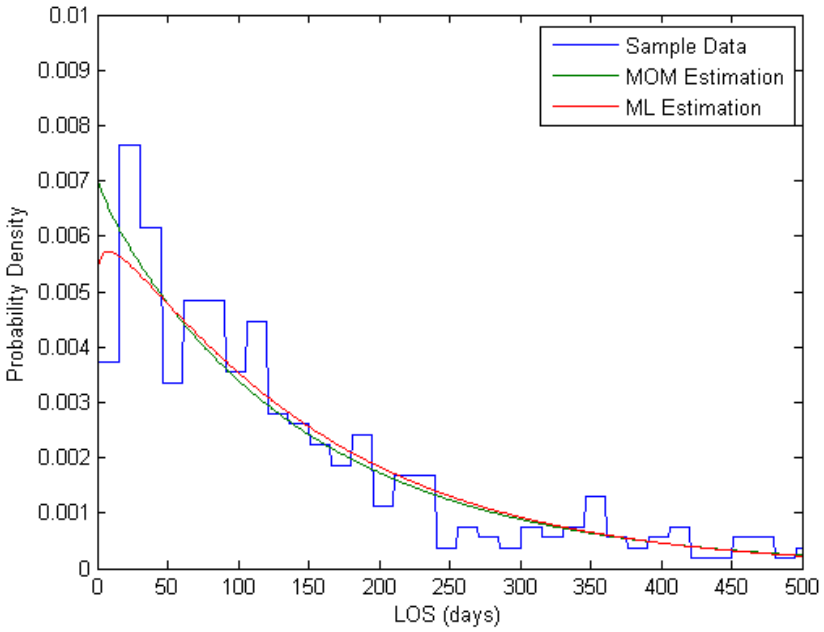


Figure 6.21 Weibull approximation to length of stay

6.5.2.4 Log-logistic

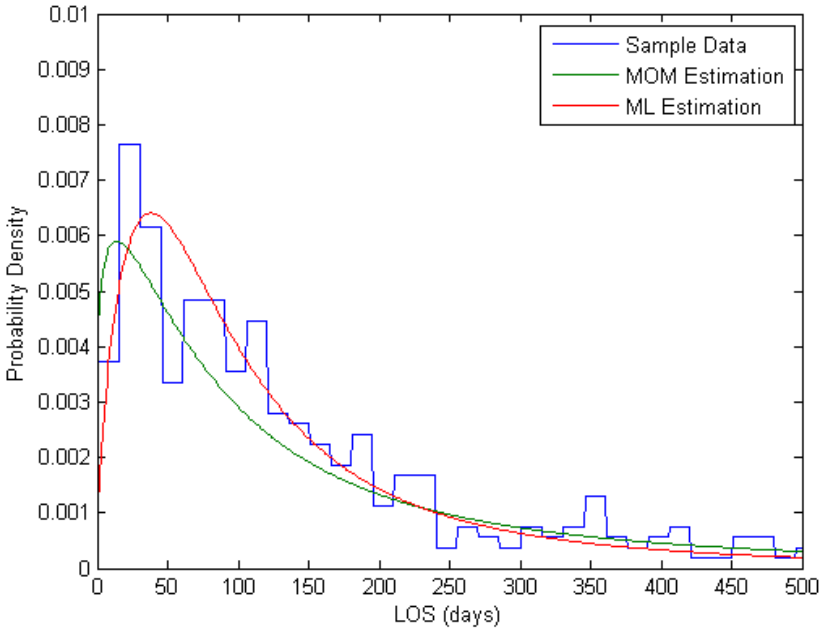


Figure 6.22 Log-logistic approximation to length of stay

6.5.3 Phase -type

The distributions considered here are fitted by the purpose-built Matlab program for phase-type distributions. The MOM estimators are derived from Ch 5.5.

6.5.3.1 Exponential

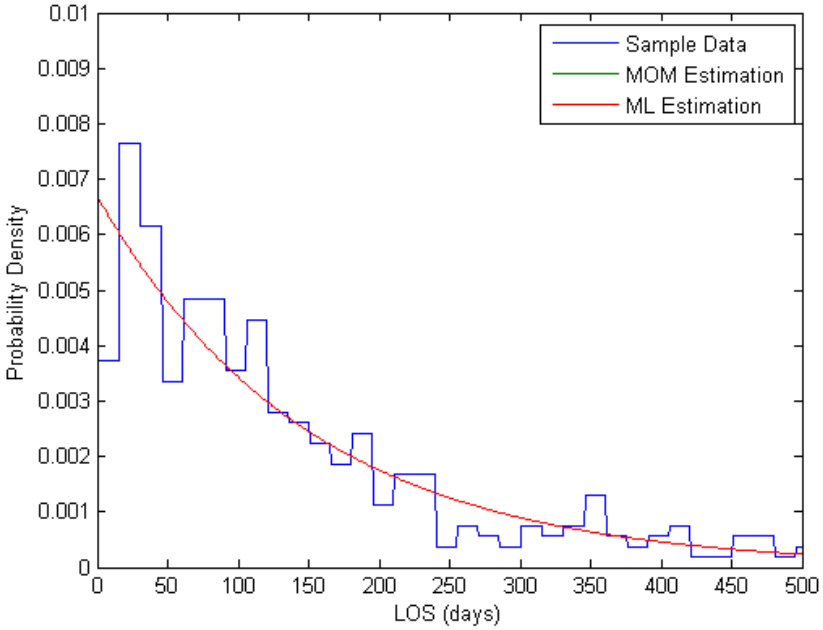


Figure 6.23 Exponential approximation to length of stay

6.5.3.2 Hyper-exponential (two-term)

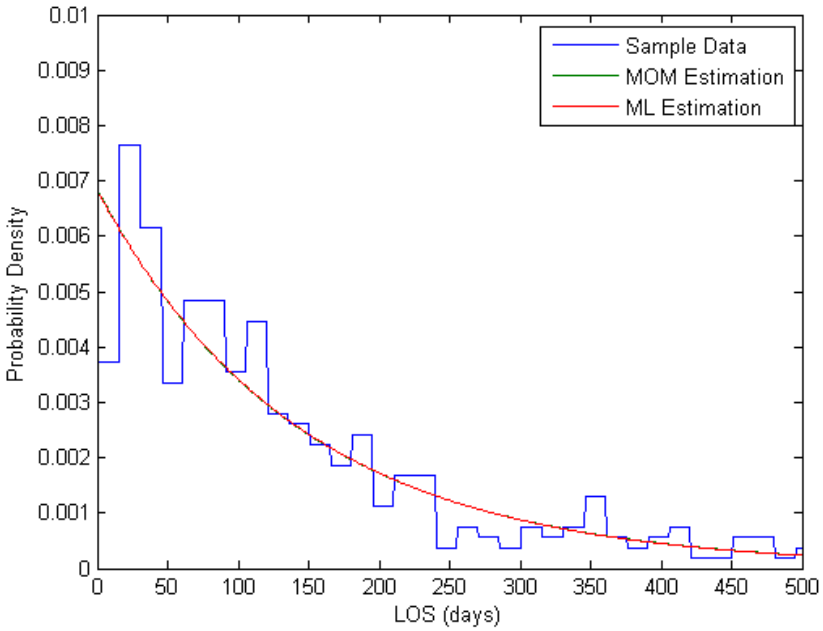


Figure 6.24 Two-term hyper-exponential approximation to length of stay

6.5.3.3 Hyper-exponential (three-term)

No analytic results are detailed in Ch 5.5 for the MOM estimators of this distribution. Instead of deriving these, the approach detailed in Ch 5.5.5.1.2 is employed, i.e. minimising the distance between the sample and theoretical moments. The *fminsearchcon*[12] algorithm in Matlab is used to this end. However, like *fminsearchbnd* this requires a specification of initial parameter values. A number of appropriate choices are experimented with including those based on the estimators of the two-term hyper-exponential distribution.

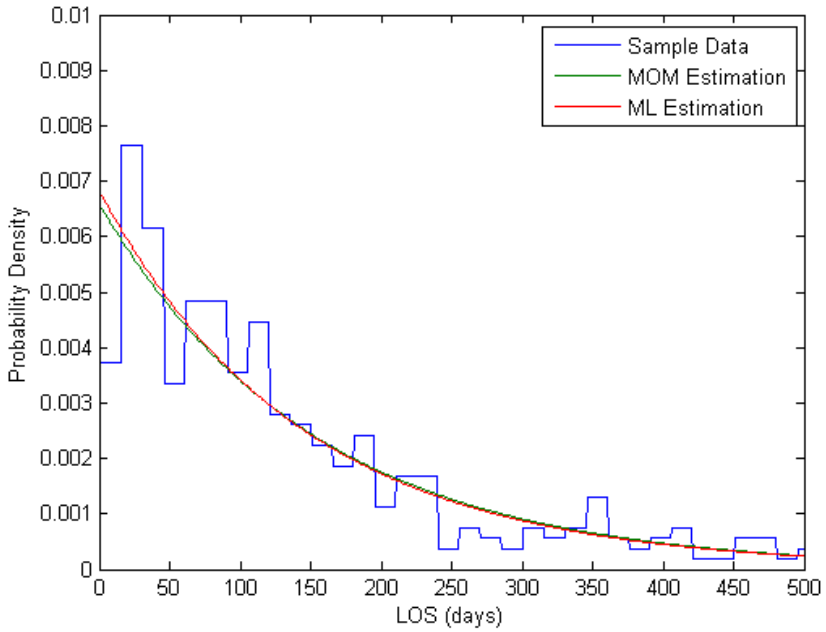


Figure 6.25 Three-term hyper-exponential approximation to length of stay

6.5.3.4 Hypo-exponential (two-term)

Since the sample coefficient of variation is greater than one the upper bound on the second sample moment (see Ch 5.5.3.1.2) is not satisfied, i.e. $m_2 > 2m_1^2$. Therefore, one of the MOM estimators is negative which is obviously invalid. However, feasible estimations are possible for ML (due to the aforementioned advantages of *fminsearchbnd*).

[12] Same as *fminsearchbnd* but allows linear constraints to be put on the parameters. In this case the value of the two parameters of the *initial probability vector* must be less than one i.e. alpha1 + alpha2 < 1.

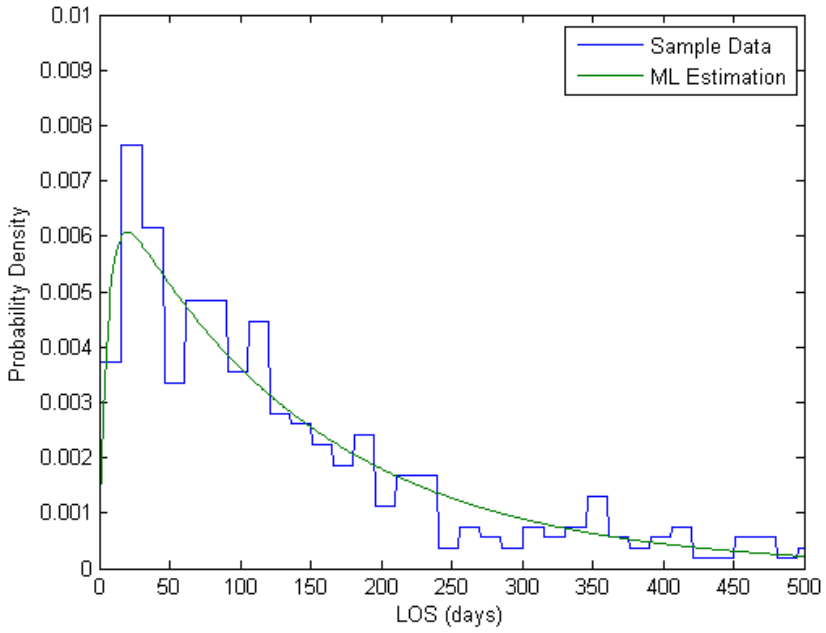


Figure 6.26 Two-term hypo-exponential approximation to length of stay

6.5.3.5 Coxian (two-term)

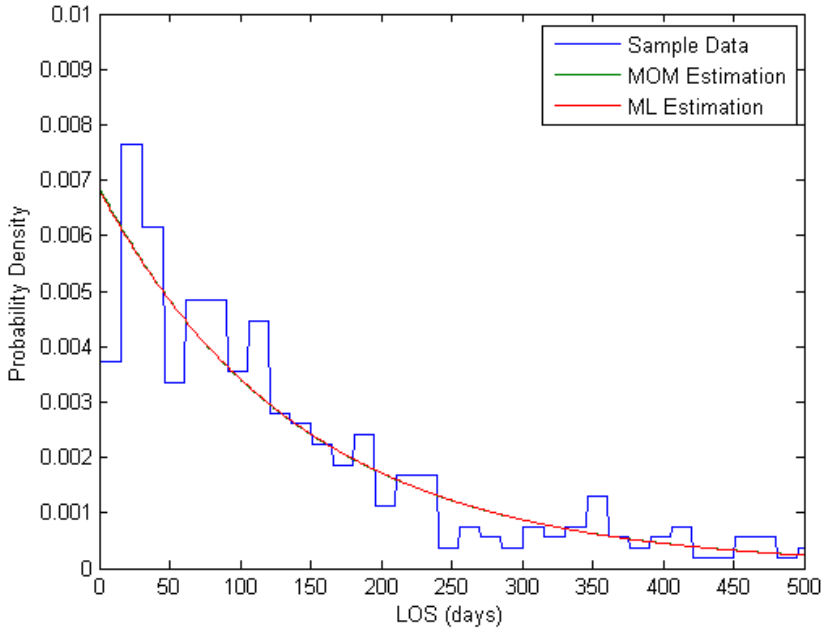


Figure 6.27 Two-term Coxian approximation to length of stay

However, it has been noticed that using different starting values from the MOM estimators results in a better fit for ML estimation.

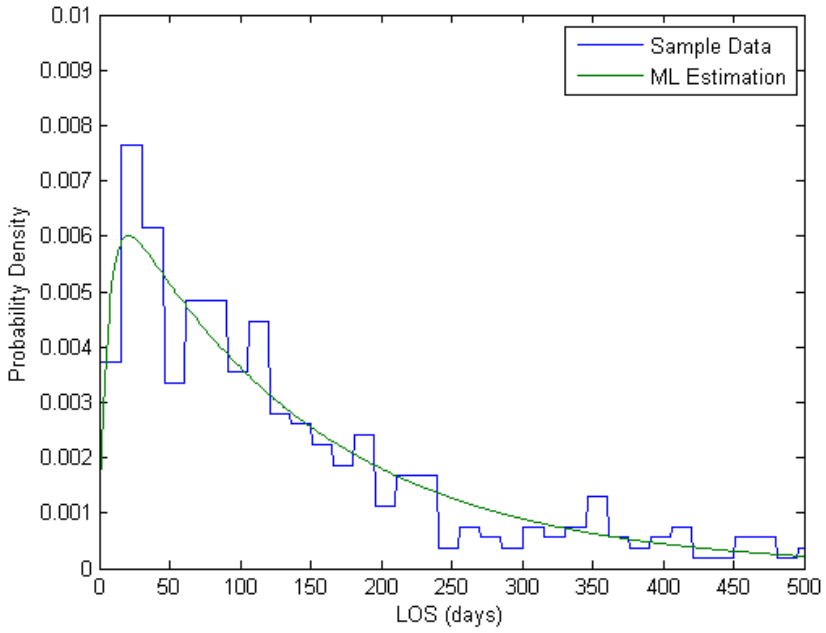


Figure 6.28 Two-term Coxian approximation to length of stay using MOM estimators

6.5.3.6 Coxian (three-term)

No analytic results are detailed in Ch 5.5 for the MOM estimators of this distribution. Therefore an equivalent approach is employed to that of the three-term hyper-exponential distribution.

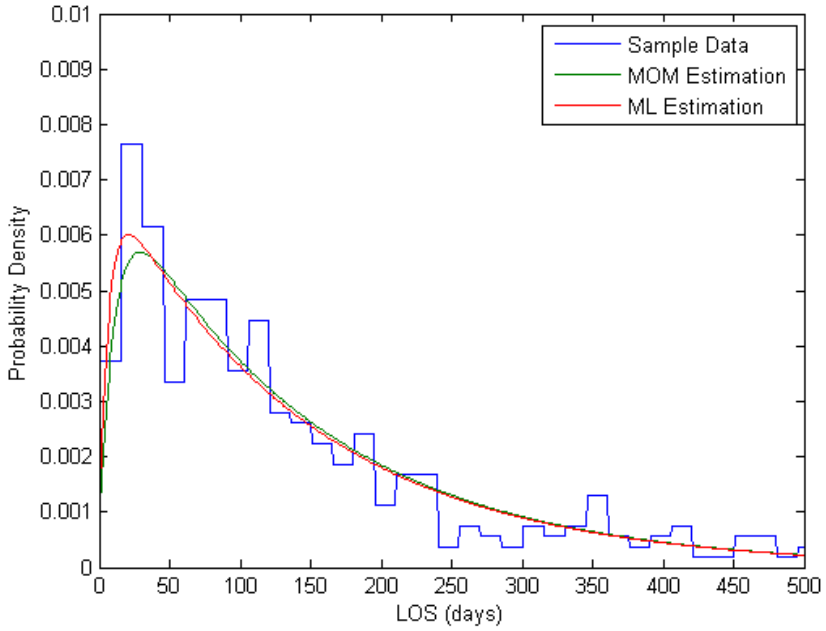


Figure 6.29 Three-term Coxian approximation to length of stay

6.5.4 Results

Table 6.13 Goodness of fit results for approximating distributions

Est method:	MLE				MOM			
Test:	LogL	AIC	BIC	χ^2	LogL	AIC	BIC	χ^2
Log normal	–2154.8	4313.5	4321.3	2.0434	–2257.6	4519.2	4527	5.5158
Gamma	–2148.6	4301.2	4309	0.51	–2151.2	4306.5	4314.2	0.7922
Weibull	–2149.5	4303.1	4310.9	0.4845	–2150.6	4305.1	4312.9	0.6959
Log logistic	–2155.3	4314.7	4322.4	1.8324	–2182	4368.1	4375.9	1.3253
Exponential	–2150.2	4302.3	4306.2	0.607	–2150.2	4302.3	4306.2	0.607
Hyper (2)	–2150.1	4306.2	4317.8	0.6535	–2150.1	4306.3	4317.9	0.6685
Hyper (3)	–2150.1	4310.2	4329.6	0.6535	*-2150.2*	*4310.4*	*4329.8*	*0.61*
Hypo (2)	–2145.6	4295.2	4302.9	0.5984	–	–	–	–
Cox (2)	–2150.1	4306.2	4317.8	0.6535	–2150.1	4306.3	4317.9	0.6764
Cox (2) a	–2145.5	4297	4308.7	0.6136	–	–	–	–
Cox (3)	–2145.5	4301	4320.4	0.6134	*-2146.5*	*4303*	*4322.4*	*1.0961*

The distributions corresponding to the results in red are optimal with respect to the equivalent statistical test. That is, the two and three-term Coxian distributions are optimal with respect to log-likelihood value. However, it is the two-term hypo-exponential distribution that is optimal with respect to AIC and BIC value since only two parameter estimates are required. This means that the 0.1 point improvement in log-likelihood value is not worth the estimation of a further one or three parameters. The Weibull distribution is optimal with respect to chi-square test statistic. However, since these values are very sensitive to class width and number of intervals, their credibility as a performance measure is questionable.

Four observations are made forthwith. For most of the phase-type distributions the MOM results are very close to the ML results. However, in general, the performance of MOM is surpassed by ML. There are only two cases (in blue) to which the converse is true. These are results for the chi-square test. It is indeed intuitive that MOM estimates are surpassed by ML estimates in the other statistical measures since these are all based on likelihood values. Secondly, it is interesting that the hypo-exponential distribution ($CV \leq 1$) provides such a good fit to the data considering that sample CV is greater than one. Thirdly, it has been noticed that the Hyper (3) is reducible to the Hyper (2) since the rate value of two of the three nodes are equal. Finally, the Cox (3) is reducible to the Cox (2) a since the probability of transition to the third phase (under ML estimation) is zero.

The results in italics represent the three-term hyper-exponential and Coxian distribution whose MOM estimations are considered dubious and possibly not globally optimal. It is postulated that this could be due to an unstable surface of the minimisation problem.

6.6 Conclusion

This chapter commences by introducing the data requirements for models of varying complexity. The practical issues surrounding the collection of data are then discussed. This includes the presentation of a database that has been specifically designed to expand the range of data that is collected. A preliminary study of the data is then presented. Finally a number of distributions are fitted to empirical LOS data by the method of moments and maximum likelihood.

All in all 183 referrals are used for the Arrivals dataset whilst 384 patient episodes are used to populate the Service dataset (of which 358 have additional NHS Trust variables and 89 have treatment intensity/date ready for discharge). It is hard to say whether these represent a sufficient number of observations to make credible conclusions about the data, but the only other alternative would have been to include possibly unreliable and irrelevant data from further back in time. As mentioned in the outset of Ch 6.3, a balance must be struck between the quantity and quality of the data.

It is a presumption of method of moments and maximum likelihood estimation that sample observations are independent of one another. Since distributions have been fitted using these parameter estimation methods it is necessary to comment on this property of the data. Firstly, *external independence* is assumed. This means that the LOS of an episode of one patient does not affect the LOS of an episode of another. By virtue of this assumption, mutual independence is attained for the (338) observations that relate to stand-alone episodes. The question that remains is that of *internal independence* for those with multiple episodes, i.e. are the episodes of a single patient independent of one another?

It is postulated that such episodes are independent if there is no significant difference in their means. A one-way ANOVA test is therefore conducted for the data described in the third table of Ch 6.4.2. The null hypothesis is that there is no difference in mean LOS between the first, second and third episodes. The alternate hypothesis is that there is a difference in mean LOS between at least two of the episodes. The grand mean is calculated as

$\bar{X}_{GM} = \sum_{i=1}^{N} x_i \Big/ \sum_{j=1}^{k} n_j = \sum_{i=1}^{20} x_i \Big/ 20 = 107.8$ where x_i is the value of the $i-th$ observation, N is the total number of observations, n_j is the number of observations in the $j-th$ group, and k is the number of groups. The between groups variation and within groups variation are calculated as $SS(B) = \sum_{j=1}^{k} n_j \left(\bar{x}_j - \bar{X}_{GM}\right)^2 = 21232.08$ and $SS(W) = \sum_{j=1}^{k} \left(n_j - 1\right) s_j^2 = 171244$.

The ANOVA table is therefore summarised as

	SS	*d.f.*	*MS*	*F*
Between	21232.08	2	10616.04	1.054
Within	171244	17	10073.18	
Total	239322.5	19		

For the (one-tailed) test the critical region (95%) is $F_{0.05,2,19} > 3.59$. Since the test statistic is not within this region the null hypothesis is not rejected – there is insufficient evidence to suggest a difference in means.

It can therefore be postulated that internal independence is attained. There is, however, a certain amount of scepticism surrounding this conclusion due to the low sample sizes (especially with regard to the third episode). In addition, it is hard to believe that, in real-life, the LOS of earlier admissions will have no affect on the LOS of later ones.

Doubts must also be raised about the validity of the assumption of external independence. At the crux of this project is the notion that, due to constrained resources, all patients cannot receive the optimal intensity of treatment. Ergo, the allocation of treatment to one patient affects the allocation of treatment to others. Since treatment intensity is related to LOS (see Ch 2.2.3) this casts doubt over the assumption that the LOS of one patient is not affected by the LOS of another. The implications of these caveats on the validity of the parameter estimation methods are not fully understood and so it will have to suffice to simply acknowledge their presence.

The appropriateness of phase-type distributions must also be addressed. There is no question of their use in determining analytic solutions to queuing problems but how good are they at representing empirical data? It is stated in Bladt, 2005 that *'care should be taken'* in using phase-type distributions to approximate heavy-tailed data. Moreover, the author comments

that '*the approximations will always be bad in the tails'*. With this is mind the heaviness of the tail for the LOS data of Rookwood NRC is investigated.

First and foremost it is necessary to define what it means for a distribution to be heavy-tailed. According to Hardle & Simar, 2007 a distribution is heavy-tailed *'if it has higher probability density in its tail area compared with a normal distribution with the same mean and variance'*. The authors state that *'a heavy-tailed distribution has kurtosis[13] greater than three'*, i.e. greater than that of the normal distribution. However, this is not an exact measure. Ruppert, 2010 comments that *'high kurtosis is nearly synonymous with having a heavy-tailed distribution'*. Danielsson et al, 2006 are even more sceptical, claiming that *'it is straightforward to construct a distribution with truncated tails, and hence thin tails, which exhibits high kurtosis'*. Instead, they define a heavy-tailed distribution as *'one characterised by the failure of the moments* $m > 0$ *'*.

The excess kurtosis[14] for the data in which distributions are fitted to in Ch 6.5 is calculated as 5.5. According to the kurtosis definition, the sample data that is approximated is therefore heavy-tailed. Despite this, the three-term Coxian distribution provides the best fit to the data in Ch 6.5 by log-likelihood value. Furthermore, by studying the fit to the tail in the corresponding graph there is no reason to suggest a bad fit.

This concludes Chapter 6. The multi-server finite-buffer queue with Markovian arrivals and Erlang service times is investigated in the next chapter.

[13] Defined by the division of the fourth central moment by the square of the variance
[14] Kurtosis minus three

Chapter 7: Queuing with Erlang Service Times

7.1 Introduction

A queuing system with Erlang distributed service times is considered in this chapter. The Erlang distribution is a phase-type distribution (Ch 1.5) that has been briefly discussed in Chapter 5.5.3.2. The *pdf* of sojourn time is given by equation (5.5.3.2.2). Since arrivals at the system are random (Ch 6.4.1) the exponential distribution is used to model inter-arrival times. This is therefore a type of Markovian Arrival process. Moreover, it is a Poisson process (Ch 1.5) since the probability of a given number of arrivals occurring in an interval of time is deduced from the Poisson distribution. The discipline of the queue is first-in first-out.

This chapter considers a variety of scenarios relating to the number of servers in the system, phases in the service distribution, and places available for waiting. The aim is to find exact[1] results for the steady-state probabilities of the $M \mid E_k \mid r \mid N$ system being in each possible state. First, the two server case is considered with two phases (i.e. $k=r=2$) and finite capacity. The methodology is then generalised to consider a system with any number of phases, k. Finally the system with a general number of servers, phases and places available for waiting is studied.

Unless specified all results and workings are produced independently in this chapter and next.

[1] That is, not approximate (see Ch 4.3)

7.2 Two Server

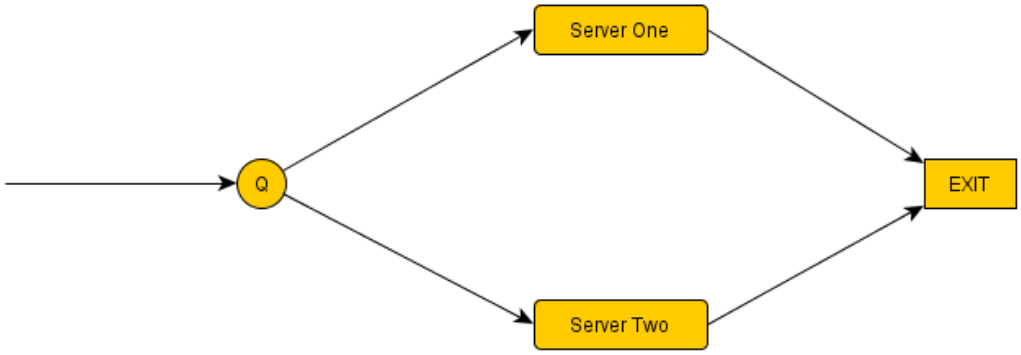


Figure 7.1 Two server queuing system

7.2.1 Two phase

Each service channel has two exponentially distributed phases. A maximum of one customer may be in service at any service channel at any time.

7.2.1.1 No waiting space

The following differential-difference equations are formed by considering the possibility of events in an arbitrarily small unit of time, ∂t. An event is an arrival (with rate λ) or a service phase completion (with rate μ). The $P_s(t)$ represent the probability of the system being in state $s \in S$ at time t where S is the state space. Each s is defined by the triple $\langle n_s, n_1, n_2 \rangle$ whose elements represent the number in system, service phase one and phase two respectively. Note that any terms of magnitude $O(\partial t)^2$ are excluded.

$$
\begin{aligned}
\frac{dP_0(t)}{dt} &= P_0(t)\left(1-\lambda\partial t\right) + P_{1,0,1}(t)\mu\partial t\left(1-\lambda\partial t\right)\\
\frac{dP_{1,1,0}(t)}{dt} &= P_{1,1,0}(t)\left(1-\mu\partial t\right)\left(1-\lambda\partial t\right) + P_0(t)\lambda\partial t + P_{2,1,1}(t)\mu\partial t\left(1-\mu\partial t\right)\left(1-\lambda\partial t\right)\\
\frac{dP_{1,0,1}(t)}{dt} &= P_{1,0,1}(t)\left(1-\mu\partial t\right)\left(1-\lambda\partial t\right) + P_{1,1,0}(t)\mu\partial t\left(1-\lambda\partial t\right) + P_{2,0,2}(t)2\mu\partial t\left(1-\lambda\partial t\right)\\
\frac{dP_{2,2,0}(t)}{dt} &= P_{2,2,0}(t)\left(1-\mu\partial t\right) + P_{1,1,0}(t)\lambda\partial t\\
\frac{dP_{2,1,1}(t)}{dt} &= P_{2,1,1}(t)\left(1-\mu\partial t\right) + P_{1,0,1}(t)\left(1-\mu\partial t\right)\lambda\partial t + P_{2,2,0}(t)2\mu\partial t\\
\frac{dP_{2,0,2}(t)}{dt} &= P_{2,0,2}(t)\left(1-2\mu\partial t\right) + P_{2,1,1}(t)\mu\partial t\left(1-\mu\partial t\right)
\end{aligned}
\qquad (7.2.1.1.1)
$$

In this case, the cardinality of S is six. Since $\frac{df(x)}{dx} := \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$ the following are obtained:

$$
\begin{aligned}
P_0'(t) &= -\lambda P_0(t) + \mu P_{1,0,1}(t)\\
P_{1,1,0}'(t) &= -\left(\lambda+\mu\right)P_{1,1,0}(t) + \lambda P_0(t) + \mu P_{2,1,1}(t)\\
P_{1,0,1}'(t) &= -\left(\lambda+\mu\right)P_{1,0,1}(t) + \mu P_{1,1,0}(t) + 2\mu P_{2,0,2}(t)\\
P_{2,2,0}'(t) &= -\mu P_{2,2,0}(t) + \lambda P_{1,1,0}(t)\\
P_{2,1,1}'(t) &= -\mu P_{2,1,1}(t) + \lambda P_{1,0,1}(t) + 2\mu P_{2,2,0}(t)\\
P_{2,0,2}'(t) &= -2\mu P_{2,0,2}(t) + \mu P_{2,1,1}(t)
\end{aligned}
\qquad (7.2.1.1.2)
$$

These are the time-dependent equations whose solution gives transient results. Steady-state results are obtained when the behaviour of the system has reached long-term equilibrium, i.e. when $P_s'(t) = 0$;

$$
\begin{aligned}
P_{1,0,1} &= \rho P_0\\
2P_{2,0,2} &= \left(\rho+1\right)P_{1,0,1}\\
P_{2,1,1} &= 2P_{2,0,2}\\
2P_{2,2,0} &= P_{2,1,1} - \rho P_{1,0,1}\\
\rho P_{1,1,0} &= P_{2,2,0}\\
\left(\rho+1\right)P_{1,1,0} &= \rho P_0 + P_{2,1,1}
\end{aligned}
\qquad (7.2.1.1.3)
$$

with ρ defined as $\lambda\mu^{-1}$. These equations have been written in such a way that each equation contains only the terms that have been previously defined (with the exception of the equation

for state 1,0,1). The final equation is superfluous since a defining equation for this state has already been written. By defining $Q_s := P_s P_0^{-1}$;

$$\begin{aligned} Q_{1,0,1} &= \rho \\ Q_{2,0,2} &= 0.5(\rho+1)\rho \\ Q_{2,1,1} &= (\rho+1)\rho \\ Q_{2,2,0} &= 0.5(\rho+1)\rho - 0.5\rho^2 \\ Q_{1,1,0} &= 0.5(\rho+1) - 0.5\rho \end{aligned} \qquad (7.2.1.1.4)$$

By applying the normalisation equation, a result for P_0 can be obtained;

$$P_0 = \left(1 + Q_{1,1,0} + Q_{1,0,1} + Q_{2,2,0} + Q_{2,1,1} + Q_{2,0,2}\right)^{-1} \qquad (7.2.1.1.5)$$

The solution of which can be substituted back into (7.2.1.1.3) to obtain the steady-state probabilities of the system being in each of the possible states.

7.2.1.2 Waiting space for one customer

An alternative way to set up the steady-state equations is to model the transitions between states by a continuous-time Markov chain (CTMC). This is appropriate since the Markov property[2] holds. Note that this is a generalisation of the *birth-death process* that is synonymous with the CTMCs of the $M|M|1$ and $M|M|C$ systems.

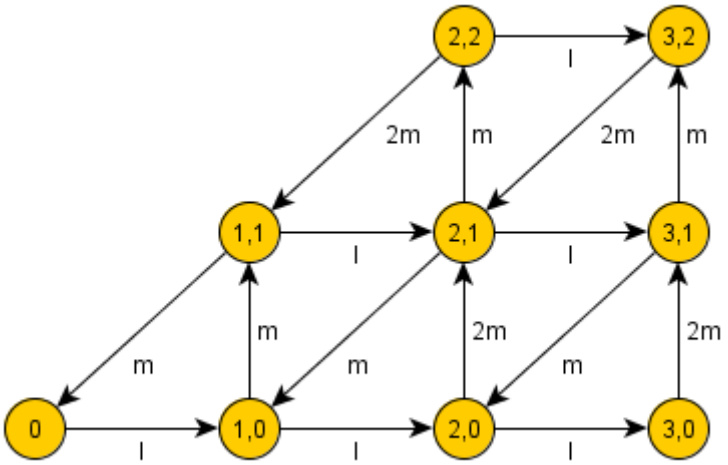


Figure 7.2 CTMC of $M/E_2/2/3$ queuing system (with $\mathrm{l}=\lambda, \mathrm{m}=\mu$)

[2] The future state is dependent upon the current state and not on previous states

Equations are derived by equating the inward and outward flow of each state. The state space has been reduced from three to two dimensions by omitting n_1. Each state is now defined by the double $\langle n_s, n_2 \rangle$. This is without loss since $n_1 = \begin{cases} n_s - n_2 & n_s \le 2 \\ 1 & n_s > 2 \end{cases}$.

$$\begin{aligned}
&\lambda P_0 = \mu P_{1,1} \\
&(\lambda+\mu)P_{1,0} = \lambda P_0 + \mu P_{2,1} \\
&(\lambda+\mu)P_{1,1} = \mu P_{1,0} + 2\mu P_{2,2} \\
&(\lambda+2\mu)P_{2,0} = \lambda P_{1,0} + \mu P_{3,1} \\
&(\lambda+2\mu)P_{2,1} = \lambda P_{1,1} + 2\mu P_{2,0} + 2\mu P_{3,2} \\
&(\lambda+2\mu)P_{2,2} = \mu P_{2,1} \\
&2\mu P_{3,0} = \lambda P_{2,0} \\
&2\mu P_{3,1} = \lambda P_{2,1} + 2\mu P_{3,0} \\
&2\mu P_{3,2} = \lambda P_{2,2} + \mu P_{3,1}
\end{aligned} \qquad (7.2.1.2.1)$$

These equations, along with the normalisation equation, can be input into a matrix as follows:

$$M = \begin{pmatrix}
-\lambda & 0 & \mu & 0 & 0 & 0 & 0 & 0 & 0 \\
\lambda & -(\lambda+\mu) & 0 & 0 & \mu & 0 & 0 & 0 & 0 \\
0 & \mu & -(\lambda+\mu) & 0 & 0 & 2\mu & 0 & 0 & 0 \\
0 & \lambda & 0 & -(\lambda+2\mu) & 0 & 0 & 0 & \mu & 0 \\
0 & 0 & \lambda & 2\mu & -(\lambda+2\mu) & 0 & 0 & 0 & 2\mu \\
0 & 0 & 0 & 0 & \mu & -(\lambda+2\mu) & 0 & 0 & 0 \\
0 & 0 & 0 & \lambda & 0 & 0 & -2\mu & 0 & 0 \\
0 & 0 & 0 & 0 & \lambda & 0 & 2\mu & -2\mu & 0 \\
0 & 0 & 0 & 0 & 0 & \lambda & 0 & \mu & -2\mu \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}$$

Note that $M = \begin{pmatrix} Q^T & 1 \end{pmatrix}^T$ where Q is the transition matrix[3] between states and 1 is a row vector of ones. If P is a column vector of $P_{s \in S}$ such that $P = \begin{pmatrix} P_0 & P_{1,0} & \cdots & P_{3,2} \end{pmatrix}^T$ and $B = \begin{pmatrix} 0 & \cdots & 0 & 1 \end{pmatrix}^T$ then the following linear system is obtained:

$$MP = B \qquad (7.2.1.2.2)$$

[3] T means transpose

This can be solved by a mathematical computer package to deduce the unknown vector *P*.

Since a nonzero queue is permitted in this scenario (*c.f.* Ch 7.2.1.1) the distribution of waiting time (*W*) can now be considered.

The probability that an arriving customer must wait is equal to the probability that there are two customers in the system (in service) at the time of arrival.

$$P\{\text{Arriving customer must wait}\} = \bar{P}_2 = P_{2,0} + P_{2,1} + P_{2,2} \tag{7.2.1.2.3}$$

where $\bar{P}_n$ is the probability of having *n* customers in the system. The waiting time distribution for an arriving customer may be written

$$P(W > t) = \sum_{\forall s \in S} P(W > t \mid \zeta = s) P_s \tag{7.2.1.2.4}$$

where ζ is the state of the system on arrival. By (7.2.1.2.3) this can therefore be reduced to

$$P(W > t) = \sum_{\eta=0}^{2} P(W > t \mid \zeta = \langle 2, \eta \rangle) P_{2,\eta} \tag{7.2.1.2.5}$$

If a customer is in phase one of service then the inverse *cdf* of time until exit (τ) is given by

$$\begin{aligned} F_1(t) &= P(\tau > t) = 1 - P(\tau < t) \\ &= 1 - \left(1 - \sum_{n=0}^{1} \frac{e^{-\mu t} (\mu t)^n}{n!} \right) = t e^{-\mu t} + \frac{1}{\mu} e^{-\mu t} \end{aligned} \tag{7.2.1.2.6}$$

If a customer is in phase two of service then the inverse cdf of time until exit is given by

$$F_2(t) = P(\tau > t) = 1 - P(\tau < t) = 1 - \left(1 - e^{-\mu t}\right) = e^{-\mu t} \tag{7.2.1.2.7}$$

Given $n_s = 2$ and in the case of $M \mid E_2 \mid 2 \mid 3$ the waiting time of an arriving customer is equal to the time until the next service completion, i.e. the minimum of the service times at each channel. This is obviously dependent on the state of the system at the arrival instance.

Note that if *X* and *Y* are independent random variables and $Z = \min\{X, Y\}$ then

$$1-F_Z(z)=P\left(Z>z\right)=P\left(\min\left\{X,Y\right\}>z\right)=P\left(X>z,Y>z\right)=P\left(X>z\right)P\left(Y>z\right) \quad (7.2.1.2.8)$$

Therefore, if the system is in state $\langle 2,0\rangle$ the distribution of time until service completion is

$$P\left(W>t\mid\zeta=\langle 2,0\rangle\right)=F_1(t)F_1(t)=\left(t+\mu^{-1}\right)^2e^{-2\mu t}$$

Similarly, (7.2.1.2.9)

$$P\left(W>t\mid\zeta=\langle 2,1\rangle\right)=F_1(t)F_2(t)=\left(t+\mu^{-1}\right)e^{-2\mu t}$$
$$P\left(W>t\mid\zeta=\langle 2,2\rangle\right)=F_2(t)F_2(t)=e^{-2\mu t}$$

It follows, (7.2.1.2.5), that the unconditional distribution of waiting time is

$$P\left(W>t\right)=\left(P_{2,2}+\left(t+\mu^{-1}\right)P_{2,1}+\left(t+\mu^{-1}\right)^2P_{2,2}\right)e^{-2\mu t} \quad (7.2.1.2.10)$$

The distribution of sojourn time (T) for an arriving customer may also be calculated. If there are fewer than two customers in the system on arrival then the distribution of sojourn time is equal to that of a two phase Erlang distribution.

$$P\left(T>t\mid\zeta=\langle 0\rangle\right)=F_1(t)$$
$$P\left(T>t\mid\zeta=\langle 1,0\rangle\right)=F_1(t) \quad (7.2.1.2.11)$$
$$P\left(T>t\mid\zeta=\langle 1,1\rangle\right)=F_1(t)$$

If the number in system is equal to two then the distribution of waiting time must be convoluted (*) with the two phase Erlang distribution;

$$LT\left\{P\left(T=t\mid\zeta=\langle 2,0\rangle\right)\right\}=LT\left\{f_1(t)\right\}*LT\left\{\left(f_1(t)\right)^2\right\}$$
$$LT\left\{P\left(T=t\mid\zeta=\langle 2,1\rangle\right)\right\}=LT\left\{f_1(t)\right\}*LT\left\{f_1(t)\,f_2(t)\right\} \quad (7.2.1.2.12)$$
$$LT\left\{P\left(T=t\mid\zeta=\langle 2,2\rangle\right)\right\}=LT\left\{f_1(t)\right\}*LT\left\{\left(f_2(t)\right)^2\right\}$$

The Laplace transform (LT) of an Erlang pdf (two phase) and an exponential pdf is[4] $LT_s\left\{f_1(t)\right\}=\left(\mu\left(s+\mu\right)^{-1}\right)^2$ and $LT_s\left\{f_2(t)\right\}=\mu\left(s+\mu\right)^{-1}$. Also, it can be shown (through integration by parts) that

[4] See Ch 5.5.3.1.2

$$
\begin{aligned}
LT_s\left\{\left(f_1(t)\right)^2\right\} &= \left(2\mu\left(s+2\mu\right)^{-1}\right)^4 \\
LT_s\left\{f_1(t)\,f_2(t)\right\} &= \mu\left(\mu\left(s+2\mu\right)^{-1}\right)^2 \\
LT_s\left\{\left(f_2(t)\right)^2\right\} &= \mu\left(\mu\left(s+2\mu\right)^{-1}\right)
\end{aligned}
\qquad (7.2.1.2.13)
$$

Therefore, and by application of the Shift theorem, the following can be derived by inverting the *LT*s of (7.2.1.2.13) and integrating over the limit (t,∞):

$$
\begin{aligned}
P\left(T>t\,|\,\zeta=\langle 2,0\rangle\right) &= e^{-2\mu t}\left(1+2\mu t+2\left(\mu t\right)^2+\frac{4}{3}\left(\mu t\right)^3\right) \\
P\left(T>t\,|\,\zeta=\langle 2,1\rangle\right) &= \frac{1}{4}\mu e^{-2\mu t}\left(1+2\mu t\right) \\
P\left(T>t\,|\,\zeta=\langle 2,2\rangle\right) &= \frac{1}{2}\mu e^{-2\mu t}
\end{aligned}
\qquad (7.2.1.2.14)
$$

In a similar manner to (7.2.1.2.4);

$$
P\left(T>t\right)=\sum_{\forall s\in S}P\left(T>t\,|\,\zeta=s\right)P_s \qquad (7.2.1.2.15)
$$

Through substitution of (7.2.1.2.14);

$$
P\left(T>t\right)=e^{-2\mu t}\left(\left(1+2\mu t+2\left(\mu t\right)^2+\frac{4}{3}\left(\mu t\right)^3\right)P_{2,0}+\frac{1}{4}\mu\left(1+2\mu t\right)P_{2,1}+\frac{1}{2}\mu P_{2,2}\right) \qquad (7.2.1.2.16)
$$

However, there is a problem here. In the formulation thus far it is conjectured that an arrival cannot be permitted to occur when the number in system is equal to three. But it is also the case that arrivals are independent of the number in system. These are obvious contradictory statements.

The effect of this inconsistency is that the waiting and sojourn time (as calculated through 7.2.1.2.4 and 7.2.1.2.15) are, in fact, underestimations of their true values.

The reason for this is that the expansions of 7.2.1.2.4 and 7.2.1.2.15 contain no terms linked to $P_{3,0}$, $P_{3,1}$ or $P_{3,2}$. Therefore, $P\left(W>t\,|\,\zeta=\langle 3,\eta\rangle\right)P_{3,\eta}=0$ and $P\left(T>t\,|\,\zeta=\langle 3,\eta\rangle\right)P_{3,\eta}=0$ for $\eta=0,1,2$. Since the solution of 7.2.1.2.2 gives $P_{3,\eta}$ as nonzero for $\eta=0,1,2$ it can be inferred that $P\left(W>t\,|\,\zeta=\langle 3,\eta\rangle\right)$ and $P\left(T>t\,|\,\zeta=\langle 3,\eta\rangle\right)$ must equal zero.

This means that arrivals instantaneously exit when the number in system is equal to three. They do not join the queue and do not enter service. This can be thought of as having a guard that only allows customers into the service part of the system if $n_s < 3$ upon arrival.

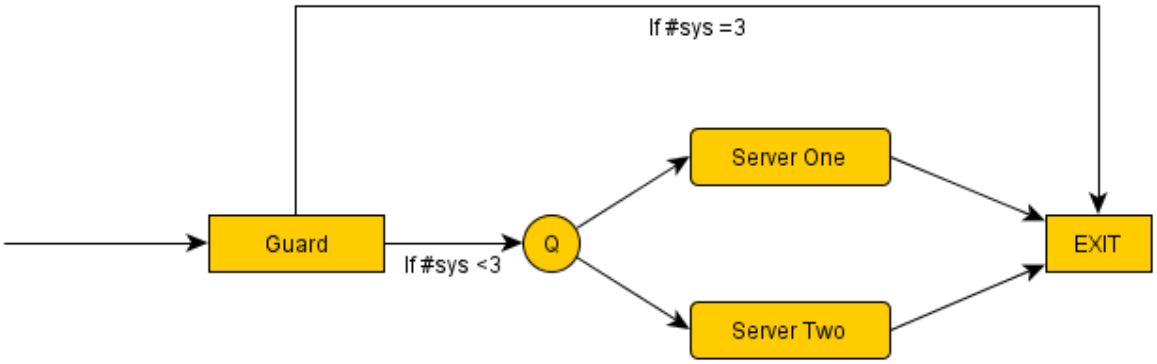


Figure 7.3 Two server queuing system with waiting space for one customer

The solution to this problem is as follows. If $n_s \leq 2$ upon arrival, the customer will spend an amount of time in service and possibly in the queue before their exit. Since waiting and service times are continuous variables, they cannot equal zero exactly. Thus, it is possible to obtain a sojourn time distribution conditional on this requisite;

$$P(T > t \mid T > 0) = \frac{\sum_{\forall s \in S} P(T > t \mid \zeta = s) P_s}{\sum_{n=0}^{2} \bar{P}_n} \qquad (7.2.1.2.17)$$

This is the sojourn time distribution for *non-trivial* customers, i.e. those that, at some point, enter service. The waiting time distribution for non-trivial customers is derived in a similar manner;

$$P(W > t \mid T > 0) = \frac{\sum_{\forall s \in S} P(W > t \mid \zeta = s) P_s}{\sum_{n=0}^{2} \bar{P}_n} \qquad (7.2.1.2.18)$$

7.2.1.3 Waiting space for *N* customers

A depiction of the CTMC for this case is given in Figure 7.4.

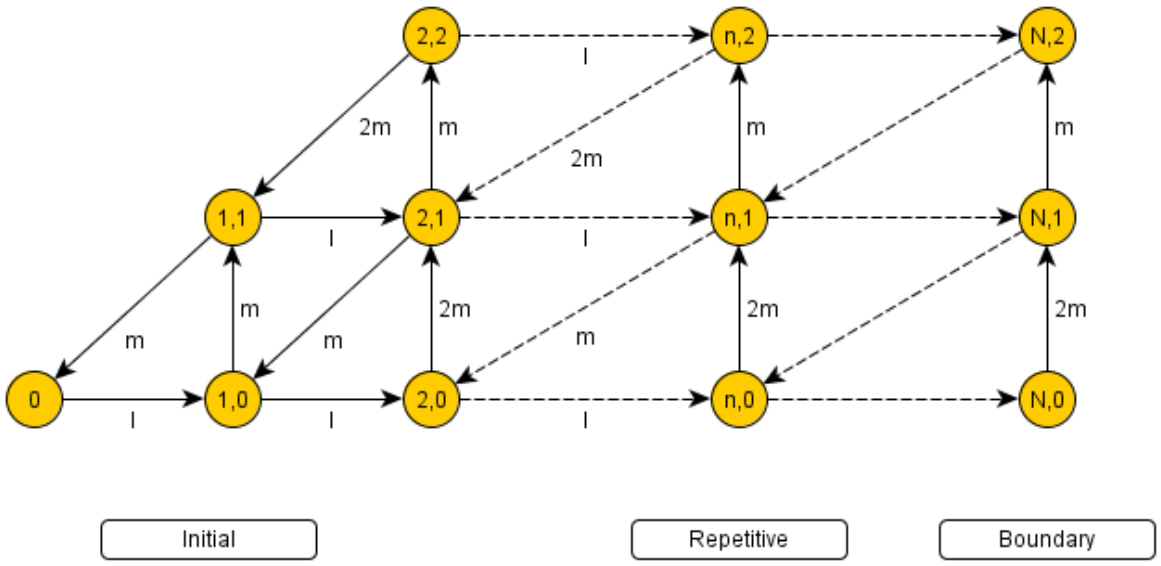


Figure 7.4 CTMC of $M / E_2 / 2 / N$ queuing system

The steady-state equations for the six 'Initial' states are equal to the first six equations of (7.2.1.2.1). The steady-state equations for the 'Repetitive' states, for $3 \le n < N$, are given as

$$\begin{aligned}
(\lambda + 2\mu) P_{n,0} &= \lambda P_{n-1,0} + \mu P_{n+1,1} \\
(\lambda + 2\mu) P_{n,1} &= \lambda P_{n-1,1} + 2\mu P_{n,0} + 2\mu P_{n+1,2} \\
(\lambda + 2\mu) P_{n,2} &= \lambda P_{n-1,2} + \mu P_{n,1}
\end{aligned} \qquad (7.2.1.3.1)$$

The 'boundary' state equations are

$$\begin{aligned}
2\mu P_{N,0} &= \lambda P_{N-1,0} \\
2\mu P_{N,1} &= \lambda P_{N-1,1} + 2\mu P_{N,0} \\
2\mu P_{N,2} &= \lambda P_{N-1,2} + \mu P_{N,1}
\end{aligned} \qquad (7.2.1.3.2)$$

Once set up, these can be solved in a similar manner to the case of one waiting space.

7.2.1.4 Infinite waiting space

Knessl et al, 1990 provide an integral equation approach to the solution of the $M|G|2|\infty$ system. In this study the authors consider the case of two-term Erlang distributed service times. The steady-state results for P_0 and the other states are given by equations (60) and (61) in this paper.

7.2.2 *k* phase

It can be seen that the dimension of the CTMC is equal to the cardinality of the system state description given by *s*. The state of the system can be described by the *k*-tuple $\langle n_s, n_2, ..., n_k \rangle$ where n_i is the number of customers in service phase *i* for $2 \le i \le k$. Therefore, the dimension of the CTMC is equal to *k*. Note that $n_1 = \begin{cases} n_s - \sum_{i=2}^{k} n_i & n_s \le 2 \\ 1 & n_s > 2 \end{cases}$. An appropriate CTMC can be constructed and solved in a method analogous to that of the two phase case for a finite or zero waiting space.

Knessl et al, 1990 use their derived integral equation (50) to explain how results can be obtained for the $M \mid E_k \mid 2 \mid \infty$ system, although they do not show any working or result.

7.3 *r* Server

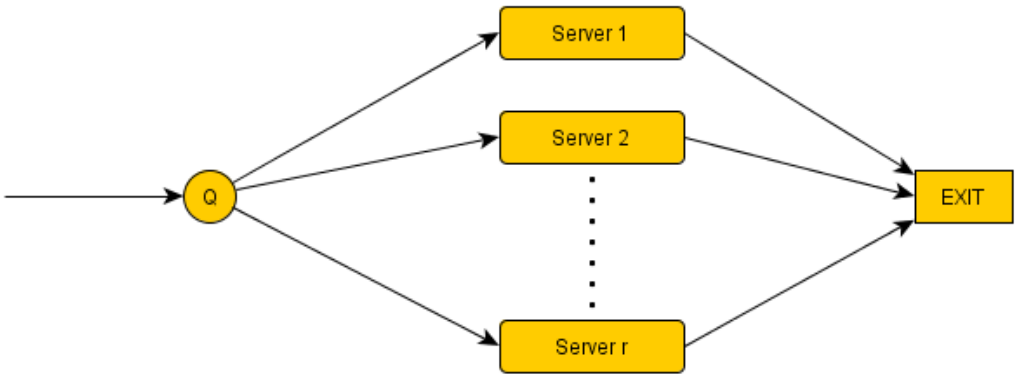


Figure 7.5 r server queuing system

7.3.1 Two phase

The $M \mid E_2 \mid r \mid \infty$ system is considered in Shapiro, 1966 with heterogeneous servers. Steady-state difference equations are set up and subdivided into two cases where $n_s < r$ and $n_s \ge r$. Since there are more unknowns than equations in case $n_s < r$ these equations are solved in terms of the unknowns from case $n_s \ge r$. The solutions for this case are obtained by an intricate method, details of which can be found in the paper.

7.3.2 k phase

The generalised system, $M \mid E_k \mid r$, is considered next.

7.3.2.1 Infinite waiting space

Perhaps the earliest paper to study the steady-state version of this system is Mayhugh & McCormick, 1968. The authors obtain the solution to 'Initial' and 'Repetitive' state equations. There are no 'Boundary' equations since there is infinite waiting space. The 'Repetitive' equations are solved by virtue of a generating function approach. A year later, a similar paper (Heffer, 1969) is published. This too uses a generating function approach but is apparently less complex *'both analytically and computationally'*.

Others (Poyntz & Jackson, 1973 and Adan et al, 1992) have investigated the further generalised $E_c \mid E_k \mid r \mid \infty$ systems.

7.3.2.2 Finite waiting space

As can be seen from the cited studies the solutions to the infinite waiting space queuing system can be, at times, abstruse and time consuming to not only comprehend, but also to implement. Since the number of beds (servers) at Rookwood NRC can vary quite substantially over time (Ch 6.4.2.2), it is necessary to produce an automated approach that is computationally fast and simple to use. Therefore, the following has been produced[5] to determine the steady-state solution of the $M \mid E_k \mid r \mid N$ system.

Computationally, the process is two-stage. The first stage is to populate the matrix, *M* (see Ch 7.2.1.2) with the appropriate transition rates. The second stage is to solve the linear system (7.2.1.2.2).

[5] Independently

7.3.2.2.1 Specification of the transpose transition matrix

$$Q^T = \begin{pmatrix} t_{0,0} & t_{0,1} & & & & & & & & \\ t_{1,0} & t_{1,1} & t_{1,2} & & & & & & & \\ & \ddots & \ddots & & & & & & & \\ & & \ddots & \ddots & & & & & & \\ & & & t_{r,r-1} & t_{r,r} & t_{r,r+1} & & & & \\ & & & & t_{r+1,r} & t_{r+1,r+1} & & & & \\ & & & & & \ddots & \ddots & & & \\ & & & & & & \ddots & \ddots & & \\ & & & & & & & t_{N-1,N-2} & t_{N-1,N-1} & t_{N-1,N} \\ & & & & & & & & t_{N,N-1} & t_{N,N} \end{pmatrix} \quad (7.3.2.2.1)$$

The dimension of this square tri-diagonal matrix is equal to the number of unique values of n_s for all $s \in S$. This is equal to *N+1* since the domain of n_s is $[0, N]$. The $t_{i,j}$ are (not necessarily square) matrices defined on $0 \le i, j \in \mathbb{Z}^+ \le N; j = i-1, i, i+1$. When $j = i-1, i, i+1$ the respective matrices are known as LEFT, MIDDLE and RIGHT. The matrix $t_{i,j}$ relates to transitions **to** states in which $n_s = i$ **from** states in which $n_s = j$. The dimension of each $t_{i,j}$ is thus dependent on the number of states for which $n_s = i$ and $n_s = j$.

The first consideration must be to determine the number of states, $\eta_{k,i}$, for which $n_s = i$ for the support given above. The table below lists the values of $\eta_{k,i}$ given a particular value of *k* and assuming $i \le r$.

Number of states		***i***				
		0	**1**	**2**	**3**	**4**
k	**1**	1	1	1	1	1
	2	1	2	3	4	5
	3	1	3	6	10	15
	4	1	4	10	20	35
	5	1	5	15	35	70

For example, if $i = 2, k = 3$ there are six possible states of the system; $\langle 2,0,0 \rangle, \langle 2,1,0 \rangle, \langle 2,0,1 \rangle, \langle 2,2,0 \rangle, \langle 2,1,1 \rangle, \langle 2,0,2 \rangle$. Clearly it can be seen from the above table

that the number of states is given by the Binomial coefficient $\binom{i+k-1}{i}$. For mathematical rigour this is proved forthwith. By considering the difference $d_{k,i}$ in the number of states for *k* and *k+1* for each *i* the following is obtained:

Difference in the number of states		***i***				
		0	**1**	**2**	**3**	**4**
k	**1**	0	1	2	3	4
	2	0	1	3	6	10
	3	0	1	4	10	20
	4	0	1	5	15	35

Note that $\eta_{k,i} = 1 + \sum_{l=1}^{k-1} d_{l,i}$. It can be seen that the series given in the above four rows can be written

$$\begin{aligned} d_{1,i} &= i \\ d_{2,i} &= \sum_{l=0}^{i} l \\ d_{3,i} &= \sum_{l_1=0}^{i} \sum_{l_2=0}^{l_1} l_2 \\ d_{4,i} &= \sum_{l_1=0}^{i} \sum_{l_2=0}^{l_1} \sum_{l_3=0}^{l_2} l_3 \end{aligned} \qquad (7.3.2.2.2)$$

Since $\sum_{l=0}^{i} l$ is a triangular number, its value can be written $\frac{i(i+1)}{2}$.

Substituting this into $d_{3,i}$ gives $d_{3,i} = \sum_{l_1=0}^{i} \frac{l_1(l_1+1)}{2} = \frac{1}{2}\sum_{l_1=0}^{i} l_1^2 + \frac{1}{2}\sum_{l_1=0}^{i} l_1$. By Faulhaber's (or Bernoulli's) formula, $\sum_{l_1=0}^{i} l_1^2 = \frac{i(i+1)(2i+1)}{6}$. Therefore, $d_{3,i} = \frac{i(i+1)(i+2)}{6}$ (the tetrahedral number).

The fourth equation of (7.3.2.2.2) can then be written

$d_{4,i} = \sum_{l_1=0}^{i} \frac{l_1(l_1+1)(l_1+2)}{6} = \frac{1}{6}\sum_{l_1=0}^{i} l_1^3 + \frac{1}{2}\sum_{l_1=0}^{i} l_1^2 + \frac{1}{3}\sum_{l_1=0}^{i} l_1$. Again, by Faulhaber's formula,

$\sum_{l_1=0}^{i} l_1^3 = \left(\frac{i(i+1)}{2}\right)^2$. This means $d_{4,i} = \frac{i(i+1)(i+2)(i+3)}{24}$ (the Pentatopic number).

It is thus postulated that the difference between states *k* and *k+1* for each *i* is

$d_{k,i} = \sum_{l_1=0}^{i} \cdots \sum_{l_{k-1}=0}^{l_{k-2}} l_{k-1} = \frac{i(i+1)(i+2)\cdots(i+k-1)}{i!}$. This is a Figurate number whose value can be represented by the Binomial coefficient $\binom{i+k-1}{i}$.

Therefore, $\eta_{k,i} = 1 + \sum_{l=1}^{k-1} \binom{i+l-1}{i}$ which simplifies to $\eta_{k,i} = \sum_{l=0}^{k-1} \binom{i+l-1}{i} = \binom{i+k-1}{i}$ as required. If $i > r$ then $\eta_{k,i} = \eta_{k,r} = \binom{r+k-1}{r}$ since the only difference in state definition is the value of n_s. This is because additional customers do not enter service if their arrival occurs when the number in system is above (or equal to) the number of service channels; they simply wait in the queue.

The total number of states in the system $M \mid E_k \mid r \mid N$ is therefore

$$\eta_{k,total} = \sum_{i=0}^{r} \eta_{k,i} + (N-r)\eta_{k,r} = \sum_{i=0}^{r} \binom{i+k-1}{i} + (N-r)\binom{r+k-1}{r} \qquad (7.3.2.2.3)$$

7.3.2.2.2 Ordering of the states

The state of the system can be described by the *k*-tuple $\langle n_s, n_2, ..., n_k \rangle$, where n_i is the number of customers in service phase *i* for $2 \le i \le k$ and $n_1 = \begin{cases} n_s - \sum_{i=2}^{k} n_i & n_s \le r \\ 1 & n_s > r \end{cases}$. The states are principally ordered by increasing value of n_s. The matrix $t_{i,j}$ specifies transition rates **from** states in which $n_s = j$ **to** states in which $n_s = i$. Since the number of states in which the

number in system is equal to *i* and *j* is $\eta_{k,i}$ and $\eta_{k,j}$ respectively then the dimension of the matrix $t_{i,j}$ is $\eta_{k,i}$ (rows) by $\eta_{k,j}$ (columns).

Each of the states within this matrix is thereafter sorted by decreasing value in the lowest-order phase. The following example of the ordering of states for $n_s = 3$ with $k = 3$ illustrates this; $\langle 3,0,0\rangle, \langle 3,1,0\rangle, \langle 3,0,1\rangle, \langle 3,2,0\rangle, \langle 3,1,1\rangle, \langle 3,0,2\rangle, \langle 3,3,0\rangle, \langle 3,2,1\rangle, \langle 3,1,2\rangle, \langle 3,0,3\rangle$. Note that there are $\eta_{3,3} = \begin{pmatrix} 3+3-1 \\ 3 \end{pmatrix} = {}^5C_3 = 10$ states here. When ordered, the latitudinal states (positioned on the horizontal of the transpose transition matrix) are known as the FROM states, whilst the longitudinal states are known as the TO states. For example, the element of the matrix Q^T given by the co-ordinate $\left(\text{col}=\langle 3,0,1\rangle, \text{row}=\langle 2,0,0\rangle\right)$ refers to the rate at which transitions occur **from** state $\langle 3,0,1\rangle$ **to** state $\langle 2,0,0\rangle$.

The justification for choosing this ordering will become clear when populating the sub-matrices of Q^T.

7.3.2.2.3 Specification of the sub-matrices for i = 0,1,…,r

The LEFT matrices

These are the matrices $t_{i,i-1}$ that are defined on $1 \le i \in \mathbb{Z}^+ \le r$. Each matrix specifies the transition rates **from** states in which there are $i-1$ customers in the system **to** states in which there are *i* customers in the system. This is only possible when an arrival occurs.

The sub-matrix $t_{i,i-1}$ contains $\eta_{k,i}$ rows and $\eta_{k,i-1}$ columns. Since an arrival from any of the $\eta_{k,i-1}$ states in this sub-matrix can take the system to one state only, there will be empty rows within this sub-matrix when $k \ge 2$. These empty rows will populate the lower end of the sub-matrix because the higher-order states in the MIDDLE matrix, $t_{i,i}$, cannot be reached by an arrival. For example, when $k = 2, r \ge 4$ the state $\langle 4,4\rangle$ cannot be reached by an arrival, but the lower-order states, $\langle 4,0\rangle, \langle 4,1\rangle, \langle 4,2\rangle, \langle 4,3\rangle$ can.

The sub-matrix $t_{i,i-1}$ can therefore be written $\left(t_{i,i-1,1} \quad 0\right)^T$ where $t_{i,i-1,1}$ is a diagonal square matrix of dimension $\eta_{k,i-1}$ whose nonzero entries are populated by the parameter, λ. The zero matrix is of dimension $\eta_{k,i-1}$ (columns) by $\eta_{k,i} - \eta_{k,i-1}$ (rows).

For example, the sub-matrix $t_{i,i-1}$ for $k = i = 3 \le r$ has dimension 10 (rows) by 6 (columns);

$$t_{3,2} = \begin{pmatrix} \lambda & & & & & \\ & \lambda & & & & \\ & & \lambda & & & \\ & & & \lambda & & \\ & & & & \lambda & \\ & & & & & \lambda \end{pmatrix}^T \qquad (7.3.2.2.4)$$

If $\tau_{i,j}$ denotes the value of the element in the *i*-th column and *j*-th row of the sub-matrix then

$$\tau_{i,j} = \begin{cases} \lambda & i = j \\ 0 & otherwise \end{cases}$$

The RIGHT matrices

These are the matrices $t_{i,i+1}$ that are defined on $0 \le i \in \mathbb{Z}^+ \le r-1$. Each matrix specifies the transition rates **from** states in which there are *i+1* customers in the system **to** states in which there are *i* customers in the system. This is only possible upon the completion of service at one of the channels, i.e. a phase completion from the *k*-th phase.

The sub-matrix $t_{i,i+1}$ contains $\eta_{k,i+1}$ columns and $\eta_{k,i}$ rows. It is not diagonal but is sparse and contains only $\eta_{k,i}$ nonzero elements. Each row of the sub-matrix contains an element since it is always a viable possibility that a service can complete, i.e. it is always possible to reach a TO state through a service completion when $0 \le i \in \mathbb{Z}^+ \le r-1$.

The sub-matrices $t_{2,3}$ for $k = 2$ and $k = 4$ with $r \ge 3$ are given below.

	3,0	**3,1**	**3,2**	**3,3**
2,0		m		
2,1			2m	
2,2				3m

			m																
						m													
								m											
									2m										
												m							
														m					
															2m				
																	m		
																		2m	
																			3m

(7.3.2.2.5)

As an example, for $k=2$ the transition rate **from** state $\langle 3,2\rangle$ **to** state $\langle 2,1\rangle$ is 2μ (in red). This is because there is twice the chance of a service completion since there are two customers in the final phase. Upon completion, there is one less in both the system and the final phase.

The identification of the nonzero elements is a more advanced task than for the LEFT matrices. So too is the determination of their values. For the example of $k=4$ above, it can be seen that the following holds:

$$\tau_{c+x_c,c} = z_c\mu \qquad 1 \le c \in \mathbb{Z}^+ \le 10$$

where z_c is the c-th term in the expansion of $\sum_{l_1=1}^{3}\sum_{l_2=1}^{l_1}\sum_{l_3=1}^{l_2}\{(l_3)\}$ and x_c is the c-th term in the expansion of $\sum_{l_1=1}^{3}\sum_{l_2=1}^{l_1}\sum_{l_3=1}^{l_2}\left\{\sum_{p=1}^{3}\binom{l_p-p+2}{3-p}\right\}$.

After careful examination of other scenarios, the general case for $r,k\in\mathbb{Z}^+, 0\le i\in\mathbb{Z}^+\le r-1$ is found;

$$\tau_{c+x_c,c} = z_c\mu \qquad 1 \le c \in \mathbb{Z}^+ \le \eta_{k,i}$$

where z_c is the c-th term in the expansion of $\sum_{l_1=1}^{i+1}\sum_{l_2=1}^{l_1}\cdots\sum_{l_{k-1}=1}^{l_{k-2}}\{(l_{k-1})\}$ and x_c is the c-th term in the expansion of $\sum_{l_1=1}^{i+1}\sum_{l_2=1}^{l_1}\cdots\sum_{l_{k-1}=1}^{l_{k-2}}\left\{\sum_{p=1}^{k-1}\binom{l_p-p+k-2}{k-p-1}\right\}$.

This can be used to verify $t_{2,3}$ for $k=2$ in the example above.

The MIDDLE matrices

These are the matrices $t_{i,i}$ that are defined on $0 \le i \in \mathbb{Z}^+ \le r$. Each matrix specifies the transition rates **from** and **to** states in which there are *i* customers in the system. This is only possible upon the completion of a service phase that is not the *k*-th phase. The state of the system is therefore changed but the number in system remains the same.

The square sub-matrix $t_{i,i}$ is of dimension $\eta_{k,i}$. The first consideration is the diagonal elements. Each row of the transpose transition matrix, Q^T, represents an equation deduced by equating the inward and outward flow for a particular state of the system. The value of the diagonal element of a particular row is therefore equal to the outward flow of that state. This is dependent only on the number in service. The diagonal elements are equal to $-(\lambda + i\mu)$. This can be easily verified by example, e.g. see Ch 7.2.1.3. Thus, $\tau_{c,c} = -(\lambda + i\mu)$ for $1 \le c \in \mathbb{Z}^+ \le \eta_{k,i}$.Note that if $r = N$ then $\lambda = 0$ when $i = r$ since no more (non-trivial) arrivals can occur.

The second consideration is that of service completion from any of the phases $1, 2, ..., k-1$. The sub-matrices $t_{2,2}$ for $k = 2$ and $k = 4$ with $r \ge 3$ are given below[6].

	2,0	**2,1**	**2,2**
2,0	-(l+2m)		
2,1	2m	-(l+2m)	
2,2		m	-(l+2m)

	2,0,0,0	**2,1,0,0**	**2,0,1,0**	**2,0,0,1**	**2,2,0,0**	**2,1,1,0**	**2,1,0,1**	**2,0,2,0**	**2,0,1,1**	**2,0,0,2**
2,0,0,0	-(l+2m)									
2,1,0,0	2m	-(l+2m)								
2,0,1,0		m	-(l+2m)							
2,0,0,1			m	-(l+2m)						
2,2,0,0		m			-(l+2m)					
2,1,1,0			m		2m	-(l+2m)				
2,1,0,1				m		m	-(l+2m)			
2,0,2,0						m		-(l+2m)		
2,0,1,1							m	2m	-(l+2m)	
2,0,0,2									m	-(l+2m)

[6] Note that by ordering the states in accordance to the policy outlined in Ch 7.3.2.2.2 a lower triangular matrix is obtained for each of the MIDDLE sub-matrices

Again, through careful assessment of these and other scenarios, the general result for the population of the non-diagonal elements of $t_{i,i}$ is found for $r,k \in \mathbb{Z}^+, 1 \le i \in \mathbb{Z}^+ \le r$. Since there are a total of $k-1$ phases whose service completion is pertinent to the MIDDLE sub-matrix, it is necessary to perform $k-1$ repetitions of the following result. For $\kappa = 1,2,...,k-1$:

$$\tau_{c+x_c,c+y_c} = z_c\mu \qquad\qquad 1 \le c \in \mathbb{Z}^+ \le \eta_{k,i-1}$$

where z_c is the *c*-th term in the expansion of $\sum_{l_1=1}^{i}\sum_{l_2=1}^{l_1}\cdots\sum_{l_\kappa=1}^{l_{\kappa-1}}\sum_{l_{\kappa+1}=1}^{\binom{l_\kappa+k-\kappa-2}{k-\kappa-1}}\left\{\left(l_{\kappa-1}-l_\kappa+1\right)\right\}$, x_c is the *c*-th term in the expansion of $\sum_{l_1=1}^{i}\sum_{l_2=1}^{l_1}\cdots\sum_{l_{\kappa-1}=1}^{l_{\kappa-2}}\sum_{l_\kappa=1}^{\binom{l_{\kappa-1}+k-\kappa-1}{k-\kappa}}\left\{\sum_{p=1}^{\kappa-1}\binom{l_p-p+k-2}{k-p-1}\right\}$, and y_c is the *c*-th term in the expansion of $\sum_{l_1=1}^{i}\sum_{l_2=1}^{l_1}\cdots\sum_{l_\kappa=1}^{l_{\kappa-1}}\sum_{l_{\kappa+1}=1}^{\binom{l_\kappa+k-\kappa-2}{k-\kappa-1}}\left\{\sum_{p=1}^{\kappa}\binom{l_p-p+k-2}{k-p-1}\right\}$.

This can be used to verify $t_{2,2}$ for $k=2$ and $k=4$ in the examples above.

7.3.2.2.4 Important patterns and trends

The identification and understanding of underlying trends in the matrix Q^T for a variety of *k* and *r* has led to the results in Ch 7.3.2.2.3. A brief description of these is now given.

The LEFT matrices

The λ 's in these matrices are contained exclusively in the rows whose corresponding TO states contain at least one customer in the first phase. For example in (7.3.2.2.4) the TO states of the rows that contain λ 's are $\langle 3,0,0\rangle, \langle 3,1,0\rangle, \langle 3,0,1\rangle, \langle 3,2,0\rangle, \langle 3,1,1\rangle, \langle 3,0,2\rangle$. This is because the arriving customer always joins the first phase of a service channel (for $i \le r$).

The RIGHT matrices

The μ 's in these matrices occur only in columns in which the corresponding FROM state contains a nonzero entry for n_k. In fact, the coefficient of μ is equal to this value of n_k. This is because the rate at which services complete at phase *k* is linearly proportionate to the number in phase *k*.

Note also, the pattern in (7.3.2.2.5). From right to left the number of consecutive columns that are populated is 3 2 1, 2 1, 1. Let the three shaded boxes denote the groups. There is a spacing of one column between elements of each group and a spacing of two between groups.

There is also a certain degree of repetition. The sub-matrix $t_{2,3}$ described in (7.3.2.2.5) is, in fact, an extension of the sub-matrix $t_{1,2}$ which contains only the first two groups of $t_{2,3}$. If $r \geq 4$ the sub-matrix $t_{3,4}$ thus contains a fourth group (4 3 2 1). The addition of a group is due to the growth in the combination of system states borne by an additional customer in service.

The MIDDLE matrices

As with the RIGHT matrices, the μ_κ 's in these matrices occur only in columns in which the corresponding FROM state contains a nonzero entry for n_κ for $\kappa = 1,2,...,k-1$. For example, the column corresponding to the FROM state $\langle 2,1,1,0 \rangle$ contains an element for μ_2 and μ_3. Thus $n_2 = n_3 = 1$ and since service rates are equal ($\mu_\kappa = \mu$) there are two separate nonzero elements in this column; both of value μ.

By studying the sub-matrices $t_{3,3}$ and $t_{4,4}$ for this value of *k* (i.e. 4) and others a pattern and level of repetition similar to that of the RIGHT sub-matrices is found. An evaluation of this here would be lengthy and digressive and is therefore omitted.

7.3.2.2.5 Specification of the sub-matrices for i = r+1, r+2,...,N

The LEFT matrices

These are the matrices $t_{i,i-1}$ that are defined on $r+1 \leq i \in \mathbb{Z}^+ \leq N$. The dimension of these is equivalent to that of $t_{r,r}$ since $\eta_{k,r} = \eta_{k,r+1} = \eta_{k,r+2} = ... = \eta_{k,N}$ (see Ch 7.3.2.2.1). Since it is possible[7] to access any state with $n_s = i$ from a state with $n_s = i-1$ the matrix is a (square) matrix with diagonal entries equal to λ. That is, $\tau_{c,c} = \lambda$ for $1 \leq c \in \mathbb{Z}^+ \leq \eta_{k,r}$.

The RIGHT matrices

These are the matrices $t_{i,i+1}$ that are defined on $r \leq i \in \mathbb{Z}^+ \leq N-1$. The dimension of these is, as before, equivalent to that of $t_{r,r}$. It is possible to write $t_{i,i+1} = \begin{pmatrix} t_{r-1,r} & 0 \end{pmatrix}^T$. The number of

[7] Because arrivals after $n_s = r$ serve only to increase the value of n_s (and not n_1)

zero row vectors is equal to $\eta_{k,r} - \eta_{k,r-1}$. The TO state corresponding to each zero row vector indicates a state that cannot be reached by a service completion. For example, if $k = 2$ and $r = 3$ then it would not be possible to reach state $\langle 3,3 \rangle$ by a service completion. However, it would be possible to reach states $\langle 3,0 \rangle, \langle 3,1 \rangle, \langle 3,2 \rangle$ since these contain at least one customer in the initial phase.

The MIDDLE matrices

These are the square matrices $t_{i,i}$ of dimension $\eta_{k,r}$ that are defined on $r+1 \le i \in \mathbb{Z}^+ \le N$. These matrices are identical in value to $t_{r,r}$. Note that any λ contained in the final MIDDLE sub-matrix (i.e. $t_{N,N}$) must be set equal to zero to indicate that a non-trivial arrival is not possible when capacity is reached.

7.3.2.2.6 Traffic intensity

This is defined by the division of the mean arrival rate by the mean service rate. The mean rate into the system is λ. To determine the mean rate of service of the system, it is first necessary to determine the service rate of each service channel. The mean service time of each service channel is $k\mu^{-1}$ since the mean service time of each service phase is μ^{-1}. Therefore the service rate per channel is $k^{-1}\mu$. The service rate of the system can thus be written $rk^{-1}\mu$ since there are *r* service channels. In a system with no upper bound on capacity it is a well known result that traffic intensity must be less than one for steady-state results to exist. Thus, the arrival rate is chosen such that $\lambda < rk^{-1}\mu$. Otherwise, the queue length would grow ever larger over time and the concept of equilibrium results would be nonsensical.

However, this is not the case for a system with limited capacity since arrivals to the queue are not possible when the maximum number in system is reached. Any arrivals that do occur when the system is at capacity are rejected. The probability that an arrival is rejected is therefore

$$P_R = \bar{P}_N = \sum_{s \in S_{n_s = N}} P_s \qquad (7.3.2.2.6)$$

The mean non-trivial arrival rate is hence defined

$$\tilde{\lambda} = (1 - P_R)\lambda \qquad (7.3.2.2.7)$$

For a stable system the arrival rate is equal to the departure rate and so

$$\tilde{\lambda} = \pi k^{-1} \mu \tag{7.3.2.2.8}$$

where π is the mean occupancy of the service channels defined as

$$\pi = \sum_{n=0}^{r} n\bar{P}_n + r \sum_{n=r+1}^{N} \bar{P}_n \tag{7.3.2.2.9}$$

The individual server utilisation, $\dot{\pi}$, is defined

$$\dot{\pi} = \pi / r \tag{7.3.2.2.10}$$

Since the maximum attainable value of π is *r* then $\tilde{\lambda} \le rk^{-1}\mu$.

Note that it is indeed possible to obtain $\tilde{\lambda} < \lambda$ with $\lambda < rk^{-1}\mu$ should $N > r$ be too small. However, if $\lambda > rk^{-1}\mu$ then it is a certainty that $\tilde{\lambda} < \lambda$.

7.3.2.2.7 Automation through computer

So far, formulae have been developed to populate the transpose transition matrix, Q^T. The next stage is to solve the linear system (7.2.1.2.2). For low-order this can be done by hand. However, as *r* and *k* are increased in value, it quickly becomes obvious that a computer is required.

A program[8] has been coded in the mathematics program Maple. There are five inputs that are required from the user. These are $k \in \mathbb{Z}^+, r \in \mathbb{Z}^+, N \in \mathbb{Z}^+ \ge r, \lambda > 0, \mu > 0$. For symbolic results, the value of λ and μ should equal *l* and *m* respectively. Otherwise, they may equal any value (expressed as a fraction or decimal) providing traffic intensity is less than one.

Steady-state probabilities for each state of the system can be deduced from a vector of results, *P*, whose dimension is equal to (7.3.2.2.3). If symbolic results are sought, each element is expressed as a function of the parameters *l* and *m*. Otherwise; they are expressed either as a fraction or a decimal.

[8] Obtainable from the author

7.3.2.2.8 Validity

For each run of the program the validity of the result can be confirmed by the normalisation equation, i.e. sum of steady-state probabilities equal to one. The summation of the elements of vector *P* is calculated automatically by the program and so validity is ensured by checking this is equal to one.

Another Maple program has been set up to test the overall validity of the methodology. This produces multiple runs of the (original) program for, initially low-order, *k, r, N* and checks that the normalisation equation holds. The program terminates when a scenario is found to which the contrary is true. After a significant amount of time no such instance was recorded. It must be noted that whilst this can give a good indication of validity, it cannot give explicit verification.

A more efficient test is to check the validity of the transpose transition matrix. This matrix is valid if the sum of each column is equal to zero. It is postulated that if this matrix is valid then the solution vector obtained by Maple is also valid. Therefore, the program does not need to solve the linear system (7.2.1.2.2). This reduces the time required to test validity, thus enabling more solutions to be checked in a given time. After a sufficient[9] amount of time with no instance of invalidity it was determined that the methodology is sound.

7.3.2.2.9 Efficiency

The following tables contain results for the computational time required to calculate the solution vector, *P*, on a modern desktop computer[10]. As part of the computer program introduced in Ch 7.3.2.2.7 the user is required to input a level of precision that Maple works to. This is expressed as a number of digits[11] that Maple uses when making calculations. A higher level of precision enables greater accuracy whilst a lower level enables greater efficiency. For symbolic results (4 digit precision):

[9] In my opinion only
[10] Windows XP, Core 2 Duo 3GHz, 3 GB RAM
[11] Default is 10

Table 7.1 Transition matrix dimension and solution time for symbolic results

r	N	k	$\eta_{k,total}$	time (s)
1	1	1	2	2
2	5	2	15	4
5	10	2	51	19
10	20	2	176	956
2	5	3	28	6
5	10	3	161	824
2	5	4	45	14

Clearly as the number of states in a system increases the time taken to solve it becomes larger. This is due to the increase in the dimension of the transition matrix. For numerical results, the values of λ and μ are both converted to fractions. For numeric results ($\lambda = 0.1, \mu = 2$, 4 digit precision):

Table 7.2 Transition matrix dimension and solution time for numerical results

r	N	k	$\eta_{k,total}$	time (s)
10	20	2	176	10
20	30	2	441	80
20	50	2	861	533
5	10	3	161	9
10	15	3	616	494
5	10	4	406	107

Whilst higher-order systems can be solved this requires a greater amount of time, RAM[12] or both.

7.4 Conclusion

A multi-server finite-buffer[13] queuing system with Markovian arrivals and Erlang distributed service times has been studied in this chapter for a number of low-order systems and for the general case. The principal objective has been to determine stationary probabilities for the system being in each possible state $s \in S$. For the general case an approach is developed to populate a transition matrix by the use of a number of specifically designed formulae. This

[12] Random access memory
[13] i.e. limited waiting space for customers

process is automated through a computer program which also solves the transition matrix providing fast and reliable results.

The use of a computer in setting up the transition matrix is essential since it is vital to be able to set up the linear system for any number of possible scenarios – not just the one that represents the current system on the ground. For example, it would be quicker to set up an $M|E_3|21|50$ system than develop and code the formulae for the general case but in doing so it would require similar systems to be set up from scratch if a different scenario is considered. This is relevant since the occupancy varies over time (Ch 6.4.2.2). Even if a computer program is not used to set up the transition matrix it would certainly be required to solve it. In the most simple of models ($k=2, r=21, N=21$) the dimension of the transition matrix would be 253; a system that could not be efficiently and effectively solved by hand.

In conclusion, whilst the Erlang distribution delivers a reasonable amount of versatility as a service time distribution it is unsuitable for representing LOS at Rookwood NRC. This is because its coefficient of variation is less than one for $k>1$ (Ch 5.5.3.2) whilst the empirical coefficient of variation is greater than one (1.018 – Ch 6.4.2.1). It is also because it does not distinguish between active and blocked LOS. What is needed is a convolution of two distributions; one fitted to active LOS and the other fitted to blocked LOS. Both of these issues are considered in the following chapter.

Chapter 8: Further Queuing Systems

8.1 Introduction

This chapter builds upon the earlier work of the previous chapter on the $M \mid E_k \mid r \mid N$ queuing system. The aim is twofold: firstly – to develop a queuing system that is more representative of the Neurological Rehabilitation Centre (NRC) at Rookwood hospital, and secondly – to derive results for the performance measures of this system.

The principle development is to partition total LOS into *active* and *blocked* LOS (see Ch 6.2.2). The Coxian distribution (see Ch 5.5.4) is used to represent active LOS since this is a highly flexible distribution that has produced a good fit to total LOS (Table 6.13). The exponential distribution (see Ch 5.5.1) is used to represent blocked LOS due to its suitability in representing discharge delay times in previous studies (Baber et al, 2008). Total LOS is therefore modelled by an acyclic phase-type distribution. The corresponding queuing system is summarised $M \mid C_{k-1} + M \mid r \mid N$.

First, steady-state results and performance measures are derived for a low-order system in addition to the distribution of waiting and sojourn time. Next, the general case is considered. The formulae for the population of the transpose transition matrix for the $M \mid E_k \mid r \mid N$ system (see Ch 7.3.2.2) are amended to reflect the change in service time distribution. Performance measures are also deduced. Following this, various types of queuing systems with heterogeneous servers are studied. The chapter concludes with an evaluation of a queuing system deemed most appropriate for the NRC at Rookwood hospital.

The notation used in this chapter is similar to that of Chapter 7.

8.2 Two Server, Three Phase

8.2.1 Waiting space for one customer

The $M\,|\,C_2 + M\,|\,2\,|\,3$ system is forthwith examined. As before (Ch 7.2.1.2) the analogy of a guard is used to discriminate between *trivial* (if $n_s = 3$) and *non-trivial* (if $n_s < 3$) arrivals.

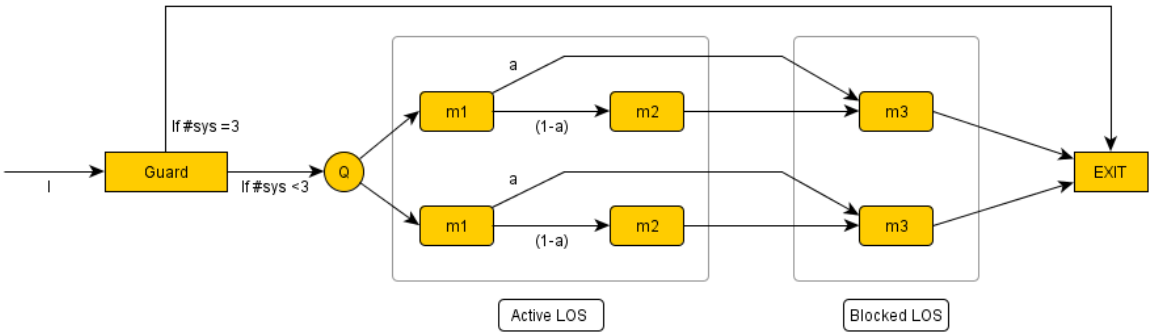


Figure 8.1 The $M/C_2 + M/2/3$ queuing system

Note that in this queuing model 'a' denotes the probability of leaving the active part of the system. This is different from the consideration of phase-type distributions in Ch 5.5.4 where this probability was represented by '$1-\beta$'.

8.2.1.1 Steady-state probabilities

Using the same definition of state space as in the previous chapter (i.e. $\langle n_s, n_2, ..., n_k \rangle$), the continuous-time Markov chain (CTMC) of the above system is depicted in Figure 8.2.

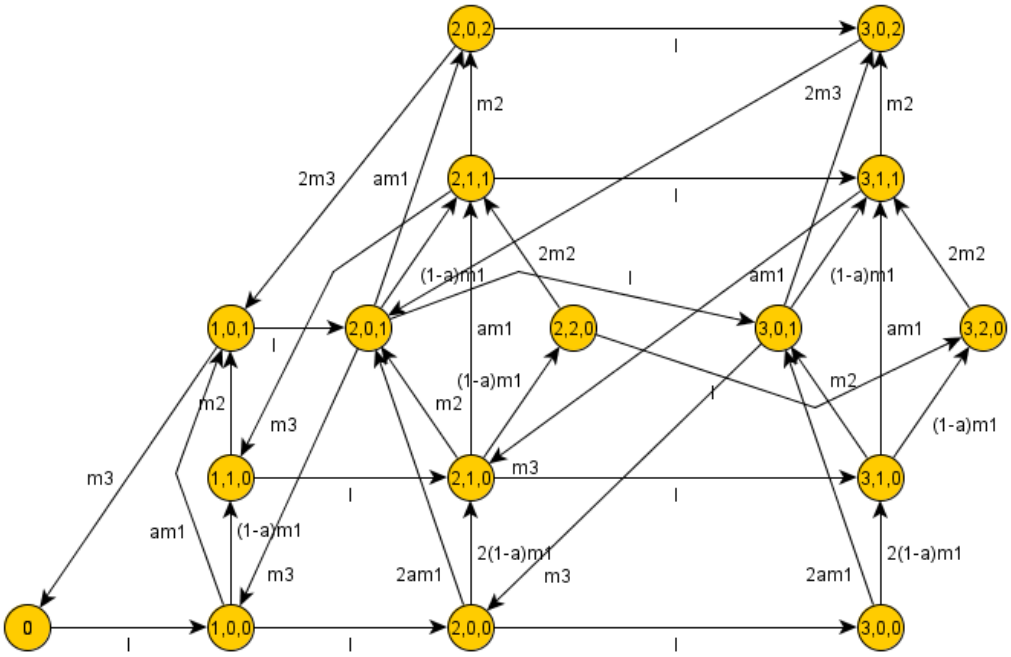


Figure 8.2 CTMC representation of the $M/C_2+M/2/3$ *queuing system*

This system has the same state space, *S*, as the $M|E_3|2|3$ system or, for that matter, any other $M|PH_3|2|3$ system[1]. The difference is in the rate of transition between states. Here, the mean service rate of each phase is assumed to be non-uniform. In addition, it is possible (with probability α) for a customer to move directly from the first service phase to the third.

By equating the inward and outward rate of each of the states the following steady-state equations are found;

[1] PH_k represents a *k*-th order phase-type distribution

$$
\begin{aligned}
&\lambda P_0 = \mu_3 P_{1,0,1}\\
&(\lambda+\mu_1)P_{1,0,0} = \lambda P_0 + \mu_3 P_{2,0,1}\\
&(\lambda+\mu_2)P_{1,1,0} = (1-\alpha)\mu_1 P_{1,0,0} + \mu_3 P_{2,1,1}\\
&(\lambda+\mu_3)P_{1,0,1} = \alpha\mu_1 P_{1,0,0} + \mu_2 P_{1,1,0} + 2\mu_3 P_{2,0,2}\\
&(\lambda+2\mu_1)P_{2,0,0} = \lambda P_{1,0,0} + \mu_3 P_{3,0,1}\\
&(\lambda+\mu_1+\mu_2)P_{2,1,0} = \lambda P_{1,1,0} + 2(1-\alpha)\mu_1 P_{2,0,0} + \mu_3 P_{3,1,1}\\
&(\lambda+\mu_1+\mu_3)P_{2,0,1} = \lambda P_{1,0,1} + 2\alpha\mu_1 P_{2,0,0} + \mu_2 P_{2,1,0} + 2\mu_3 P_{3,0,2}\\
&(\lambda+2\mu_2)P_{2,2,0} = (1-\alpha)\mu_1 P_{2,1,0}\\
&(\lambda+\mu_2+\mu_3)P_{2,1,1} = \alpha\mu_1 P_{2,1,0} + (1-\alpha)\mu_1 P_{2,0,1} + 2\mu_2 P_{2,2,0}\\
&(\lambda+2\mu_3)P_{2,0,2} = \mu_2 P_{2,1,1} + \alpha\mu_1 P_{2,0,1}\\
&(2\mu_1)P_{3,0,0} = \lambda P_{2,0,0} + \mu_3 P_{3,0,1}\\
&(\mu_1+\mu_2)P_{3,1,0} = \lambda P_{2,1,0} + 2(1-\alpha)\mu_1 P_{3,0,0}\\
&(\mu_1+\mu_3)P_{3,0,1} = \lambda P_{2,0,1} + 2\alpha\mu_1 P_{3,0,0} + \mu_2 P_{3,1,0}\\
&(2\mu_2)P_{3,2,0} = \lambda P_{2,2,0} + (1-\alpha)\mu_1 P_{3,1,0}\\
&(\mu_2+\mu_3)P_{3,1,1} = \lambda P_{2,1,1} + \alpha\mu_1 P_{3,1,0} + (1-\alpha)\mu_1 P_{3,0,1} + 2\mu_2 P_{3,2,0}\\
&(2\mu_3)P_{3,0,2} = \lambda P_{2,0,2} + \mu_2 P_{3,1,1} + \alpha\mu_1 P_{3,0,1}
\end{aligned}
\qquad (8.2.1.1.1)
$$

As before (Ch 7.2.1.2) these equations can be used to form a linear system of equations (7.2.1.2.2). This can thereafter be solved to obtain the solution vector of steady-state probabilities. The transpose transition matrix (deduced from 8.2.1.1.1) can be written

$$
\left(\begin{array}{c|cc|ccccc|ccccccc}
-\lambda & & & \mu_3 & & & & & & & & & & & \\
\lambda & -(\lambda+\mu_1) & & & \mu_3 & & & & & & & & & & \\
 & (1-\alpha)\mu_1 & -(\lambda+\mu_2) & & & & & \mu_3 & & & & & & & \\
 & \alpha\mu_1 & \mu_2 & -(\lambda+\mu_3) & & & & & 2\mu_3 & & & & & & \\
\hline
 & \lambda & & -(\lambda+2\mu_1) & & & & & & & & \mu_3 & & & \\
 & & \lambda & 2(1-\alpha)\mu_1 & -(\lambda+\mu_1+\mu_2) & & & & & & & & \mu_3 & & \\
 & & & 2\alpha\mu_1 & \mu_2 & -(\lambda+\mu_1+\mu_3) & & & & & & & & 2\mu_3 & \\
 & & & & (1-\alpha)\mu_1 & & -(\lambda+2\mu_2) & & & & & & & & \\
 & & & & \alpha\mu_1 & (1-\alpha)\mu_1 & 2\mu_2 & -(\lambda+\mu_2+\mu_3) & & & & & & & \\
 & & & & & \alpha\mu_1 & & \mu_2 & -(\lambda+2\mu_3) & & & & & & \\
\hline
 & & & \lambda & & & & & & -2\mu_1 & & & & & \\
 & & & & \lambda & & & & & 2(1-\alpha)\mu_1 & -(\mu_1+\mu_2) & & & & \\
 & & & & & \lambda & & & & 2\alpha\mu_1 & \mu_2 & -(\mu_1+\mu_3) & & & \\
 & & & & & & \lambda & & & & (1-\alpha)\mu_1 & & -2\mu_2 & & \\
 & & & & & & & \lambda & & & \alpha\mu_1 & (1-\alpha)\mu_1 & 2\mu_2 & -(\mu_2+\mu_3) & \\
 & & & & & & & & \lambda & & & \alpha\mu_1 & & \mu_2 & -2\mu_3
\end{array}\right)
\qquad (8.2.1.1.2)
$$

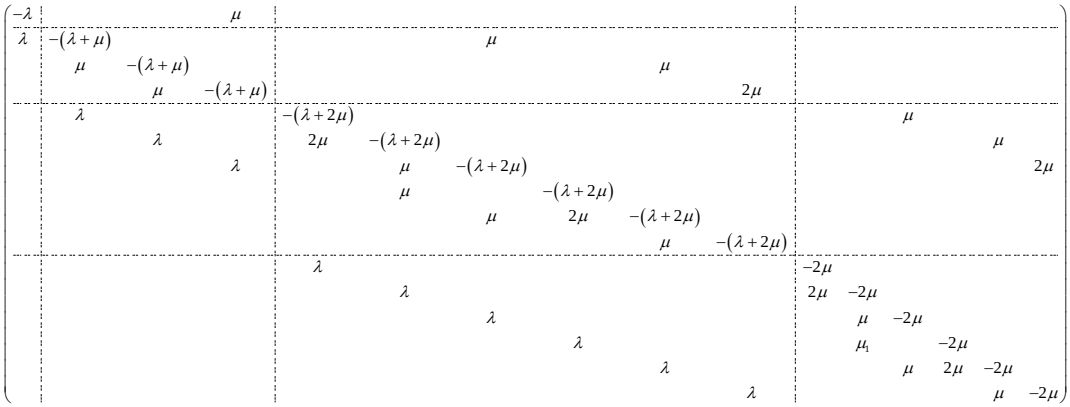
The dashed lines are used to represent the sub-matrices of the system. (8.2.1.1.2) is therefore equivalent to

$$Q^T = \begin{pmatrix} t_{0,0} & t_{0,1} & 0 & 0 \\ t_{1,0} & t_{1,1} & t_{1,2} & 0 \\ 0 & t_{2,1} & t_{2,2} & t_{2,3} \\ 0 & 0 & t_{3,2} & t_{3,3} \end{pmatrix} \tag{8.2.1.1.3}$$

where the dimension of the diagonal square sub-matrices is equal to 1,3,6,6 respectively.

8.2.1.2 Comparison with Erlang service times

It can be seen that transpose transition matrix for the $M\,|\,C_2 + M\,|\,2\,|\,3$ system shares much in common with that of the $M\,|\,E_3\,|\,2\,|\,3$ system given below.



(8.2.1.2.1)

As previously mentioned, there are two critical differences. First is the heterogeneity of service rates between phases (not servers). The coefficients of the service rates (μ_1, μ_2, μ_3) in (8.2.1.1.2) are dependent on the values of n_1, n_2, n_3 in the corresponding FROM state. For example, consider the ninth column. The FROM state that corresponds to this column is $\langle 2,1,1 \rangle$, i.e. $n_1 = 0$ and $n_2 = n_3 = 1$. Any elements in this column that contain a phase service rate have a multiplier given by the corresponding value of n_i for $i = 1,2,3$. Therefore, there are no nonzero entries for μ_1 in this column. See also the tenth column. Since this corresponds to the FROM state $\langle 2,0,2 \rangle$, the only nonzero service rate multiplier is for service phase three. This multiplier is equal to two because there are two customers in this service phase and so the rate of service completion from this phase is double the service node rate.

The second difference is the possibility to bypass the second phase. Since this transition does not lead to a change to the number in system it affects only the MIDDLE sub-matrices. For

example, take the sub-matrix $t_{1,1}$. For Erlang service times there is a zero probability of transition from state $\langle 1,0,0 \rangle$ to state $\langle 1,0,1 \rangle$. Therefore, the lower-left element in this sub-matrix is equal to zero. However, in this case, where such a transition is possible (with probability α), the value of this element is equal to $\alpha\mu_1$. Since the probability of progressing from phase one to phase two is equal to $1-\alpha$, the value of the element immediately above the lower-left element is equal to $(1-\alpha)\mu_1$. This represents the rate of transition from state $\langle 1,0,0 \rangle$ to state $\langle 1,1,0 \rangle$.

8.2.1.3 Distribution of waiting time

The waiting time (*W*) distribution (for non-trivial customers) is derived using a method similar to that used to derive the waiting time distribution for the $M \mid E_2 \mid 2 \mid 3$ system (Ch 7.2.1.2). Therefore, the following is written:

$$P\{\text{Arriving customer must wait}\} = \bar{P}_2 = P_{2,0,0} + P_{2,1,0} + ... + P_{2,0,2} \tag{8.2.1.3.1}$$

$$P(W > t \mid T > 0) = \frac{\sum_{\forall s \in S} P(W > t \mid \zeta = s) P_s}{\sum_{n=0}^{2} \bar{P}_n} \tag{8.2.1.3.2}$$

Since it is known that $P(W > t \mid \zeta = s) = 0$ for $s \in S_{n_s \neq 2}$, (8.2.1.3.2) is simplified to

$$P(W > t \mid T > 0) = \frac{\sum_{\forall s \in S_{n_s=2}} P(W > t \mid \zeta = s) P_s}{\sum_{n=0}^{2} \bar{P}_n} \tag{8.2.1.3.3}$$

The P_s are known (solution of 8.2.1.1.2) so all that remains is to determine $P(W > t \mid \zeta = s)$ for $s \in S_{n_s=2}$. Let $f_i(t)$ represent the *pdf* of time until service completion for a customer in service at phase *i*;

$$
\begin{aligned}
f_3(t) &= LT_s^{-1}\left\{\frac{\mu_3}{\mu_3+s}\right\} = \mu_3 e^{-\mu_3 t} \\
f_2(t) &= LT_s^{-1}\left\{\frac{\mu_2}{\mu_2+s}\frac{\mu_3}{\mu_3+s}\right\} = \frac{\mu_1\mu_3}{\mu_1-\mu_3}\left(e^{-\mu_3 t}-e^{-\mu_1 t}\right) \\
f_1(t) &= LT_s^{-1}\left\{\frac{\mu_1}{\mu_1+s}\frac{\mu_3}{\mu_3+s}\left(a+(1-a)\frac{\mu_2}{\mu_2+s}\right)\right\} \\
&= \frac{\mu_1\mu_3}{(\mu_1-\mu_2)(\mu_1-\mu_3)(\mu_2-\mu_3)}\begin{pmatrix}(\mu_2-\alpha\mu_1)(\mu_2-\mu_3)e^{-\mu_1 t}+\\ \mu_2(\alpha-1)(\mu_1-\mu_3)e^{-\mu_2 t}+\\ (\mu_2-\alpha\mu_3)(\mu_1-\mu_2)e^{-\mu_3 t}\end{pmatrix}
\end{aligned}
\qquad (8.2.1.3.4)
$$

By integration over the limit t,∞ the following are obtained:

$$
\begin{aligned}
F_3(t) &= e^{-\mu_3 t} \\
F_2(t) &= \frac{\mu_1\mu_3}{\mu_1-\mu_3}\left(\mu_3 e^{-\mu_3 t}-\mu_1 e^{-\mu_1 t}\right) \\
F_1(t) &= \frac{1}{(\mu_1-\mu_2)(\mu_1-\mu_3)(\mu_2-\mu_3)}\begin{pmatrix}\mu_3(\mu_2-\alpha\mu_1)(\mu_2-\mu_3)e^{-\mu_1 t}+\\ \mu_1\mu_3(\alpha-1)(\mu_1-\mu_3)e^{-\mu_2 t}+\\ \mu_1(\mu_2-\alpha\mu_3)(\mu_1-\mu_2)e^{-\mu_3 t}\end{pmatrix}
\end{aligned}
\qquad (8.2.1.3.5)
$$

Therefore, by (7.2.1.2.8);

$$
\begin{aligned}
P\left(W>t\,|\,\zeta=\langle 2,0,0\rangle\right) &= F_1(t)F_1(t) \\
P\left(W>t\,|\,\zeta=\langle 2,1,0\rangle\right) &= F_1(t)F_2(t) \\
P\left(W>t\,|\,\zeta=\langle 2,0,1\rangle\right) &= F_1(t)F_3(t) \\
P\left(W>t\,|\,\zeta=\langle 2,2,0\rangle\right) &= F_2(t)F_2(t) \\
P\left(W>t\,|\,\zeta=\langle 2,1,1\rangle\right) &= F_2(t)F_3(t) \\
P\left(W>t\,|\,\zeta=\langle 2,0,2\rangle\right) &= F_3(t)F_3(t)
\end{aligned}
\qquad (8.2.1.3.6)
$$

The distribution of waiting time for non-trivial customers is now calculated through (8.2.1.3.3). The conditional distribution for waiting time for non-trivial customers who have had to wait is found by

$$
P\left(W>t\,|\,W>0\right) = \frac{\sum_{\forall s\in S_{n_s=2}} P\left(W>t\,|\,\zeta=s\right)P_s}{\bar{P}_2} \qquad (8.2.1.3.7)
$$

8.2.1.4 Distribution of sojourn time

The sojourn time (T) distribution is derived by the following formula:

$$P(T>t\,|\,T>0)=\frac{\sum_{\forall s\in S}P(T>t\,|\,\zeta=s)P_s}{\sum_{n=0}^{2}\bar{P}_n} \tag{8.2.1.4.1}$$

Since it is known that $P(T>t\,|\,\zeta=s)=0$ for $s\in S_{n_s>2}$, (8.2.1.4.1) is simplified to

$$P(T>t\,|\,T>0)=\frac{\sum_{\forall s\in S_{n_s\le 2}}P(T>t\,|\,\zeta=s)P_s}{\sum_{n=0}^{2}\bar{P}_n} \tag{8.2.1.4.2}$$

where

$$P(T=t\,|\,\zeta=s)=\begin{cases} f_1(t) & n_s<2 \\ LT^{-1}\left\{LT\{f_1(t)\}*LT\{P(W=t\,|\,\zeta=s)\}\right\} & n_s=2 \\ 0 & n_s>2 \end{cases} \tag{8.2.1.4.3}$$

The results for $P(T>t\,|\,\zeta=s)$ and hence (8.2.1.4.2) follow when (8.2.1.4.3) are integrated over the limit t,∞.

8.2.1.5 Performance measures

There are four main performance summary measures in queuing theory that are inextricably linked. They are the mean waiting time $E[W]$, the mean sojourn time $E[T]$, the mean number in queue $E[L_q]$, and the mean number in system $E[L]$.

The non-trivial mean waiting time can be derived from (8.2.1.3.3) as follows.

$$E[W\,|\,T>0]=\int_0^\infty tP(W=t\,|\,T>0)dt=\int_0^\infty t\left(\frac{\sum_{\forall s\in S_{n_s=2}}P(W=t\,|\,\zeta=s)P_s}{\sum_{n=0}^{2}\bar{P}_n}\right)dt \tag{8.2.1.5.1}$$

This is evaluated using the solution of (8.2.1.1.2). The solutions to (8.2.1.3.6) must also be used, but first they must be deducted from one and differentiated in order to obtain the $P(W=t\,|\,\zeta=s)$ for $s\in S_{n_s=2}$. The output is in terms of the parameters $\lambda,\mu_1,\mu_2,\mu_3,\alpha$.

Upon the determination of this, or any other of the performance measures, the remainder can be found using two important results. The first states that the mean sojourn time is a sum of the mean waiting time and the mean service time;

$$E[T]=E[W]+E[\Upsilon] \tag{8.2.1.5.2}$$

where Υ is defined as the service time. The pdf of service time is $\Upsilon(t)=P(\Upsilon=t)=f_1(t)$ with $f_1(t)$ defined in (8.2.1.3.4). Therefore, $E[\Upsilon]=\int_0^\infty t\Upsilon(t)\,dt$. The second is a well known result in queuing theory. Little's law states that *'the long-term average number of customers in a stable system is equal to the long-term arrival rate multiplied by the long-term average time a customer spends in the system'* (Little, 1961). Typically this is written $E[L]=\lambda E[T]$. However, λ is the long-term arrival rate of the system and hence contains trivial arrivals. The non-trivial long-term (or mean) arrival rate is given by (7.3.2.2.7). Therefore

$$E[L]=\tilde{\lambda}E[T] \tag{8.2.1.5.3}$$

This result can also be applied to the queue;

$$E[L_q]=\tilde{\lambda}E[W] \tag{8.2.1.5.4}$$

Therefore, $E[T],E[L],E[L_q]$ can be respectively deduced through (8.2.1.5.2), (8.2.1.5.3), (8.2.1.5.4) given that $E[W]$ is known.

Note also that the non-trivial mean sojourn time can be determined through (8.2.1.4.1);

$$E[T\,|\,T>0]=\int_0^\infty tP(T=t\,|\,T>0)dt=\int_0^\infty t\left(\frac{\sum_{\forall s\in S_{n_s=2}}P(T=t\,|\,\zeta=s)P_s}{\sum_{n=0}^{2}\bar{P}_n}\right)dt \tag{8.2.1.5.5}$$

This requires an evaluation of the solution to (8.2.1.1.2) and (8.2.1.4.3). As before, the remainder of performance measures can be derived following the determination of this one result.

8.2.2 Infinite waiting space

Knessl et al, 1990 consider the steady-state solution of the $M|G|2|\infty$ queuing system through an integral equation approach. The service channels need not be homogeneous but Markovian densities are required. The approach within this study is to make a number of consecutive transformations that *'reduce the problem of determining the marginal probabilities of the number of customers present to the solution of a pair of coupled integral equations'*. If the servers are homogeneous then it is only necessary to solve the single integral equation

$$sT(s,t)-\int_0^{\infty} e^{\lambda(s-1)z}\left[T(s,z)\Upsilon(z+t)+T(s,z+t)\Upsilon(z)\right]dz=\int_t^{\infty}\Upsilon(z)dz \qquad (8.2.2.1)$$

where *s* is the generating function variable, $\Upsilon(t)$ is the service density , and $T(s,z)$ is the generating function to be found. The solution can then be used in the identities (51), (52) and (53) to obtain steady-state results.

For the $M|C_2+M|2|\infty$ queuing system the service density can be written more compactly as $\Upsilon(t)=\sum_{i=1}^{3}\beta_i e^{-\mu_i t}$ where the β_i's are known constants involving only the parameters μ_1,μ_2,μ_3,α. On substitution into (8.2.2.1)

$$sT(s,t)-\int_0^{\infty} e^{\lambda(s-1)z}\left[T(s,z)\sum_{i=1}^{3}\beta_i e^{-\mu_i(z+t)}+T(s,z+t)\sum_{i=1}^{3}\beta_i e^{-\mu_i z}\right]dz=\sum_{i=1}^{3}\frac{\beta_i}{\mu_i}e^{-\mu_i t} \quad (8.2.2.2)$$

which can be expanded to give

$$sT(s,t)-\sum_{i=1}^{3}\beta_i e^{-\mu_i t}\int_0^{\infty} e^{(\lambda(s-1)-\mu_i)z}T(s,z)dz-\sum_{i=1}^{3}\beta_i\int_0^{\infty} e^{(\lambda(s-1)-\mu_i)z}T(s,z+t)dz-\sum_{i=1}^{3}\frac{\beta_i}{\mu_i}e^{-\mu_i t}=0$$

A change of variable $(u=z+t)$ yields

$$sT(s,t)-\sum_{i=1}^{3}\beta_i e^{-\mu_i t}\int_0^{\infty} e^{(\lambda(s-1)-\mu_i)z}T(s,z)dz-\sum_{i=1}^{3}\beta_i e^{-(\lambda(s-1)-\mu_i)t}\int_t^{\infty} e^{(\lambda(s-1)-\mu_i)u}T(s,u)du-\sum_{i=1}^{3}\frac{\beta_i}{\mu_i}e^{-\mu_i t}=0$$

(8.2.2.3)

This is a second-order mixed integral equation. The integral equation to the left-hand side is Fredholm type whilst the other is Volterra type (since the lower limit of integration is a function of the variable *t*). Unfortunately, this equation does not fit into a common classification to which a direct method of solution is available. Attempts are therefore made to convert this to a differential equation.

After much deliberation a particular solution for $T(s,t)$ could not be found. Problems of a similar nature were also encountered when exponentially distributed service times were considered. For brevity, the process of finding a general solution is explained for this case.

Firstly, with $\Upsilon(t)=\mu e^{-\mu t}$, (8.2.2.3) can be rewritten (with $\beta_1=\mu, \beta_2=\beta_3=0$) as

$$sT(s,t)-\mu e^{-\mu t}\int_0^{\infty} e^{(\lambda(s-1)-\mu)z}T(s,z)dz-\mu e^{-(\lambda(s-1)-\mu)t}\int_t^{\infty} e^{(\lambda(s-1)-\mu)u}T(s,u)du-e^{-\mu t}=0$$

The integral on the left-hand side may be eliminated on application of the operator $\mu+\frac{\partial}{\partial t}$;

$$sT'(s,t)+\mu(s+1)T(s,t)+\mu(\lambda(s-1)-2\mu)e^{-(\lambda(s-1)-\mu)t}\int_t^{\infty} e^{(\lambda(s-1)-\mu)u}T(s,u)du=0$$

The remaining integral may now be eliminated on application of the operator $\lambda(s-1)-\mu+\frac{\partial}{\partial t}$;

$$sT''(s,t)+(\lambda s(s-1)+\mu)T'(s,t)+(\mu s(\lambda(s-1)-\mu)+\mu^2)T(s,t)=0 \qquad (8.2.2.4)$$

This is a second-order homogenous differential equation with constant coefficients. With appropriate substitutions this equation can be expressed as

$$a_2T''(s,t)+a_1T'(s,t)+a_0T(s,t)=0 \qquad (8.2.2.5)$$

By making the substitution $T(s,t)=e^{vt}$ the characteristic equation is found to be

$$a_2v^2+a_1v+a_0=0 \qquad (8.2.2.6)$$

The roots of this equation are deduced by the quadratic formula;

$$\nu_{1,2} = \frac{-a_1 \pm \sqrt{a_1^2 - 4a_2 a_0}}{2a_2}$$

Therefore

$$\nu_1 = -\mu \qquad \nu_2 = \left(1 - s^{-1}\right)\left(\mu - \lambda s\right)$$

The general solution can now be written

$$T\left(s,t\right) = C_1 e^{\nu_1 t} + C_2 e^{\nu_2 t} \tag{8.2.2.7}$$

The authors suggest that the unknown constants can be obtained by substituting (8.2.2.7) into (8.2.2.2). However, this approach was unsuccessful in determining a particular solution. Fortunately, the $M \mid M \mid 2 \mid \infty$ system has been studied as an example within the paper and the solution is written, without working, as

$$T\left(s,t\right) = \frac{2\mu - \lambda\left(s-1\right)}{\left(s-1\right)\left(2\mu - \lambda s\right)} e^{-\mu t} \tag{8.2.2.8}$$

This result is achieved if C_2 is assumed to equal zero and (8.2.2.7) is substituted into (8.2.2.2). However, after studying the methodology of this paper in detail, no boundary conditions could be found that support this assumption. Regrettably, this approach has not yielded the steady-state solution to the $M \mid C_2 + M \mid 2 \mid \infty$ system.

8.3 *k* Phase, *r* Server, Waiting Space for *N*-*r* Customers

The general case, $M \mid C_{k-1} + M \mid r \mid N$, is now considered. This system shares much in common with the system $M \mid E_k \mid r \mid N$. The similarities are discussed in Ch 8.2.2 for case $k = 3, r = 2, N = 3$.

8.3.1 Steady-state probabilities

The long-term results for the probability of the system being in each state, $s \in S$, are forthwith determined.

8.3.1.1 Specification of the transpose transition matrix

The transpose transition matrix, Q^T, is equivalent to that defined by (7.3.2.2.1). That is, the matrix is tri-diagonal with LEFT, MIDDLE, and RIGHT sub-matrices. Since the number of

accessible service phases (*k*) is the same, the cardinality of the state space, *S*, is consistent with that of Erlang service. The number of states within each sub-matrix is also equivalent. Therefore, the dimension of the sub-matrices is unchanged.

8.3.1.2 Ordering of the states

The TO and FROM states of the $M|C_{k-1}+M|r|N$ and $M|E_k|r|N$ systems are identical in both value and order (see Ch 7.3.2.2.2).

8.3.1.3 Specification of the sub-matrices for *i=0,1,…,r*

The LEFT matrices

These sub-matrices represent the rates of transition that result in an additional customer being present in system. This is only possible with an arrival. The sub-matrices are equivalent to those defined in Ch 7.3.2.2.3.

The RIGHT matrices

Rates of transition that result in one fewer customer present in the system are represented by these sub-matrices. This is only possible through the completion of service. Since exit from the system is (still) only possible through the final phase there is no difference between these sub-matrices and those defined in Ch 7.3.2.2.3 (apart from the replacement of μ by μ_k).

The MIDDLE matrices

These sub-matrices contain transition rates that do not affect the number in system. They share some similarity with the MIDDLE sub-matrices of the system with Erlang based service times. There is still a probability attached to the progression from phase i to $i+1$ for $1\le i\in\mathbb{Z}^+\le k-1$. For Erlang service this probability is equal to one (and so the rate is simply μ). Whilst this remains the case for $i=k-1$, when $i<k-1$ the probability is equal to $1-\alpha_i$ (and so the rate is $(1-\alpha_i)\mu_i$). The general result for the population of such elements of $t_{i,i}$ for $r,k\in\mathbb{Z}^+,1\le i\in\mathbb{Z}^+\le r$ is therefore as follows. Since there are a total of $k-1$ phases whose service completion is pertinent to the MIDDLE sub-matrix, it is necessary to perform $k-1$ repetitions of the following result. For $\kappa=1,2,...,k-1$:

$$\tau_{c+x_c,c+y_c}=z_c(1-\alpha_\kappa)\mu_\kappa \quad 1\le c\in\mathbb{Z}^+\le\eta_{k,i-1} \quad 1\le\kappa\in\mathbb{Z}^+\le k-2$$
$$\tau_{c+x_c,c+y_c}=z_c\mu_\kappa \quad 1\le c\in\mathbb{Z}^+\le\eta_{k,i-1} \quad \kappa=k-1$$

with z_c, x_c, y_c given by the appropriate formulae in Ch 7.3.2.2.3. Recall $\tau_{i,j}$ denotes the value of the *i*-th column and *j*-th row of the sub-matrix in question.

It will also be noticed that for $1 \le \kappa \in \mathbb{Z}^+ \le k-2$ there is an alternate transition to phase *k* that bypasses any intermediate phases. This occurs with probability equal to α_κ (rate: $\alpha_\kappa \mu_\kappa$). The following result describes this. For $\kappa = 1, 2, ..., k-2$:

$$\tau_{c+x_c, c+y_c} = z_c \alpha_\kappa \mu_\kappa \quad 1 \le c \in \mathbb{Z}^+ \le \eta_{k,i-1}$$

Consider the pair of alternate elements in the column. Since the FROM state is unchanged, the values of x_c are also unchanged, i.e. x_c is the *c*-th term in the expansion of $\sum_{l_1=1}^{i}\sum_{l_2=1}^{l_1}\cdots\sum_{l_{\kappa-1}=1}^{l_{\kappa-2}}\sum_{l_\kappa=1}^{\binom{l_{\kappa-1}+k-\kappa-1}{k-\kappa}}\left\{\sum_{p=1}^{\kappa-1}\binom{l_p - p + k - 2}{k-p-1}\right\}$. The TO state, however, is not the state that represents movement to the next phase, but the state that represents movement to the final phase. This TO state is equivalent to the TO state of the $(k-1)$-th μ_κ. That is, the element is on the same row as the μ_{k-1}. This can be seen in (8.2.1.2) where the $\alpha\mu_1$'s are on the same row as the μ_2's. Therefore, y_c is the *c*-th term in the expansion of $\sum_{l_1=1}^{i}\sum_{l_2=1}^{l_1}\cdots\sum_{l_{k-1}=1}^{l_{k-2}}\sum_{l_k=1}^{\binom{l_\kappa+1}{0}}\left\{\sum_{p=1}^{k-1}\binom{l_p - p + k - 2}{k-p-1}\right\}$. Since the number in each phase of service is equivalent to that of its alternate element, the rate multiplier is the same. Therefore, z_c is the *c*-th term in the expansion of $\sum_{l_1=1}^{i}\sum_{l_2=1}^{l_1}\cdots\sum_{l_\kappa=1}^{l_{\kappa-1}}\sum_{l_{\kappa+1}=1}^{\binom{l_\kappa+k-\kappa-2}{k-\kappa-1}}\left\{\left(l_{\kappa-1} - l_\kappa + 1\right)\right\}$.

Whilst it is entirely possible to obtain a formula for the population of the diagonal elements, it is more convenient to make use of the condition that column values must sum to zero. Since, at this stage, the other column values are known, all that remains is to subtract these from zero to obtain the value of the diagonal elements.

8.3.1.4 Specification of the sub-matrices for *i=r+1,r+2,...,N*

See Ch 7.3.2.2.5.

8.3.1.5 Traffic intensity

Applying the same rationale of Ch 7.3.2.2.6 the following is obtained:

$$\tilde{\lambda} = \pi E\left[\Upsilon\right]^{-1} \tag{8.3.1.5.1}$$

Using (7.3.2.2.10) the individual server utilisation is therefore

$$\dot{\pi} = \tilde{\lambda} E\left[\Upsilon\right] / r \tag{8.3.1.5.2}$$

The mean service time per channel, $E\left[\Upsilon\right]$, can be derived from the pdf of service time per channel through its Laplace transform;

$$LT_s\left\{\Upsilon(t)\right\} = \frac{\mu_1}{\mu_1+s}\frac{\mu_k}{\mu_k+s}\left[\left(\alpha_1+(1-\alpha_1)\frac{\mu_2}{\mu_2+s}\left(\alpha_2+(1-\alpha_2)\frac{\mu_3}{\mu_3+s}\left(\cdots\cdots\left(\alpha_{k-2}+(1-\alpha_{k-2})\frac{\mu_{k-1}}{\mu_{k-1}+s}\right)\right)\right)\right)\right] \tag{8.3.1.5.3}$$

Note that when $k=3$ this result is analogous to $LT\left(f_1(t)\right)$ in (8.2.1.3.4) where $\Upsilon(t) = f_1(t)$. The following is found by expanding the result contained in the square bracket:

$$\alpha_1 + \alpha_2 x_2 + \alpha_3 x_2 x_3 + \cdots\cdots + \alpha_{k-3} x_2 x_3 \ldots x_{k-2} + \alpha_{k-2} x_2 x_3 \ldots x_{k-1} \tag{8.3.1.5.4}$$

where $x_l := (1-\alpha_{l-1})\frac{\mu_l}{\mu_l+s}$. This series can be summarised within a more convenient formulation of (8.3.1.5.3) given below.

$$LT_s\left\{\Upsilon(t)\right\} = \frac{\mu_1}{\mu_1+s}\frac{\mu_k}{\mu_k+s}\left[\sum_{i=1}^{k-1}\alpha_i\prod_{j=2}^{i}(1-\alpha_{j-1})\frac{\mu_j}{\mu_j+s}\right] \tag{8.3.1.5.5}$$

Note that $\alpha_{k-1} = 0$ as the customer is always transferred to the final phase after a service completion in the $(k-1)$-th phase. The mean service time per channel can be calculated through

$$E[\Upsilon] = \int_0^\infty t\Upsilon(t)\,dt$$
$$= \int_0^\infty tLT_s^{-1}\left\{\frac{\mu_1}{\mu_1+s}\frac{\mu_k}{\mu_k+s}\left[\sum_{i=1}^{k-1}\alpha_i\prod_{j=2}^{i}\left(1-\alpha_{j-1}\right)\frac{\mu_j}{\mu_j+s}\right]\right\}dt \qquad (8.3.1.5.6)$$

For the $M|C_2+M|2|3$ case considered in Ch 8.2 the value of (8.3.1.5.6) is found to be

$$E[\Upsilon] = \frac{1}{\mu_1}+(1-\alpha)\frac{1}{\mu_2}+\frac{1}{\mu_3} \qquad \alpha := \alpha_1 \qquad (8.3.1.5.7)$$

There is, however, an alternative method of determining $E[\Upsilon]$ that is both more efficient and intuitive. Since a customer must pass through phases one and *k* their mean service time is at least the mean of these two phases. Furthermore, the customer passes through phase two with probability $(1-\alpha_1)$ and phase three with probability $(1-\alpha_1)(1-\alpha_2)$ and so on. It is therefore possible to write

$$E[\Upsilon] = \frac{1}{\mu_1}+\frac{1}{\mu_k}+(1-\alpha_1)\frac{1}{\mu_2}+(1-\alpha_1)(1-\alpha_2)\frac{1}{\mu_3}+\ldots\ldots \qquad (8.3.1.5.8)$$

which simplifies to

$$E[\Upsilon] = \frac{1}{\mu_1}+\frac{1}{\mu_k}+\sum_{i=2}^{k-1}\frac{1}{\mu_i}\sum_{j=1}^{i-1}\left(1-\alpha_j\right) \qquad (8.3.1.5.9)$$

This verifies (8.3.1.5.7) when *k* is set equal to three.

8.3.1.6 Computational efficiency

The Maple program of the previous chapter has been amended[2] to incorporate the differences outlined above. The following tables contain results for the computational time required to calculate the solution vector, *P*, on a modern desktop computer. For symbolic results (4 digit precision):

[2] Whilst retaining the original (Erlang service times)

Table 8.1 Transition matrix dimension and solution time for symbolic results

r	N	k	$\eta_{k,total}$	time (s)
2	5	2	15	8
5	10	2	51	69
10	20	2	176	Not solvable
2	5	3	28	85
5	10	3	161	Not solvable
2	5	4	45	Not solvable

The results for solution time are in stark contrast to those obtained with Erlang service times (Table 7.1). This is not because of a difference in transition matrix dimension since the $\eta_{k,total}$ are identical for each of the six comparable examples. Instead it is because of an increase in matrix density (measured by the proportion of the transition matrix populated by nonzero elements) in which nonzero elements contain a greater number of unassigned variables (e.g. $\lambda, \alpha_1, \mu_1, \mu_2$ etc instead of just λ and μ). In three of the six examples this has led to Maple not being able to solve the system (error message: object too large) whilst in others the solution time is greatly inflated. For numerical results ($\lambda = 0.1, \mu_1 = 2, \mu_2 = 3, \mu_3 = 4, \mu_4 = 5, \alpha_1 = 0.2,$ $\alpha_2 = 0.3$, 4 digit precision):

Table 8.2 Transition matrix dimension and solution time for numerical results

r	N	k	$\eta_{k,total}$	time (s)
10	20	2	176	12
20	30	2	441	80
20	50	2	861	542
5	10	3	161	11
10	15	3	616	550
5	10	4	406	107

It is interesting that, on comparison with the numerical results with Erlang service times (Table 7.2), there is little difference in solution time. The differences that do exist could be put down to a difference in transition matrix density. This would also explain why the solution time of a system with 616 states is greater than that of a system with 861 states. Through other trials, it has also been noticed that the number of digits precision can have a significant effect on the solution time for numeric results (since more RAM is required).

8.3.2 Performance measures

It would be very difficult to derive the distribution of waiting or sojourn time for the general case considered here. However, it is still possible to obtain expected results for the performance measures W, T, L_q, L (see Ch 8.2.1.5). It is only necessary to calculate the expected value for one of these measures since the remainder can be derived using equations (8.2.1.5.2), (8.2.1.5.3), and (8.2.1.5.4).

First considered is a *direct*[3] approach for the $M|M|1|\infty$ system. *Mean value analysis* makes use of Little's result (8.2.1.5.3) in addition to the *arrival relation*;

$$E[T] = E[W] + E[\Upsilon] = E[L]E[\Upsilon] + E[\Upsilon] \tag{8.3.2.1}$$

i.e. mean sojourn time equals mean waiting time plus mean service time. These give

$$E[L] = \lambda E[L]E[\Upsilon] + \lambda E[\Upsilon] \tag{8.3.2.2}$$

With $E[\Upsilon]$ known (i.e. μ^{-1}) the value of $E[L]$ is found to be $E[\Upsilon]\left(\lambda^{-1} - E[\Upsilon]\right)^{-1} = \lambda(\mu - \lambda)^{-1}$. Note that the equation (8.3.2.1) requires the application of the Poisson arrivals see time averages (PASTA) property. This infers that the probability of the state as seen by an outside observer is the same as the probability of the state seen by an arriving customer. Moreover, the mean number of customers in system seen by an arriving customer is equal to the mean number seen by an outside observer, $E[L]$. This property is only applicable to systems with Poisson arrivals.

The result (8.3.2.2) cannot readily be used for the system under consideration in this chapter. First, the result must be adjusted to take account of more general (Markovian) service times. To do this the mean waiting time given by $E[L] \cdot E[\Upsilon]$ in (8.3.2.2) is redefined. This is to reflect the difference in time spent in service for the active customer[4] and those in the queue. Let the expected (remaining) time in service for the active customer be defined as $E[\Upsilon_S] = E[R]$. The expected service time for the customers in the queue is obviously equal

[3] Does not require a determination of steady-state probabilities
[4] At time of arrival

to $E\left[\Upsilon_Q\right]=E\left[\Upsilon\right]$. Next, Little's law for both the system (8.2.1.5.3) and the queue (8.2.1.5.4) is applied to equation (8.2.1.5.2) to give

$$E\left[L\right]=E\left[L_q\right]+\lambda E\left[\Upsilon\right] \tag{8.3.2.3}$$

Therefore, the expected number in service and in the queue is $E\left[L_S\right]=\lambda E\left[\Upsilon\right]$ and $E\left[L_Q\right]=E\left[L_q\right]$ respectively. It follows that:

$$\begin{aligned} E\left[W\right]&=E\left[L\right]E\left[\Upsilon\right]=E\left[L_S\right]E\left[\Upsilon_S\right]+E\left[L_Q\right]E\left[\Upsilon_Q\right] \\ &=\lambda E\left[\Upsilon\right]E\left[R\right]+E\left[L_q\right]E\left[\Upsilon\right]=\lambda E\left[\Upsilon\right]E\left[R\right]+\lambda E\left[W\right]E\left[\Upsilon\right] \end{aligned} \tag{8.3.2.4}$$

The mean waiting time can therefore be written

$$E\left[W\right]_{M|G|1|\infty}=\frac{\rho}{1-\rho}E\left[R\right] \tag{8.3.2.5}$$

where ρ, the non-trivial traffic intensity, is defined as $\tilde{\lambda}E\left[\Upsilon\right]$. The Pollaczek-Khintchine formula states that the mean waiting time for the $M|G|1|\infty$ system is

$$E\left[W\right]=E\left[\Upsilon\right]\frac{\rho}{1-\rho}\frac{1+C^2}{2} \tag{8.3.2.6}$$

where $C^2=E\left[\Upsilon\right]^{-2}\left(E\left[\Upsilon^2\right]-E\left[\Upsilon\right]^2\right)$ is the coefficient of variation of the service time distribution. Therefore, the expected remaining time in service for the active customer is

$$E\left[R\right]=\frac{E\left[\Upsilon^2\right]}{2E\left[\Upsilon\right]} \tag{8.3.2.7}$$

Secondly, the generalisation of the number of service channels is considered. More of an intuitive approach is required here. In the $M|G|1|\infty$ system detailed above, $\lambda E\left[\Upsilon\right]$ in (8.3.2.4) is defined as $E\left[L_S\right]$, the mean number of customers in service. It does, however, have another meaning; the probability that there is a customer in service. Put alternatively, this is the probability that a customer will have to wait, P_W. In the $M|G|r|\infty$ system this probability is defined when all service channels are occupied.

Care must also be taken in the interpretation of $E\left[\Upsilon_Q\right]=E\left[\Upsilon\right]$. In the above workings $E\left[\Upsilon\right]$ is used to represent the mean service time (per channel) when in actual fact, what is of interest is the mean time between service completions, δ. In the $M|G|1|\infty$ system these results are equivocal, but this is not the case in the multi-server system. In the $M|G|r|\infty$ system the service rate is the product of *r* and the service rate per channel, $E\left[\Upsilon\right]^{-1}$. Therefore, the mean time between service completions is given by

$$\delta \approx \frac{1}{r}E\left[\Upsilon\right] \tag{8.3.2.8}$$

This result is approximate since the time until a departure is not dependent on the state of the system before the departure. In other words, the actual positions of the *r* customers in service are not taken into account. By taking into account these two differences, (8.3.2.4) can be rewritten as

$$E\left[W\right] \approx P_W E\left[R\right] + \lambda E\left[W\right]\frac{1}{r}E\left[\Upsilon\right] \tag{8.3.2.9}$$

where $E\left[R\right]$ is the expected remaining time until an active service completes. Therefore

$$E\left[W\right]_{M|G|r|\infty} \approx \frac{rP_W E\left[R\right]}{r-\lambda E\left[\Upsilon\right]} \tag{8.3.2.10}$$

For exponential service times the derivation of an exact result is possible. This is because $\delta = r^{-1}E\left[\Upsilon\right] = \left(r\mu\right)^{-1}$ since there is only one configuration for the positions of the *r* customers in service (as there is only one phase for each channel). In addition, P_W can be determined analytically[5] to be $\frac{\left(r\rho\right)^r}{r!}\left(\left(1-\rho\right)\sum_{n=0}^{r-1}\frac{\left(r\rho\right)^n}{n!}+\frac{\left(r\rho\right)^r}{r!}\right)^{-1}$. Therefore

$$E\left[W\right]_{M|M|r|\infty} = \frac{P_W}{r\mu-\lambda} \tag{8.3.2.11}$$

For $C_{k-1}+M$ distributed service time $E\left[R\right]$ can be calculated (through similar equations to (8.2.1.3.6)) by

[5] Proof omitted: standard result

$$E[R]=\frac{\sum_{\forall s\in S_{n_S=r+1}} P_s E[R\mid \zeta=s]}{\sum_{\forall s\in S_{n_S=r+1}} P_s} \tag{8.3.2.12}$$

where ζ is the state of the system upon arrival, and P_W can be calculated by

$$P_W=\sum_{n=r}^{N}\bar{P}_n=1-\sum_{n=0}^{r-1}\bar{P}_n \tag{8.3.2.13}$$

However, since these results require a determination of the steady-state probabilities, they cannot be used as part of a direct approach. After deliberation it is thought that a method to obtain a truly analytic solution to (8.3.2.10) is of substantial difficulty. This being said, there are many studies that approximate this performance measure.

The following approximation has been suggested in Maaloe, 1973 and Nozaki & Ross, 1978:

$$E[W]_{M|G|r|\infty}\approx\frac{E[W]_{M|G|1|\infty}}{E[W]_{M|M|1|\infty}}E[W]_{M|M|r|\infty} \tag{8.3.2.14}$$

where $E[W]_{M|M|1|\infty}=\lambda\left(\mu(\mu-\lambda)\right)^{-1}$, $E[W]_{M|G|1|\infty}=$(8.3.2.5), and $E[W]_{M|M|r|\infty}=$(8.3.2.11). Cosmetatos, 1976 obtains the result

$$E[W]_{M|G|r|\infty}\approx \mathrm{var}[\Upsilon]E[W]_{M|M|r|\infty}+\left(1-\mathrm{var}[\Upsilon]\right)E[W]_{M|D|r|\infty} \tag{8.3.2.15}$$

stating that *'relative percentage errors incurred seem to be below 2% for most practical purposes'*. It is reported in Boxma et al, 1979 that this is a *'much sharper approximation, which is also particularly good in the heavy traffic case'*, thus making it appropriate for Rookwood hospital. The authors to this paper produce their own approximation which is *'slightly better'* than that of Cosmetatos, 1976. They also provide an approximate result for $E[W]_{M|D|r|\infty}$.

A more efficient and effective method is to make use of the steady-state probabilities determined in Ch 8.3.1. For the $M\mid M\mid 1$ system it is a well known result that $E[L]=\sum_{\forall n} nP_n$ where *n* is the number in system and P_n is the probability that there are *n* in the system. On extending this to the $M\mid C_{k-1}+M\mid r\mid N$ system;

$$E[L]=\sum_{n=0}^{N} n\bar{P}_n \tag{8.3.2.16}$$

If the steady-state probabilities have been determined by the Maple program then this result can be deduced with ease. Let the solution vector of steady-state probabilities equal Φ. The dimension of Φ is equal to the cardinality of the state space, $\eta_{k,total}$, which can be deduced through equation (7.3.2.2.3). By letting $\Phi(j)$ be the *j*-th element of the vector Φ, and $\eta_{k,i}$ be the number of states for which $n_s = i$ (see Ch 7.3.2.2.1),

$$\begin{aligned} \bar{P}_n &= \sum_{i=1}^{\eta_{k,n}} \Phi\left(i+\sum_{p=0}^{n-1}\eta_{k,p} \right) && n=0,1,...,r \\ \bar{P}_n &= \sum_{i=1}^{\eta_{k,n}} \Phi\left(i+\sum_{p=0}^{r}\eta_{k,p}+\left(n-(r+1)\right)\eta_{k,r} \right) && n=r+1,r+2,...,N \end{aligned} \tag{8.3.2.17}$$

The Maple program automatically calculates (8.3.2.16) in addition to $\tilde{\lambda}, P_R, P_W, \pi, E[W], E[T], E[L_q], C.V.[L_q]$, the coefficient of variation of the number in queue, and $\bar{P}_n$. Also calculated is the annual throughput,

$$Tp = \pi \cdot 365 / E[\Upsilon] \tag{8.3.2.18}$$

with $E[\Upsilon]$ deduced through (8.3.1.5.9). The output is given on the Maple script directly below the solution vector of steady-state probabilities, Φ. A graph of $\bar{P}_n$ is automatically plotted against *n*.

8.4 Heterogeneous Servers

The queuing systems that have been studied thus far have all involved homogenous servers. Now, the case of non-uniform service rates is considered.

At the NRC at Rookwood hospital there are many factors[6] that affect patient LOS (service time). If modelled by a homogenous server system then a significant standard deviation of service time would result. This has the effect of reducing the accuracy of the queuing theoretic model and LOS predictions since the individuality of patients are neglected. The

[6] Treatment control factors (Ch 6.2.6); treatment intensity (Ch 6.2.4); discharge destination (Ch 6.2.5)

influences of the aforementioned factors can be incorporated within the model by allowing the service rates to differ between the service channels.

8.4.1 Integrated system

An integrated system is defined as a queuing system that contains a single queue. It is analogous to all systems considered thus far. Setting up such a system with heterogeneous servers is far from a trivial exercise; a point illustrated in the following example.

8.4.1.1 Erlang service, two-server, no waiting space

This queuing system is identical to that of Ch 7.2.1.1 with the exception of uniform phase service rates. Specifically, the service rate of a phase in service channel one is μ_1 whilst the service rate of a phase in service channel two is μ_2. A state of this system is described by the double $\langle z_1, z_2 \rangle$ where z_j is the phase occupied by the customer in service channel *j* (if the service channel is empty then the value is zero). The CTMC for this system is

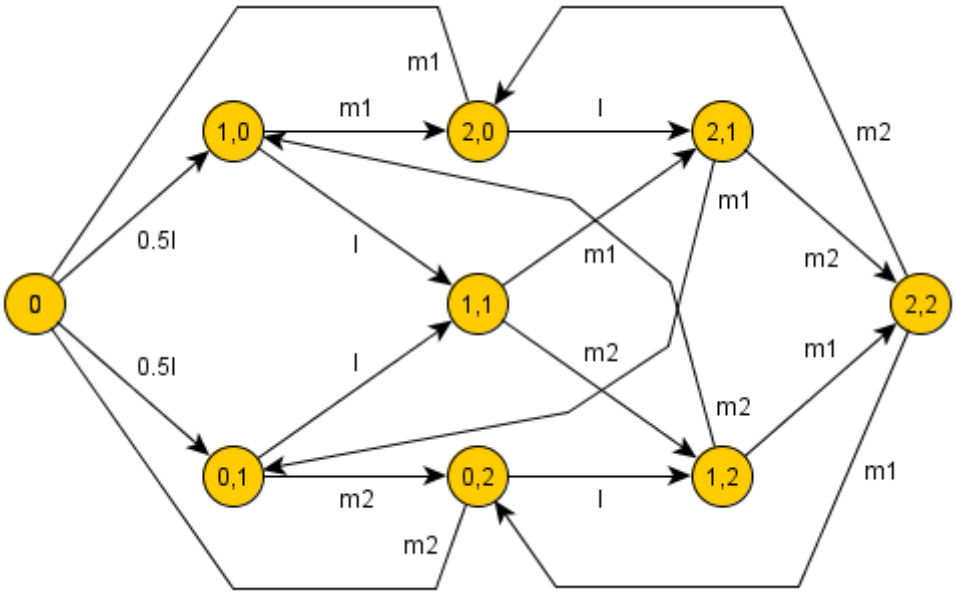


Figure 8.3 The $M / E_2 / 2 / 2$ queuing system with heterogeneous servers

Recall that the dimension of the CTMC for homogenous servers is equal to the number of phases of the service time distribution, *k*. In the case of heterogeneous servers, however, the minimal dimension is equivalent to the number of servers, *r*, for the case of no waiting space or *r+1* otherwise. This is due to the alternate definition of system state.

In addition, since the state of the system contains information on individual servers it becomes necessary to assign probabilities that an arrival will enter service at each of the service channels. In the above example, there is no preference and so each probability is 0.5.

It will also be noticed that there are more states in the system with heterogeneous servers. Note that the number of states depends (only) on the order of the phase-type distribution (in this case, two), the number of servers, and the number of waiting spaces.

8.4.1.2 The number of states in a *k*-th order phase-type service system

The following matrices contain a description of the state space for the systems $M \mid PH_2 \mid 1 \mid 1$ and $M \mid PH_2 \mid 2 \mid 2$:

$$S_{2,1,1} = \left(0 \;\vdots\; 1 \quad 2\right) \qquad S_{2,2,2} = \left(\begin{array}{c:cc} 0,0 & 0,1 & 0,2 \\ \hdashline 1,0 & 1,1 & 1,2 \\ 2,0 & 2,1 & 2,2 \end{array}\right)$$

The dotted lines are used to separate areas of the matrices that contain differing numbers of customers in the system. For example, in $S_{2,2,2}$ the upper left and lower right sub-matrices refer to the states of the system in which there are zero and two customers present whilst the upper right and lower left sub-matrices both refer to states in which only one service channel is occupied. The number in system is denoted by *i*. The total number of states for these systems are $\left(k+1\right)=3$ and $\left(k+1\right)^2=9$ respectively.

For the $M \mid PH_2 \mid 3 \mid 3$ system

$$S_{2,3,3} = \left(\begin{array}{c:cc:cc:cc} 0,0,0 & 0,0,1 & 0,0,2 & 1,0,0 & 1,0,1 & 1,0,2 & 2,0,0 & 2,0,1 & 2,0,2 \\ \hdashline 0,1,0 & 0,1,1 & 0,1,2 & 1,1,0 & 1,1,1 & 1,1,2 & 2,1,0 & 2,1,1 & 2,1,2 \\ 0,2,0 & 0,2,1 & 0,2,2 & 1,2,0 & 1,2,1 & 1,2,2 & 2,2,0 & 2,2,1 & 2,2,2 \end{array}\right)$$

The total number of states is $\left(k+1\right)^3=27$. By defining the symbol $\wedge$ as the concatenation of elements, $S_{2,3,3}$ can be expressed as

$$S_{2,3,3} = \left(0 \wedge S_{2,2,2} \;\vdots\; 1 \wedge S_{2,2,2} \quad 2 \wedge S_{2,2,2}\right) = S_{2,1,1} \wedge S_{2,2,2}$$

Moreover, $S_{k,3,3} = S_{k,1,1} \wedge S_{k,2,2}$ with

$$S_{k,1,1} = \left(0 \;\vdots\; 1 \quad 2 \quad \cdots \quad k\right) \qquad S_{k,2,2} = \left(\begin{array}{c:ccc:c} 0,0 & 0,1 & 0,2 & \cdots & 0,k \\ \hdashline 1,0 & 1,1 & 1,2 & & \\ 2,0 & 2,1 & 2,2 & & \\ \vdots & & & \ddots & \\ k,0 & & & & k,k \end{array}\right)$$

and through a similar argument $S_{k,4,4} = S_{k,2,2} \wedge S_{k,2,2}$.

By further generalising these results the following is obtained

$$S_{k,r,r} = \begin{cases} S_{k,1,1}^{\frac{r-1}{2}} \wedge S_{k,2,2} & r \text{ odd} \\ S_{k,2,2}^{\frac{r}{2}} \wedge S_{k,2,2} & r \text{ even} \end{cases}$$

where, for example, $S_{k,1,1}^{\frac{r-1}{2}} \wedge S_{k,2,2}$ with $r = 5$ is $S_{k,1,1} \wedge S_{k,2,2} \wedge S_{k,2,2}$.

The following table displays the number of states, η_i, for which the number in system is equal to a particular value, *i*, for the matrices $S_{k,1,1}$ and $S_{k,2,2}$:

η_i	$S_{k,1,1}$	$S_{k,2,2}$
$i = 0$	1	1
$i = 1$	k	$2k$
$i = 2$		k^2

Consider the matrix $S_{k,3,3} = S_{k,1,1} \wedge S_{k,2,2} = \left(0 \wedge S_{k,2,2} \;\vdots\; 1 \wedge S_{k,2,2} \quad 2 \wedge S_{k,2,2} \quad \cdots \quad k \wedge S_{k,2,2}\right)$;

	$0 \wedge S_{k,2,2}$	$1 \wedge S_{k,2,2}$	$\cdots$	$k \wedge S_{k,2,2}$
$i = 0$	1	0		0
$i = 1$	$2k$	1	$\cdots$	1
$i = 2$	k^2	$2k$		$2k$
$i = 3$	0	k^2	$\cdots$	k^2

The following is obtained by summing the values contained in the *k+1* columns for each row:

η_i	$S_{k,3,3}$
$i=0$	1
$i=1$	$3k$
$i=2$	$3k^2$
$i=3$	k^3

Next consider the matrix

$$S_{k,4,4} = \dot{S}_{k,2,2} \wedge S_{k,2,2} = \left(\begin{array}{c|ccccc} 0,0 \wedge S_{k,2,2} & 1,0 \wedge S_{k,2,2} & 2,0 \wedge S_{k,2,2} & \cdots & k,0 \wedge S_{k,2,2} \\ \hline 0,1 \wedge S_{k,2,2} & 1,1 \wedge S_{k,2,2} & 2,1 \wedge S_{k,2,2} & & \\ 0,2 \wedge S_{k,2,2} & 1,2 \wedge S_{k,2,2} & 2,2 \wedge S_{k,2,2} & & \\ \vdots & & & \ddots & \\ 0,k \wedge S_{k,2,2} & & & & k,k \wedge S_{k,2,2} \end{array}\right)$$

where $\dot{S}_{k,2,2} = S_{k,2,2}$. Note that the values to the left of $\wedge$ are the values of z_1 and z_2 whilst the values in $S_{k,2,2}$ are the values of z_3 and z_4. There is only one state, $\langle 0,0,0,0 \rangle$, that satisfies $i=0$. Now consider the states that satisfy $i=1$. There are $2k$ states in $\dot{S}_{k,2,2}$ for which $\tilde{z}_1 + \tilde{z}_2 = 1$ where $\tilde{z}_j = \begin{cases} 0 & z_j = 0 \\ 1 & z_j > 0 \end{cases}$ and one state for which $\tilde{z}_3 + \tilde{z}_4 = 0$. There is also one state in $\dot{S}_{k,2,2}$ for which $\tilde{z}_1 + \tilde{z}_2 = 0$ and $2k$ states in $S_{k,2,2}$ for which $\tilde{z}_3 + \tilde{z}_4 = 1$.

Therefore, $\eta_1 = (2k \cdot 1) + (1 \cdot 2k) = 4k$. Through a similar argument,

$\eta_2 = (k^2 \cdot 1) + (2k \cdot 2k) + (1 \cdot k^2) = 6k^2$, $\eta_3 = (k^2 \cdot 2k) + (2k \cdot k^2) = 4k^2$, and $\eta_4 = (k^2 \cdot k^2) = k^4$.

The following table displays further results:

η_i	$S_{k,1,1}$	$S_{k,2,2}$	$S_{k,3,3}$	$S_{k,4,4}$	$S_{k,5,5}$	$S_{k,6,6}$
$i=0$	1	1	1	1	1	1
$i=1$	k	$2k$	$3k$	$4k$	$5k$	$6k$
$i=2$		k^2	$3k^2$	$6k^2$	$10k^2$	$15k^2$
$i=3$			k^3	$4k^3$	$10k^3$	$20k^3$
$i=4$				k^4	$5k^4$	$15k^4$
$i=5$					k^5	$6k^5$
$i=6$						k^6

It is now clear that the total number of states for the $M \mid PH_k \mid r \mid r$ system is

$$\eta_{total} = \sum_{i=0}^{r} \eta_i = \sum_{i=0}^{r} \binom{r}{i} k^i \qquad (8.4.1.2.1)$$

This is equivalent to the binomial series expansion of $(k+1)^r$ – a result that has previously been acknowledged for $k = 2, r = 1,2,3$. For the $M \mid PH_k \mid r \mid N$ system

$$\eta_{total} = \sum_{i=0}^{r} \binom{r}{i} k^i + (N-r)k^r \qquad (8.4.1.2.2)$$

since $\eta_{r+1} = \eta_{r+2} = \ldots = \eta_N = \eta_r$ (similar argument to that found in Ch 7.3.2.2.1).

8.4.1.3 Comparison between homogenous and heterogeneous server systems

The dimension of the transpose transition (square) matrix is equivalent to the total number of states in the queuing system. A matrix of greater dimension is harder to solve *ceteris paribus*[7]. The complexity, measured only by the total number of states, is forthwith compared[8] for a number of queuing systems involving homogeneous and heterogeneous servers.

Table 8.3 Number of states for equivalent homogeneous and heterogeneous queuing systems

η_{total}	Homogenous	Heterogeneous
$M \mid PH_k \mid 1 \mid N$	$1+Nk$	$1+Nk$
$M \mid PH_2 \mid 3 \mid 5$	18	43
$M \mid PH_2 \mid 10 \mid 15$	121	64,169
$M \mid PH_2 \mid 20 \mid 30$	441	3,497,270,161
$M \mid PH_2 \mid 30 \mid 50$	1,116	205,912,606,931,129
$M \mid PH_3 \mid 10 \mid 15$	616	1,343,821
$M \mid PH_3 \mid 20 \mid 30$	1,606	2,229,556
$M \mid PH_3 \mid 30 \mid 50$	15,376	1,157,039,327,248,739,956
$M \mid PH_4 \mid 10 \mid 15$	2,431	15,008,505
$M \mid PH_5 \mid 10 \mid 15$	8,008	109,294,301
$M \mid PH_{10} \mid 10 \mid 15$	646,646	75,937,424,601

It can be seen that the number of states for any heterogeneous multi-server system exceeds the number required to represent the corresponding homogeneous server system. In low-order

[7] i.e. if of similar, recurrent structure with similar density
[8] Using (7.3.2.2.3) and (8.4.1.2.2)

systems this difference is relatively small (e.g. 25 states for $M|PH_2|3|5$) but extends rapidly as parameter values are increased (e.g. 109,286,293 states for $M|PH_5|10|15$).

What is of particular concern, however, is the substantial increase in η_{total} when the number of service channels is increased. For example, when the values of *r* and *N* are increased from three and five to 20 and 30 for the system with $k=2$ the value of η_{total} increases by 2,350% for homogeneous servers and $8\times10^{9}\%$ for heterogeneous servers. Such a level of additional complexity is clearly unacceptable when modelling a facility which typically maintains around 21 service channels (beds).

8.4.1.4 Multiple classes of server

It is perhaps more appropriate to consider the integrated system with a number of classes of homogeneous servers. The number of servers and associated rates are permitted to vary between the classes. Customers are assigned to classes based on the value of certain personal attributes. This is well suited to healthcare facilities such as Rookwood hospital where patients are often grouped by characteristics such as age and gender.

However, such a system is practically infeasible. It can be seen that the cardinality of the state space of the system with multiple classes of servers is bounded below by the cardinality of the state space of the equivalent homogeneous server system. Equality is only achieved when the number of classes is one. Even if a simple model (20 beds, 10 waiting spaces, three phase servers) is used to represent Rookwood hospital then the cardinality of the state space would easily exceed 1,606 (assuming there are at least two classes of server). This is unacceptable considering the processing limitations of a desktop computer.

8.4.2 Disconnected system

A disconnected system is defined as a queuing system that contains multiple integrated systems. Since the constituent systems are disconnected there exists independence between them. The order of the system is said to represent the number of constituent systems.

8.4.2.1 Multiple classes of server

Figure 8.4 depicts the disconnected system of order *p* in which each constituent system represents a different class of server. Therefore, there are *p* classes of server. Each

class/constituent system, $i = 1, 2, .., p$, has r_i homogenous servers. The system can be expressed $M, M, ..., M \mid PH_{k_1}, PH_{k_2}, ..., PH_{k_p} \mid r_1, r_2, ..., r_p \mid N_1, N_2, ..., N_p$.

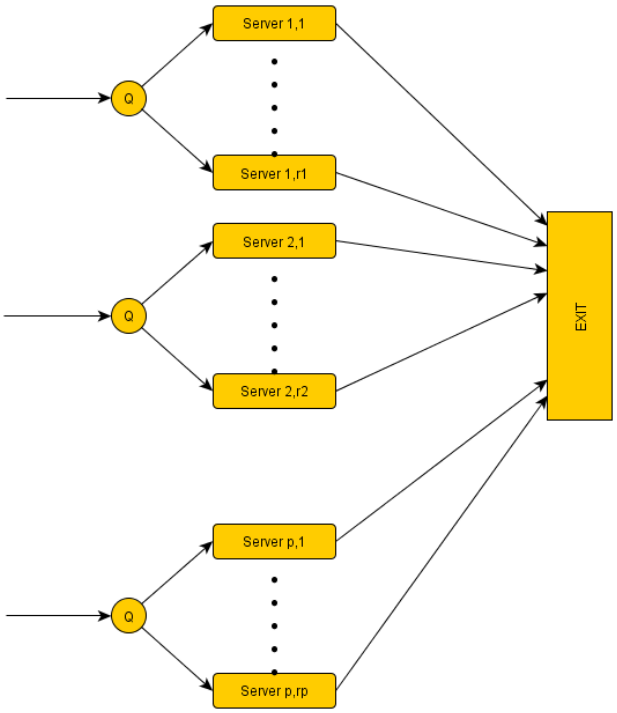


Figure 8.4 The disconnected system with multiple classes of server

The most significant difference between the integrated and disconnected systems is the routing of customers. In the integrated system customers arrive at a single queue. If, on arrival, there is more than one service channel available then the customer is routed probabilistically (see Ch 8.4.1.1). These probabilities can be manipulated in order to affect the throughput in each class of server. If only one service channel is available then the customer immediately enters service. Otherwise, they must wait. Therefore, there is no possibility of an unoccupied service channel when there are customers in the queue. This is not so in the disconnected system where customers arrive for service at a specific class of server. Instead of routing probabilities, the arrival rates into each class can be adjusted in order to affect class throughput and other performance measures.

The major advantage of modelling a multiple class system by a disconnected system is the computational ease with which steady-state probabilities can be obtained. For the integrated system it is necessary to solve a square matrix of least dimension given by (7.3.2.2.3). For the disconnected system it is necessary to solve *p* matrices each of dimension (7.3.2.2.3) with

$r = r_i$, N the maximum capacity of the *i*-th constituent system and *k* the number of phases in the *i*-th class of server. The steady-state probabilities for the system can be deduced by the multiplication of the steady-state probabilities for each constituent system (due to independence). Therefore

$$\begin{aligned} &P\left(\left\langle n_{s_1}, n_{2_1}, ..., n_{r_1}\right\rangle \cup \left\langle n_{s_2}, n_{2_2}, ..., n_{r_2}\right\rangle \cup ... \cup \left\langle n_{s_p}, n_{2_p}, ..., n_{r_p}\right\rangle\right) \\ &= P\left(\left\langle n_{s_1}, n_{2_1}, ..., n_{r_1}\right\rangle\right) \cdot P\left(\left\langle n_{s_2}, n_{2_2}, ..., n_{r_2}\right\rangle\right) \cdot ... \cdot P\left(\left\langle n_{s_p}, n_{2_p}, ..., n_{r_p}\right\rangle\right) \end{aligned} \qquad (8.4.2.2.1)$$

In addition, the performance measures can easily be determined for each class (see Ch 8.3.2) or for the system as a whole. If P is used to denote either *W* or *T* then

$$\begin{aligned} E\left[\mathrm{P}_{sys}\right] &= \frac{1}{\tilde{\lambda}_1 + \tilde{\lambda}_2 + ... + \tilde{\lambda}_p}\left(\tilde{\lambda}_1 E\left[\mathrm{P}_1\right] + \tilde{\lambda}_2 E\left[\mathrm{P}_2\right] + ... + \tilde{\lambda}_p E\left[\mathrm{P}_p\right]\right) \\ &= \left(\sum_{i=1}^{p} \tilde{\lambda}_i\right)^{-1}\left(\sum_{i=1}^{p} \tilde{\lambda}_i E\left[\mathrm{P}_i\right]\right) \end{aligned} \qquad (8.4.2.2.2)$$

However, if P is used to denote either *L* or L_q then

$$E\left[\mathrm{P}_{sys}\right] = E\left[\mathrm{P}_1\right] + E\left[\mathrm{P}_2\right] + ... E\left[\mathrm{P}_p\right] \qquad (8.4.2.2.3)$$

8.5 Conclusion

This chapter has built upon the work of the previous chapter in developing a more representative queuing model of the NRC at Rookwood hospital. The most significant advancement is the replacement of Erlang service times by Coxian (active) and exponential (blocked) service times. Steady-state results and performance measures are derived analytically for the low-order system. This is thereafter compared with the three phase Erlang system to evaluate any differences in the transition matrices. The Maple program of the previous chapter is then amended to incorporate these differences for the general case. Following this, direct approaches to the solution of the performance measures are considered. Indirect approaches are also considered and are embedded within the new Maple program. It is then explained why it is necessary to consider a level of heterogeneity with regard to the service channels.

An integrated system with heterogeneous servers is rejected due to the substantial size of the associated transition matrix – it would be unreasonable to expect a desktop computer to solve

this. Another limitation of this system is the inability to obtain performance measures for patients of a particular group (since all patients are considered equal before they enter service). This is a drawback when considering the effects of *what-if* type scenarios on a specific patient demographic.

The disconnected system is therefore studied. Firstly, on the topic of computational efficiency, this system is far more favourable than its alternative. For example, to model an integrated $M \mid PH_3, PH_3, PH_3 \mid 8,8,8 \mid 48$ system requires a transition matrix of dimension at least 10,725. However, for an equivalent third-order disconnected system, $M, M, M \mid PH_3, PH_3, PH_3 \mid 8,8,8 \mid 16,16,16$, the matrix dimension for each constituent system is only 525. It is possible to achieve numerical results for such a system in a reasonable amount of time (Ch 8.3.1.6).

Secondly, the arrival rate of patients into each class of server can be individually manipulated. Since each class of server represents a particular group of patients this allows a more flexible and detailed approach to what-if type scenarios. For example, what if there was an increase in the number of referrals for elderly patients (due to an ageing population)?

Thirdly, expected values for the performance measures can be readily calculated for each class of server. This enables a more detailed analysis for each group of patients associated with each class. This is particularly advantageous when considering the repercussions of what-if type scenarios. For example, what is the effect of increasing the number of beds for young male patients with TBI?

Finally, some limitations must be acknowledged. Admission policy at Rookwood hospital NRC is made largely on an impromptu basis due to the many factors that must be taken into account when answering questions such as who to admit or how many beds to make available. There are no clear, coherent rules to answer these questions and it is therefore difficult to fully evaluate the extent to which reality is represented by the model. For example, the model allows for the possibility that a server can be unoccupied whilst patients are waiting. Whilst this can realistically occur due to a deficiency in employee hours or experience it is undesirable since the model does not take into account such variables. The prevalence of this can be minimised by ensuring the stationary probabilities of states in which a server is empty are negligible. Also, the model assumes that the partition of servers (i.e. the number of service channels in each class) is fixed over time. This is unlikely to be a realistic

assumption in the short-term since the number of servers in each class would depend on the demographic of the queue.

In conclusion, the disconnected system with Coxian and exponential service time has been solved for steady-state probabilities and performance measures. In the next chapter this project concludes with the presentation of a model based on this system that is used to model activities at the NRC at Rookwood hospital.

Chapter 9: The Rookwood Model

9.1 Introduction

This chapter is concerned with the production of a queuing model that is representative of the major activities at the Neurological Rehabilitation Centre (NRC) at Rookwood hospital – a task that involves combining the various work contained in this thesis thus far. The first phase is to produce an initial, representative model of the unit which is described in the following four chronological stages.

First, historical data is used to determine the characteristics of a small number of homogeneous patient groups. The objective is to deduce the range of values for appropriate branching variables (e.g. age, gender, diagnosis – see Ch 6.2.6) that return a low variance of *active* length of stay (LOS) for each patient group.

In the second stage the scheduling program (Chapters 3 and 4) is used to produce treatment timetables by fitting patient demand to staff supply. The aim is to use the program to produce average values of treatment intensity for each patient group that are approximately equal to their empirical counterparts. This is achieved by initialising the program with variable values that are equivalent to the situation on the ground at Rookwood NRC. That is, the number of beds for each group, number and band level of staff, and number of hours worked each day should be realistic.

Thirdly, these average values of treatment intensity are converted to average values of active LOS through a two-dimensional line graph for each patient group. These graphs are produced using historical data to examine the relationship between treatment intensity and active LOS

(Ch 6.4.2.3). Assuming that these graphs and the average values of treatment intensity obtained through stage two are correct then the average values of active LOS should match their empirical counterparts (for each group).

In the fourth and final stage, the queuing model is constructed. This consists of a number of disconnected homogeneous server queuing systems; one for each patient group (Ch 8.4.2.1). The distribution of referral arrivals for each queuing system is (justifiably – Ch 6.4.1) assumed Poisson and rates are derived from historical data (Ch 6.4.1). A Coxian phase-type and exponential distribution are used to model active and *blocked* LOS respectively for each system. These are fitted using the results of Chapter 5, the Matlab program introduced in Ch 6.5.1 and the average values of active LOS of stage three. Each queuing system is equivalent to that specified in Ch 8.3 and can so be solved for steady-state probabilities and performance measures using the Maple program mentioned in Ch 8.3.1.6. Ultimately, steady-state probabilities, performance measures and costs are output for the holistic queuing model (Ch 8.4.2.1).

This initial model is, in fact, equivalent to Route 4 (1-2-4-6) of Figure 6.8. That is, there is a queuing model (1) that has LOS partitioned to its active and blocked component (2) with a service rate dependent on treatment intensity (4) and subjects grouped by various predictor variables (6). The second phase is to extrapolate the model through a number of hypothetical *what-if* type scenarios that relate to the *major policy decisions* (who at admit, what care to give, when to discharge). These involve a modification to either the demand-side or the supply-side of treatment provision at the unit. For example, the effect of an ageing population can be modelled by increasing the arrival rate for the patient group that contains older patients (assuming age is a branching variable in stage one). An example of a supply-side modification is the employment of additional therapists.

Contrary to the above, this chapter contains not just one model but three. The first subchapter introduces a very simple queuing model equivalent to Route 1. This is further developed in the following subchapter (model equivalent to Route 3). The chapter concludes with the final model under consideration which is equivalent to that mentioned above (Route 4).

9.2 Model One

The most simplistic of models (analogous to that of Ch 6.2.1) is considered here. Treatment intensity is not a consideration at present and so stages two and three are omitted. Stage one

is described as follows. The arrival rate (as determined in Ch 6.4.1) is 0.36364, i.e. roughly five arrivals every fortnight. The mean LOS for the 358 patient episodes that contain additional NHS Trust variables (Ch 6.4.2.1) is 149.3 days. The most appropriate service distribution was found to be the three term Coxian distribution (Ch 6.5.4). However, since this is intended to be an elementary model the exponential distribution (rate 0.006698) is used. The number of service channels is 21 since this is the mean number of beds occupied in recent years (Ch 6.4.2.2). No efforts are made to partition the 358 episodes into patient groups. This model is therefore analogous to Route 1 of Figure 6.8.

9.2.1 Initial version

Stage four therefore requires the solution of the $M|M|21$ queuing system. This system is easily solved for an unbounded queue size (a result for P_W, the probability that a customer must wait, is provided in Ch 8.3.2). Whilst analytic solutions can also be found by hand for a bounded queue size, it is simpler to use the computer program of Chapter 7. Results are derived for the $M|E_1|21|50$ system, i.e. 21 beds and space in the queue for 29 referrals[1].

Instead of listing all 51 probabilities the graphical output of the Maple program is given.

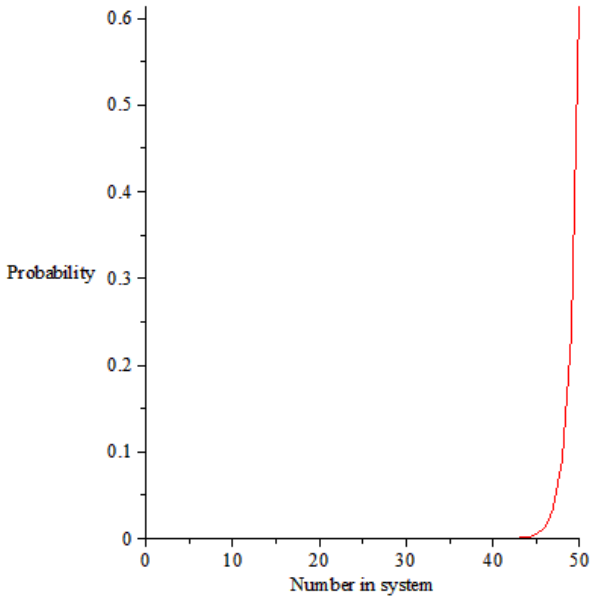


Figure 9.1 Probability density graph for number in system

[1] The number of referrals in the queue from January 2010 to April 2011 does not exceed 25 (see Ch 6.4.1)

Some summary statistics including the probability of rejection and the *effective arrival rate*[2] are deduced[3];

$$E[L] = 49.369 \qquad P_W = 1$$
$$E[L_q] = 28.369 \qquad P_R = 0.6132$$
$$E[T] = 350.987 \qquad \lambda_{eff} = 0.1407$$
$$E[W] = 201.689 \qquad \pi = 21$$
$$Tp = 51.34$$

Clearly, this cannot be representative of reality. The steady-state probabilities are skewed so far to the right since the rate of arrival (0.36364) is much greater than the (maximal) rate of departure (21×0.006698). The effective arrival rate attains this departure rate since $\tilde{\lambda} = \pi\mu$ under steady-state and in this case, $\pi = r$ (see Ch 7.3.2.2.6). The term *divergent* is used here to describe the limiting behaviour toward the *x*-axis as $n \to N$. To digress, the shape of the graphs corresponding to the systems where the arrival rate is equal (0.1407) and less than the rate of departure (say 0.12) are of interest here.

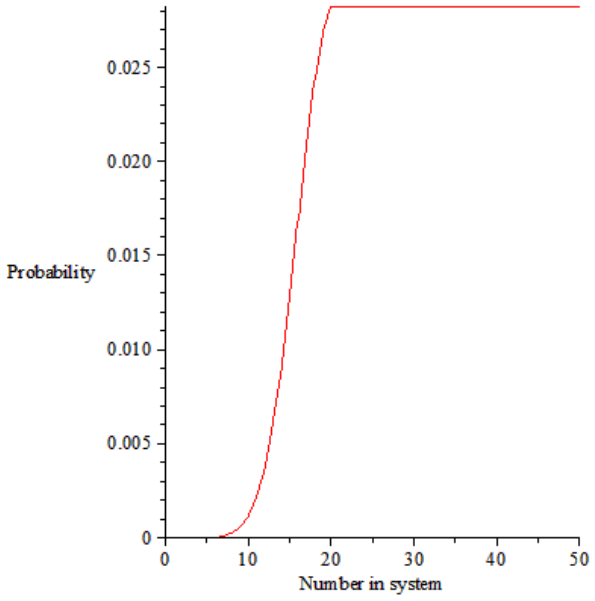


Figure 9.2 Arrival rate equal to rate of departure

[2] The arrival rate of referrals that enter the queue (previously referred to as the *non-trivial* arrival rate)
[3] Automatically by the Maple program

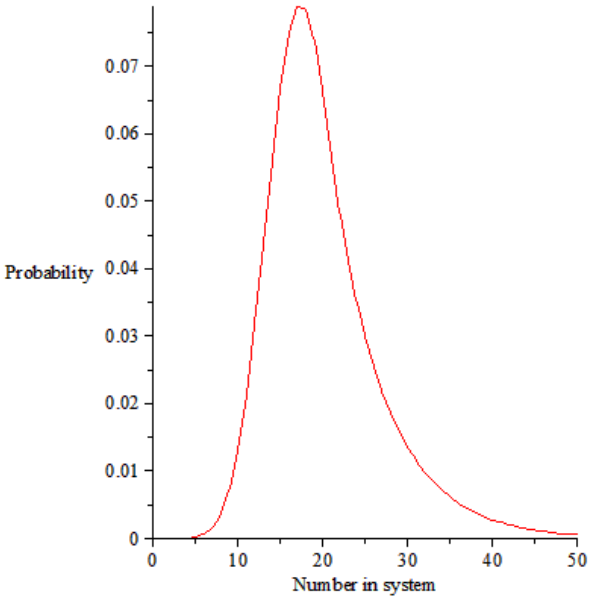


Figure 9.3 Arrival rate less than rate of departure

The behaviour displayed in Figure 9.2 is described as *neutral* since it is neither convergent nor divergent to the *x*-axis. Obviously, the most realistic interpretation of Rookwood NRC is provided by the *convergent* behaviour depicted in Figure 9.3 corresponding to traffic intensity less than one. Of course, it would be nonsensical to arbitrarily set the arrival rate at a value less than the departure rate. Instead, there are two types of realistic customer behaviour that prevent divergent behaviour. These are *balking* and *reneging*.

9.2.2 Balking

Balking is a concept in which arriving customers do not join the queue. This could be for a number of reasons but is usually because of the size of the queue (e.g. too many customers at a supermarket checkout) or the perceived waiting time (e.g. each customer in the queue has a full trolley). Obviously balking for such reasons requires the arriving customer to have knowledge of the state of the queue, i.e. a *visible queue*. This does not exist in Rookwood NRC because of the mental disposition of prospective patients. Instead, the decision to balk is made by clinicians who have an insight into the state of the queue. Note that balking can be prevalent for both types of arrival, i.e. the submission of referrals for patients ready for transfer and the instance at which non-ready patients on the waiting list become ready (see Ch 1.2).

Balking is technically part of the finite buffer system[4] of Ch 9.2.1 since when capacity is reached arrivals are lost. The effective arrival rate when the system is at capacity is therefore zero, otherwise, it is λ. This is intuitively unrealistic since, for this system, balking should be progressive. That is, the arrival rate of referrals should decreases as the queue length becomes larger, rather than just cutting-off at a certain value. This is because less of an effort is made to proactively 'scout' patients (see Ch 1.2) and prospective referrers are put off by the expected waiting time. This concept can be modelled by introducing an arrival rate that is dependent on the state of the system.

The Maple programs developed to produce steady-state results for the $M|E_k|r|N$ and $M|C_{k-1}+M|r|N$ systems have been amended to include, as an option, a state-dependent arrival rate. There are many options that exist regarding the actual dependence of the arrival rate on the state of the system. For this model the following is used to define λ_n, the arrival rate when *n* customers are present in the system:

$$\lambda_n = \lambda\left(1-Pb_n\right) \tag{9.2.2.1}$$

where Pb_n is the probability of an arrival balking given *n* in system, defined as

$$Pb_n = \begin{cases} 0 & 0 \le n < r \\ \left(\dfrac{n-r+1}{N-r}\right)^{\delta} & r \le n \le N-1 \\ 1 & n \ge N \end{cases} \tag{9.2.2.2}$$

That is, the arrival rate is as prescribed (i.e. λ) when there are fewer customers in the system than servers. Otherwise, the arrival rate decreases polynomially (quadratically for $\delta = 2$) as the number in queue (given by $n-r$) increases. The denominator, in (9.2.2.2), is chosen so that the arrival rate into the state(s) with $n = N$ is zero provided $N > r$. This means that the probability of the system being in these state(s) is zero and so the probability of rejection is also zero. The convergence of the arrival rate for $r \le n \le N-1$ to zero is dependent on the parameter δ.

The expected number of customers in the system is derived through (8.3.2.16) using (8.3.2.17). The results (8.2.1.5.2), (8.2.1.5.3), and (8.2.1.5.4) are used to produce results for

[4] A fixed number of waiting spaces in the queue

the other performance measures, albeit, with a different interpretation of $\tilde{\lambda}$. In the case of a fixed arrival rate (i.e. Chapters 7 and 8) this is defined as the mean non-trivial arrival rate given by (7.3.2.2.7). It is clear to see how this is formulated since $P_R\lambda$ is the rate at which customers are lost to the system. However, this result can be derived through a more comprehensive definition. The mean arrival rate is defined as

$$\tilde{\lambda} = \lambda\left(1-\tilde{P}b\right) \tag{9.2.2.3}$$

where $\tilde{P}b$ is defined as the mean balking probability, given as

$$\tilde{P}b = \sum_{\forall n} Pb_n \bar{P}_n \tag{9.2.2.4}$$

where, as before, $\bar{P}_n$ is defined as $\sum_{\forall s\in S_{n_s=n}} P_s$. The non-trivial arrival rate is defined in Chapters 7 and 8 as $\lambda_n = \begin{cases}\lambda & 0\le n<N\\ 0 & n\ge N\end{cases}$ from which it can be inferred, through (9.2.2.1), that $Pb_n = \begin{cases}0 & 0\le n<N\\ 1 & n\ge N\end{cases}$. Therefore, $\tilde{P}b = 1\cdot\bar{P}_N = P_R$ and so $\tilde{\lambda} = \lambda\left(1-P_R\right)$, *c.f.* (7.3.2.2.7), as required. For the case of the state dependent arrival rate defined above in (9.2.2.1)

$$\tilde{P}b = \sum_{n=0}^{r-1} 0\cdot\bar{P}_n + \sum_{n=r}^{N-1} Pb_n\cdot\bar{P}_n + \sum_{n=N}^{\infty} 1\cdot 0 = \sum_{n=r}^{N-1} Pb_n\cdot\bar{P}_n$$

since $Pb_n = 0$ for $0\le n\le r-1$ and $\bar{P}_n = 0$ for $n\ge N$ (as $Pb_{N-1}=1$). Thus

$$\tilde{\lambda} = \lambda\left(1-\sum_{n=r}^{N-1} Pb_n\bar{P}_n\right) \tag{9.2.2.5}$$

This can also be derived through the (perhaps) more intuitive relation

$$\tilde{\lambda} = \sum_{\forall n}\lambda_n\bar{P}_n \tag{9.2.2.6}$$

which gives

$$\tilde{\lambda} = \lambda\sum_{n=0}^{r-1}\bar{P}_n + \sum_{n=r}^{N-1}\lambda_n\bar{P}_n + \sum_{n=N}^{\infty} 0\cdot 0 = \lambda\sum_{n=0}^{r-1}\bar{P}_n + \sum_{n=r}^{N-1}\lambda_n\bar{P}_n$$

since λ_n and P_n are equal to zero for $n \geq N$. By (9.2.2.1) this equals

$$\tilde{\lambda} = \lambda \sum_{n=0}^{r-1} \bar{P}_n + \lambda \sum_{n=r}^{N-1} (1 - Pb_n) \bar{P}_n = \lambda \sum_{n=0}^{r-1} \bar{P}_n + \lambda \sum_{n=r}^{N-1} \bar{P}_n - \lambda \sum_{n=r}^{N-1} Pb_n \bar{P}_n$$
$$= \lambda \sum_{n=0}^{N-1} \bar{P}_n - \lambda \sum_{n=r}^{N-1} Pb_n \bar{P}_n = \lambda \cdot 1 - \lambda \sum_{n=r}^{N-1} Pb_n \bar{P}_n = \lambda \left(1 - \sum_{n=r}^{N-1} Pb_n \bar{P}_n \right)$$

as required. Note that $Pb_{N-1} = 1$ and $\lambda_{N-1} = 0$.

Like the two original Maple programs (Chapters 7 and 8) the amended versions also include an automated derivation of steady-state probabilities, performance measures and a graph depicting the relationship between the number in system and the corresponding probability. In addition, graphs are also produced for the probability of balking versus the number in system and the arrival rate versus the number in system. These graphs provide much insight into the behaviour of the queuing system when studied together. The mean balking probability is also determined which can be compared with its empirical equivalent.

However, no data on the prevalence of balking at Rookwood NRC is available and furthermore it has been made clear through discussions with clinicians that balking is not a significant issue on the ground. However, it is not without mathematical benefit since it can reduce what would otherwise be divergent or neutral behaviour to the more favourable convergent behaviour. The clinicians point to reneging as the dominant type of behaviour that prevents queue length from becoming unmanageable.

9.2.3 Reneging

Reneging occurs when queuing customers leave before entering service. As with balking this could be for a number of reasons but it is typically because of the perceived waiting time. At Rookwood NRC it is not the patient that decides if and when to renege but the clinician. Reneging occurs if a patient is discharged from their current facility (either back home or to a non-NHS facility e.g. nursing home), transferred to another NHS facility, or dies.

Again, the original Maple programs have been amended to include, as an option, the concept of reneging. Like balking, this does not involve changing the model when there are fewer customers in the system than service channels. However, whereas balking affects the LEFT matrices for $r \leq n \leq N-1$ (i.e. transitions from *n* in system to *n+1* – see Ch 7.3.2.2.5), reneging affects the RIGHT matrices for $r+1 \leq n \leq N$ (i.e. transitions from *n* in system to

$n-1$). Consider the transpose transition matrix for the system $M\,|\,E_2\,|\,2\,|\,3$ with balking and reneging.

	$\langle 0\rangle$	$\langle 1,0\rangle$	$\langle 1,1\rangle$	$\langle 2,0\rangle$	$\langle 2,1\rangle$	$\langle 2,2\rangle$	$\langle 3,0\rangle$	$\langle 3,1\rangle$	$\langle 3,2\rangle$
$\langle 0\rangle$	$-\lambda_0$		μ						
$\langle 1,0\rangle$	λ_0	$-(\lambda_1+\mu)$			μ				
$\langle 1,1\rangle$		μ	$-(\lambda_1+\mu)$			2μ			
$\langle 2,0\rangle$		λ_1		$-(\lambda_2+2\mu)$			$\Re_3$	μ	
$\langle 2,1\rangle$			λ_1	2μ	$-(\lambda_2+2\mu)$			$\Re_3$	2μ
$\langle 2,2\rangle$					μ	$-(\lambda_2+2\mu)$			$\Re_3$
$\langle 3,0\rangle$				λ_2			$-(2\mu+\Re_3)$		
$\langle 3,1\rangle$					λ_2		2μ	$-(2\mu+\Re_3)$	
$\langle 3,2\rangle$						λ_2		μ	$-(2\mu+\Re_3)$

Balking is incorporated by defining the arrival rates (in blue) by (9.2.2.1). Reneging is incorporated by populating the diagonal entries (in red) in the appropriate (square) RIGHT matrices with the rates of reneging. Should the customer in the queue renege then this does not affect the state of the system with regard to those in service, i.e. upon reneging the value of n_s is reduced by one but the value of n_2, the number of customers in phase two of service, remains unchanged. Thus, the rate of reneging given by the $\Re_3$ correspond to transitions from $<3,n_2>$ to $<2,n_2>$ for $n_2=0,1,2$. It is stated in Pla et al, 2004 that the mean probability that an arriving customer will renege before entering service is calculated from the mean reneging rate divided by the mean arrival rate. For the system considered here, this is equal to

$$\tilde{Pr}=\frac{1}{\tilde{\lambda}}\sum_{\forall n}\Re_n\bar{P}_n=\frac{1}{\tilde{\lambda}}\sum_{n=0}^{r}0\cdot\bar{P}_n+\frac{1}{\tilde{\lambda}}\sum_{n=r+1}^{N}\Re_n\bar{P}_n+\frac{1}{\tilde{\lambda}}\sum_{n=N+1}^{\infty}0\cdot 0=\frac{1}{\tilde{\lambda}}\sum_{n=r+1}^{N}\Re_n\bar{P}_n \qquad (9.2.3.1)$$

The rates of reneging are determined as follows. When the system is in a particular state, there are a number of events that can occur. For example, if the system is in state $<2,1>$ a service completion in either of the phases could take the system to state $<1,0>$ or $<2,2>$ or an arrival could take the system to state $<3,1>$. The probability of these events occurring are $\mu/2\mu+\lambda$, $\mu/2\mu+\lambda$, and $\lambda/2\mu+\lambda$. Moreover, a transition rate matrix can easily be converted to a (in this case, left[5]) stochastic transition probability matrix as follows:

[5] Column values must sum to one

	$\langle 0\rangle$	$\langle 1,0\rangle$	$\langle 1,1\rangle$	$\langle 2,0\rangle$	$\langle 2,1\rangle$	$\langle 2,2\rangle$	$\langle 3,0\rangle$	$\langle 3,1\rangle$	$\langle 3,2\rangle$
$\langle 0\rangle$			$\mu/\mu+\lambda_1$						
$\langle 1,0\rangle$	1				$\mu/2\mu+\lambda_2$				
$\langle 1,1\rangle$		$\mu/\mu+\lambda_1$				$2\mu/2\mu+\lambda_2$			
$\langle 2,0\rangle$		$\lambda_1/\mu+\lambda_1$					$\mathfrak{R}_{3,1}/2\mu+\mathfrak{R}_{3,1}$	$\mu/2\mu+\mathfrak{R}_{3,2}$	
$\langle 2,1\rangle$			$\lambda_1/\mu+\lambda_1$	$2\mu/2\mu+\lambda_2$				$\mathfrak{R}_{3,2}/2\mu+\mathfrak{R}_{3,2}$	$2\mu/2\mu+\mathfrak{R}_{3,3}$
$\langle 2,2\rangle$					$\mu/2\mu+\lambda_2$				$\mathfrak{R}_{3,3}/2\mu+\mathfrak{R}_{3,3}$
$\langle 3,0\rangle$				$\lambda_2/2\mu+\lambda_2$					
$\langle 3,1\rangle$					$\lambda_2/2\mu+\lambda_2$		$2\mu/2\mu+\mathfrak{R}_{3,1}$		
$\langle 3,2\rangle$						$\lambda_2/2\mu+\lambda_2$		$\mu/2\mu+\mathfrak{R}_{3,2}$	

For example, the probability of the customer in the queue reneging when in state $<3,1>$ is $\frac{\mathfrak{R}_{3,2}}{2\mu+\mathfrak{R}_{3,2}}\Big/\left(\frac{\mu}{2\mu+\mathfrak{R}_{3,2}}+\frac{\mathfrak{R}_{3,2}}{2\mu+\mathfrak{R}_{3,2}}+\frac{\mu}{2\mu+\mathfrak{R}_{3,2}}\right)=\frac{\mathfrak{R}_{3,2}}{2\mu+\mathfrak{R}_{3,2}}$. It is clearly more intuitive to ask the user to stipulate such probabilities than the actual transition rates. For the columns that involve reneging

$$\mathfrak{R}\big/\big(\mathfrak{R}+(\text{sum of all other non-negative column values})\big)=p \qquad (9.2.3.2)$$

and so these probabilities are converted to rates by the formula

$$\mathfrak{R}=\big(p\times(\text{sum of all other non-negative column values})\big)\big/1-p \qquad (9.2.3.3)$$

In total, $N-r$ values of p, or Pr_n, the probability of reneging when there are $r+1\le n\le N$ customers in the system, are required. There are two options to define these probabilities. The first is perhaps the most intuitive. The Pr_n are calculated by summing the individual probabilities of reneging for each customer in the queue. The individual probability of reneging, ρ_{n-r}, is the probability of the $(n-r)$-th customer in the queue reneging. It is obvious that this should be a non-decreasing function and, as with balking, there are many ways to define this. The following function is chosen:

$$\rho_{n-r}=\gamma\left(\frac{n-r}{N-r}\right)^{\theta} \qquad n=r+1,\ldots,N \qquad (9.2.3.4)$$

This is similar to (9.2.2.2) with the addition of a second parameter, γ, that acts as a scaling factor. If there is only one customer in the queue then the probability of reneging is simply

$Pr_{r+1} = \rho_1$. If there are two customers in the queue then the probability of reneging is $Pr_{r+2} = \rho_1 + \rho_2$ and so on. Hence

$$Pr_n = \sum_{i=1}^{n-r} \rho_i \qquad n = r+1,...,N \qquad (9.2.3.5)$$

Note that the two parameters must be chosen such that $Pr_N < 1$. However, such a description of reneging probabilities is not shared by the data (see the graph corresponding to the probability of reneging based on number in queue – Ch 6.4.1).

A more flexible approach is offered as an alternative. This does not take into account the reneging probabilities of individuals and is defined

$$Pr_n = \gamma\left(\frac{n-r}{N-r}\right)^{\theta} \qquad n = r+1,...,N \qquad (9.2.3.6)$$

Note Pr_N is always equal to the scaling parameter, $0 < \gamma < 1$. When $0 < \theta < 1$ the curve of (9.2.3.6) is concave downward, when $\theta > 1$ it is concave upward and when $\theta = 1$ it is linear.

Similar to the original Maple programs steady-state results and performance measures are output. It is important to note, however, that these performance measures are pertinent to all customers and not just those that ultimately enter service. Thus the expected waiting time, $E[W]$, is actually defined as the mean time until a referral is removed (either because the patient is admitted or because they have reneged). A formula for the expected waiting time for those that are ultimately admitted has not been found. The interpretation of the expected total time in system, $E[T]$, has also changed to include reneged referrals. This means that it is possible that $E[T] < E[\Upsilon]$, the expected service time. These changes limit the usefulness of these performance measures and care must be taken in their interpretation. Unlike balking a nonzero probability of rejection is attainable since it is possible to reach the states corresponding to $n = N$. The program outputs two graphs; one, as before, plotting the number in system against probability, and the other plotting the reneging probability over the same range ($n = 0,1,...N$). The mean reneging probability (9.2.3.1) is also output.

Upon testing the program it was noticed that the addition of many more nonzero fields in the transition matrix led to increased computational time. To get around this an optional

parameter, $r+1 \leq \varphi \leq N$, is introduced that can be defined as the minimum number in system at which reneging occurs. Therefore, instead of having many negligible values for the $\tilde{P}r_n$ when the number in queue is small, these values are zero. This obviously has consequence for the function definitions above; however, a commentary of this is omitted.

It can be seen from Ch 6.4.2.2 that average occupancy in recent years is about 21; whilst from January 2010 there have been about 25 available beds at the unit. Since there has always been a queue (from January 2010 to April 2011 – Ch 6.4.1) it can be surmised that the major limiting factor of occupancy is the skill mix of the staff in relation to the demands of the patients. Since the queuing model is not capable of representing such behaviour the approach taken is to set the number of service channels at the average occupancy (i.e. $r = 21$) in order to ensure the properties of the queue are well represented. The probability of having fewer patients in the system than beds should be very small indeed.

With a given service rate and number of service channels, this can only be achieved by setting the arrival rate at a sufficiently high value. However, even when the arrival rate is set equal to the maximal rate of departure (see second graph of this chapter) there is still a 15% chance that the system is in a state in which there is at least one bed available. Since this occurs when the traffic intensity is unitary this represents the lowest possible probability of such an event (as traffic intensity cannot exceed for stationary results). Even if this probability is acceptable, the associated characteristic of the queue is certainly not. It has been surmised from the referrals dataset that there are about sixteen ready referrals in the queue on average (min 10, max 25), which is obviously not attained by the initial model (Ch 9.2.1). It would appear that the third graph of this chapter describes a system which attains realistic properties of the queue. However, the probability of at least one empty bed is a wholly unacceptable 62%.

It is indeed true that a sufficiently high arrival rate is the only way to achieve a negligible probability of having fewer patients in the system than beds *ceteris paribus*. However, reneging or balking must also be included to control for the limiting behaviour of the queue.

9.2.4 Balking and reneging model

The two original Maple programs have been amended to include, as options, balking and reneging.

9.2.5 Second version

The $M|E_1|21|50$ system of Ch 9.2.1 is now revisited to incorporate the concepts of balking and reneging. According to the Arrivals dataset (Ch 6.4.1) there is always a queue for the NRC. In the model, this would be put down to 100% bed occupancy (despite the possibility that there could well be empty beds due to insufficient human resources). It is therefore essential to ensure that the empirical and model properties of the queue are comparable. The metrics that will be assessed are: mean effective arrival rate, mean waiting time, mean number in queue, coefficient of variation of number in queue, and mean reneging probability. Note that the mean balking probability, $\tilde{P}b$, is excluded because empirical results are not known.

The empirical result for $\tilde{\lambda}$ has been determined using the 168 non-censored and right-censored observations of the referrals dataset. The empirical result for $E[W]$ has been calculated using all 183 observations but since 27 of these are either left or right-censored the result is a lower bound on the true value[6]. An explanation of the process deriving empirical results for $E[L_q]$ and $C.V.[L_q]$ can be found in Ch 6.4.1. Finally, $\tilde{P}r$ is determined from the table of results in Ch 6.4.1.

The process of fitting the model is as follows. The service rate and number of servers are fixed at the values ascribed at the outset of this subchapter. The arrival rate, balking parameter, δ, and the two reneging parameters, γ and θ, are then varied in order to attain results for the aforementioned queue properties congruent to their empirical counterparts. The probabilities of reneging are given by the alternative description, i.e. (9.2.2.3) since the approach involving the addition of individual reneging probabilities retuned unrealistic results. The remaining questions are therefore: how are the parameters varied and how is congruency measured?

There are many options available. An exhaustive search (see Ch 4.3) of the parameter values (to a reasonable number of decimal points) is an option. However, this would be inappropriate for more advanced models that take significant time to evaluate. In such cases, a heuristic method could well be employed. Both of these approaches rely on the specification of soft and hard constraints. An example of a hard constraint is $Pr_N < 1$. The

[6] Since they represent longer admissions (see Ch 6.4.1)

(weighted) difference between the model and empirical queue properties would form the objective function, the aim of which is to minimise.

Since this is meant to be a simple model a straightforward alternative is favoured. The five queue properties have been compared for a number of sensible values of $\lambda, \delta, \gamma, \theta$. After a few experiments good results were found for $\lambda = 0.3885, \delta = 6, \gamma = 0.702, \theta = 1.432$.

Table 9.1 Summary of empirical and model performance measures

	Empirical	Model
$\tilde{\lambda}$	0.36364	0.36363
$E[W]$	41.04	43.55
$E[L_q]$	15.83	15.83
$C.V.[L_q]$	0.217	0.217
$\tilde{P}r$	0.75	0.613

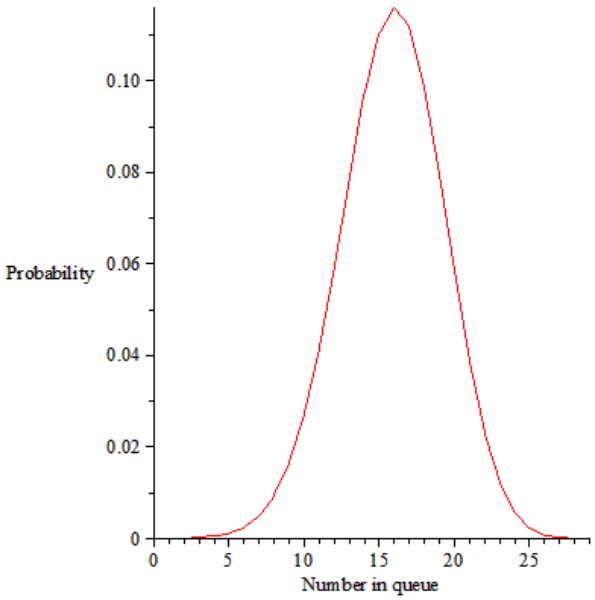


Figure 9.4 Probability density graph for number in queue with reneging and balking

As can be seen the shape of the graph looks accaptable *c.f.* Figure 6.13. This is supported by the congruency of measurements based on the first two moments ($E[L_q]$ and $C.V.[L_q]$) for the empirical and model results. The model $E[W]$ attains the lower bound of its empirical counterpart, but little else can be surmised from this value. What is of particular concern is the disparity in values for the mean reneging probability. This translates to an annual difference of about 19 reneged referrals (annual demand is 0.36364*365). Clearly something is fundamentally wrong.

Consider the notion that under steady-state conditions the long term arrival rate is equivalent to the long term departure rate (see (7.3.2.2.7)). That is

$$\tilde{\lambda} = \pi \cdot E[\Upsilon]^{-1} + \tilde{\lambda} \cdot \tilde{P}r \qquad (9.2.5.1)$$

where π is the mean occupancy of service channels defined as (7.3.2.2.9) and, in this case, $E[\Upsilon] = \mu^{-1}$.

Note that for all systems considered by the computer program the equation (9.2.5.1) is always balanced. If $\tilde{P}r$ is adjusted then either $\tilde{\lambda}$ and/or π will adjust accordingly to enforce the equality of (9.2.5.1). The probability of reneging, given by (9.2.3.1), can be adjusted by altering the reneging rates which, in turn, are adjusted by altering the reneging parameters γ and θ through (9.2.3.3) and (9.2.3.6). There is not a limitless range of values within $[0,1]$ that can be attained for $\tilde{P}r$. This is because the steady-state probabilities in (9.2.3.1) are, of course, dependent on the reneging rates and act as a stabiliser.

For example, if balking is not employed and $\sum_{n=r}^{N-1} \bar{P}_n \approx 1$ (i.e. $\pi = r, \bar{P}_N = 0, \tilde{\lambda} = \lambda$) both before and after a change to the reneging rates then this change is absorbed by the consequent change in $\bar{P}_n$ without altering the values of $\tilde{P}r$. However, if through an adjustment to the reneging rates the mean occupancy, π, has decreased then $\tilde{P}r$ must have increased. This is understandable since if there are more reneged referrals then there are fewer customers that make it to service. If, instead, $\bar{P}_N = P_R$ has increased (above zero) then $\tilde{\lambda} < \lambda$ which means that $\tilde{P}r$ must have decreased. Again this is understandable, since the length of the queue would increase if fewer customers renege. Indeed balking also has an effect on $\tilde{\lambda}$ which affects the $\bar{P}_n$.

To digress further, the effect of a change to the scaling factor, γ, and the exponent, θ, on both the expectation and variance of queue length is investigated. Changing the scaling factor has the effect of linearly stretching or contracting the Pr_n defined in (9.2.3.6) whilst altering the exponent changes the concavity of the Pr_n (note that $Pr_n = 0$ for $n \le r$ and $Pr_N = \gamma$). First, the exponent is fixed (at one) and the scaling factor varied ($\lambda = 0.38, \mu = 0.0067522$, no balking).

$\theta = 1; \gamma =$	0.6	0.7	0.8	0.9	0.9999
$E\left[L_q\right]$	14.72	12.55	10.92	9.65	8.64
$s.d.\left[L_q\right]$	4.04	3.73	3.49	3.28	3.11

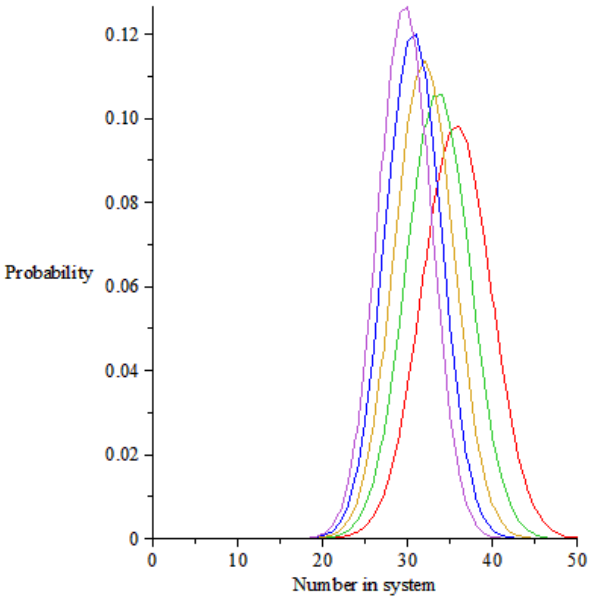


Figure 9.5 Probability density graph for number in system with varying scaling factor of reneging

When the scaling factor is increased the likelihood of reneging is greater. Therefore the mean number in system/queue is reduced. Since, in this case, it is improbable for the system to be in a state in which $n_s < r$ (because $\lambda >> r\mu$) this has the effect of bunching up the probabilities; thus reducing the variance. Next, the exponent is varied whilst the scaling factor is held constant at 0.8.

$\gamma = 0.8; \theta =$	0.25	0.5	1.25	3	6
$E\left[L_q\right]$	1.54	4.63	13.14	20.38	23.89
$s.d.\left[L_q\right]$	1.73	2.93	3.44	2.81	2.23

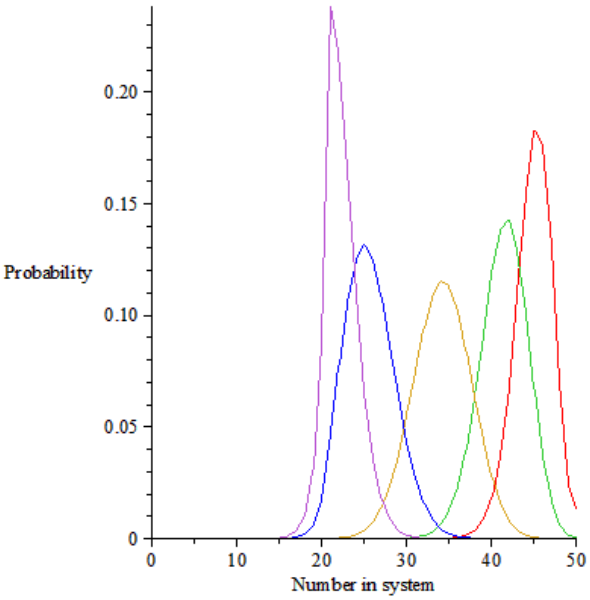


Figure 9.6 Probability density graph for number in system with varying exponent of reneging

There are two types of behaviour present in this graph that relate to $\theta<1$ and $\theta>1$. As θ is increased from zero to one the graph of the Pr_n becomes less concave downward and more linear. This has the effect of decreasing the rate at which the Pr_n increase for $n>r$ close to r but increasing it for $n\le N$ close to N up until, when $\theta=1$, the rate of increase is uniform over n. This leads to a larger mean number in system/queue since there is less difference between the Pr_n for small $n>r$ and so states which correspond to a larger number in system become more accessible. This also has the effect of increasing the variance since more states are frequented by the system.

As θ is increased from one the graph of the Pr_n becomes less linear and more concave upward. When $\theta\to\infty$ not only does the numerical difference between the Pr_n become negligible for the majority of values of $n<<R$ but so too do their actual values. This causes the mean number in system/queue to grow further. However, as $n\to R$ the rate of increase grows rapidly, thus reducing the steady-state probability density beyond this point, and obtaining a graph shape more like a spike (lower variance). These two forms of limiting behaviour are evident in the example above.

To return to the issue of a lack of congruency in $\tilde{P}r$ for empirical and model results, it may be stated that a higher value of $\tilde{P}r$ is only attainable at the expense of other performance measures. For example, the $\Re_n$ could be chosen such that π is lower, and thus $\tilde{P}r$ is larger (to a limit) but then $E\left[L_q\right]$ and $E\left[W\right]$ become negligible. The only way[7] to increase the congruency of all four performance measures is to change the fundamental parameters of the system, i.e. λ,μ,r,N. But surely it is wrong to meddle with measures that have been directly obtained through raw data? Whilst this is true, it is important to point out that the datasets these have been obtained through are representative of different timescales. λ,r and N have all been deduced from the Arrivals dataset (January 2010 to April 2011) whilst μ is deduced from the Service dataset (January 2003 to May 2011). Upon discussions with clinicians it was

[7] Assuming the nature of the system remains, i.e. Markovian arrivals and services, first-in first-out etc

learnt that there was a gut feeling[8] that in recent months the LOS may have been unusually high due to increased bed-blocking.

9.2.6 Third version

Obviously (9.2.5.1) does not hold with the empirical results $\tilde{\lambda} = 0.36364, \mu = 0.006698,$ $\pi = 21$ and $\tilde{P}r = 0.75$ but using only $\tilde{\lambda}, \pi$ and $\tilde{P}r$ the equation is solved for μ;

$$\mu = \frac{\tilde{\lambda}\left(1-\tilde{P}r\right)}{\pi} = \frac{4329}{999989} = 0.004329 \qquad (9.2.6.1)$$

Therefore, it is inferred that the average LOS for patients within the sixteen months from January 2010 to April 2011 is roughly 231 days – about 80 days more than the eight year average obtained from the Service dataset. After some[9] fine-tuning a good fit was obtained with parameter values $\lambda = 0.3879, \delta = 6.1, \gamma = 0.70405, \theta = 1.073$.

Table 9.2 Summary of empirical and model performance measures

	Empirical	Model
$\tilde{\lambda}$	0.36364	0.36364
$E[W]$	41.04	43.54
$E[L_q]$	15.83	15.83
$C.V.[L_q]$	0.217	0.217
$\tilde{P}r$	0.75	0.75

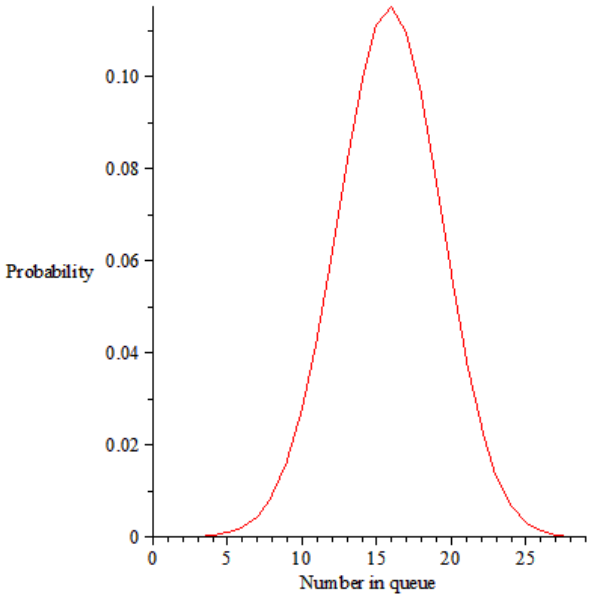


Figure 9.7 Probability density graph for number in queue with reneging and balking

Note that this model is constructed only from the Arrivals dataset; taken from a period of time in which LOS is *'unusually high'*. The model can now be used for what-if analyses.

[8] Hard to verify since many patients admitted in 2010 and 2011 were excluded from the NHS Trusts data request since at the end of January 2011 they were still inpatients

[9] Took approximately 30 minutes

However, before any hypothetical scenarios can be considered it is necessary to produce a cost-benefit measure in order to compare various configurations.

This measure could be in the form of a single-valued cost function, whose minimum value is sought. The cost, per unit time, (in this case, days) is given by the formula

$$\begin{aligned}&\text{Cost function}_{/day} = \pi \cdot \text{Bed cost}_{/day} \\ &+ \left(1-\tilde{P}r\right) \cdot E\left[L_q\right] \cdot \text{Deterioration cost}_{/day} + \tilde{P}r \cdot \tilde{\lambda} \cdot \text{Reneging cost}\end{aligned}$$

The (average) bed cost per day for an occupied bed is £480 (2011). The (average) deterioration cost per day is much harder to quantify. This represents the cost to restore the functional ability of patients to the level at when their (ready) referral was received. By and large, the longer a patient has queued the more they have deteriorated, and thus, the more therapy is needed to restore functionality. Without an in-depth study it is impossible to quantify this cost. The same applies for the reneging cost. This represents the cost of a patient reneging from the queue. Since the patient is not admitted to the NRC the cost is not borne directly by Rookwood hospital but by the wider community (social services, other NHS facilities etc) as a result of the patient not receiving the specialist rehabilitation. The deterioration and reneging costs serve only to act as a penalty to discourage excessive waiting time and reneging and do not form part of operating costs. However, due to the complexity associated with the determination of these costs an alternative measure of cost-benefit is favoured.

The financial operating costs are simply compared with the performance measures for a particular configuration. The performance measures are given in Table 9.2 for this version of model one and the annual operating cost of the NRC is calculated as

$$21 \times £480 \times 365 = £3.679\text{m}$$

9.2.6.1 Hypothetical Scenario One: Number of beds (2010-2011)

It is mentioned in the outset of Ch 9.2 that the (simple) model under consideration here is analogous to Route 1 of Figure 6.8. In the corresponding subchapter to this route (Ch 6.2.1) it is stated that the only reasonable variable in this model is the number of service channels. A cursory example of a what-if type analysis in this variable is considered using the Maple program (Ch 7.3.2.2.7) with parameters of the third version but with arrival rate varied such that effective arrival rate is approximately equal to its empirical counterpart (0.36364).

Table 9.3 Performance measures and costs for varying number of beds

Number of beds:	19	20	21	22	23
$\tilde{\lambda}$	0.36362	0.36364	0.36364	0.36363	0.36364
π	19	20	21	22	23
$\tilde{P}r$	0.774	0.762	0.75	0.738	0.726
Tp	30.02	31.6	33.18	34.76	36.34
$E[W]$	48.06	45.77	43.54	41.35	39.21
$E[L_q]$	17.47	16.64	15.83	15.04	14.26
Cost pa	£3.329m	£3.504m	£3.679m	£3.854m	£4.030m

So, is the 3.2% reduction in the probability of reneging and 3.16 patient increase in throughput worth spending an extra £350k to raise capacity from 21 to 23? Or, in times of austerity, is the decrease in throughput to 30 patients per year acceptable if the budget is to be reduced to three and a third million?

9.2.6.2 Evaluation

What would appear to be a good fit has been obtained for the third version (Ch 9.2.6). Four of the five measurements assessed were accurate to at least four significant figures. It is not easy to assess the accuracy of the mean waiting time since only a lower bound on this value was attainable from the referrals dataset. The accuracy is in fact dependent on the extent to which the assumptions of this model are correct. The validity of the assumptions made by the model is forthwith assessed.

Arrival distribution

Assumption: Random (Poisson) arrivals. *Validity*: Justified. The inter-arrival times were shown to be exponentially distributed in Ch 6.4.1.

Assumption: Balking. *Validity*: Unknown. Real-life prevalence not known.

Service distribution

Assumption: Homogenous servers. *Validity:* Unjustified. There are significant differences in LOS for patients of differing attribute (Ch 6.4.2.1).

Assumption: Exponential service distribution. *Validity*: Unknown. Whilst the exponential distribution provides a reasonable fit to the Service dataset LOS (Ch 6.5.3.1) there was no LOS data available for the period of time considered here.

Assumption: Service rate = 0.004329. *Validity*: Unknown. Real-life data not available.

Number of service channels

Assumption: # servers = # beds. *Validity*: Unjustified. Occupancy depends on skill mix of staff and needs of patients (see Ch 1.2). The number of beds is an upper bound on occupancy.

Capacity of system

Assumption: Capacity = 50. *Validity*: Justified. Just because the number in system did not exceed fifty in the sixteen months of the Arrivals dataset this does not mean that it never does. However, most of the time the number in system has been well below this value and if it is indeed true that patients stayed longer during this time then this would only serve to reduce this number further.

Queue discipline

Assumption: First-in first-out. *Validity*: Unjustified. Queue is priority-based at Rookwood NRC (see Ch 1.2).

9.2.7 Hypothetical Scenario Two: Number of beds (2003-2011)

A what-if scenario similar to that of Ch 9.2.6.1 is now studied but for the period of time from 2003 to 2011. In this case (9.2.5.1) is solved for the probability of reneging, $\tilde{P}r$, since an empirical measure of this is not available for the timescale considered here. It can sensibly be assumed that the (effective) arrival rate during the sixteen month window of the Arrivals dataset is indifferent from its value over the eight years of the Service dataset. This assumption is valid since balking is not *'a significant issue on the ground'* and it is used only to promote convergent behaviour of the queue. The mean LOS for this eight year period has been found to equal 149.3 days. It is assumed that the mean occupancy of the system is equivalent to the number of service channels, 21. Using these it is found that $\tilde{P}r = 0.614$. It is impossible to determine numerical targets for the remaining three summary measures $E[W], E[L_q]$ and $CV[L_q]$ because these are not available for the 2003-2011 timescale. It would be foolhardy to simply state that because LOS is, for example, 36% less (i.e. 231 to 149 days) that $E[W]$ and $E[L_q]$ target values should be 36% less than those given in Table 9.2.

There is therefore a fair amount of flexibility with regard to choosing the controls $\lambda, \delta, \gamma, \theta$ as it is only required to match two summary measures ($\tilde{\lambda} = 0.36364$ and $\tilde{P}r = 0.614$). It is

therefore essential for the user to inspect $E[W], E[L_q]$ and the probability graph of number in system to ensure that sensible values of the controls are selected. Appropriate values are found to be $\lambda = 0.3642, \delta = 10, \gamma = 0.9, \theta = 1.2$. In an identical approach to before (Ch 9.2.6.1) the number of beds is varied whilst ensuring the effective arrival rate is (approximately) equal to 0.36364.

Table 9.4 Performance measures and costs for varying number of beds

Number of beds:	19	20	21	22	23
$\tilde{\lambda}$	0.36364	0.36364	0.36364	0.36364	0.36364
π	19	20	21	22	23
$\tilde{P}r$	0.65	0.632	0.613	0.595	0.576
Tp	46.45	48.89	51.34	53.78	56.23
$E[W]$	34.89	32.86	30.89	28.96	27.09
$E[L_q]$	12.69	11.95	11.23	10.53	9.85
Cost pa	£3.329m	£3.504m	£3.679m	£3.854m	£4.030m

Perhaps of most interest is the mean probability of reneging and the throughput since the expected waiting time and length of queue include patients who will ultimately renege. Questions similar to those posed in Ch 9.2.6.1 can be asked.

In changing the number of service channels it is assumed that mean LOS, and so treatment intensity, is unaltered. Therefore, the amount of treatment provided must increase or decrease by the same proportion as the number of beds increase or decrease. If the number of beds is increased then the additional cost of higher occupancy is borne from requiring more staff to achieve the same intensity of treatment for each patient.

9.3 Model Two

Whilst the third version of model one provides what would appear to be a very good fit (by the comparison of performance measures) it lacks the complexity to represent the various components of the system that exist in real-life at Rookwood NRC and limits the scope of a what-if analysis. Such a trade-off between simplicity and realism has been introduced in the outset of Ch 6.2. This subchapter introduces the various components of the system that are eminent at Rookwood NRC and culminates in a summary (Ch 6.2.7) which illustrates a number of possible combinations of components that can be used as part of a queuing model.

The model of this subchapter is a midpoint between the relatively simple model one and the more complex model three. It is analogous to Route 3 of Figure 6.8.

That is, LOS is partitioned to active and blocked (Ch 6.2.2) and patients are grouped by treatment control factors (Ch 6.2.6). As before, treatment intensity is not (yet) a consideration and so stage two and three are excluded. The Coxian and exponential distributions are fitted to active and blocked LOS respectively (Chapter 8) whilst a disconnected system (Ch 8.4.2) with multiple classes of server (Ch 8.4.2.1) is used to account for the patient groups.

9.3.1 Phase One; Stage One: Patient groups

In the first stage an appropriate number, and criteria, of patient groups are determined for which active (not total) LOS is similar. This is done using regression tree analysis. However, there is a problem here as only 89 of the 358 patient episodes have a date ready for discharge (Ch 6.4.2.3). To overcome this, the total LOS is used in the regression tree analysis. Using total, and not active, LOS in this analysis means that patient groups are deduced for which total LOS is similar. By assuming homogeneity of blocked LOS within these groups it can be inferred that active LOS is also similar.

Treeworks[10] is used for the regression tree analysis. The dependent variable is total LOS whilst the independent variables are age on admission, gender, local health board, primary diagnosis, admission source, admission method, and discharge destination (see Ch 6.4.2.1). First, the data is split into two sets based on the branching[11] of a certain variable that produces the largest reduction in variance.

[10] A CART (classification and regression tree) analysis program written by Evandro Leite and Paul Harper
[11] Partition of data into sets based on variable value

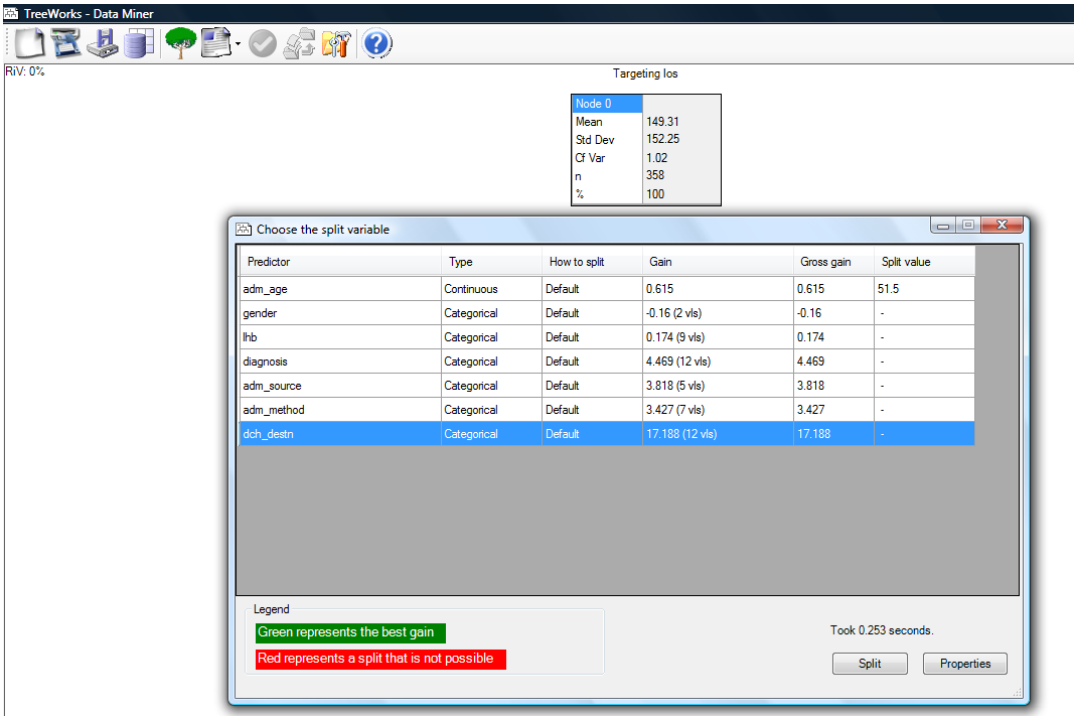


Figure 9.8 Choice of branching variables in Treeworks for first split

Clearly, discharge destination is the variable that, if branched on, would create the largest reduction in variance (17%). However, this is not the sole determining factor of which variable to branch upon. There are two conditions that must be met. First, the number in each group, $n_i, i = 1,..., p$ where p is the number of groups after the split (at this stage $p = 2$), must be sufficiently large (since the accuracy of a fitted distribution depends on the sample size). Second, the number of service channels assigned to the integrated system of each group (see Ch 8.4.2), r_i, must be sufficiently large so that the likelihood of an empty bed in one group at the same time as a queue in another is very low indeed. The proportion of service channels assigned to the integrated system of each group is equal to the proportion of bed days occupied by patients of that group. Therefore

$$r_i = \frac{n_i \cdot \widehat{LOS}_i^{total}}{\sum_{t=1}^{p} n_t \cdot \widehat{LOS}_t^{total}} \cdot r \qquad (9.3.1.1)$$

must be large enough (note r is the total number of service channels). Branching on discharge destination yields $n_1 = 223, \widehat{LOS}_1^{total} = 99.86, n_2 = 136, \widehat{LOS}_2^{total} = 231$ which gives $r_1 = 8.71, r_2 = 12.29$. Here, both n_i and r_i are sufficiently large.

After much deliberation the following splits have been made (21% reduction in variance) based on the aforementioned criteria.

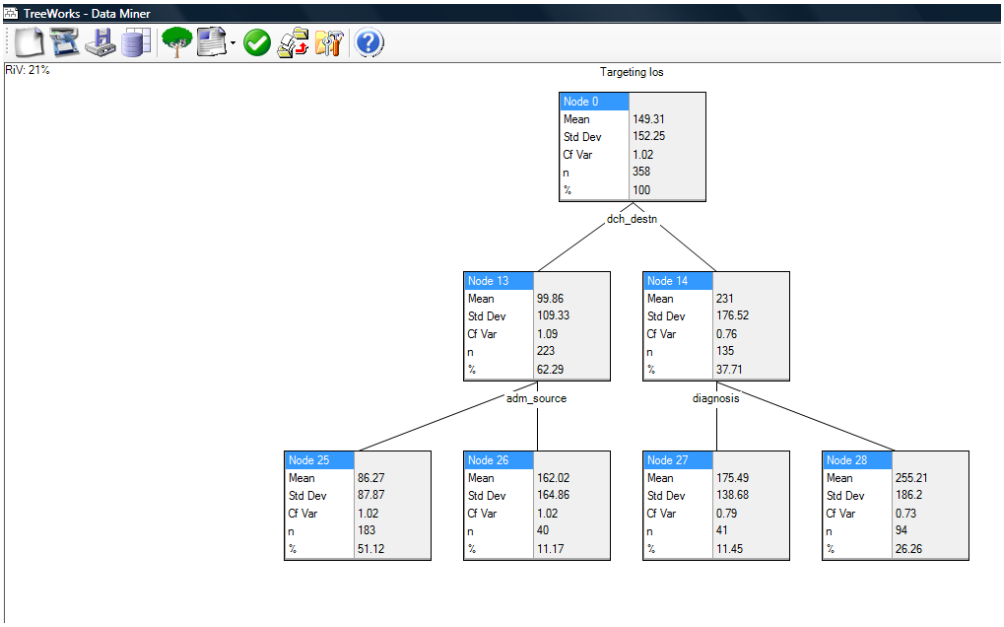


Figure 9.9 Final tree with four patient groups

Table 9.5 and Table 9.6 summarise some details of each patient group.

Table 9.5 Attributes of appropriate branching variables for each patient group

Group	Discharge destination	Admission source	Diagnosis
1	Usual residence, Temporary residence	Cardiff & Vale hosp, Usual residence	
2	Usual residence, Temporary residence	NHS hosp	
3	Internal transfer (gen), Other Trust (gen), Non-NHS nursing home		GBS, MS, Anoxia, MS &other neuropathies
4	Internal transfer (gen), Other Trust (gen), Non-NHS nursing home		ABI, TBI, other

Table 9.6 Sample size, average length of stay and number of beds for each patient group

Group,i	n_i	$\widehat{LOS}_i^{total}$	r_i
1	183	86.27	6.2
2	40	162.02	2.55
3	41	175.49	2.83
4	94	255.21	9.42

The r_i must obviously be integer[12] and are thus rounded such that

$$r_1 = 6, r_2 = 3, r_3 = 3, r_4 = 9 \quad (9.3.1.2)$$

The effect of rounding up and down the number of beds for a group is that constituent patients are over and under-represented in the model (albeit not by that much). Note that whilst discharge destination, a retrospective predictor of LOS, can be used in the queuing model, it cannot be used in a prognostic model (Ch 6.2.6). A regression tree excluding this variable is produced (with nine leaves) and is available in Appendix 9.1. However, the reduction in variance here is only 12%. If discharge destination can be estimated pre or on-arrival then a tree with eight leaves can be produced with a reduction in variance of 23% (Appendix 9.2).

Note that the number of patients from each group at Rookwood NRC varies from week to week in real-life. However, since the model is constructed under the steady-state condition these partitions are assumed to be fixed in the long run.

9.3.2 Phase One; Stage Four: Queuing model

9.3.2.1 Task One: Determine target arrival rates

It is known (Ch 6.4.1) that the (effective) arrival rate of ready referrals to Rookwood NRC is 0.36364. The task now is to determine what the effective arrival rate into each of the four disconnected queuing systems (that represent patient groups – see Ch 8.4.2) should be. On the face of it, such a task would appear trivial – first, simply find the proportion of the 168 patient episodes of the Arrivals dataset (Ch 6.4.1) that belong to each of the patient groups. Then, if half of the patient episodes are from group three then it can be surmised that the effective arrival rate for the third queuing system should be 0.18182 (note that reneging from this group would have to be very high since the throughput, with only three servers and a LOS of 175 days, is very low indeed).

However, since the (NHS Trust) attributes (age, gender, diagnosis etc) of referrals are unknown, an alternative approach is necessary. Essentially, the proportion of total arrivals to each patient group is set equal to their proportion of total throughput. That is, the arrival rate into the system that represents patient group *i* is given by

[12] Since whilst in the model the number of beds in each group must be fixed, in reality it varies over time depending on the demographic of the waiting list

$$\tilde{\lambda}_i = \frac{\frac{r_i}{\widehat{LOS}_i^{total}}}{\sum_{t=1}^{p} \frac{r_t}{\widehat{LOS}_t^{total}}} \cdot \tilde{\lambda} \qquad (9.3.2.1.1)$$

This yields ($p = 4$)

$$\tilde{\lambda}_1 = 0.18010, \tilde{\lambda}_2 = 0.04795, \tilde{\lambda}_3 = 0.04427, \tilde{\lambda}_4 = 0.09132 \qquad (9.3.2.1.2)$$

The drawback to this approach is that the mean probability of an arriving customer reneging before entering service is assumed to be equivalent across all queuing systems (patient groups). The proof of this is as follows.

Extending (9.2.5.1) to each group *i* (note: $E\left[\Upsilon_i\right] = \widehat{LOS}_i^{total}$);

$$\tilde{\lambda}_i = \pi_i \cdot E\left[\Upsilon_i\right]^{-1} + \tilde{\lambda}_i \cdot \tilde{P}r_i \qquad (9.3.2.1.3)$$

Assuming that the model λ_i s are chosen such that the effective mean arrival rates, $\tilde{\lambda}_i$, into each group are equal to (9.3.2.1.2) and that the mean occupancy of each disconnected system is approximately equal to the number of servers yields

$$\tilde{\lambda}_i \approx r_i \cdot E\left[\Upsilon_i\right]^{-1} + \tilde{\lambda}_i \cdot \tilde{P}r_i \qquad (9.3.2.1.4)$$

and substituting (9.3.2.1.1) gives

$$\frac{r_i \cdot E\left[\Upsilon_i\right]^{-1}}{\sum_{t=1}^{p} r_t \cdot E\left[\Upsilon_t\right]^{-1}} \tilde{\lambda}\left(1 - \tilde{P}r_i\right) \approx r_i \cdot E\left[\Upsilon_i\right]^{-1} \qquad (9.3.2.1.5)$$

which simplifies to

$$\tilde{P}r_i \approx 1 - \frac{\sum_{t=1}^{p} r_t \cdot E\left[\Upsilon_t\right]^{-1}}{\tilde{\lambda}} \qquad (9.3.2.1.6)$$

Note the absence of dependence of the $\tilde{P}r_i$ on *i*. With $\tilde{\lambda} = 0.36364$

$$\tilde{P}r_i \approx 0.614 \qquad (9.3.2.1.7)$$

Therefore, the mean probability of an arriving customer reneging before entering service is the same for each of the four queuing systems. Moreover, and as is to be expected, the value of this probability is approximately equal to that of the second version of model one (*c.f.* $\tilde{P}r$ in first table of Ch 9.2.5).

Without the appropriate data it is not possible to assess the validity of the result (9.3.2.1.7). It will suffice to say that in real-life it is entirely plausible that the diagnosis and admission source (or perhaps other attributes such as age) could influence the probability of reneging.

9.3.2.2 Task Two: Determine buffer sizes

Buffer size (or maximal number of customers in queue) for each patient group, b_i, is deduced by the proportion of beds allocated to each group multiplied by the total buffer size, $b = 29$.

Thus, $b_1 = \frac{r_1}{r} \cdot b = \frac{6}{21} \cdot 29 = 8.29, b_2 = b_3 = 4.14, b_4 = 12.43$. Of course the b_i must be integer and so, taking

$$b_i = \left\lfloor \frac{r_i}{r} \cdot b \right\rfloor \qquad (9.3.2.2.1)$$

yields $b_1 = 8, b_2 = b_3 = 4, b_4 = 12$. Note that here *b* has been reduced to 28.

9.3.2.3 Task Three: Fit service distributions

The third task is to fit appropriate distributions to active and blocked LOS for each patient group. However, since not all patient episodes have a specified date ready for discharge there is incomplete information. Of the 183, 40, 41 and 94 observations in the patient groups only 46, 13, 7 and 23 patient episodes have a specified date ready for discharge. This problem is dealt with as follows.

For each patient group the observations with complete information are used to estimate the date ready for discharge for those with empty fields. For the $\hat{n}_i \le n_i$ observations that have complete information a mean proportion, $\hat{P}_i^{active} \le 1$, of active to total LOS is determined. If $LOS_{i,j}$ represents the LOS of the *j*-th observation with complete information of patient group *i* then,

$$\hat{P}_i^{active} = \sum_{j=1}^{\hat{n}_i} \frac{LOS_{i,j}^{active}}{LOS_{i,j}^{total}} \Big/ \hat{n}_i \qquad (9.3.2.3.1)$$

Therefore, the mean proportion of blocked to total LOS, $\hat{P}_i^{blocked}$, is given by

$$\hat{P}_i^{blocked} = 1 - \hat{P}_i^{active} \tag{9.3.2.3.2}$$

Since the exponential distribution is used to model blocked LOS (Chapter 8),

$$\frac{\hat{LOS}_i^{blocked}}{\hat{LOS}_i^{total}} \sim \exp\left(\frac{1}{\hat{P}_i^{blocked}}\right) \tag{9.3.2.3.3}$$

This distribution is now sampled to obtain exponential variates, $s_k, k = 1,2,..,n_i - \tilde{n}_i$, for the observations with incomplete information using the relation[13]

$$s_k = -\hat{P}_i^{blocked} \cdot \ln(rnd) \tag{9.3.2.3.4}$$

where $0 < rnd < 1$ is a randomly generated number. Therefore, through (9.3.2.3.3),

$$\frac{LOS_{i,k}^{blocked}}{LOS_{i,k}^{total}} = s_k \tag{9.3.2.3.5}$$

and since the $LOS_{i,k}^{total}$ are known for the $n_i - \tilde{n}_i$ observations with incomplete information, the generated values for blocked and active LOS are

$$\begin{aligned} LOS_{i,k}^{blocked} &= s_k \cdot LOS_{i,k}^{total} \\ LOS_{i,k}^{active} &= LOS_{i,k}^{total} - LOS_{i,k}^{blocked} \end{aligned} \qquad k = 1,2,..,n_i - \tilde{n}_i \tag{9.3.2.3.6}$$

However, the mean proportion of blocked to active LOS for each patient group is not necessarily the same for the $\hat{n}_i$ episodes that have information on date ready for discharge as it is for the n_i episodes that include generated values. Defining

$$\begin{aligned} \bar{P}_i^{blocked} &= \sum_{l=1}^{n_i} \frac{LOS_{i,l}^{blocked}}{LOS_{i,l}^{total}} \Big/ n_i = \left(\sum_{j=1}^{\hat{n}_i} \frac{LOS_{i,j}^{blocked}}{LOS_{i,j}^{total}} + \sum_{k=1}^{n_i - \hat{n}_i} \frac{LOS_{i,k}^{blocked}}{LOS_{i,k}^{total}} \right) \Big/ n_i \\ &= \frac{\hat{n}_i}{n_i} \hat{P}_i^{blocked} + \frac{1}{n_i} \sum_{k=1}^{n_i - \hat{n}_i} \frac{LOS_{i,k}^{blocked}}{LOS_{i,k}^{total}} = \frac{\hat{n}_i}{n_i} \hat{P}_i^{blocked} + \frac{1}{n_i} \sum_{k=1}^{n_i - \hat{n}_i} s_k \end{aligned} \tag{9.3.2.3.7}$$

[13] Easily derived using inverse transform sampling

using (9.3.2.3.5), it can be ensured that $\bar{P}_i^{blocked}$ is closer to the value of $\hat{P}_i^{blocked}$ if a linear scaling function, a_i, is attached to the exponential variates such that

$$\hat{P}_i^{blocked} = \frac{\hat{n}_i}{n_i}\hat{P}_i^{blocked} + \frac{1}{n_i}\sum_{k=1}^{n_i-\hat{n}_i} a_i \cdot s_k \qquad (9.3.2.3.8)$$

It follows that

$$a_i = \frac{\left(n_i - \hat{n}_i\right)\cdot \hat{P}_i^{blocked}}{\sum_{k=1}^{n_i-\hat{n}_i} s_k} \qquad (9.3.2.3.9)$$

Here

$$\bar{P}_1^{blocked} = 0.15797, \bar{P}_2^{blocked} = 0.16263, \bar{P}_3^{blocked} = 0.30271, \bar{P}_4^{blocked} = 0.30208 \quad (9.3.2.3.10)$$

Using Table 9.6, (9.3.2.3.10) and the formulae $\hat{LOS}_i^{blocked} = \bar{P}_i^{blocked} \cdot \hat{LOS}_i^{total}$ and $\hat{LOS}_i^{active} = \hat{LOS}_i^{total} - \hat{LOS}_i^{blocked}$;

$$\hat{LOS}_1^{active} = 72.64, \hat{LOS}_2^{active} = 135.67, \hat{LOS}_3^{active} = 122.37, \hat{LOS}_4^{active} = 178.12 \quad (9.3.2.3.11)$$

$$\hat{LOS}_1^{blocked} = 13.63, \hat{LOS}_2^{blocked} = 26.35, \hat{LOS}_3^{blocked} = 53.12, \hat{LOS}_4^{blocked} = 77.09 \quad (9.3.2.3.12)$$

Since $LOS_{i,k}^{active}$ and $LOS_{i,k}^{blocked}$ are now known $\forall k \in \left(1,2,...,n_i\right)$, appropriate distributions can now be fitted to the (empirical and generated) data for active and blocked LOS. This is done using a similar approach to that employed in Ch 6.5. For active LOS the aim is not to find the best fitting distribution but to find the most appropriate Coxian distribution (for each group). That is, log-likelihood value can always be improved upon[14] with an additional phase, but this increases the size of the transition matrix to be solved (Chapter 8) and so increases complexity. An appropriate distribution must balance goodness of fit with tractability (AIC and BIC can be used to measure this).

[14] Or strictly speaking – can never be worsened

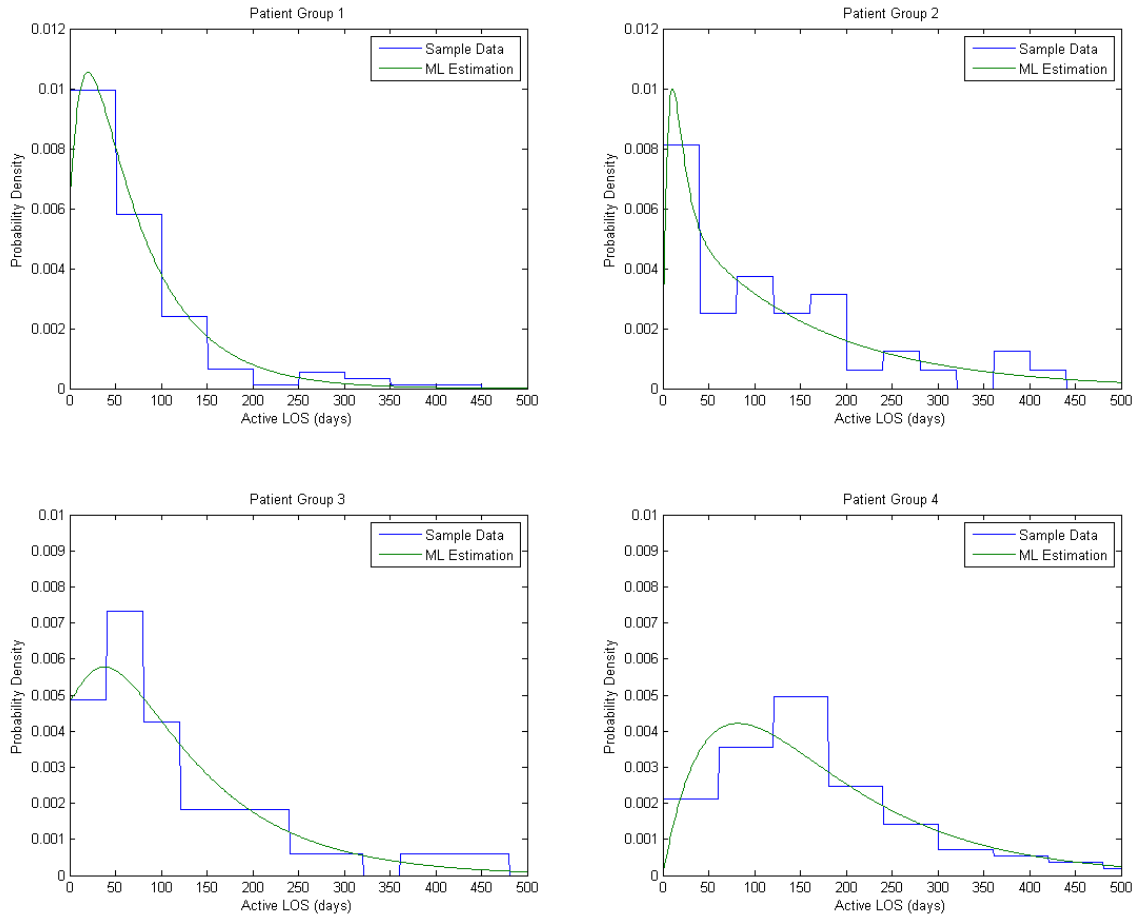


Figure 9.10 Coxian approximations to active length of stay for each patient group

P.G. 1 (2 phase): $\mu_{1,1} = 0.0734, \mu_{1,2} = 0.0157, \alpha_{1,1} = 0.08$

P.G. 2 (3 phase): $\mu_{2,1} = 0.1238, \mu_{2,2} = 0.1238, \mu_{2,3} = 0.0068, \alpha_{2,1} = 0, \alpha_{2,2} = 0.1813$

P.G. 3 (3 phase): $\mu_{3,1} = 0.0097, \mu_{3,2} = 0.0328, \mu_{3,3} = 0.0648, \alpha_{3,1} = 0.5, \alpha_{3,2} = 0.6721$

P.G. 4 (2 phase): $\mu_{4,1} = 0.0171, \mu_{4.2} = 0.0084, \alpha_{4,1} = 0$ (9.3.2.3.13)

It would appear that a good fit to active LOS is attained for each patient group (respective chi square test statistics: 0.46, 0.17, 0.54, 1.21). The only fit that is not so good is that of patient group four. Clearly additional phases are required to better represent this data but, with nine service channels and a buffer size of twelve, an extra phase would incur a transition matrix of dimension 3355. In fact, to aid tractability, $\alpha_{4,1}$, the probability of transition to the blocked phase, has been fixed at zero to reduce the number of nonzero entries in the transition matrix (see, for example, (8.2.1.1.2)). The results of Ch 8.3.1.6 show how it is not just the dimension of the transition matrix but also its density that determines the speed of solution. This effective reduction to a two term hypo-exponential distribution (Ch 5.5.3.1) has done little in reducing the quality of the fit.

Fitting the exponential distribution to blocked LOS is a far easier task as the two parameter estimation methods (ML and method of moments) agree on the value of the single parameter to be estimated. Therefore, by (5.5.1.1.2),

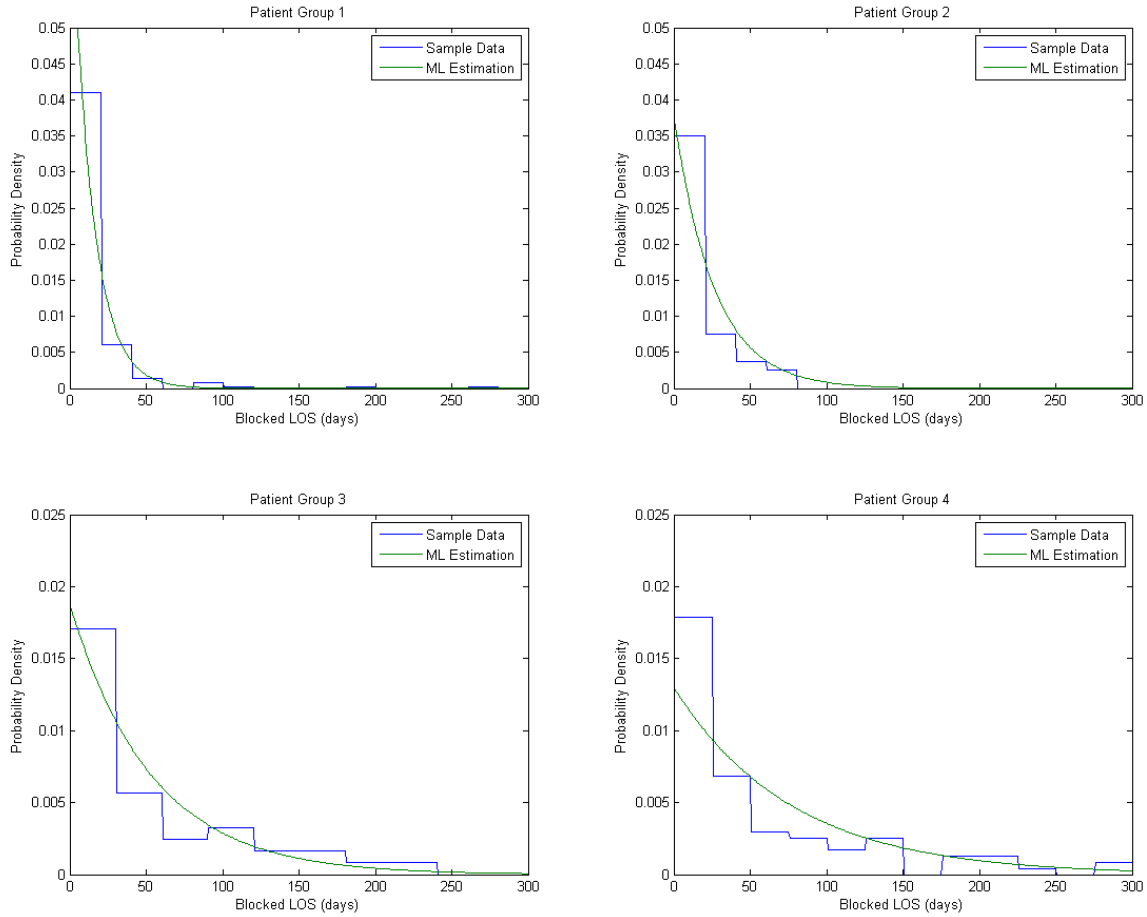


Figure 9.11 Exponential distributions to blocked length of stay for each patient group

$$\mu_{1,3} = 0.0734, \mu_{2,4} = 0.0380, \mu_{3,4} = 0.0181, \mu_{4,x} = 0.0130 \qquad (9.3.2.3.14)$$

Again, what would appear to be a good fit has been attained for each patient group. It is conjectured that a hyper-exponential distribution may well be better suited to these data; particularly patient group four.

By definition the mean of each approximating distribution for blocked LOS is equivalent to the mean of the (empirical and generated) data. That is, one over the results of (9.3.2.3.14) returns the values (9.3.2.3.12). However, this is not the case for active LOS. Therefore, using (8.3.1.5.9) with (9.3.2.3.13) gives

$$\widehat{LOS}_1^{active} = 72.22, \widehat{LOS}_2^{active} = 136.55, \widehat{LOS}_3^{active} = 120.87, \widehat{LOS}_4^{active} = 177.53 \quad (9.3.2.3.15)$$

Note these are similar but not equal in value to (9.3.2.3.11).

9.3.2.4 Task Four: Construct queuing model

The final task of phase one is to solve the system

$$M,M,M,M\,|\,C_2+M,C_3+M,C_3+M,C_2+M\,|\,6,3,3,9\,|\,8,4,4,12$$

Service rates are fixed at (9.3.2.3.13) and (9.3.2.3.14); arrival rates are not known; and the balking and reneging parameters are also not known. As before (Model One; Third version – Ch 9.2.7) the values of these are determined by targeting the effective arrival rate and mean reneging probability but this time, for each patient group – (9.3.2.1.2) and (9.3.2.1.7).

As before, there is a fair amount of flexibility with regard to choosing the controls $\lambda,\delta,\gamma,\theta$ as it is only required to match two summary measures (for each patient group). It is therefore essential for the user to inspect $E[W], E[L_q]$ and the probability graph of number in system to ensure that sensible values of the controls are selected. The expected waiting time and length of queue for the entire disconnected system can be calculated through (8.4.2.2.2) and (8.4.2.2.3).

The result (8.4.2.2.2) can be used to deduce a system average of either *W* or *T* denoted by P. This formula obtains a system average, $E[\mathrm{P}_{sys}]$, by summing the product of the class performance measure, $E[\mathrm{P}_i]$, and the respective probability that a customer on exit has undergone service at that class[15], $\tilde{\lambda}_i \Big/ \sum_{j=1}^{p} \tilde{\lambda}_j$, for each class, $i=1,\ldots,p$. However, with the introduction of reneging[16] these probabilities require a different formula. Since $1-\tilde{P}r_i$ represents the probability that an arriving customer at class *i* enters service, the probability that a customer on exit of the system has undergone service at that class is $\tilde{\lambda}_i\left(1-\tilde{P}r_i\right)\Big/\sum_{j=1}^{p}\tilde{\lambda}_j\left(1-\tilde{P}r_j\right)$. Therefore

$$E[\mathrm{P}_{sys}]=\left(\sum_{i=1}^{p}\tilde{\lambda}_i\left(1-\tilde{P}r_i\right)\right)^{-1}\left(\sum_{i=1}^{p}\tilde{\lambda}_i\left(1-\tilde{P}r_i\right)E[\mathrm{P}_i]\right) \qquad (9.3.2.4.1)$$

[15] Equivalent to the respective probability that an arriving customer enters that class
[16] Not balking as the effective arrival rate is, of course, dependent on the balking parameter through (9.2.2.1)

Note that the target for each $\tilde{\lambda}_i$ and $\tilde{Pr}_i$ is given by (9.3.2.1.2) and (9.3.2.1.7) respectively. The result (8.4.2.2.3) is unchanged with the introduction of reneging and balking.

The probability graph of the (total) number in system is determined through the $\bar{P}_n$. To calculate this for $n = 0,1,....,49$ the (readily available) $\bar{P}_{i,n}$ [17] are used alongside (8.4.2.2.1) such that

$$\bar{P}_n = \sum_{\forall n_1,n_2,....,n_p} \bar{P}_{1,n_1} \cdot \bar{P}_{2,n_2} \cdot .... \cdot \bar{P}_{p,n_p} \qquad \begin{matrix} n_i = 0,1,....,r_i \\ \sum_{i=1}^{p} n_i = n \end{matrix} \qquad (9.3.2.4.2)$$

So, for the system described above with, for example, $n = 1$;

$$\bar{P}_1 = \bar{P}_{1,1} \cdot \bar{P}_{2,0} \cdot \bar{P}_{3,0} \cdot \bar{P}_{4,0} + \bar{P}_{1,0} \cdot \bar{P}_{2,1} \cdot \bar{P}_{3,0} \cdot \bar{P}_{4,0} + \bar{P}_{1,0} \cdot \bar{P}_{2,0} \cdot \bar{P}_{3,1} \cdot \bar{P}_{4,0} + \bar{P}_{1,0} \cdot \bar{P}_{2,0} \cdot \bar{P}_{3,0} \cdot \bar{P}_{4,1}$$

The Maple program automatically computes (9.3.2.4.1) and (9.3.2.4.2) for the one or more disconnected queuing systems. For clarity a summary of the Maple program is provided. The inputs must be provided in a text file named ‘data.txt’. The first five lines provide input details for the first system to be solved, whilst the second five provide information for the second, etc. The input details required for each system are

line 1:	misc	$k-1$	r	N	λ	
line 2:	service rates	μ_1	μ_2	μ_3	...	μ_k
line 3:	transition probabilities	α_1	...	α_{k-2}		
line 4:	balking	b_{yn}	δ			
line 5:	reneging	r_{yn}	r_{st}	γ	θ	r_{ip}

If balking is to be included (for a system) then $b_{yn} = 1$, otherwise $b_{yn} = 0$. Similarly if reneging is to be included then $r_{yn} = 1$, otherwise $r_{yn} = 0$. The variable r_{st} is used to define the number in system at which reneging occurs (clearly $r+1 \le r_{st} \in \mathbb{Z} \le N$). Finally, if the Pr_n are calculated by summing the individual probabilities of reneging for each customer in the queue then $r_{ip} = 1$, otherwise $r_{ip} = 0$. A simple modification to the code allows the determination of symbolic results with some or all of the following as free (undefined) variables: $\lambda, \mu_l (l = 1,...,k), \alpha_l (l = 1,...,k-2), \delta, \gamma, \theta$.

[17] The probability that there is *n* in the individual system of class *i*

Outputs are provided for each class of server as well as for the holistic system. First, for each class the following results are output:

$$\tilde{\lambda}, P_R, P_W, \pi, \tilde{P}r, \tilde{P}b, Tp, E[L], E[T], E[L_q], E[W], CV[L_q], \bar{P}_n$$

in addition to the graphs:

$x-axis$	$y-axis$
# in system	probability
# in queue	probability
# in system	balking probability
# in system	arrival rate
# in system	reneging probability

For the holistic system the following results are output:

$$\tilde{\lambda}, \pi, \tilde{P}r, Tp, E[L], E[T], E[L_q], E[W]$$

as well as the graph of total number in system against probability. Note that the inputs and outputs for the Maple program for Erlang service are similar but are input directly to the maple worksheet. Obviously, there is only one value attributed to μ and there are no α's. Note also that this program only solves one system at a time.

A reasonable fit has been obtained for the following control values:

Pat group :	1	2	3	4
λ	0.1902	0.058	0.052	0.925
δ	10	8	10	10
γ	0.75	0.9	0.85	0.75
θ	1.5	1.32	1.2	1.5

The full (20 line) input file is given in Appendix 9.3. The queuing systems corresponding to each of the four patient groups are detailed below in addition to that of the holistic system.

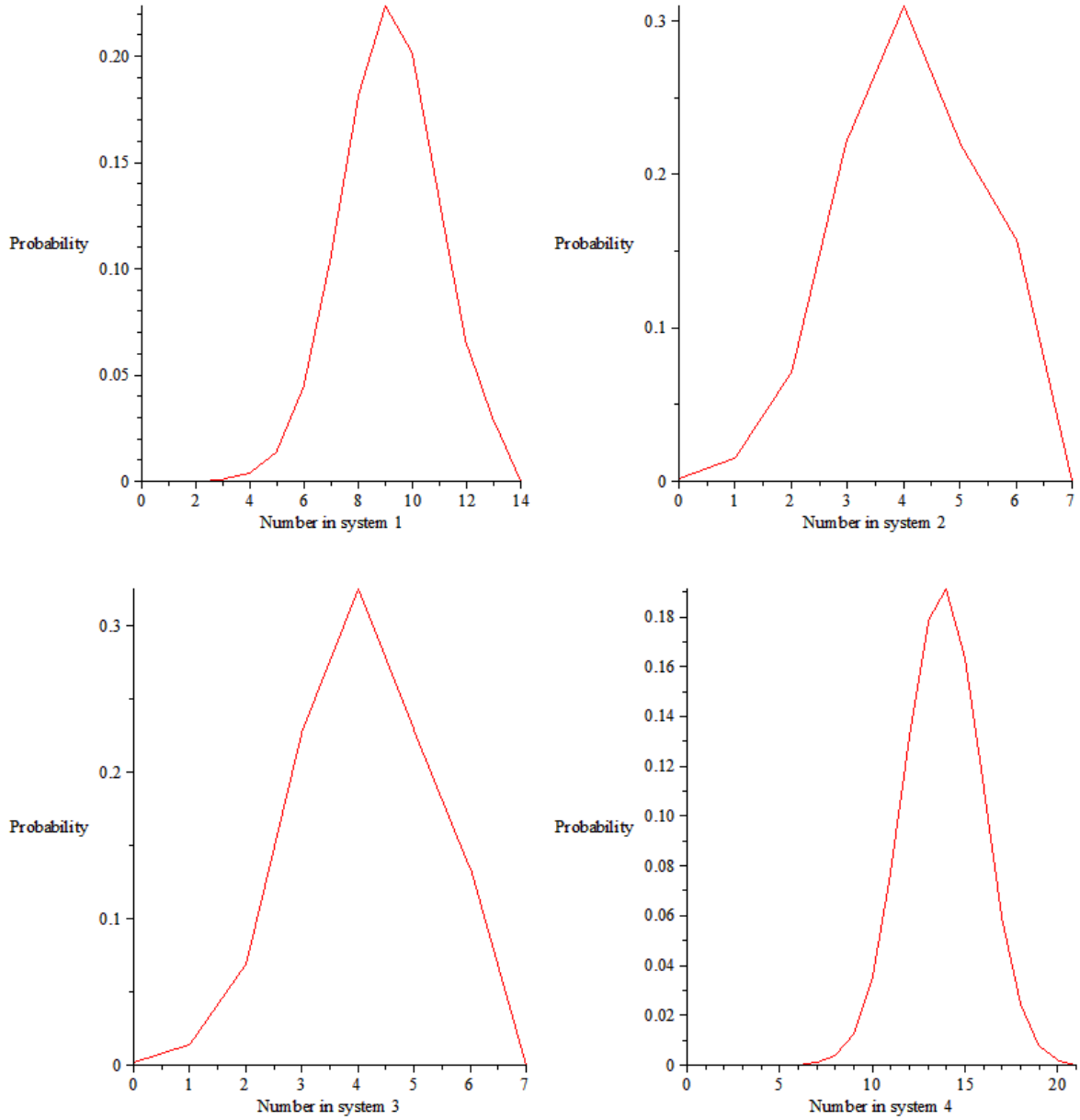


Figure 9.12 Probability density graphs for number in system for each patient group

Table 9.7 Empirical (target) and model performance measures for each patient group

Pat group:	1		2		3		4	
	Target	Model	Target	Model	Target	Model	Target	Model
$\tilde{\lambda}_i$	0.1801	0.1800	0.04795	0.04747	0.04427	0.4431	0.09132	0.09109
$P_{R,i}$	–	0	–	0	–	0	–	0
$P_{W,i}$	–	0.9806	–	0.9128	–	0.9163	–	0.9949
π_i	–	5.974	–	2.894	–	2.899	–	8.993
$\tilde{P}r_i$	0.613	0.6141	0.613	0.626	0.613	0.628	0.613	0.6114
$\tilde{P}b_i$	–	0.05375	–	0.1816	–	0.1478	–	0.01524
Tp	–	25.4	–	6.486	–	6.004	–	12.892
$E[L_i]$	–	9.15	–	4.121	–	4.088	–	13.8
$E[T_i]$	–	50.83	–	86.81	–	92.96	–	151.5
$E[L_{q,i}]$	–	3.175	–	1.227	–	1.189	–	13.8
$E[W_i]$	–	17.64	–	25.84	–	26.82	–	52.77
$CV[L_{q,i}]$	–	0.535	–	0.8598	–	0.860	–	0.4254

Table 9.8 Empirical (target) and model performance measures for the holistic system

	Target	Model
$\tilde{\lambda}_{sys}$	0.36364	0.3629
π_{sys}	–	20.76
$\tilde{P}r_{sys}$	0.613	0.617
Tp	–	50.79
$E[L_{sys}]$	–	31.16
$E[T_{sys}]$	–	85.89
$E[L_{q,sys}]$	–	10.4
$E[W_{sys}]$	–	28.66

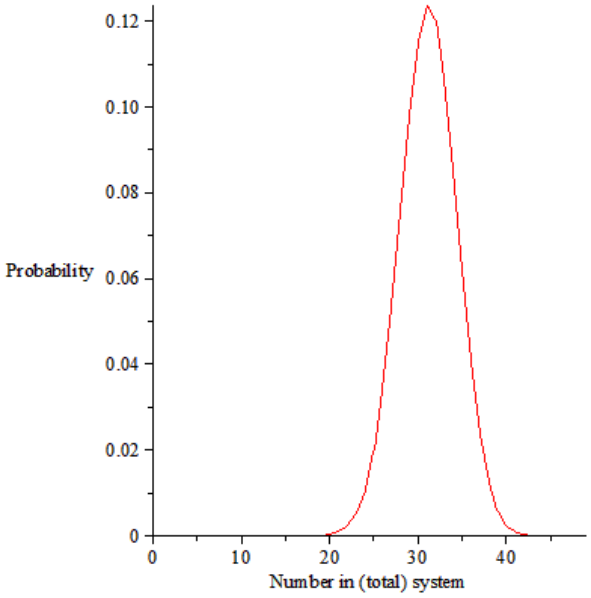


Figure 9.13 Probability density graph for number in system

As can be seen from the results, this is what would appear to be a reasonable[18] representation of reality. It will be noticed, however, that the model $\tilde{P}r_i$ are fairly larger than their target values for groups two and three whilst the model and target $\tilde{\lambda}_i$ are approximately equal. This

[18] A more accurate fit could no doubt be obtained given more trials

is because the rounding of number of beds (9.3.1.2) was more significant for these groups as the unrounded r_i were small. Table 9.9 provides information on the dimension of the transition matrix, number of digits (Ch 7.3.2.2.9) used in Maple calculations and the time taken (seconds) to solve each queuing systems.

Table 9.9 Transition matrix dimension and solution time of each queuing system

Pat grp:	Dimension	Digits	Time taken
1	308	4	136
2	115	4	10
3	115	4	10
4	880	4	5164

The annual running cost is estimated at $20.76 \times £480 \times 365 = £3.637\text{m}$.

9.3.3 Phase Two: Hypothetical scenarios

A greater number of what-if questions can be asked of this model than that of Ch 9.2 since there are more parameters owing to a higher design complexity. Two examples are given.

9.3.3.1 Hypothetical Scenario Three: Blocked LOS

What if the blocked component of LOS is reduced by 50% (for all patient groups)?

Table 9.10 Performance measures and costs for 50% reduction in blocked length of stay

$\tilde{\lambda}$	π	$\tilde{P}r$	Tp	$E[W]$	$E[L_q]$	Cost pa
0.3664	20.65	0.576	56.68	25.71	9.42	£3.618m

On comparison with the system results of Ch 9.3.2.4 it can be seen that expected waiting time (before a referral is admitted or reneges) has been reduced by about three days, the length of the queue is, on average, one referral fewer, and prospective patients are 7% less likely to renege. In addition, throughput has been increased by (almost) six patients per year. With the cost of an admission well into the tens of thousands of pounds (Ch 1.2) this represents significant financial gain. By spending less time in the queue it is also psychologically better for the patient as there is less waiting about. There are physical benefits too. If a patient is not in a facility where physiotherapy, for example, is administered then their condition deteriorates. The longer they are in such a state the more time is needed at Rookwood NRC to restore motor function. Perhaps most importantly, however, is the knock-on effect upstream. If a patient is ready for discharge at Rookwood NRC but for some reason this is delayed they

are said to be bed-blocking - the negative effects of which are documented in Ch 6.2.2. This is not just a problem for Rookwood NRC but for the facilities that discharge patients to the unit. Reducing the waiting time for the NRC reduces the problem of bed-blocking upstream at these facilities.

But how can bed-blocking be reduced at Rookwood NRC in order to bring about the halving of blocked LOS? The reasons for bed-blocking are laid out in a table towards the end of Ch 1.2. Clearly, the majority of delays to discharge are caused by either waiting for an assessment of future care needs or waiting for a suitable discharge destination. If an estimated date of readiness could be provided to social services and the (expected) discharge destination then preparations could be made well in advance of the date the patient is actually ready for discharge. According to the model a patient can expect to spend

$$\hat{LOS}^{active} = \frac{\sum_{i=1}^{4} \tilde{\lambda}_i \left(1 - \tilde{Pr}_i\right) \hat{LOS}_i^{active}}{\sum_{i=1}^{4} \tilde{\lambda}_i \left(1 - \tilde{Pr}_i\right)} = 113 \qquad (9.3.3.1.1)$$

days in Rookwood NRC before they are ready for discharge. However, this single figure does not account for the heterogeneity of patients. Treeworks has been used to create similar trees for active LOS as have been created for total LOS in Appendices 9.1 and 9.2. These give therapists increased predictability of active LOS based on the values of certain patient attributes (age, gender, diagnosis etc).

But what if blocked LOS is reduced to zero?

Table 9.11 Performance measures and costs for 100% reduction in blocked length of stay

$\tilde{\lambda}$	π	$\tilde{Pr}$	Tp	$E[W]$	$E[L_q]$	Cost pa
0.3674	20.4	0.452	59.54	22.33	8.21	£3.574m

It will be noticed that these interesting and very useful scenarios could not have been considered if LOS had not been partitioned into its active and blocked components.

9.3.3.2 Hypothetical Scenario Four: Arrival rates

The second example of an appropriate hypothetical scenario is to increase the arrival rate for older patients. This is studied due to the premise of an ageing population in the UK. But in the initial construction of the model (Ch 9.3.1, Ch 9.3.2) age is not a branching variable for

any of the four patient groups. In order to vary the arrival rate for patients of an older age it is necessary for age to be a branching variable for the patient groups. Therefore, stage one is recreated with this in mind;

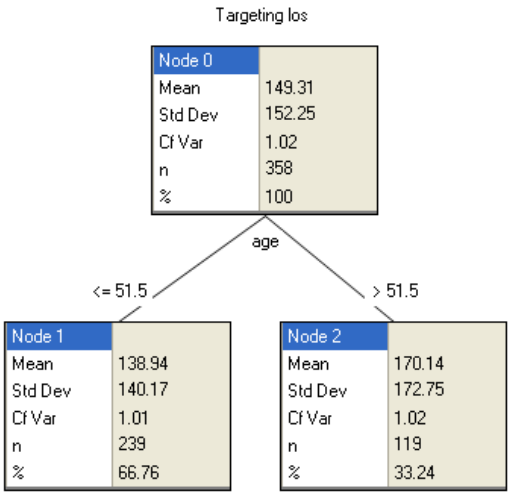


Figure 9.14 Final tree with two patient groups

Only one split is made to keep things simple. The largest reduction in variance (just 1%) is attained when age on admission below and above 51.5 years is considered. The scale of the reduction in variance is on-the-whole unimportant here since the objective of the split is to obtain two patient groups whose arrival rates can be independently varied. Using (9.3.1.1), (9.3.2.1.1) and (a variation of) (9.3.2.2.1) yields $r_1 = 13, r_2 = 8, \tilde{\lambda}_1 = 0.242, \tilde{\lambda}_2 = 0.1216$ and $b_1 = 18, b_2 = 11$. Employing an approach similar to that of Ch 9.3.2.3 returns

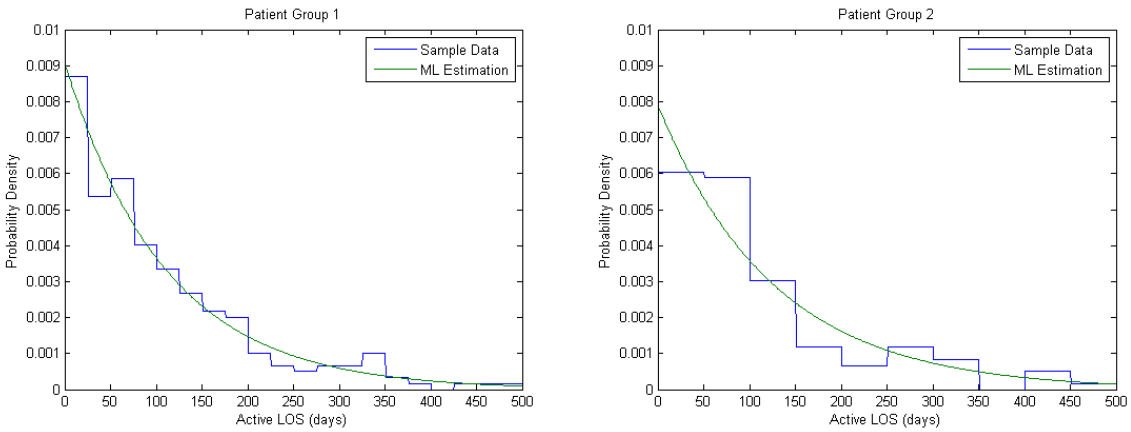


Figure 9.15 Exponential approximations to active length of stay for each patient group

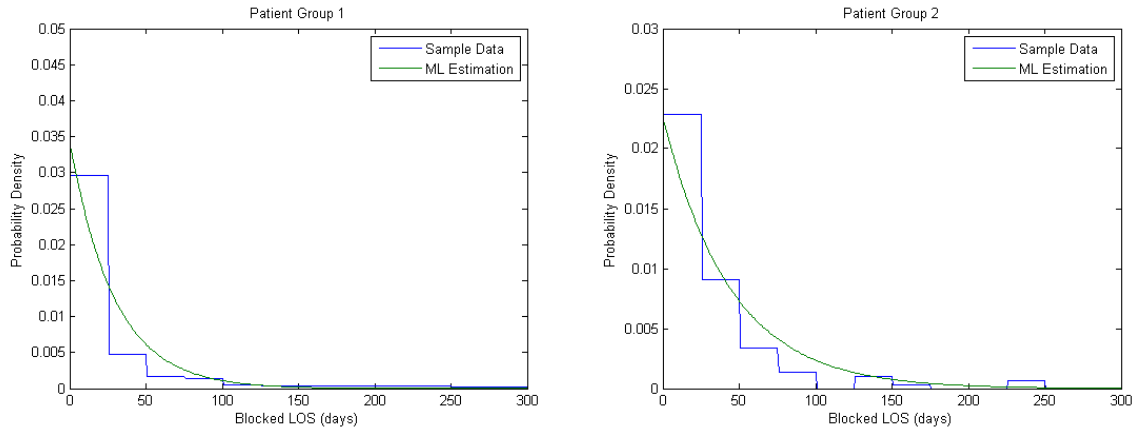


Figure 9.16 Exponential approximations to blocked length of stay for each patient group

Here

$$\begin{aligned}&\text{P.G. 1 (1 phase active): } \mu_{1,1} = 0.0091, \mu_{1,2} = 0.0345\\ &\text{P.G. 2 (1 phase active): } \mu_{2,1} = 0.0079, \mu_{2,2} = 0.0227\end{aligned} \qquad (9.3.3.2.1)$$

Note the poor fit of the exponential distribution to active LOS of patient group two – this could be averted by using a higher-order approximating distribution. However, this would require the solution of a transition matrix of least dimension 2,450.

The queuing system $M, M \mid M+M, M+M \mid 13,8 \mid 31,19$ is now constructed and a reasonable fit has been obtained.

Table 9.12 Empirical (target) and model performance measures for the holistic system

	Target	Model
$\tilde{\lambda}$	0.36364	0.36367
π	–	20.98
$\tilde{P}r$	0.613	0.614
Tp	–	51.23
$E[W]$	–	28.52
$E[L_q]$	–	10.37

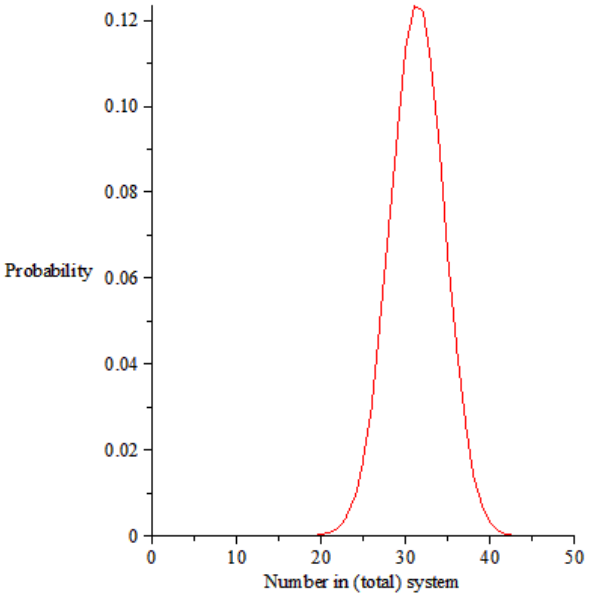


Figure 9.17 Probability density graph for number in system

The control values are

Pat group:	λ	δ	γ	θ
1	0.2465	6	0.87	1.3
2	0.1255	6	0.87	1.3

Now that phase one is complete, appropriate what-if questions may be asked of the model. Such as what if the (effective) arrival rate of older (age on admission > 51.5 years) patients (patient group two) is increased by one third?

Table 9.13 Performance measures and costs for a one-third increase in the arrival rate of older patients

$\tilde{\lambda}$	π	$\tilde{P}r$	Tp	$E[W]$	$E[L_q]$	Cost pa
0.4045	20.99	0.653	51.27	27.44	11.1	£3.677m

Therefore the (overall) effective arrival rate target is $\tilde{\lambda}_1 + 4\tilde{\lambda}_2/3 = 0.4047$. The probability of an older patient reneging has increased to 0.711 bringing the system average to 0.653. Since the system was operating at (almost) full capacity beforehand there is negligible difference in throughput. It may also be seen that the expected length of the queue has increased by about 7%. What may appear confusing, however, is the drop in expected waiting time. This is because this does not represent the (average) time spent in queue before a referral is admitted but the (average) time spent in queue before a referral is removed whatever the outcome – admitted or reneged. Since the arrival rate for patient group two has increased by one third this has led to steady-state probabilities being denser for a greater number in system where the probability of reneging is at its highest. Thus, when an arrival occurs it is much more likely to renege than before, and so the referral spends less time in the queue. Since 71.1% of older patients renege these low durations have a profound impact on the value of $E[W]$. A formula for the expected waiting time for those who are ultimately admitted has not been deduced. However, this value will not have changed (from before the increase in arrival rate) since the system was operating at (near) full capacity anyway.

9.4 Model Three

Until now the effect of treatment intensity on a patient's LOS has not been a consideration. In this, the final model, the relationship between the intensity of treatment and the active

component of LOS is realised by including stages two and three. This model is the most complex under consideration within this project and is analogous to Route 4 of Figure 6.8.

Since data is not available on treatment intensity (and readiness date) from departments other than physiotherapy (see Ch 6.4.2.3) it is the *physiotherapy treatment intensity* and *physiotherapy active LOS* that are used to develop the relationship outlined in Ch 6.2.4. Even if the other departments had such data it would be little use without a department specific automated scheduling program (Ch 3.3.3) that can be used to evaluate the effect of hypothetical scenarios on treatment intensity (Ch 3.3.2).

9.4.1 Phase One: Construction

The classification and size of each patient group is identical to that of Model Two. That is, stage one is unchanged. So too is stage four since the average active LOS for each of the four patient groups must remain the same. This is because the objective at phase one is to obtain a representation of reality – and so the typical treatment intensities (deduced by the scheduling program in stage two[19]) give rise to the typical active LOSs (stage three) that are, of course, equivalent in value to the means of the approximating distributions, i.e. (9.3.2.3.15). The question is how to obtain the treatment intensities in stage two and how are these converted to active LOS in stage three?

9.4.1.1 Task One: Stage Three: Treatment intensity and average active LOS

For reasons that will become clear stage three is considered before stage two. Here the relationship between intensity of treatment and average active LOS are determined for each patient group. A scatter plot of observations is made of treatment intensity (x-axis) against average active LOS (y-axis) for the 46, 13, 7, and 23 patient episodes of the patient groups that contain the relevant information[20]. A curve is then fitted to each of these plots by least squares regression (Ch 5.2.1). Whilst many studies reviewed in Ch 2.2.3 suggest a relationship between treatment intensity and LOS, not one of these provide either an empirical finding or hypothesis on the actual dependence of LOS on treatment intensity for a range of values. Therefore, there is a fair amount of flexibility in determining the equations of the curves. There are three conditions that must however be met for each patient group. First, average active LOS is inversely proportional to treatment intensity. Second, there are diminishing returns, i.e. the positive effect of treatment intensity on average active LOS

[19] See Ch 9.1 for an overview of the four stages
[20] Date ready for discharge and treatment intensity

diminishes beyond a certain point. And third, the average active LOS, given by (9.3.2.3.15) must correspond to the average level of treatment intensity, defined shortly. A 'one over x' type curve can satisfy these conditions and is applied to the scatter plots.

It is stated in Ch 5.2.1 that the objective is to minimise (5.2.1.1), i.e. $\sum_{\forall j} r_j^2$, with, in this case

$$r_j = y_j - \beta_1 - \frac{\beta_2}{x_j} \tag{9.4.1.1.1}$$

since

$$f\left(x_j;\beta\right) = y_j = \beta_1 + \frac{\beta_2}{x_j} \tag{9.4.1.1.2}$$

where the x_j and the y_j relate to the two values of the *j*-th observation. It is possible to linearise the function but owing to the drawbacks alluded to in Ch 5.2.1 a different approach is favoured. Since $y \Leftrightarrow \widehat{LOS}^{active}$ and $x \Leftrightarrow \widehat{INT}$ the y_j and x_j represent the active LOS and treatment intensity of patient episode j of class i where $j = 1, 2, ..., \hat{n}_i$. These values are inserted to an MS Excel worksheet and solver is used to minimise $\sum_{j=1}^{\hat{n}_i} r_j^2$ given two *sensible* starting values of β_1 and β_2. The following constraint is added to satisfy the third condition (mentioned above):

$$\widehat{LOS}_i^{active} = \beta_1 + \frac{\beta_2}{\widehat{INT}_i} \tag{9.4.1.1.3}$$

where $\widehat{INT}_i$ is the average treatment intensity (hours/day) of patient group *i*. The $\widehat{LOS}_i^{active}$ are given by (9.3.2.3.15). The most simple formulation of the $\widehat{INT}_i$ is

$$\widehat{INT}_i = \frac{\sum_{k=1}^{\hat{n}_i} INT_{i,k}}{\hat{n}_i} \tag{9.4.1.1.4}$$

where $INT_{i,k}$ represents the treatment intensity of the *k*-th patient episode with relevant information of patient group *i*. However, a much more accurate formula is achieved by weighting each value of treatment intensity by their active LOS;

$$\widehat{INT}_i = \frac{\sum_{k=1}^{\hat{n}_i} LOS_{i,k}^{active} \cdot INT_{i,k}}{\sum_{k=1}^{\hat{n}_i} LOS_{i,k}^{active}} \qquad (9.4.1.1.5)$$

This formula yields

$$\widehat{INT}_1 = 0.44881, \widehat{INT}_2 = 0.40178, \widehat{INT}_3 = 0.60486, \widehat{INT}_4 = 0.52381 \qquad (9.4.1.1.6)$$

This constraint is added since it is known that, typically, the active LOSs are equal to (9.3.2.3.15) and that treatment intensities are equal to (9.4.1.1.6).

Consider first patient group two. The least squares fit is depicted in Figure 9.18.

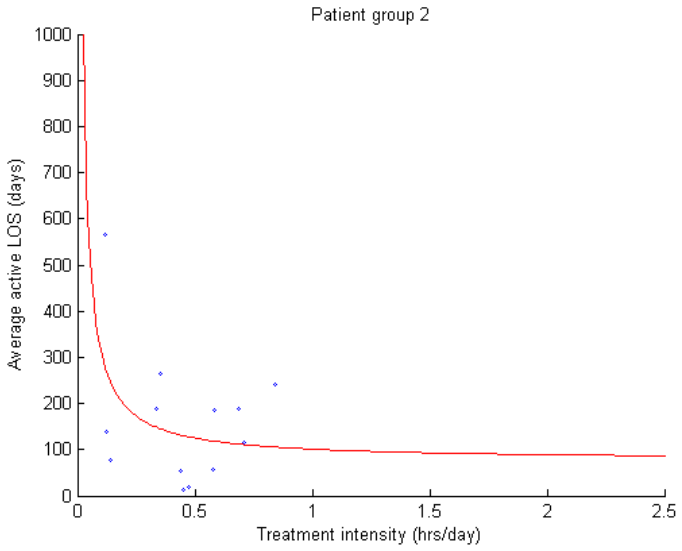


Figure 9.18 Relationship between treatment intensity and active length of stay for patient group two

with $\beta_1 = 77.3$ and $\beta_2 = 23.81$, i.e. the lowest average active LOS possible is 77.3 days. For patient group one the value of the shape parameter β_2 is found to be negative. Clearly this is not appropriate. A non-negativity constraint was added but this yielded $\beta_2 = 0$; a horizontal line. Finally, the following fit was obtained:

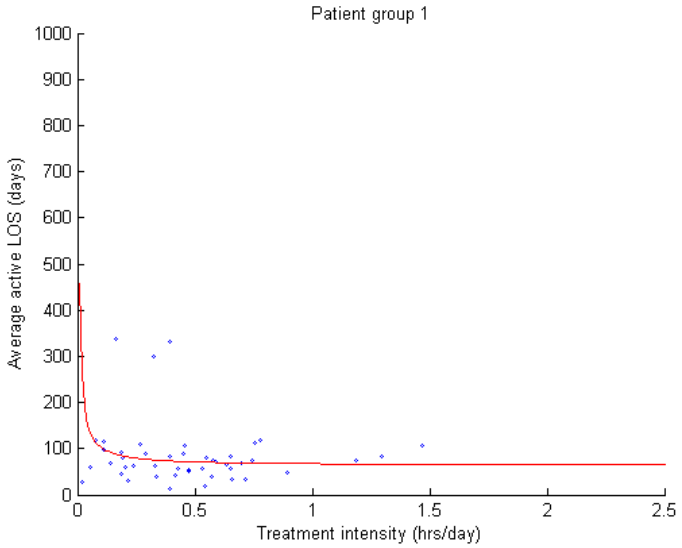


Figure 9.19 Relationship between treatment intensity and active length of stay for patient group one

with $\beta_1 = 63.43$ and $\beta_2 = 3.95$. This was as a result of a modification to (9.4.1.1.1) designed to pull the curve closer to the three observations with $LOS^{active} \approx 300$ by setting $r_j = \left(y_j - \beta_1 - \frac{\beta_2}{x_j} \right)^3$. Observe that average active LOS appears to be far less responsive to changes in treatment intensity for patients of this group than those of patient group two. This is shown by the flatness of the curve borne from a smaller shape parameter. For patient group three with $\beta_1 = 84.22$ and $\beta_2 = 22.17$

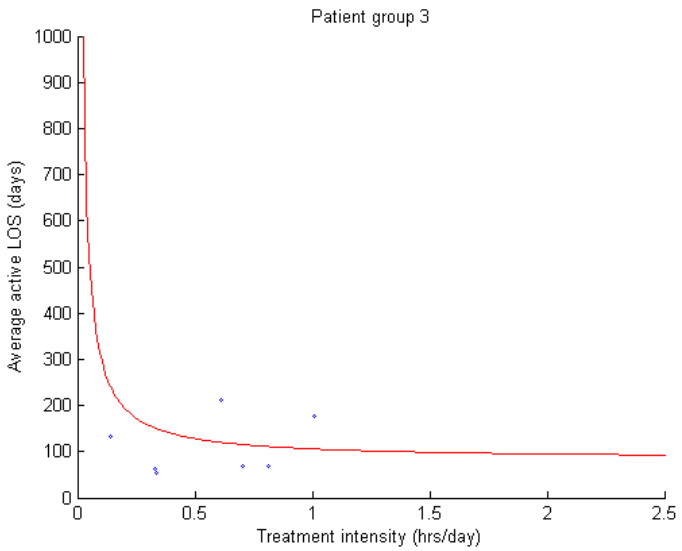


Figure 9.20 Relationship between treatment intensity and active length of stay for patient group three

Further modifications to r_j were required to yield a positive shape parameter, $\beta_2 = 24.85$, for patient group four which with $\beta_1 = 130.09$ produced

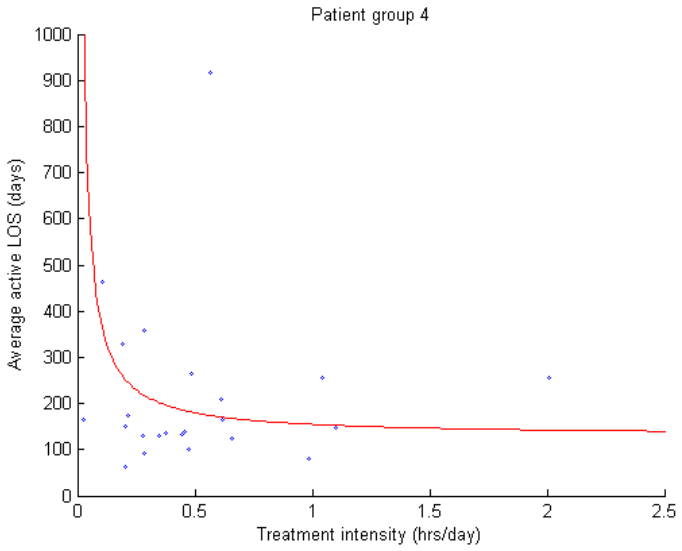


Figure 9.21 Relationship between treatment intensity and active length of stay for patient group four

It will be observed that the 'one over x' type curve provides a not entirely convincing fit to the data. That being said however, it is hard to see what other type of curve would better fit the little data that is available for (some more than others of) the patient groups.

9.4.1.2 Task Two: Stage Two: Determine treatment intensities

The typical levels of treatment intensity for each patient group are known from the previous subchapter. The task now is to initialise the scheduling program with appropriate input values such that the outputs (i.e. patient group treatment intensity) correspond to (9.4.1.1.6).

Firstly it is necessary to detail a modification to the scheduling program that enables it to deal with patient groups as opposed to individual patients. For the weekly scheduling of physiotherapy treatment to individual patients the original version (Chapter 4) is of course used. However, patients are grouped in the model and so (to avoid the repetition and complication of providing individual levels of demand, details and priorities) typical values for these are required only for each group in addition to the number of beds assigned to that group. Therefore, the 'Input patient details' section of the 'Create New' menu is amended.

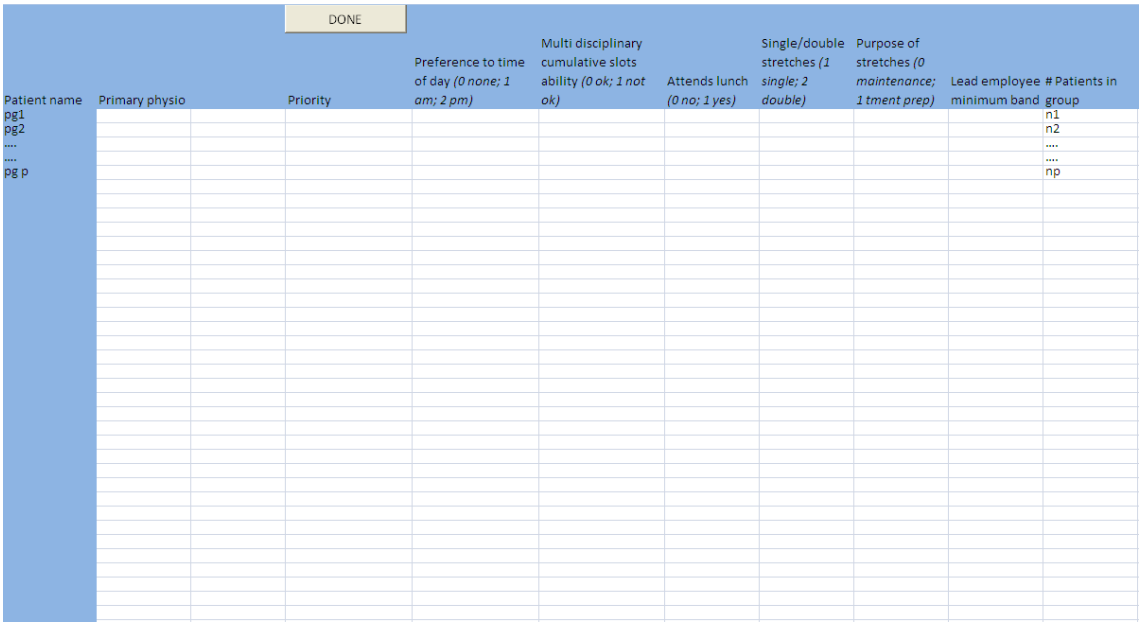


Figure 9.22 'Input patient details' screen of amended scheduling program

The 'Input patient demand' section of the 'Create New' menu is also amended.

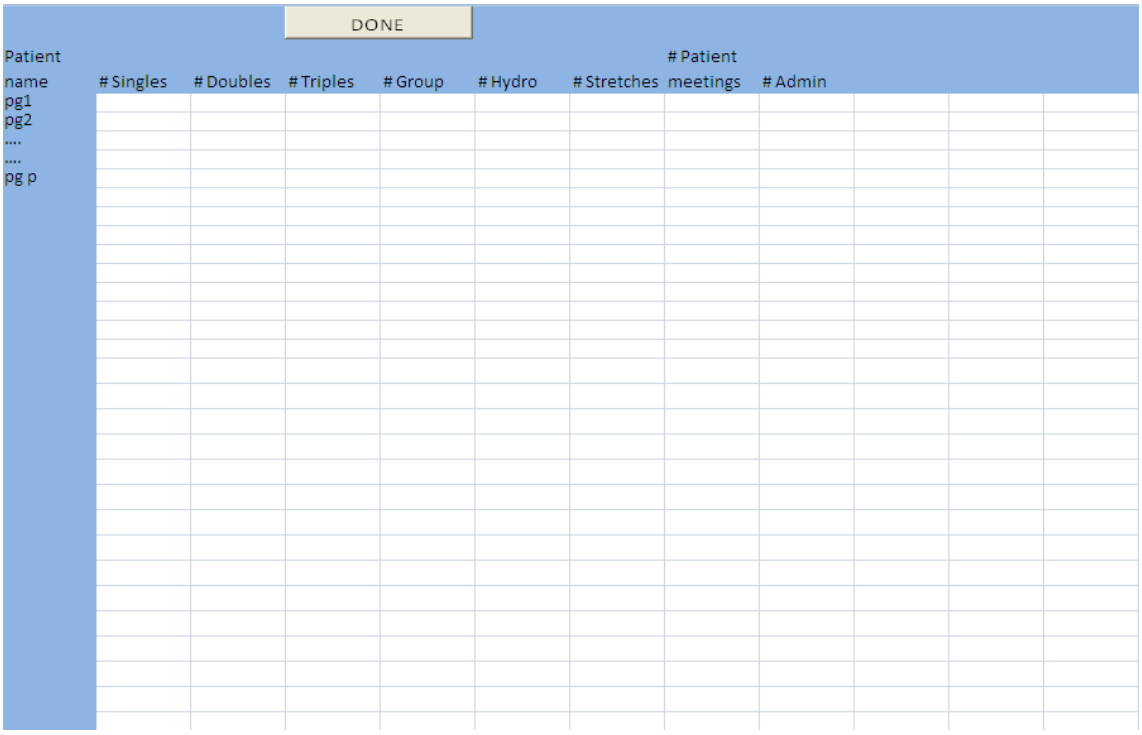


Figure 9.23 'Input patient demand' screen of amended scheduling program

The major difference to the output is in the ‘View other details’ section of the ‘View New’ menu. Here, as well as having individual details for each patient of each group, there are averages (as well as variances) for each patient group.

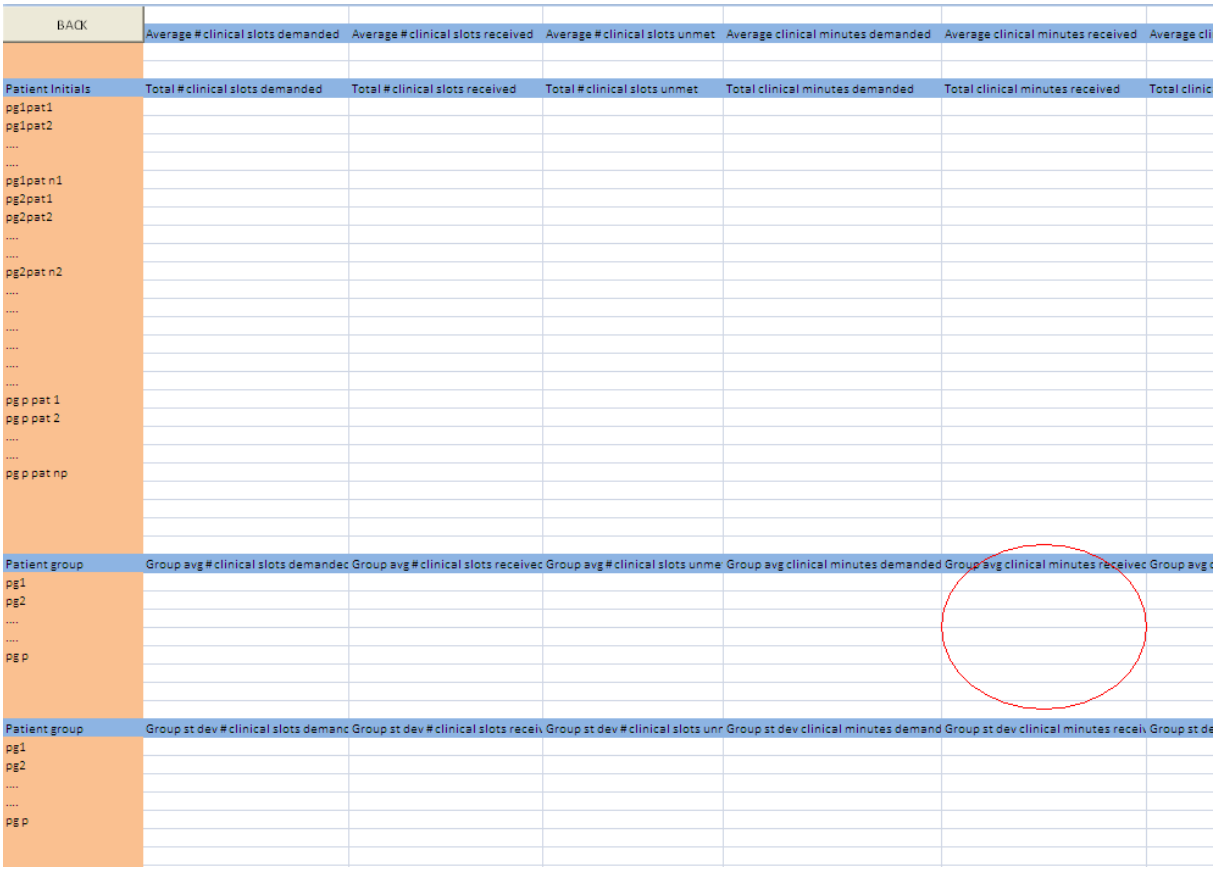


Figure 9.24 'View other details' screen of amended scheduling program

Most important is the (highlighted) 'Group avg clinical minutes received'. Dividing these (weekly) figures by seven (number of days in a week) and then by sixty (number of minutes in an hour) gives the group average treatment intensity in hours per day. Note that, due to simplicity, there is no PTO to detail patient unavailability and so it is assumed that, upon commencement of the scheduling, all patients of each group are available at any time.

Table 9.14 and Table 9.15 display typical values for patient details and demand that have been deduced through discussions with clinicians and by examining historic records.

Table 9.14 Typical patient details

Patient group:	Primary/ secondary band level	Priority	Preference to time of day	Multi-disciplinary cumulative slots	Attends lunch	Single/ double stretch	Purpose of stretch	Lead employee minimum band
1	5, 6, 7, 8	3	–	–	1	S	Tment prep	5
2	5, 6, 7, 8	2	–	–	1	S	Tment prep	5
3	5, 6, 7, 8	4	–	1	–	D	Tment prep	5
4	5, 6, 7, 8	3	–	–	–	D	Tment prep	5

Table 9.15 Typical patient demand

Patient group:	Single	Double	Triple	Group	Hydro	Stretch	Patient meeting	Patient admin
1	1	1		2				
2	1	1		2				
3		4				3		1
4		4				3		1

Patient groups three and four relate to patients with a higher severity of injury since these patients are too ill to return to usual residence or temporary residence (Ch 9.3.1). Such patients typically require two therapists for a treatment session (i.e. double and double stretch). Now that the inputs relating to the patients have been determined it is necessary to look at inputs relating to the therapists. Again through discussion and examining historic records the staff attributes and working hours have been deduced.

Table 9.16 Typical staff attributes and working hours

Role	Band level	Lead	Group	Stretches 0-cannot 1-assist 2-lead	Patient meetings	Goal planning meetings	M	T	W	T	F
Physio	9	Y	Y	2	Y	Y	Y	N	N	N	Y
	8	Y	Y	2	Y	Y	Y	Y	N	N	am
	7	Y	Y	2	Y	Y	Y	Y	Y	Y	Y
	6	Y	Y	2	Y	Y	Y	Y	N	Y	N
	6	Y	Y	2	Y	Y	N	Y	Y	N	Y
	6	Y	Y	2	Y	Y	Y	Y	Y	Y	Y
	6	Y	Y	2	Y	Y	Y	Y	Y	Y	Y
	6	Y	Y	2	Y	Y	Y	N	Y	N	pm
	5	Y	Y	2	Y	Y	Y	Y	Y	Y	Y
Assistant	3	N	Y	2	N	N	N	8 – 2.30	8 – 2.30	8 – 2.30	Y
	2	N	Y	2	N	N	Y	Y	Y	Y	Y
Student	3	N	Y	1	N	N	0.5	0.5	0.5	0.5	0.5

The next consideration is staff availability. Whilst typical values could be deduced from historic records this is a time consuming task and inherent inaccuracies could also lead to

unsuitable results for treatment intensities. Staff availability is therefore used as a free variable that is manipulated to ensure that, with the above patient and staff details, appropriate values of treatment intensity are output.

After varying therapist availabilities (i.e. adding meetings, annual leave etc on top of standard working hours) an appropriate configuration was found. Table 9.17 displays values for group average treatment intensity over five runs (time allowance: two hours).

Table 9.17 Group average treatment intensity for each patient group

Run:	1	2	3	4	5	Avg
Patient Group 1	203	175	178	180	173	182
Patient Group 2	145	160	160	165	190	164
Patient Group 3	250	275	250	265	275	263
Patient Group 4	215	210	222	227	222	219

Dividing the averages by seven and then by sixty returns treatment intensity in hours per day;

$$\widehat{INT}_1 = 0.43333, \widehat{INT}_2 = 0.39048, \widehat{INT}_3 = 0.62619, \widehat{INT}_4 = 0.52143 \qquad (9.4.1.2.1)$$

These are all within 5% of the empirical values given in (9.4.1.1.6). Despite this encouraging congruence the variances are rather large for some patient groups, particularly patient group two (range: 145 - 190 minutes per week). On the other hand patient group four has a range of only twelve minutes across the five runs. The reason for the variance is, of course, the random numbers that are used in the scheduling program (see Ch 4.5.1).

Note that here it has been assumed that each of the 21 patients is in their *rehabilitative* phase of treatment, i.e. they are not yet ready for discharge. In reality, some of these patients would be in the *maintenance* phase and would require less treatment.

9.4.2 Phase Two: Hypothetical scenarios

What-if questions relating to treatment intensity are now studied. The automated scheduling program has been constructed specifically in order to efficiently evaluate these scenarios (Ch 3.3.2). Without it, a schedule would have to be produced by hand (8 hours) for each scenario in question. Two examples are considered.

9.4.2.1 Hypothetical Scenario Five: Additional physiotherapists

The employment of additional physiotherapists is firstly considered. For this example, two full-time band sixes and one full-time band three join the roster of Ch 9.4.1.2.

Table 9.18 Group average treatment intensity for each patient group

Run:	1	2	3	Avg
Patient Group 1	248	253	255	252
Patient Group 2	205	211	205	207
Patient Group 3	345	325	320	330
Patient Group 4	297	298	305	300

Thus

$$\widehat{INT}_1 = 0.6, \widehat{INT}_2 = 0.49286, \widehat{INT}_3 = 0.78571, \widehat{INT}_4 = 0.71429 \qquad (9.4.2.1.1)$$

Using (9.4.1.1.3) alongside the β_1 and β_2 obtained in Ch 9.4.1.1 for each patient group returns

$$\widehat{LOS}_1^{active} = 70.01, \widehat{LOS}_2^{active} = 125.61, \widehat{LOS}_3^{active} = 112.44, \widehat{LOS}_4^{active} = 164.88 \qquad (9.4.2.1.2)$$

So active LOS has been reduced by approximately 2, 11, 8 and 13 days respectively. Clearly, increasing treatment intensity for patient group one has had little impact on active LOS. This is due to the flatness of the approximating curve borne by having a low-valued shape parameter ($\beta_2 = 3.95$). Perhaps additional employee time can be better spent on patients of other groups.

The active LOS distributions of Ch 9.3.2.3 are now amended such that their means are congruent to (9.4.2.1.2). This is done by making proportionate adjustments to the (active) service rates whilst holding constant the transition probabilities. So for a two phase distribution if service rate one is increased by 10% then service rate two must also be increased by 10%. This ensures that the underlying shape of the approximating distribution remains the same but the variable (patient group active LOS) values are appropriately scaled (see Ch 6.2.4). It is done by solving the following equation[21] for $\hat{p}_i$, the scalar by which each service phase rate is multiplied by:

[21] Based on (8.3.1.5.9)

$$\hat{LOS}_i^{active} = \frac{1}{\hat{p}_i \cdot \mu_{i,1}} + \sum_{l=2}^{k_i-1} \frac{1}{\hat{p}_i \cdot \mu_{i,l}} \sum_{j=1}^{l-1} \left(1-\alpha_j\right) \qquad (9.4.2.1.3)$$

where k_i is the number of active service phases in patient group *i*. The new $\mu_{i,j}$ $(i=1,...,4;\ j=1,...,k_i)$ are deduced by multiplying each of the old $\mu_{i,j}$ by the respective $\hat{p}_i$. Solving (9.4.2.1.3) gives $\hat{p}_1 = 1.0316, \hat{p}_2 = 1.0767, \hat{p}_3 = 1.0871, \hat{p}_4 = 1.0749$. Thus (9.3.2.3.13) can be rewritten

P.G. 1 (2 phase): $\mu_{1,1} = 0.0757, \mu_{1,2} = 0.0162$
P.G. 2 (3 phase): $\mu_{2,1} = 0.1333, \mu_{2,2} = 0.1333, \mu_{2,3} = 0.0073$
P.G. 3 (3 phase): $\mu_{3,1} = 0.0105, \mu_{3,2} = 0.0357, \mu_{3,3} = 0.0704$
P.G. 4 (2 phase): $\mu_{4,1} = 0.0184, \mu_{4.2} = 0.0090$ (9.4.2.1.4)

Replacing these values in the queuing system of Ch 9.3.2.4 and solving yields

Table 9.19 Performance measures for additional therapists

$\tilde{\lambda}$	π	$\tilde{P}r$	Tp	$E[W]$	$E[L_q]$
0.3651	20.73	0.604	52.78	27.62	10.08

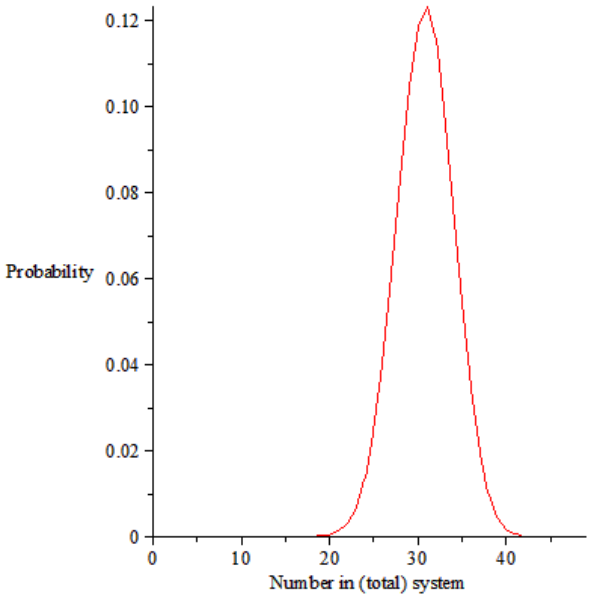


Figure 9.25 Probability density graph for number in system

To summarise: throughput has increased by two patients per year, the probability of a prospective patient reneging has fallen by just over 2%, the expected queue length has decreased; and so too has waiting time by one day. But in order to achieve this an additional one band three and two band six therapists are required at an annual cost of £17k and 2 x £30k respectively.

As mentioned in the outset of this subchapter, it is supposed that either assumption in the final paragraph of Ch 6.4.2.3 is valid. The first assumption is that other departments must replicate any change in treatment intensity in order to have the desired effect on active LOS. Here, it is assumed this is done by replicating any changes in personnel. The physiotherapy department has 6.4 qualified therapist[22] FTEs and two non-qualified assistant FTEs. The two other departments with more than one FTE are Occupational Therapy (3.5 qualified FTEs and 1.5 assistant FTEs) and Speech and Language Therapy (1.5 qualified FTE and 1 assistant FTEs). The additional employments in physiotherapy represent an increase in qualified and assistant FTEs by approximately one third each. Replicating this change in the other (two) departments means increasing the number of qualified therapists by 1.2 and 0.5 FTEs and assistants by 0.5 and 0.3 FTEs. This means that the annual operating cost is an additional $(2+1.2+0.5)\times £30,000+(1+1+0.3)\times £17,000 = £150\text{k}$ above the standard cost of $20.73\times £480\times 365 = £3.632\text{m}$. Thus total annual cost is £3.782m.

The second assumption is that the readiness of the patient with respect to physiotherapy only dictates the (overall) date ready for discharge. In this case it is not necessary to alter the personnel of the other departments and so the total annual cost is £3.709m.

9.4.2.2 Hypothetical Scenario Six: Group sessions

In what is the final hypothetical scenario under consideration the effect of increasing the number of group sessions is studied. The typical levels of patient demand (Table 9.15) are replaced by those of Table 9.20.

Table 9.20 Typical levels of demand after session reconfiguration

Patient group:	Single	Double	Triple	Group	Hydro	Stretch	Patient meeting	Patient admin
1	1			5				
2	1			5				
3		3		3		3		1
4		3		3		3		1

That is, one double session has been removed and three group sessions have been added for each patient group. The composition of staffing availabilities has not changed but more group

[22] Band level of at least five

session availability times (see Ch 3.4.2) have been added in order to cope with the additional demand.

Table 9.21 Group average treatment intensity for each patient group

Run:	1	2	3	Avg
Patient Group 1	318	313	325	319
Patient Group 2	320	310	300	310
Patient Group 3	380	375	370	375
Patient Group 4	350	335	343	343

Note how the average number of minutes received by each group is much greater than the amount typically provided (Ch 9.4.1.2) or the amount provided when the number of FTEs is increased by one third (Ch 9.4.2.1). This is because a group session is provided by only one employee to about three patients (Ch 3.4.2). It may also be noted that the averages for the first two groups are similar despite patient group one being a higher priority than group two. This is because, over the three runs, all group sessions were allocated and the level of demand for these groups was equivalent. There did, however, still exist some unmet needs with respect to the double sessions of groups three and four. In hours per day

$$\widehat{INT}_1 = 0.75952, \widehat{INT}_2 = 0.7381, \widehat{INT}_3 = 0.89286, \widehat{INT}_4 = 0.81667 \qquad (9.4.2.2.1)$$

Using (9.4.1.1.3) alongside the β_1 and β_2 obtained in Ch 9.4.1.1 for each patient group returns

$$\widehat{LOS}_1^{active} = 68.63, \widehat{LOS}_2^{active} = 109.56, \widehat{LOS}_3^{active} = 109.05, \widehat{LOS}_4^{active} = 160.52 \qquad (9.4.2.2.2)$$

So active LOS has been reduced by approximately 4, 27, 12 and 17 days respectively. Again, increasing treatment intensity for patient group one has had little impact on active LOS.

As before (Ch 9.4.2.1), the active LOS distributions of Ch 9.3.2.3 are now amended such that their means are (approximately) congruent to (9.4.2.2.2). Solving (9.4.2.1.3) gives $\hat{p}_1 = 1.0523, \hat{p}_2 = 1.106, \hat{p}_3 = 1.2464, \hat{p}_4 = 1.1084$. Thus (9.3.2.3.13) can be rewritten

$$\begin{aligned}
&\text{P.G. 1 (2 phase): } \mu_{1,1} = 0.0772, \mu_{1,2} = 0.0165 \\
&\text{P.G. 2 (3 phase): } \mu_{2,1} = 0.1543, \mu_{2,2} = 0.1543, \mu_{2,3} = 0.0085 \\
&\text{P.G. 3 (3 phase): } \mu_{3,1} = 0.0108, \mu_{3,2} = 0.0364, \mu_{3,3} = 0.078 \\
&\text{P.G. 4 (2 phase): } \mu_{4,1} = 0.0189, \mu_{4.2} = 0.0093
\end{aligned} \qquad (9.4.2.2.3)$$

Replacing these values in the queuing system of Ch 9.3.2.4 and solving yields

Table 9.22 Performance measures for session reconfiguration

$\tilde{\lambda}$	π	$\tilde{P}r$	Tp	$E[W]$	$E[L_q]$
0.3673	20.69	0.594	54.4	26.81	9.85

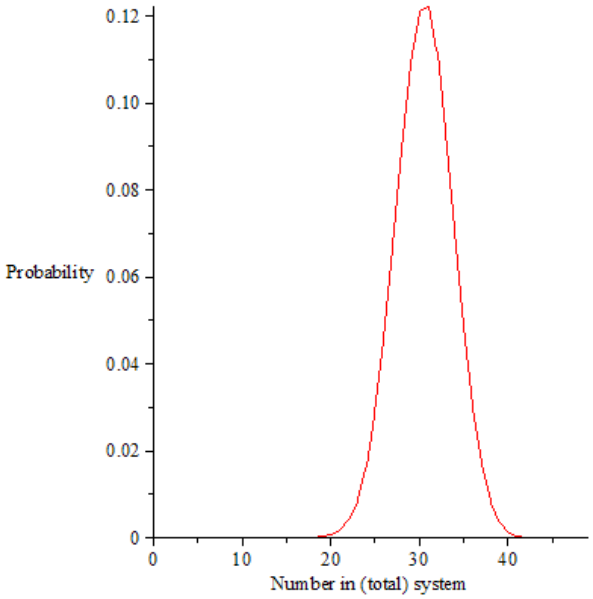


Figure 9.26 Probability density graph for number in system

So compared to the (phase one) model of the current situation at Rookwood NRC (Ch 9.3.2.4) the mean probability of reneging has decreased by (almost) 4% and throughput has increased by 3.61 patients per year. The annual cost is estimated at $20.69 \times £480 \times 365 = £3.625\text{m}$ – a saving of £12k. In addition, since the average LOS under this scenario is 138 days (see (9.3.3.1.1)) it would cost $3.61 \times £480 \times 138 = £239\text{k}$ for the admission of the 3.61 extra patients each year. Also, the figure of £480 could be an over-estimate since group sessions can be provided by an employee of any band level (Ch 3.4.2). Costs would be significantly lower if the skill mix is skewed toward the less expensive and less qualified therapists (whilst retaining sufficient numbers of qualified staff for stretches, goal planning meetings, hydro sessions etc).

However, this example highlights an obvious but important drawback to the model. Whilst quantity of treatment is indeed a factor that influences LOS and patient outcome, so too is quality. Unfortunately, the quality of treatment sessions is not as easy to quantify as the intensity of treatment and so is not a consideration within this project. Therefore, the example given here must be treated with caution. As mentioned before (Ch 9.4.1.2) patients from groups three and four are more severely injured and require two therapists for treatment sessions. They are thus unlikely to perform in, or be able to receive, group sessions which involve only one therapist and two or more patients.

There are a number of other hypothetical scenarios that relate to changes in treatment quality. For example, what if the same number of therapists work the same number of hours but over evenings and weekends as well as during normal working hours? Therefore, the treatment intensity would not change but the quality of the treatment would since the sessions can be more evenly spread throughout the week. This would have a significant effect on *SCON2* (Ch 3.5.2) and *MCON2* (Ch 3.5.3) but would not alter the output of the queuing model (as treatment intensity is unchanged). It is hence useful to check the performance measures of the scheduling program (*SCON* value and *MCON* violations) in addition to the performance measures of the queuing model.

9.5 Conclusion

In this chapter three models of varying complexity have been analysed. For each of these, an initial, representative model has first been constructed (phase one) and subsequently used to study the effect of hypothetical what-if type scenarios (phase two). The first model (Ch 9.2) is the most simplistic fitting an exponential distribution to total LOS and not taking into account patient heterogeneity, treatment intensity, and the active and blocked components of LOS. The second model (Ch 9.3) makes the distinction between active and blocked LOS; fitting two and three-term Coxian distributions to active and exponential to blocked LOS. Patients are also split into four patient groups based on the effect of certain factors on LOS. In the final model (Ch 9.4), treatment intensity is incorporated through the use of the specially designed automated scheduling program (Chapters 3 and 4).

For each of these models a number of performance measures are output in addition to an expected value for the annual running cost. These are summarised in table 9.23. The expected waiting time and expected length of queue are not displayed since these include details for reneged patients. Hypothetical scenario one (Ch 9.2.6.1) is not included since this is pertinent only to data from the sixteen months from January 2010 to April 2011.

Table 9.23 Summary of performance measures and costs for typical and hypothetical scenarios

Model type	Modification	π	$\tilde{Pr}$	Tp	Cost pa	Cost per episode
Typical (Model One)		21	0.613	51.34	£3.679m	£72k
Scenario Two	beds ↓ 2	19	0.65	46.45	£3.329m	£72k
	beds ↓ 1	20	0.632	48.89	£3.504m	£72k
	beds ↑ 1	22	0.595	53.78	£3.854m	£72k
	beds ↑ 2	23	0.576	56.23	£4.030m	£72k
Typical (Model Two)		20.76	0.617	50.79	£3.637m	£72k
Scenario Three	blocked LOS ↓ 50%	20.65	0.576	56.68	£3.618m	£64k
	blocked LOS ↓ 100%	20.4	0.452	59.54	£3.574m	£60k
Scenario Four	arrival rate ↑ 33% for older patients	20.99	0.653	51.27	£3.677m	£72k
Scenario Five	physiotherapists ↑ 33%	20.73	0.604	52.78	see below	see below
Scenario Six	group sessions ↑ 3 double sessions ↓ 1	20.69	0.594	54.4	£3.625m	£67k

The cost per episode is simply calculated by dividing the annual cost by the throughput. This gives a single measure of financial value that is easily comparable. However, this does not take into account the financial implications of lost admissions. It can be seen from scenario two that the cost per episode is equal regardless of the number of beds. However, the mean probability of reneging varies considerably from 0.576 to 0.65. A higher value means that fewer patients ultimately receive the specialist care that is required at Rookwood NRC. Although this has no direct financial implications for the unit it does affect the healthcare system and wider economy. If patients are denied opportunities for the specialist rehabilitation programme then their ability to regain motor and cognitive function may not only be slower but may never reach the levels otherwise attainable (see Ch 2.2.4). This has a doubly negative effect since if the patient is too ill to re-enter the workforce then not only does the local economy suffer but social services must also commit resources to care for the patient. It is therefore crucial that the mean probability of reneging is examined alongside the expected cost per episode.

Clearly, a reduction in blocked LOS is of substantial benefit. This reduces LOS and so increases the throughput of the unit without having to employ additional resources. Improved throughput is not just of great benefit to Rookwood NRC but also to other facilities upstream

(neuro ward and ICU – see Ch 1.2). Since the bed-cost per day is greater in these units than at the NRC this offers significant savings, not to mention the aforementioned benefits with reducing patient waiting time (Ch 9.3.3.1). However, just as the neuro wards rely on Rookwood for this, Rookwood must rely on social services and the other discharge destinations. An idea to make these more efficient has been put forward in Ch 9.3.3.1. There is, however, one problem with decreasing blocked LOS. If therapists are no longer working on maintenance patients, who typically require less treatment than rehabilitative patients, then are more therapists needed to deal with the additional treatment requirements?

The fourth scenario considers the impact of a one-third increase in demand from older patients. If the number of beds (and/or treatment intensity) is not appropriately increased then this would push occupancy to capacity and significantly increase the mean probability of these patients reneging. Clearly more beds (and/or therapists) are needed to deal with the increased demands of an ageing population.

The size of the physiotherapy workforce is increased by one third in scenario five. This reduces active LOS and increases throughput by two patients per year. If it is assumed that any change to the physiotherapy department must be replicated by the other departments (see Ch 6.4.2.3) then the average cost per episode would stand at £72k. Otherwise, if it is the readiness of the patient with respect to physiotherapy only that dictates the (overall) date ready for discharge then the average cost per episode would be £70k (Ch 6.4.2.1).

Finally, a different composition of physiotherapy treatment sessions is studied. It is shown that if one less double session and three more group sessions are demanded for each patient then this can greatly increase treatment intensity. This leads to a significant reduction in active LOS in patient groups two, three and four that accounts for increased throughput. Since the (physiotherapy) workforce has not changed, savings are approximately £5k for each patient episode. Again, depending on the validity of either assumption in Ch 6.4.2.3 this change may or may not need to be replicated across the other departments. The principle problem with this scenario is the efficacy of group sessions for more severely injured patients.

All of these what-if questions relate to the major policy decisions (Ch 1.7.2). Changes to the number of beds and arrival rates relate to *who to admit* whilst changes to blocked LOS relate to *when to discharge*. Changes to the composition of the workforce or therapy relate to *what care to give*. This thesis is concluded in the next, and final, chapter.

Chapter 10: Conclusion

10.1 Objective

The aim of this project is to model the activities at the Neurological Rehabilitation Centre (NRC) at Rookwood hospital. This is done using Queuing Theory (Ch 1.4) alongside phase-type distributions (Ch 1.5) and a consideration of treatment intensity (Ch 1.6). The first phase is to construct an initial, representative model of the unit. The second phase is to use this model to evaluate the impact of various hypothetical *what-if* type scenarios. These relate to the *major policy decisions* (who to admit, what care to give, when to discharge), not for individual patients, but on a macro level. Performance measures and costs can then be used to compare different configurations.

10.2 Approach

There are four stages to the construction of the model (phase one). A description of these is available in Ch 1.7.1. A number of relevant hypothetical scenarios have been identified through discussions with clinicians at Rookwood NRC. These are evaluated by making appropriate modifications to some (or all) of the four stages of phase one. However, some scenarios can be more complex to evaluate than others. For example, if a scenario is tested that just changes the arrival rates of a pre-existing patient group then this can easily be evaluated by making a simple modification to stage four. But, if a new composition of patient groups is sought then changes need to be made as far back as stage one. Because the other stages have a sequential dependence then stages two, three and four all need to be modified as well. In general, the lower the stage that is affected, the more complex the evaluation.

10.3 Results

The main results of the project are summarised (Table 9.23) and discussed in Ch 9.5.

10.4 Evaluation

In my opinion the greatest emphasis should be placed on reducing blocked length of stay (LOS). This would have outstanding effects on throughput and lost admissions without the need to employ additional resources (beds or therapists). However, how this can be achieved is another matter (although a possible solution has been suggested in Ch 9.3.3.1). Increasing the number of therapists could be another option to increase throughput and reduce lost admissions although the reduction in cost per episode would be small. The same applies for increasing the number of beds but in doing this there is no reduction in cost per episode at all.

If the available budget for the NRC at Rookwood hospital increases by about £150k to £3.782m it is possible to increase throughput to 52.78 and reduce the mean probability of reneging to 0.604 by increasing the number of therapists by one third (across all departments). However, by spending an additional £70k the unit could, instead, open one more bed which would increase throughput to 53.78 and decrease the probability of reneging to 0.595. If this option is favoured then the number of FTEs would have to increase by 22 over 21. Many other scenarios can be evaluated in this manner.

It should also be mentioned that whilst, in these example scenarios, only one variable is changed and all others are held constant this is by no means essential. The model is perfectly capable of evaluating a scenario in which many simultaneous changes are made to the number of beds, intensity of treatment, blocked LOS, arrival rates etc.

10.5 Limitations

Many of the model limitations have been discussed along the way throughout this thesis (e.g. Ch 9.2.6.2). A handful of those that I feel are most significant are now summarised.

In stage one CART analysis is used to determine a number of patient groups. The target variable here should be active LOS. However, since only 89 of the 358 patient episodes have data on their date ready for discharge, total LOS is used. This produces a number of patient groups for which total LOS is homogenous. The episodes within these groups that contain a date ready for discharge are then used to estimate such dates for those without this information. This makes the assumption that groups with homogeneity with respect to total LOS also have homogeneity with respect to active and blocked LOS.

In stage two the automated scheduling program is used to generate average values of treatment intensity for patients of each group. However, these values represent the amount of treatment that is scheduled whilst the relationship between treatment intensity and active LOS in the next stage is constructed using data on treatment that is actually received. It is therefore assumed by the model that the amount of treatment scheduled is equivalent to the amount that is received. Clearly, this is not always the case.

Since only the scheduling of physiotherapy is studied it is necessary to make assumptions about the treatment intensity of other departments. As mentioned in Ch 6.4.2.3 it could be assumed that any change to treatment intensity (in stage two) must be replicated by the other departments. In this case the treatment intensity of stage three is an average over the departments. Otherwise, it could be assumed that the readiness of the patient with respect to physiotherapy only dictates the (overall) date ready for discharge. In this case it is not necessary for the treatment intensity of other departments to change. The validity of either of these assumptions is unknown. Another limitation in stage three is the amount of data used to construct the graphs for patient groups two and three (see Ch 9.4.1.1).

One of the many assumptions of stage four is that of the queue discipline first-in first-out. Realistically it is known that the positions of referrals in the queue is based on patient priority which is calculated from injury severity and time spent in queue. However, not only would this be tricky to model but it would also require additional data. The full effects of this incorrect assumption on the model are not entirely understood. An additional limitation has been the inability to derive an expected waiting time for customers that ultimately enter service when reneging is employed. This would be very useful as it could be used to measure the cost of a patient's deterioration in functional ability whilst waiting in the queue. This could be used alongside the mean occupancy and the mean reneging probability to form a single-valued cost function (see Ch 9.2.6).

Generally speaking, the lack of data has been the largest limitation of this project. Whilst some data was readily-available (i.e. NHS Trust dataset) this did not include variables relating to treatment intensity. Neither did it include details that relate to the state of the queue over time. Such data was eventually obtained although this was a time-consuming task for both myself and the technician at Rookwood hospital. Even when the (Arrivals and Service) datasets were populated the number of observations was low. From the Service dataset, only 358 patient episodes were available from 2003 to 2011 of which only 89 of

these had details for treatment intensity and date ready for discharge. In addition, the Arrivals dataset only covered sixteen months of data and so there was a lack of congruency in the timescales of the two datasets. This was problematic since performance measures based on the state of the queue (i.e. mean length of queue) were not relevant for the period of time covered by the Service dataset. Finally, if more data had been available then the model could have been internally validated using a separate cohort.

10.6 Further Work

There are many matters that can be investigated based on the foundations of this project. Some of this work could relate to the shortfalls of the model identified in the previous subchapter. For example, the treatment intensity and active LOS of other departments could be integrated within the model (Ch 6.2.3). This would permit a wider range of what-if type scenarios that could be used to better co-ordinate the readiness of patients from each department; thus reducing maintenance treatment (see Ch 6.2.3). The model described by Route 6 (Figure 6.8) would be possible if discharge attributes (Ch 6.2.5) are also considered. Whilst discharge destination is an important branching variable in the determination of patient groups it does not include the level of family support. A consultant in rehabilitation at Rookwood hospital has said that this variable is crucial in predicting LOS, treatment intensity or outcome.

As part of the trilateral relationship (see Ch 2.2.3), outcome is an important metric that has been held constant in this project. That is, changes to treatment intensity affect active LOS such that patient outcome is unaltered. The ability to have this as a free variable would permit the evaluation of some interesting hypothetical scenarios. The regional rehabilitation unit at Northwick Park hospital has produced a number of studies (Ch 2.2.4) in which outcome is varied and the effect on continuing care is evaluated. Basically, the notion is that more functional ability on discharge equates to less cost per unit time post-discharge. Typically, they investigate the time required to offset the additional costs associated with providing a higher outcome from rehabilitation. However, they do not have a model that can tell them how, exactly, a higher outcome can be achieved. According to the trilateral relationship this can be achieved through increased LOS or treatment intensity or both. If included in the queuing model then the consequent effect of such alterations on outcome could be evaluated. If the relationship between outcome and cost per unit time post-discharge is known then a

number of valuable what-if questions can be asked, such as, is it more cost-effective for younger patients to have a higher rehabilitation outcome?

Clearly, the integration of Rookwood NRC with its related components (in this case, downstream) presents some interesting opportunities. The upstream facilities are now considered. It is known that the bed cost per day at a neuro ward (see Ch 1.2) is higher than at the NRC. Surely it would be better to spend more money at Rookwood NRC to reduce blocking (by increasing number of beds and/or therapists) in order to reduce the blockages at these more expensive units? An integrated model could evaluate this. Such a policy would also serve to reduce blockages at the ICU ward which is even more expensive. It is postulated that to increase the overall fluidity of a patient flow system it is essential that investments are made downstream where the costs are lowest. System dynamics (Ch 1.3) could well be a powerful tool for investigating this point in addition to the overall flow of ABI patients through the various healthcare facilities.

10.7 Legacy

At the time of writing and to the best knowledge of the author this is the first queuing theory study of a specialist neurological rehabilitation unit. In particular, a novel part of this project has been to condition the service rates on the treatment intensity received by patients. To this end an **automated scheduling program** (Chapters 3 and 4) has been developed. The objective of this was to mimic the, at the time, by-hand scheduling procedure so that what-if type questions pertinent to treatment intensity could be quickly evaluated. However, since the program was so much better than the manual alternative it was adopted in January 2011 and has been used weekly up to the time of writing. It is expected that this will be used for the foreseeable future. Note that it was also necessary to develop a new **patient timetable overview** (Ch 4.4.1) for use with the program. This too enjoys problem-free use by all departments in the unit.

Treatment intensity has been incorporated within the model since many studies (Ch 2.2.3) have found that patient LOS in a neurological rehabilitation setting is affected by the intensity of treatment received. It has been found in this project that LOS is indeed dependent on treatment intensity but not for all patients (Ch 9.4.1.1). In order to deduce the relationship between treatment intensity and active LOS it is necessary to capture data on the amount of therapy received and the date ready for discharge. Since historical data did not exist a **database** was built in MS Access to prospectively store data for these variables. Whilst some

data was input by the clinicians this did not include all patients since its inauguration. It is hoped that, in time, as therapists get more comfortable with using the software its usage will improve.

The data that has been used in this project has been amalgamated into two datasets. The **Arrivals dataset** (Ch 6.3.1) contains sixteen months of data on the state of the queue at Rookwood NRC. From this it is possible to automatically derive a number of summary measures (see Ch 6.4.1). If Rookwood NRC continues to populate this dataset then updated values of these can be conveniently output. The same is true for the **Service dataset** (Ch 6.3.2) that contains approximately eight years of data on patient episodes. The program Treeworks uses this dataset to produce a **CART analysis**; the results of which can be used by the unit to accurately predict discharge readiness dates. If Rookwood NRC continues to populate this dataset then the accuracy and relevancy of such predictions will only increase.

Distributions are fitted to active LOS data from the Service dataset using the purpose-built **Matlab programs**. One of the programs is for phase-type approximating distributions and the other is for non phase-type approximating distributions. It is necessary to provide the method of moments parameter estimates (which can be calculated through the formulae developed in Chapter 5) which are used as starting values for the maximum likelihood approach. Graphs for both approximating distributions are displayed against the sample data. Due to the flexibility and usefulness of these programs I fully expect to use them in future work unrelated to this project.

The parameter estimates output from the Matlab program are thereafter used in either of the **Maple programs** (depending on whether the service distribution is Erlang or Coxian and exponential). Balking and reneging are two types of customer behaviour that are optional in each of these programs. To the best knowledge of the author the approach employed to populate the transition matrix using the formulae developed in Chapters 7 and 8 is new and does not appear in the literature. Following the determination of steady-state probabilities many performance measures and graphs are automatically output. These computer programs are not consigned for use exclusively with healthcare systems and can be used for a variety of applications.

As a final note, it is hoped that the research carried out in this project will be developed in years to come and lead to further advancements in the management of patients with acquired brain injury.

Appendix 3.1

How to Timetable for Neuro-Physiotherapy Team

The timetable for physio is written on a Thursday for the following Monday to Friday, psychology, speech therapy and occupational therapy should already have done theirs. The individual timetables should be printed and given to the patients on a Friday afternoon and therapy overviews given to each ward.

Timetable Prep (should be done before Thursday):

- Check names and dates at the top of the sheets and that the hours are correct for the part-time staff.
- Jakko's Timetable – accessed from Outlook Calender by Pat Grundon or Sarah Duthie. Fill in any clinical time he has.
- Fill in relevant information from diary:
    - Annual leave
    - Study leave
    - Staff meetings (check wall in staff room)
    - Goal planning meetings (stripey folder on ward 7)
    - Check ward diaries for patients appointments
    - Hydrotherapy
    - Group session Monday/Wednesday/Friday 11.15
    - Timetabling all day on a Thursday x1 person
    - Short term goals meeting Wednesday 3.15
    - Rookwood ward round Tuesday 8.30 – 11.15
    - Admissions meeting Thursday 9.00 – 10.00 Jo PJ or Sarah D.
    - UHW Ward round Wednesday 9-2 Jo PJ (Then Lieu)
    - In service training (Check diary)
    - Ward round feedback Tuesday 12.00
    - Slots preparation Tuesday 1 slot (Sarah D, Sarah A or Jo H)
    - Timetable prep 1 slot before Thurs (Any PTA or Tech)
    - Supervision time as requested
    - Student sessions with clinical educator – 2 slots
    - Admin x2 half days for Senior 1
    - Admin x2 per week for Static Senior II as requested.
    - Friday – splinting forms – I slot
    - Clean Gym – last slot Friday pm.
    - Slots for updating Clinical Workstation x1 slot for each physio at 12.00 or 3.15 (can't be at the same time as any one else)

Getting Started – What you need:

- Timetable sheets should be in the yellow timetable folder
- List of primary and secondary physiotherapists (black folder)
- Request list from short term goals meeting
- Departmental Diary
- List of joint sessions with OT
- Pencil and Rubber
- Access to a computer

Prioritisation of patients:

Prioritise patients on the request sheet as follows:

Score 1 point for:	Recent Admission (Less than 2 weeks)
	Discharge due (Less than 2 weeks)
	Poor Chest
	Improving with physio input
	Likely to deteriorate without physio input

The scores should be written against the patients name on the timetable request form for ease of reference when timetabling.

The highest scoring patients get timetabled first as they are the highest priority. Therefore unmet needs should only affect the low priority patients.

Filling in the timetable:

Log on to the computer using department log on and go to Rookwood neuro rehab on the S drive which should be found on the desk top. Open the file New Timetable and then Timetables therapies. This gives you the overview so you can see what other therapies have already put on the timetable.

Check the date on the top is the correct one for the week you are timetabling for.

Check for any joint sessions that have already been put in and write them on the timetable sheets.

Start by filling in the hydrotherapy slots first, then group, then morning sessions (either stretches or joint sessions with nurses).

Continue with physio sessions for the high priority patients first by checking the request list and slotting patients in with their primary or secondary physiotherapist, using the PTA's and Tech's for doubles and triples. (Make sure that the patients put in to the 11.35 slot are peg fed patients)

Each time you fill in a session make sure you enter it on the overview as well as the paper physiotherapy timetable sheets

Finally list how many sessions you were not able to book, record them as unmet needs on the bottom of the timetable sheet for the relevant day.

Make sure you save all the changes you have made to the overview.

Printing the Timetables:

Timetables should be printed out on a Friday afternoon.

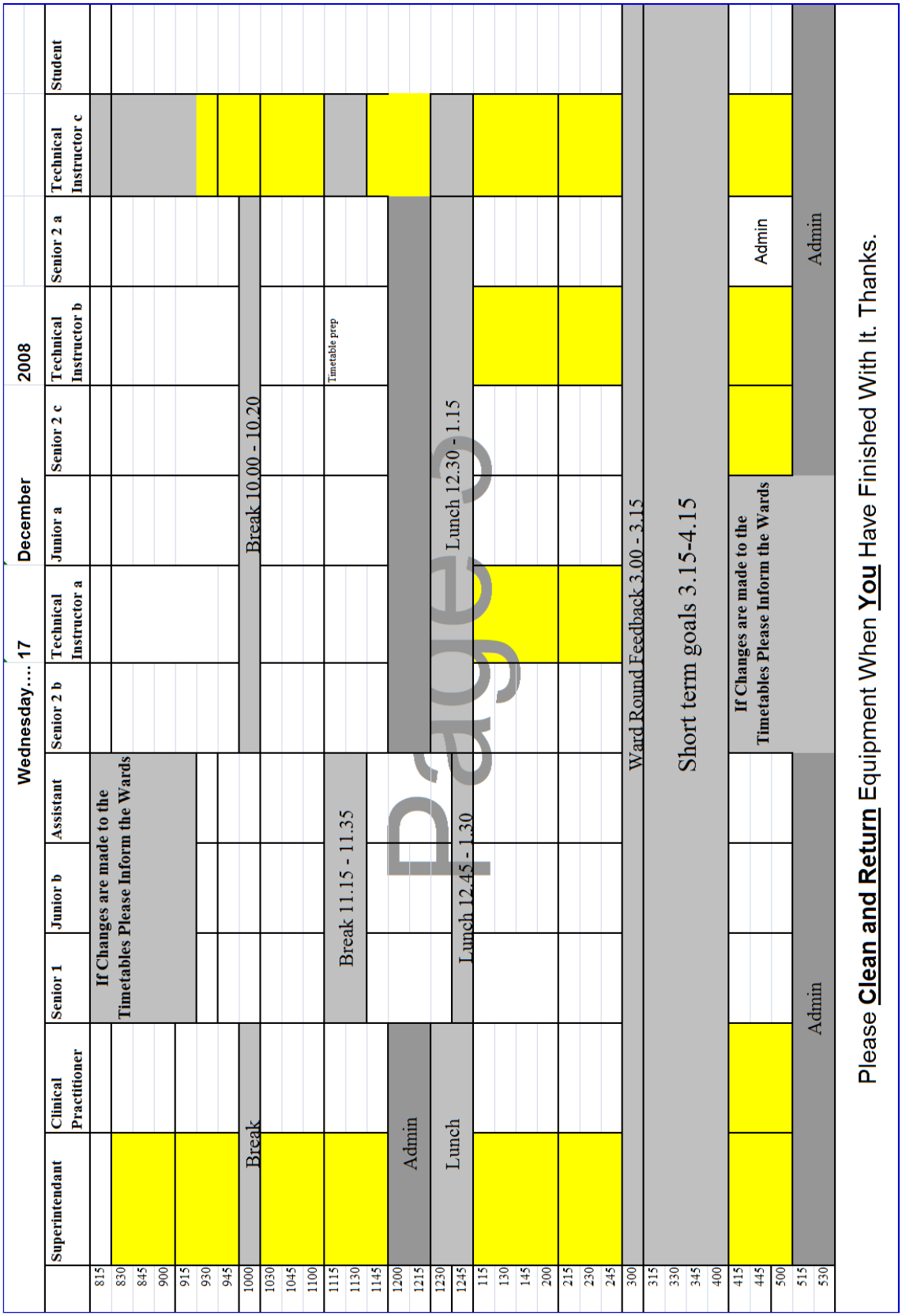
Print out the overview and the individual timetables.

The overview needs to be photocopied and given to each ward. (They are put in the folders labelled patient therapies, red folder on ward 7 and black folder on ward 8)

The individual timetables should be given to the patient or put on their wall or in their rehab diary.

Appendix 3.2

Even Week Wednesday



Appendix 3.3

Syntax

Affiliation	Name	Dependence	Description
Patient	*ps*	*i,d,s*	Start time of patient session *s* on day *d* for patient *i*
	pe	*i,d,s*	End time
	nps	*i,d*	Number of patient sessions on day *d* for patient *i*
	pps	*i,d,s*	Start time of patient physiotherapy session *s* on day *d* for patient *i*
	ppe	*i,d,s*	End time
	ppea	*i,d,s*	List of employees assigned
	ppst	*i,d,s*	Type of physiotherapy session
	npps	*i,d*	Number of physiotherapy patient sessions on day *d* for patient *i*
Patient Demand	*d**	*i*	Demand for patient session type *
	d^{*0}	*i*	Remaining demand for patient session type * after manual input
	d^{*1}	*i*	Remaining demand for patient session type * after automation
	d^{*2}	*i*	Number of sessions scheduled of type *
Patient Attributes	*ast*	*i*	Returns value 1 if patient requires one employee for stretches or 2 if patient requires two employees
	aal	*i*	Returns value 1 if patient attends lunch
	amb	*i*	Minimum band therapist for patient *i*
	apr	*i*	Returns value 0 if patient has no preference; 1 if patient has preference to morning; 2 if patient has preference to afternoon
	amdc	*i*	Returns value 1 if patient does not respond well to multidisciplinary cumulative sessions; 0 otherwise
	app	*i*	Primary therapist for patient *i*
	asp	*i*	Secondary therapist for patient *i*
	cap	*i*	Priority level of patient *i*
Staff	*sss*	*j,d,l*	Start time of staff slot *l* on day *d* for employee *j*
	sse	*j,d,l*	End time
	ssr	*j,d,l*	List of recipients
	sst	*j,d,l*	Type of staff slot

Staff attributes	*tb*	*j*	Band level of therapist *j*
	tl	*j*	Returns value 1 if therapist *j* can lead patient sessions
	tg	*j*	Returns value 1 if therapist *j* can provide group sessions
	tst	*j*	Returns value 1 if therapist *j* can assist in stretches or 2 if can lead stretches
	tpm	*j*	Returns value 1 if therapist *j* can provide patient meetings
	tgm	*j*	Returns value 1 if therapist *j* can provide goal planning meetings
Lunch	*ls*	*d*	Start time of lunch on day *d*
	le	*d*	End time
Groups	*gs*	*r,d*	Start time of group slot availability period *r* on day *d*
	ge	*r,d*	End time
	minngsr		Minimum number of patients in a group slot
	maxngsr		Maximum
Stretches	*es*	*f,d*	Start time of stretch availability period *f* on day *d*
	ee	*f,d*	End time
Weight	*ẁ*	*	Weight value of patient priority level *
	w	*,**	Weight value ** of SCON *
Misc	*cci**		Coded colour index of employee slot *
	*md**		Day of patient session *
	ma	*i,k,x*	The actual day/time of patient session *x* of session group *k* for patient *i*
	mo	*i,k,x*	The optimal day/time
	T	*i,k*	The day/time of the last occurring patient session of session group *k* for patient *i* in the previous week

Appendix 3.4

Hard Constraints

Patient related

HCON 1: More than one treatment session cannot be received at any time by the patient

$$\begin{aligned} ps_{i,d,s_a} &\notin \left[ps_{i,d,s_b}, pe_{i,d,s_b} \right) \wedge pe_{i,d,s_a} \notin \left(ps_{i,d,s_b}, pe_{i,d,s_b} \right] \\ ps_{i,d,s_b} &\notin \left[ps_{i,d,s_a}, pe_{i,d,s_a} \right) \wedge pe_{i,d,s_b} \notin \left(ps_{i,d,s_a}, pe_{i,d,s_a} \right] \end{aligned} \qquad \forall s_b \neq s_a; \forall s_a; \forall d; \forall i$$

HCON 2: A physiotherapy session must have at least one therapist assigned

$$\left| ppea_{i,d,s} \right| \geq 0 \qquad \forall s; \forall d; \forall i$$

HCON 3: The start time of the session must be before the end time

$$pe_{i,d,s} > ps_{i,d,s} \qquad \forall s; \forall d; \forall i$$

HCON 4: The number of sessions manually scheduled must be less than or equal to that demanded by each patient

$$ds_i^0, dd_i^0, dt_i^0, dg_i^0, dh_i^0, dst_i^0, dj_i^0, dpm_i^0, dgm_i^0, dop_i^0, dad_i^0 \geq 0 \qquad \forall s; \forall d; \forall i$$

HCON 5: The number of sessions computationally scheduled must be less than or equal to that demanded by each patient

$$ds_i^1, dd_i^1, dt_i^1, dg_i^1, dh_i^1, dst_i^1, dj_i^1, dpm_i^1, dgm_i^1, dop_i^1, dad_i^1 \geq 0 \qquad \forall s; \forall d; \forall i$$

HCON 6: Group slots must occur within permissible times

$$\exists! r : pps_{i,d,s} \geq gs_{r,d} \wedge ppe_{i,d,s} \leq ge_{r,d} \qquad \forall s : ppst_{i,d,s} = ccig; \forall d; \forall i$$

HCON 7: Stretches must occur within permissible times

$$\exists! f : pps_{i,d,s} \geq es_{f,d} \wedge pe_{i,d,s} \leq ee_{f,d} \qquad \forall s : ppst_{i,d,s} = ccist; \forall d; \forall i$$

HCON 8: No physiotherapy treatment may occur during the lunch break of patients who attend lunch

$$pps_{i,d,s_a} \notin \left[ls_{i,d,s_b}, le_{i,d,s_b} \right) \wedge ppe_{i,d,s_a} \notin \left(ls_d, le_d \right] \quad \forall s; \forall d; \forall i$$
$$ls_d \notin \left[pps_{i,d,s}, ppe_{i,d,s} \right) \wedge le_d \notin \left(pps_{i,d,s}, ppe_{i,d,s} \right]$$

HCON 9: A single session must be provided by one therapist to one patient

$$\left| ppea_{i,d,s} \right| = 1 \qquad \forall s : ppst_{i,d,s} = ccis; \forall d; \forall i$$

HCON 10: A double session must be provided by two therapists to one patient

$$\left| ppea_{i,d,s} \right| = 2 \qquad \forall s : ppst_{i,d,s} = ccid; \forall d; \forall i$$

HCON 11: A triple session must be provided by three therapists to one patient

$$\left| ppea_{i,d,s} \right| = 3 \qquad \forall s : ppst_{i,d,s} = ccit; \forall d; \forall i$$

HCON 12: A single stretch must be provided by one therapist to one patient

$$\left| ppea_{i,d,s} \right| = 1 \qquad \forall s : ppst_{i,d,s} = ccist; \forall d; \forall i : ast_i = 1$$

HCON 13: A double stretch must be provided by two therapists to one patient

$$\left| ppea_{i,d,s} \right| = 2 \qquad \forall s : ppst_{i,d,s} = ccist; \forall d; \forall i : ast_i = 2$$

HCON 14: A single/double/triple/group/patient meeting must be of duration 45 or 60 minutes

$$ppe_{i,d,s} - pps_{i,d,s} = 45 \vee 60 \qquad \forall s : ppst_{i,d,s} = ccis \vee ccid \vee ccit \vee ccig \vee ccipm; \forall d; \forall i$$

HCON 15: A stretch must be of duration 30 minutes

$$ppe_{i,d,s} - pps_{i,d,s} = 30 \qquad \forall s : ppst_{i,d,s} = ccist; \forall d; \forall i$$

HCON 16: The end time of any patient meetings cannot be before the start time of any goal planning meetings (if goal planning meeting scheduled)

$$md^p \le md^g \qquad \forall md^p : \exists s : ppst_{i,d_p,s} = ccipm; \forall md^g : \exists s : ppst_{i,d_g,s} = ccigm; \forall i$$

$$ppe_{i,d,s_p} \le pps_{i,d,s_g} \qquad \forall s_p, s_g : \exists d : ppst_{i,d,s_p} = ccipm \wedge ppst_{i,d,s_g} = ccigm; \forall i$$

HCON 17: The band level of the lead employee of a session must be greater than or equal to the patient's minimum permissible band level

$$\exists j \in ppea_{i,d,s} : tb_j \geq amb_i \qquad \forall s : ppst_{i,d,s} = ccis \vee ccid \vee ccit; \forall d; \forall i$$

HCON 18: The sessions contained on the PTO must correspond to the sessions contained on the PPTO

$$\exists s_p : pps_{i,d,s_p} = ps_{i,d,s} \wedge ppe_{i,d,s_p} = pe_{i,d,s} \qquad \forall s; \forall d; \forall i$$

$$\exists s : ps_{i,d,s} = pps_{i,d,s_p} \wedge pe_{i,d,s} = ppe_{i,d,s_p} \qquad \forall s_p; \forall d; \forall i$$

HCON 19: The sessions contained on the PPTO must correspond to the sessions contained on the (SDTs)

$$\exists l : i \in ssr_{j,d,l} \wedge sss_{j,d,l} = pps_{i,d,s} \wedge sse_{j,d,l} = ppe_{i,d,s} \wedge sst_{j,d,l} = ppst_{i,d,s}$$
$$\forall j \in ppea_{i,d,s}; \forall s; \forall d; \forall i$$

HCON 20: At least one therapist providing a stretch must be a lead

$$\exists j \in ppea_{i,d,s} : tst_j = 2 \qquad \forall s : ppst_{i,d,s} = ccist; \forall d; \forall s$$

Employee related

HCON 21: An employee cannot provide more than one treatment session at any time

$$\begin{matrix} sss_{j,d,l_a} \notin \left[sss_{j,d,l_b}, sse_{j,d,l_b} \right) \wedge sse_{j,d,l_a} \notin \left(sss_{j,d,l_b}, sse_{j,d,l_b} \right] \\ sss_{j,d,l_b} \notin \left[sss_{j,d,l_a}, sse_{j,d,l_a} \right) \wedge sse_{j,d,l_b} \notin \left(sss_{j,d,l_a}, sse_{j,d,l_a} \right] \end{matrix} \qquad \forall l_b \neq l_a; \forall l_a; \forall d; \forall j$$

HCON 22: An employee must provide treatment to at least one patient in a patient session

$$\left| ssr_{j,d,l} \right| \geq 1 \qquad \forall l; \forall d; \forall j$$

HCON 23: A group session must be provided to no more than the maximum or no fewer than the minimum number of patients

$$minngsr \leq \left| ssr_{j,d,l} \right| \leq maxngsr \qquad \forall l : sst_{j,d,l} = ccig; \forall d; \forall j$$

HCON 24: The sessions contained on the SDTs must correspond to the sessions contained on the PPTO

$$\exists s : j \in ppea_{i,d,s} \wedge pps_{i,d,s} = sss_{j,d,l} \wedge ppe_{i,d,s} = sse_{j,d,l} \wedge ppst_{i,d,s} = sst_{j,d,l}$$

$$\forall i \in ssr_{j,d,l}; \forall d; \forall j$$

HCON 25: The therapist(s) assigned to patient sessions must be suitable

$$\exists j \in ppea_{i,d,s} : tl_j = 1 \qquad \forall s : ppst_{i,d,s} = ccis; \forall d; \forall i$$

$$\exists j \in ppea_{i,d,s} : tg_j = 1 \qquad \forall s : ppst_{i,d,s} = ccig; \forall d; \forall i$$

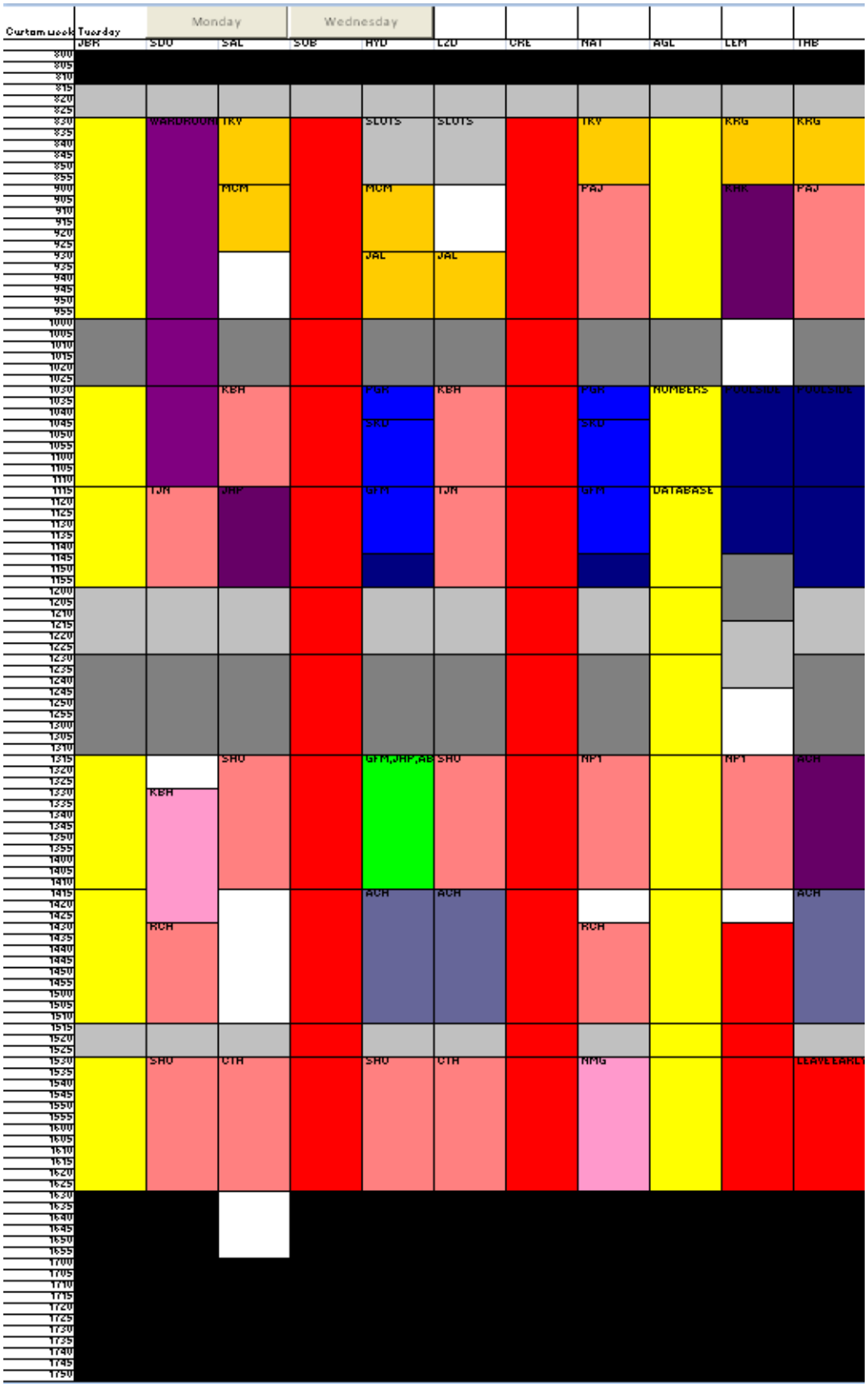
$$\exists j \in ppea_{i,d,s} : tst_j = 2 \qquad \forall s : ppst_{i,d,s} = ccist; \forall d; \forall i$$

$$\exists j \in ppea_{i,d,s} : tpm_j = 1 \qquad \forall s : ppst_{i,d,s} = ccipm; \forall d; \forall i$$

$$\exists j \in ppea_{i,d,s} : tgm_j = 1 \qquad \forall s : ppst_{i,d,s} = ccigm; \forall d; \forall i$$

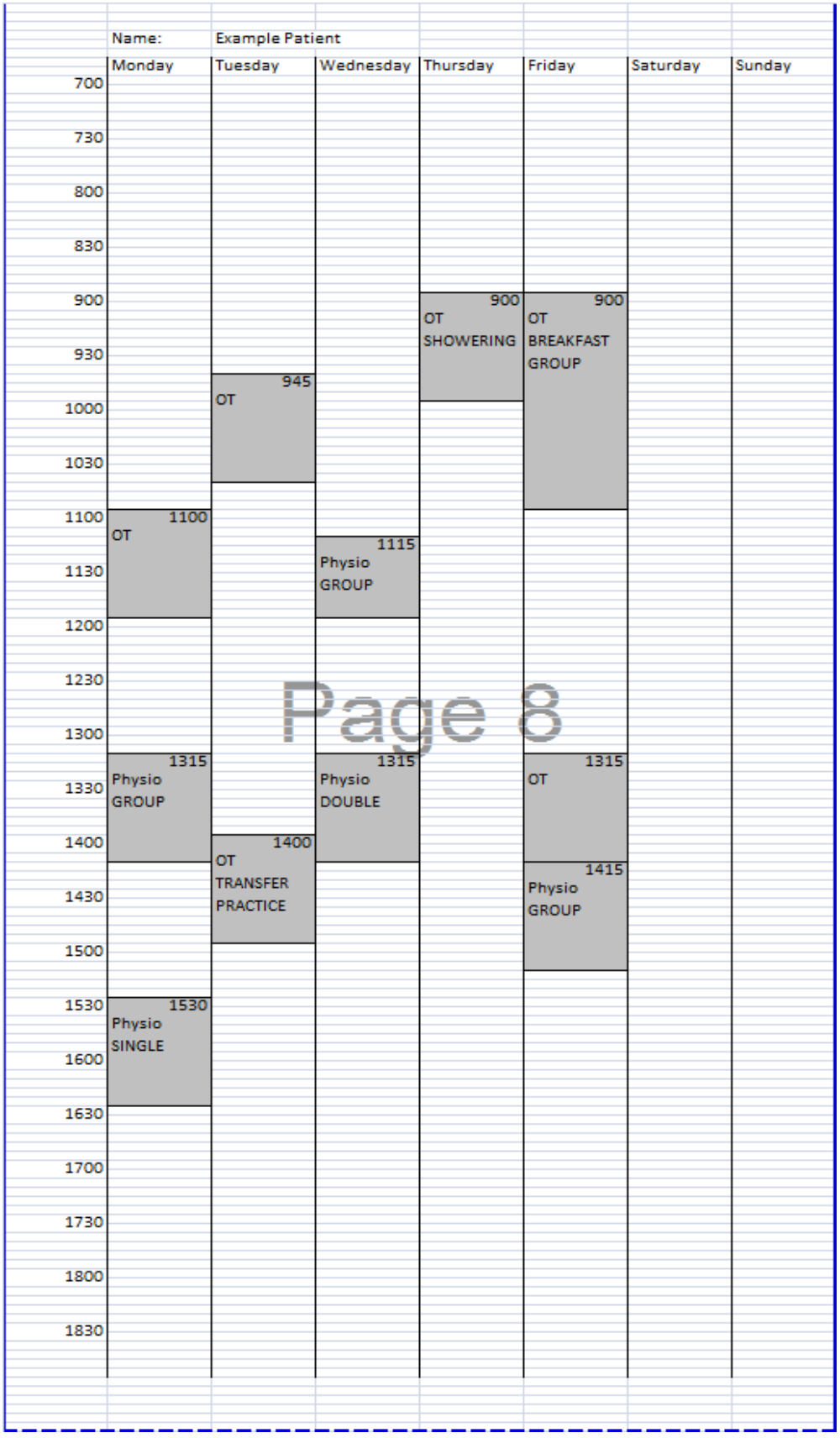
Appendix 4.1

Example: Computer Generated Solution (Tuesday)



Appendix 4.2

Example: Patient Handout



Appendix 4.3

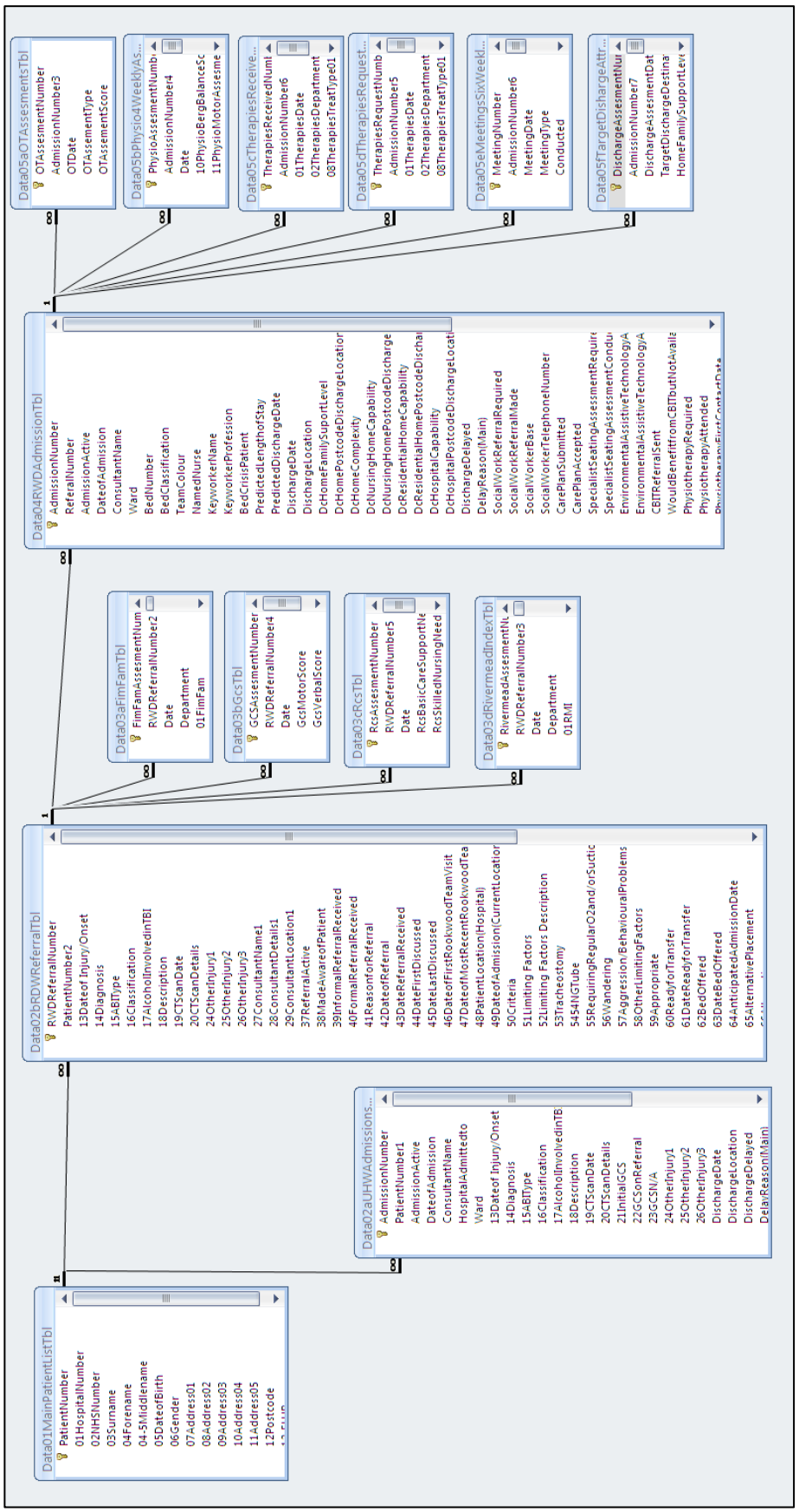
Approximate Values of Weights used in Automated Scheduling Program

The weights given below have been deduced following trials of the automated scheduling program. For each treatment week during the trials, schedules were created with a number of different weight values. These were presented to clinicians whose response determined whether values were increased or decreased.

Weight	Value
W_1	100
W_2	10
W_3	5
W_4	40
W_5	10
W_6	10
$\dot{w}_0$	5
$\dot{w}_1$	10
$\dot{w}_2$	15
$\dot{w}_3$	20
$\dot{w}_4$	25
$\dot{w}_5$	30
w^1_{a1}	10
w^1_{a2}	10
w^1_{a3}	10
w^1_{a4}	10
w^1_{a5}	10
w^1_{a6}	10
w^1_{a7}	10
w^1_{a8}	10
w^1_{a9}	10
w^1_{a10}	10
w^2_1	10
w^2_2	10
w^2_3	10
w^2_α	2
w^3	50
w^4_1	5
w^4_2	20
w^5_α	12

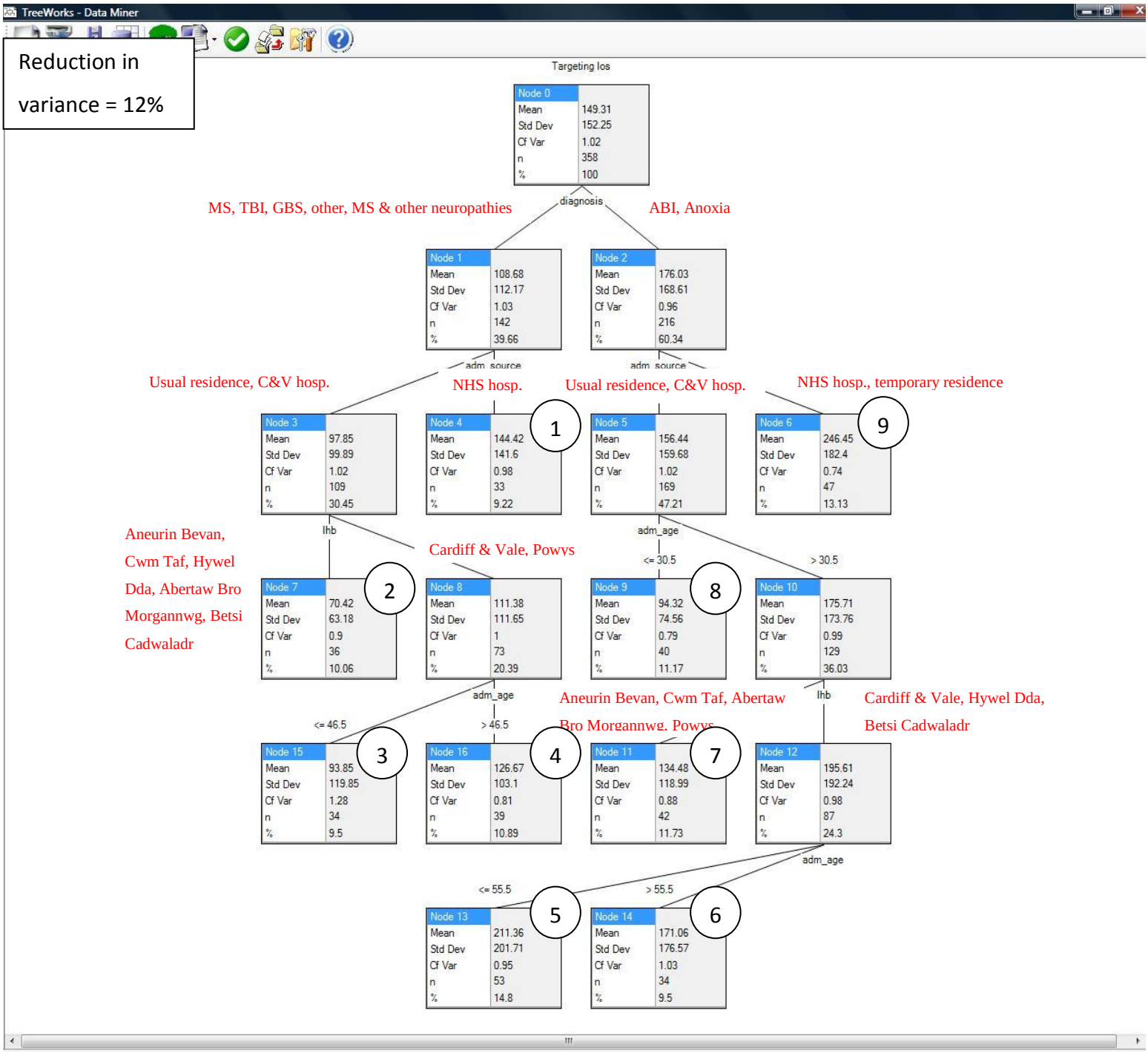
Appendix 6.1

Database – Relationships



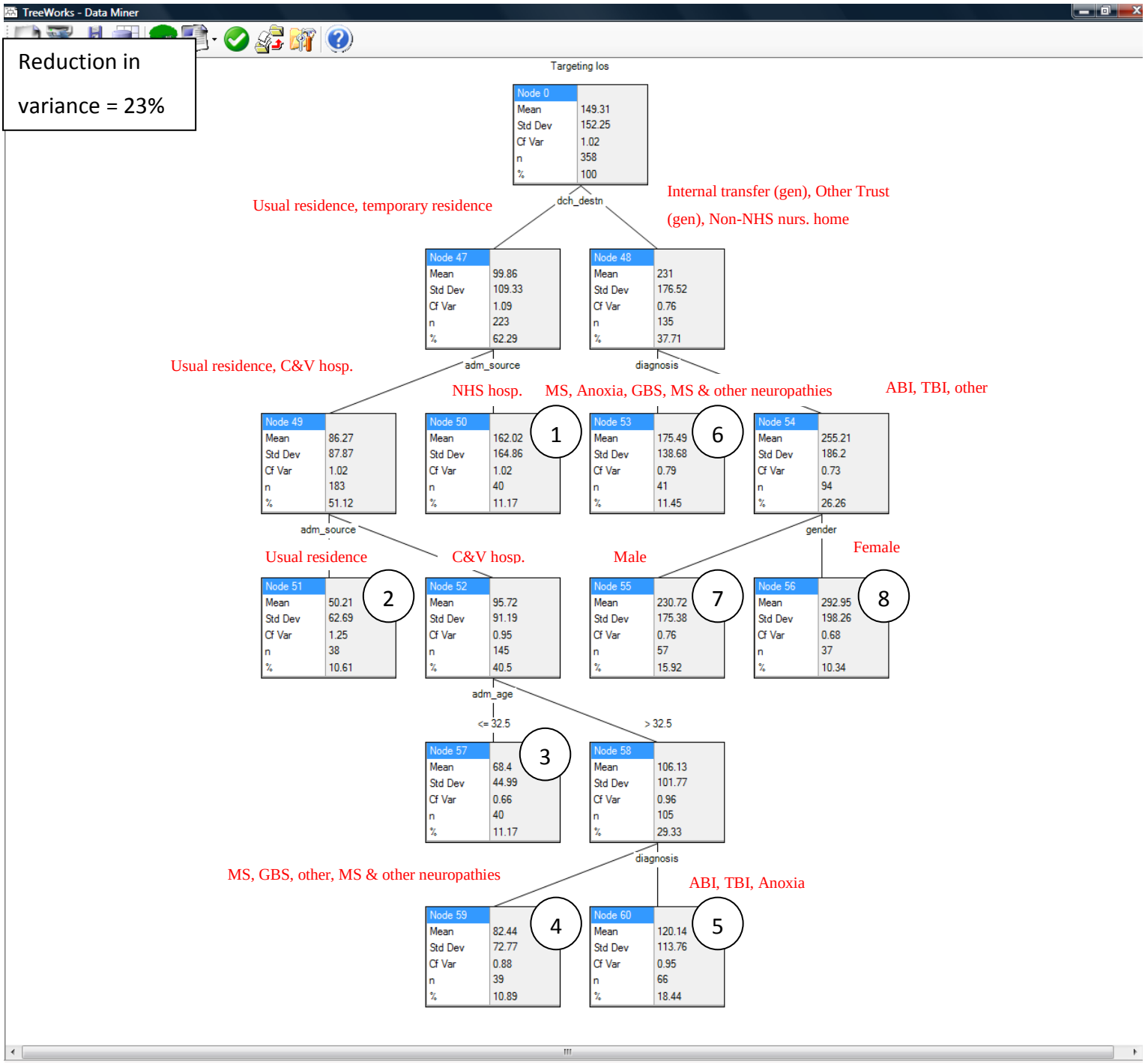
Appendix 9.1

CART Tree (Excl. Discharge Destination)



Appendix 9.2

CART Tree (Incl. Discharge Destination)



Appendix 9.3

Model Two Input File

2 6 14 0.1902
0.0734 0.0157 0.0734
0.08
1 10
1 7 0.75 1.5 0
3 3 7 0.058
0.1238 0.1238 0.0068 0.0380
0 0.1813
1 8
1 4 0.9 1.32 0
3 3 7 0.052
0.0097 0.0328 0.0648 0.0181
0.5 0.6721
1 10
1 4 0.85 1.2 0
2 9 21 0.0925
0.0171 0.0084 0.013
0
1 10
1 10 0.75 1.5 0

References

Aarts, E and Lenstra, JK (2003), *Local Search in Combinatorial Optimization* (Princeton University Press).

Abujudeh, H, Vuong, B, and Baker, SR (2005), 'Quality and operations of portable X-ray examination procedures in the emergency room: queuing theory at work', *Emergency Radiology,* 11 (5), 262-66.

Adan, I., Vandewaarsenburg, WA, and Wessels, J. (1992), 'Analysing E (sub k)/E (sub r)/c queues', *Unpublished Work*.

Ahmad, MI, Sinclair, CD, and Werritty, A (1988), 'Log-logistic flood frequency analysis', *Journal of Hydrology,* 98 (3-4), 205-24.

Aitchison, J and Brown, JAC (1957), *The lognormal distribution* (Cambridge University Press).

Akaike, H. (1974), 'A new look at the statistical model identification', *Automatic Control, IEEE Transactions on,* 19 (6), 716-23.

Al-Fawzan, MA (2000), 'Methods for estimating the parameters of the Weibull distribution', (King Abdulaziz City for Science and Technology, Saudi Arabia Unpublished Work).

Aldous, D and Shepp, L (1987), 'The least variable phase type distribution is Erlang', *Stochastic Models,* 3 (3), 467-73.

Altiok, T (1985), 'On the phase-type approximations of general distributions', *IIE Transactions,* 17 (2), 110-16.

Ashkar, F and Mahdi, S (2006), 'Fitting the log-logistic distribution by generalised moments', *Journal of Hydrology,* 328 (3-4), 694-703.

Asmussen, S, Nerman, O, and Olsson, M (1996), 'Fitting phase-type distributions via the EM algorithm', *Scandinavian Journal of Statistics,* 23 (4), 419-41.

Atkinson, AC (1970), 'A method for discriminating between models', *Journal of the Royal Statistical Society. Series B (Methodological),* 32 (3), 323-53.

Augustin, R and Büscher, KJ (1982), 'Characteristics of the COX-distribution', *ACM Sigmetrics Performance Evaluation Review,* 12 (1), 22-32.

Baber, J, Griffiths, JD, Williams, JE (2008), 'Queues in Series with Blocking', (Cardiff University PhD Thesis).

Badsberg, JH (1992), 'A guide to CoCo–an environment for graphical models', (Aalborg University, Denmark).

Bailey, NTJ (1952), 'Operational research in medicine', *Operational Research Quarterly*, 3 (2), 24-30.

Bailey, NTJ (1954), 'Queuing for medical care', *Applied Statistics*, 3 (3), 137-45.

Balakrishnan, N and Kateri, M (2008), 'On the maximum likelihood estimation of parameters of Weibull distribution based on complete and censored data', *Statistics & Probability Letters*, 78 (17), 2971-75.

Beecham, J, et al. (2009), 'Treatment paths and costs for young adults with acquired brain injury in the United Kingdom', *Brain Injury*, 23 (1), 30-38.

Beggs, S, et al. (1971), 'Evaluation of a system for on-line computer scheduling of patient care activities', *Computers and Biomedical Research*, 4 (6), 634-54.

Black, K, et al. (2000), 'Sitting balance following brain injury: does it predict outcome?', *Brain Injury*, 14 (2), 141-52.

Blackerby, WF (1990), 'Intensity of rehabilitation and length of stay', *Brain Injury*, 4 (2), 167-73.

Bladt, M. (2005), 'A review on phase-type distributions and their use in risk theory', *Astin Bulletin*, 35 (1), 145-61.

Blazewicz, J, et al. (1997), 'Scheduling computer and manufacturing processes', *Journal of the Operational Research Society*, 48 (6), 659-59.

Blazewicz, J, et al. (2007), Handbook on scheduling: from theory to applications (Springer Verlag).

Bobbio, A, Horváth, A, and Telek, M (2005), 'Matching three moments with minimal acyclic phase type distributions', *Stochastic Models*, 21 (2), 303-26.

Boldy, D (1981), Operational research applied to health services (Taylor & Francis).

Bondy, JA and Murty, US (2008), 'Graph theory', (Berlin: Springer).

Bourque, M (1980), 'Patient scheduling in a therapeutic clinic', (American Medical Informatics Association Conference Proceedings), 5 (2), 781.

Boxma, O.J., Cohen, JW, and Huffels, N. (1979), 'Approximations of the mean waiting time in an M/G/s queuing system', *Operations Research*, 27 (6), 1115-27.

Broder, S (1964), 'Final examination scheduling', *Communications of the ACM*, 7 (8), 498.

Brown, L, et al. (2005), 'Statistical analysis of a telephone call centre', *Journal of the American Statistical Association*, 100 (469), 36-50.

Bugnon, B, Stoffel, K, and Widmer, M (1995), 'FUN: A dynamic method for scheduling problems', *European Journal of Operational Research*, 83 (2), 271-82.

Burke, DC (1995), 'Models of brain injury rehabilitation', *Brain Injury,* 9 (7), 735-43.

Burke, E, et al. (1996), 'Examination timetabling in British universities: A survey', *Proceedings of the first PATAT international conference* (Springer), 76-90.

Burke, EK and Kendall, G (2005), Search methodologies: introductory tutorials in optimization and decision support techniques (Springer Verlag).

Burke, EK and Petrovic, S (2002), 'Recent research directions in automated timetabling', *European Journal of Operational Research,* 140 (2), 266-80.

Bush, RN, et al. (1961), 'Using machines to make the high-school schedule', *The School Review,* 69 (1), 48-59.

Campbell, M. (2000), *Rehabilitation for traumatic brain injury: physical therapy practice in context* (Elsevier Health Sciences).

Chakravarthy, SR and Alfa, AS (1997), *Matrix-analytic methods in stochastic models* (Marcel Dekker).

Cheng, J and Greiner, R (2001), 'Learning bayesian belief network classifiers: Algorithms and system', *Advances in Artificial Intelligence,* 141-51.

Chien, CF, Tseng, FP, and Chen, CH (2008), 'An evolutionary approach to rehabilitation patient scheduling: A case study', *European Journal of Operational Research,* 189 (3), 1234-53.

Cifu, DX, et al. (2003), 'The relationship between therapy intensity and rehabilitative outcomes after traumatic brain injury: A multicentre analysis', *Archives of Physical Medicine and Rehabilitation,* 84 (10), 1441-48.

Cochran, JK and Roche, K (2007), 'A queuing-based decision support methodology to estimate hospital inpatient bed demand', *Journal of the Operational Research Society,* 59 (11), 1471-82.

Cohen, AC (1965), 'Maximum likelihood estimation in the Weibull distribution based on complete and on censored samples', *Technometrics,* 7 (4), 579-88.

Collaborators, MRCCT (2008), 'Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients', *British Medical Journal,* 336 (7641), 425.

Colm Art, OC (1999), 'Phase-type distributions: open problems and a few properties', *Comm. Statist. Stochastic Models,* 15 (4), 731-57.

Commault, C and Mocanu, S (2003), 'Phase-type distributions and representations: some results and open problems for system theory', *International Journal of Control,* 76 (6), 566-80.

Cope, DN (1995), 'The effectiveness of traumatic brain injury rehabilitation: a review', *Brain Injury,* 9 (7), 649-70.

Cope, DN and Hall, K (1982), 'Head injury rehabilitation: benefit of early intervention', *Archives of Physical Medicine and Rehabilitation,* 63 (9), 433.

Cosmetatos, G.P. (1976), 'Some approximate equilibrium results for the multi-server queue (M/G/r)', *Operational Research Quarterly,* 27 (3), 615-20.

Cousineau, D, Brown, S, and Heathcote, A (2004), 'Fitting distributions using maximum likelihood: Methods and packages', *Behavior Research Methods, Instruments, & Computers,* 36 (4), 742.

Cowen, TD, et al. (1997), 'Influence of early variables in traumatic brain injury on functional independence measure scores and rehabilitation length of stay and charges', *The Journal of Head Trauma Rehabilitation,* 12 (3), 97.

Cox, DR (1955), 'A use of complex probabilities in the theory of stochastic processes', (Mathematical Proceedings of the Cambridge Philosophical Society), 51, 313-19.

Cox, DR (1962), 'Further results on tests of separate families of hypotheses', *Journal of the Royal Statistical Society. Series B (Methodological),* 24 (2), 406-24.

Cullen, N, et al. (2007), 'The efficacy of acquired brain injury rehabilitation', *Brain Injury,* 21 (2), 113-32.

Cumani, A (1982), 'On the canonical representation of homogeneous Markov processes modelling failure-time distributions', *Microelectronics and reliability,* 22 (3), 583-602.

D'Agostino, RB and Stephens, MA (1986), *Goodness-of-fit techniques* (CRC Press).

Dahiya, RC (1981), 'An improved method of estimating an integer-parameter by maximum likelihood', *American Statistician,* 35 (1), 34-37.

Danielsson, J., et al. (2006), 'Comparing downside risk measures for heavy tailed distributions', *Economics letters,* 92 (2), 202-08.

de Bruin, AM, et al. (2007), 'Modelling the emergency cardiac in-patient flow: an application of queuing theory', *Health Care Management Science,* 10 (2), 125-37.

Dehon, M and Latouche, G (1982), 'A geometric interpretation of the relations between the exponential and generalised Erlang distributions', *Advances in Applied Probability,* 14 (4), 885-97.

Department of Health (1998), 'Report of the national traumatic brain injury study', (London).

Dey, AK and Kundu, D (2010), 'Discriminating between the log-normal and log-logistic distributions', *Communications in Statistics-Theory and Methods,* 39 (2), 280-92.

Dumitrescu, I and Stützle, T (2003), 'Combinations of local search and exact algorithms', *Applications of Evolutionary Computing*, 57-68.

Ehrgott, M and Gandibleux, X (2000), 'A survey and annotated bibliography of multiobjective combinatorial optimization', *OR Spectrum*, 22 (4), 425-60.

Elmaghraby, SE and Park, SH (1974), 'Scheduling jobs on a number of identical machines', *IIE Transactions*, 6 (1), 1-13.

Embling, S (1995), 'Exploring multidisciplinary teamwork', *British Journal of Therapy and Rehabilitation*, 2 (3), 142–44.

Erlang (1917), 'Solution of some probability problems of significance for automatic telephone exchanges', *Elektroteknikeren, Copenhagen*, 13.

Erlang, AK (1909), 'The theory of probabilities and telephone conversations', *Nyt Tidsskrift for Matematik B*, 20, 33.

Ernst, AT, et al. (2004), 'Staff scheduling and rostering: A review of applications, methods and models', *European Journal of Operational Research*, 153 (1), 3-27.

Fackrell, M (2009), 'Modelling healthcare systems with phase-type distributions', *Health care management science*, 12 (1), 11-26.

Faddy, M, Graves, N, and Pettitt, A (2009), 'Modelling length of stay in hospital and other right skewed data: comparison of phase-type, gamma and log-normal distributions', *Value in Health*, 12 (2), 309-14.

Faddy, MJ (1994), 'Examples of fitting structured phase-type distributions', *Applied Stochastic Models and Data Analysis*, 10 (4), 247-55.

Faddy, MJ (1995), 'Phase-type distributions for failure times', *Mathematical and Computer Modelling*, 22 (10-12), 63-70.

Faddy, MJ (1998), 'On inferring the number of phases in a Coxian phase-type distribution', *Stochastic Models*, 14 (1), 407-17.

Faddy, MJ and McClean, SI (1999), 'Analysing data on lengths of stay of hospital patients using phase-type distributions', *Applied Stochastic Models in Business and Industry*, 15 (4), 311-17.

Faddy, MJ and McClean, SI (2005), 'Markov chain modelling for geriatric patient care', *Methods of information in medicine*, 44 (3), 369.

Feigenson, JS, et al. (1977), 'Factors influencing outcome and length of stay in a stroke rehabilitation unit. Part 1. Analysis of 248 unscreened patients – medical and functional prognostic indicators', *Stroke*, 8 (6), 651.

Ferguson, TS (1996), *A course in large sample theory* (Chapman & Hall/CRC Press).

Finch, M, et al. (1997), 'Admission examination factors predicting cognitive improvement during acute brain injury rehabilitation', *Brain Injury*, 11 (10), 713-22.

Fischer, MJ, et al. (2001), 'Analyzing the waiting time process in internet queuing systems with the transform approximation method', *The Telecommunications Review*, 12, 21-32.

Fomundam, S and Herrmann, JW (2007), 'A survey of queuing theory applications in healthcare', (University of Maryland Technical Report).

Fraile, R and García-Ortega, E (2005), 'Fitting an exponential distribution', *Journal of Applied Meteorology*, 44 (10), 1620-25.

Garg, L, et al. (2009), 'A phase type survival tree model for clustering patients' hospital length of stay', *ASMDA conference proceedings*.

Garthwaite, PH, Jolliffe, IT, and Jones, B (2002), *Statistical inference* (Oxford University Press, USA).

Ghalebsaz-Jeddi, B, Donohue, GL, and Shortle, JF (2009), 'A statistical analysis of the aircraft landing process', *Journal of Industrial and Systems Engineering*, 3 (3), 152-69.

Glover, F (1990), 'Tabu search: A tutorial', *Interfaces*, 20 (4), 74-94.

Gonzalez, T and Sahni, S (1976), 'Open shop scheduling to minimise finish time', *Journal of the ACM (JACM)*, 23 (4), 665-79.

Goodeve, C (1953), 'Operational research as a science', *Journal of the Operations Research Society of America*, 1 (4), 166-80.

Gove, JH (2003), 'Moment and maximum likelihood estimators for Weibull distributions under length-and area-biased sampling', *Environmental and Ecological Statistics*, 10 (4), 455-67.

Greiner, M, Jobmann, M, and Lipsky, L (1999), 'The importance of power-tail distributions for modelling queuing systems', *Operations Research*, 47 (2), 313-26.

Griffiths, JD, Lawson, ZF, and Williams, JE (2006), 'Modelling treatment effects in the HIV/AIDS epidemic', *Journal of the Operational Research Society*, 57 (12), 1413-24.

Griffiths, JD, et al. (2006), 'A queuing model of activities in an intensive care unit', *IMA Journal of Management Mathematics*, 17 (3), 277.

Gross, D and Juttijudata, M (1997), 'Sensitivity of output performance measures to input distributions in queuing simulation modelling', (Proceedings of the 1997 Winter Simulation Conference), 296-302.

Hall, KM, et al. (1996), 'Functional measures after traumatic brain injury: ceiling effects of FIM, FIM+ FAM, DRS, and CIQ', *The Journal of Head Trauma Rehabilitation,* 11 (5), 27.

Hansen, P and Mladenovic, N (2006), 'First vs. best improvement: An empirical study', *Discrete Applied Mathematics,* 154 (5), 802-17.

Hanson, BL (1973), 'A statistical model for length of stay in a mental hospital', *Health Services Research,* 8 (1), 37.

Harper, PR (2002), 'A framework for operational modelling of hospital resources', *Health Care Management Science,* 5 (3), 165-73.

Harrison, GW and Millard, PH (1991), 'Balancing acute and long-term care: the mathematics of throughput in departments of geriatric medicine', *Methods of Information in Medicine,* 30 (3), 221.

Heffer, JC (1969), 'Steady-state solution of the M/Ek/C (FIFO) queuing system', *CORS Journal,* 7, 16-30.

Heinemann, AW, et al. (1995), 'Functional status and therapeutic intensity during inpatient rehabilitation', *American Journal of Physical Medicine & Rehabilitation,* 74 (4), 315.

Horváth, A and Telek, M (2002), 'Phfit: A general phase-type fitting tool', *Computer Performance Evaluation: Modelling Techniques and Tools*, 1-14.

Horváth, A and Telek, M (2007), 'Matching more than three moments with acyclic phase type distributions', *Stochastic Models,* 23 (2), 167-94.

Härdle, W. and Simar, L. (2007), *Applied multivariate statistical analysis* (Springer Verlag).

Irvine, V, McClean, SI, and Millard, P (1994), 'Stochastic models for geriatric in-patient behaviour', *Mathematical Medicine and Biology,* 11 (3), 207.

Jennett, B and Bond, M (1975), 'Assessment of outcome after severe brain damage: A practical scale', *The Lancet,* 305 (7905), 480-84.

Johnson, MA and Taaffe, MR (1989), 'Matching moments to phase distributions: Mixtures of Erlang distributions of common order', *Stochastic Models,* 5 (4), 711-43.

Johnson, MA and Taaffe, MR (1990), 'Matching moments to phase distributions: Nonlinear programming approaches', *Stochastic Models,* 6 (2), 259-81.

Johnson, MA and Taaffe, MR (1991), 'An investigation of phase-distribution moment-matching algorithms for use in queuing models', *Queuing Systems,* 8 (1), 129-47.

Karlin, S and Studden, WJ (1966), Tchebycheff systems: with applications in analysis and statistics (Wiley New York).

Keith, RA, et al. (1987), 'The functional independence measure: a new tool for rehabilitation', *Advances in clinical rehabilitation,* 1, 6.

Kelley, CT (1999), *Iterative methods for optimization* (Society for Industrial Mathematics).

Kendall, DG (1953), 'Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain', *The Annals of Mathematical Statistics,* 24 (3), 338-54.

Kendall, G (2007), 'Scheduling English football fixtures over holiday periods', *Journal of the Operational Research Society,* 59 (6), 743-55.

Kiviat, P and Colker, A (1964), 'GASP – A general activity simulation program', (RAND Technical Report).

Knessl, C., et al. (1990), 'An integral equation approach to the M/G/2 queue', *Operations Research,* 38 (3), 506-18.

Kolker, A (2008), 'Process modelling of emergency department patient flow: Effect of patient length of stay on ED diversion', *Journal of Medical Systems,* 32 (5), 389-401.

Lan, S, Clarke, JP, and Barnhart, C (2006), 'Planning for robust airline operations: Optimising aircraft routings and flight departure times to minimise passenger disruptions', *Transportation Science,* 40 (1), 15-28.

Lawrence, RJ (1984), 'The lognormal distribution of the duration of strikes', *Journal of the Royal Statistical Society. Series A (General),* 147 (3), 464-83.

Lee, D and Hahn, SH (1970), 'An analysis of short-stay hospital records and measurement of the probability discharged as cured from the severance hospital, 1967~ 1969', *Yonsei Medical Journal,* 11 (1).

Lehmkuhl, DI, et al. (1993), 'Factors that influence costs and length of stay of persons with traumatic brain injury in acute care and inpatient rehabilitation', *The Journal of Head Trauma Rehabilitation,* 8 (2), 88.

Little, J.D.C. (1961), 'A proof of the queuing formula L= lambda W', *Operations Research,* 9 (3), 383-87.

Lo, VM (1988), 'Heuristic algorithms for task assignment in distributed systems', *IEEE Transactions on Computers,* 37 (11), 1384-97.

Maaløe, E. (1973), 'Approximation formulae for estimation of waiting-time in multiple-channel queuing system', *Management Science,* 19 (6), 703-10.

Mackay, LE, et al. (1992), 'Early intervention in severe head injury: Long-term benefits of a formalised program', *Archives of Physical Medicine and Rehabilitation,* 73 (7), 635.

Mahoney, FI and Barthel, DW (1965), 'Functional evaluation: the Barthel index', *Maryland State Medical Journal,* 14, 61.

Mandy, P (1996), 'Interdisciplinary rather than multidisciplinary or generic practice', *British Journal of Therapy and Rehabilitation,* 3, 110-12.

Marazzi, A, et al. (1998), 'Fitting the distributions of length of stay by parametric models', *Medical Care,* 36 (6), 915-27.

Marshall, AH and McClean, SI (2003), 'Conditional phase-type distributions for modelling patient length of stay in hospital', *International Transactions in Operational Research,* 10 (6), 565-76.

Marshall, AH, et al. (2001), 'Developing a Bayesian belief network for the management of geriatric hospital care', *Health Care Management Science,* 4 (1), 25-30.

Marshall, AH and Zenga, M (2009), 'Recent developments in fitting coxian phase-type distributions in healthcare', *ASMDA conference proceedings*.

Marshall, AH and Zenga, M (2010), 'Experimenting with the Coxian phase-type distribution to uncover suitable fits', *Methodology and Computing in Applied Probability*, 1-16.

Max, W, MacKenzie, EJ, and Rice, DP (1991), 'Head injuries: costs and consequences', *The Journal of Head Trauma Rehabilitation,* 6 (2), 76.

Mayhew, L and Smith, D (2008), 'Using queuing theory to analyse the government's 4-h completion time target in accident and emergency departments', *Health Care Management Science,* 11 (1), 11-21.

Mayhugh, JO and McCormick, RE (1968), 'Steady state solution of the queue M/E k/r', *Management Science,* 14 (11), 692-712.

McClean, S and Millard, P (1993), 'Patterns of length of stay after admission in geriatric medicine: an event history approach', *The Statistician,* 42 (3), 263-74.

McClean, S and Millard, P (2007), 'Where to treat the older patient? Can Markov models help us better understand the relationship between hospital and community care?', *Journal of the Operational Research Society,* 58 (2), 255-61.

Millard, PH (1988), 'Geriatric medicine A new method of measuring bed usage and a theory for planning'.

Miller, GK (1999), 'Maximum likelihood estimation for the Erlang integer parameter', *Statistics & Probability Letters,* 43 (4), 335-41.

Miller, GK and Bhat, UN (1997), 'Estimation for renewal processes with unobservable gamma or Erlang interarrival times', *Journal of Statistical Planning and Inference,* 61 (2), 355-72.

Miller, HE, Pierskalla, WP, and Rath, GJ (1976), 'Nurse scheduling using mathematical programming', *Operations Research,* 24 (5), 857-70.

Minka, TP (2000), 'Beyond Newton's method', (Microsoft Research Technical Report).

Minka, TP (2002), 'Estimating a Gamma distribution', (Microsoft Research Technical report).

Nance, RE (1993), ‘A history of discrete event simulation programming languages’, (Technical report, Virginia Polytechnic Institute)

National Institute for Clinical Excellence (2003) ‘NICE 2003/ 024a Press Release’, (NICE)

Nozaki, S.A. and Ross, S.M. (1978), 'Approximations in finite-capacity multi-server queues with Poisson arrivals', *Journal of Applied Probability*, 15 (4), 826-34.

Ogulata, SN, Koyuncu, M, and Karakas, E (2008), 'Personnel and patient scheduling in the high demanded hospital services: a case study in the physiotherapy service', *Journal of Medical Systems*, 32 (3), 221-28.

Olsson, M (1998), 'The empht-programme', (Gothenburg University, Gothenburg, Sweden).

Osborne, MR (1992), 'Fisher's method of scoring', International Statistical Review, 60 (1), 99-117.

Osman, IH and Laporte, G (1996), 'Metaheuristics: A bibliography', *Annals of Operations Research*, 63 (5), 511-623.

Osogami, T and Harchol-Balter, M (2003), 'A closed-form solution for mapping general distributions to minimal PH distributions', *Computer Performance*, 200-17.

O’Cinneide, CA (1989), 'On non-uniqueness of representations of phase-type distributions', *Stochastic Models*, 5 (2), 247-59.

Ozgun, O and Barlas, Y (2009), ‘Discrete vs continuous simulation: when does it matter?’, *International Conference of the System Dynamics Society*

Papadopoulos, HT (1998), 'Analysis of production lines with Coxian service times and no intermediate buffers', *Naval Research Logistics*, 45 (7), 669-85.

Pawitan, Y (2001), In all likelihood: statistical modelling and inference using likelihood (Oxford University Press, USA).

Perel, P, et al. (2006), 'Systematic review of prognostic models in traumatic brain injury', *BMC Medical Informatics and Decision Making*, 6 (1), 38.

Perez-Hoyos, S, et al. (2000), 'Length of stay in a hospital emergency room due to asthma and chronic obstructive pulmonary disease: Implications for air pollution studies', *European Journal of Epidemiology*, 16 (5), 455-63.

Petrovic, S and Berghe, GV (2008), 'Comparison of algorithms for nurse rostering problems (*Procedings of the 7th International Conference on the Practice and Theory of Automated Timetabling)*.

Pitsoulis, LS and Resende, MGC (2002), 'Greedy randomised adaptive search procedures', *Handbook of Applied Optimisation* (AT+T Labs), 168–83.

Pla, V., Casares-Giner, V., and Martínez, J. (2004), 'On a multiserver finite buffer queue with impatient customers', *ITC Specialist Seminar on Access Networks*

Poyntz, CD and Jackson, RRP (1973), 'The steady-state solution for the queuing process Ek/Em/r', *Operational Research Quarterly*, 24 (4), 615-25.

Prabhakar Murthy, DN, Bulmer, M, and Eccleston, JA (2004), 'Weibull model selection for reliability modelling', *Reliability Engineering & System Safety*, 86 (3), 257-67.

Puchinger, J and Raidl, GR (2005), 'Combining metaheuristics and exact algorithms in combinatorial optimization: A survey and classification', *Artificial Intelligence and Knowledge Engineering Applications: a Bioinspired Approach*, 41-53.

Pulungan, R and Hermanns, H (2008), 'Effective minimization of acyclic phase-type representations', *Analytical and Stochastic Modelling Techniques and Applications*, 128-43.

Rappaport, M, et al. (1982), 'Disability rating scale for severe head trauma: coma to community', *Archives of Physical Medicine and Rehabilitation*, 63 (3), 118.

Reeves, CR (1993), *Modern heuristic techniques for combinatorial problems* (John Wiley & Sons, Inc. New York, NY, USA).

Resende M and Ribeiro C (2003), 'Greedy randomised adaptive search procedures', *Operations Research and Management Science*, 57, 219-249.

Rider, PR (1961), 'The method of moments applied to a mixture of two exponential distributions', *The Annals of Mathematical Statistics*, 32 (1), 143-47.

Ruppert, D. (2010), Statistics and Data Analysis for Financial Engineering (Springer Verlag).

Sandhaug, M, et al. (2010), 'Functional level during sub-acute rehabilitation after traumatic brain injury: Course and predictors of outcome', *Brain Injury*, 24 (5), 14-447.

Schaerf, A (1999), 'A survey of automated timetabling', *Artificial Intelligence Review*, 13 (2), 87-127.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The annals of Statistics*, 6 (2), 461-64.

Scott, D. W. (1979), 'On optimal and data-based histograms', *Biometrika*, 66 (3), 605.

Select Committee on Health (2000-2001), 'Third Report', (House of Commons)

Select Committee on Health (2001-2002), 'Third Report', (House of Commons)

Severini, TA (2000), Likelihood methods in statistics (Oxford University Press, USA).

Shapiro, S (1966), 'The M-server queue with Poisson input and gamma-distributed service of order two', *Operations Research*, 14 (4), 685-694.

Shen, M, Tzeng, GH, and Liu, DR (2003), 'Multi-criteria task assignment in workflow management systems', Hawaii International Conference on System Sciences, 9.

Shiel, A, et al. (2001), 'The effects of increased rehabilitation therapy after brain injury: results of a prospective controlled trial', *Clinical Rehabilitation,* 15 (5), 501.

Shortle, JF, et al. (2003), 'Using the transform approximation method to analyse queues with heavy-tailed service', *Journal of Probability and Statistical Science,* 1 (1), 15-27.

Shoukri, MM, Mian, IUH, and Tracy, DS (1988), 'Sampling properties of estimators of the log-logistic distribution with application to Canadian precipitation data', *Canadian Journal of Statistics,* 16 (3), 223-36.

Singh, VP and Guo, H (1995), 'Parameter estimation for 2-parameter log-logistic distribution (LLD2) by maximum entropy', *Civil Engineering and Environmental Systems,* 12 (4), 343-57.

Sivenius, J, et al. (1985), 'The significance of intensity of rehabilitation of stroke – a controlled trial', *Stroke,* 16 (6), 928.

Slade, A, Tennant, A, and Chamberlain, MA (2002), 'A randomised controlled trial to determine the effect of intensity of therapy upon length of stay in a neurological rehabilitation setting', *Journal of Rehabilitation Medicine,* 34 (6), 260-66.

Smith, DS, et al. (1981), 'Remedial therapy after stroke: a randomised controlled trial', *British Medical Journal,* 282 (6263), 517.

Sobieraj, JE (2006), 'Queuing analysis of a general medical practice', (Boston University Unpublished Work).

Spivack, G, et al. (1992), 'Effects of intensity of treatment and length of stay on rehabilitation outcomes', *Brain Injury,* 6 (5), 419-34.

Srivastava, PW and Shukla, R (2008), 'A log-logistic step-stress model', *Reliability, IEEE Transactions on,* 57 (3), 431-34.

Steadman, HJ, et al. (2001), 'Assessing the New York City involuntary outpatient commitment pilot program', *Psychiatric Services,* 52 (3), 330.

Stewart, WJ (2009), Probability, Markov chains, queues, and simulation: the mathematical basis of performance modelling (Princeton University Press).

Sturges, H. A. (1926), 'The choice of a class interval', *Journal of the American Statistical Association,* 21 (153), 65-66.

Stützle, DWTG (1998), 'Local search algorithms for combinatorial problems', (Darmstadt University of Technology PhD Thesis).

Svolba, G (2006), Data preparation for analytics using SAS (SAS Publishing).

Taylor, GJ, McClean, SI, and Millard, PH (1998), 'Using a continuous-time Markov model with Poisson arrivals to describe the movements of geriatric patients', *Applied Stochastic Models and Data Analysis,* 14 (2), 165-74.

Teasdale, G and Jennett, B (1974), 'Assessment of coma and impaired consciousness: A practical scale', *The Lancet,* 304 (7872), 81-84.

Teasell, R, et al. (2007), 'A systematic review of the rehabilitation of moderate to severe acquired brain injuries', *Brain Injury,* 21 (2), 107-12.

Telek, M and Heindl, A (2002), 'Matching moments for acyclic discrete and continuous phase-type distributions of second order', *International Journal of Simulation,* 3 (3-4), 47.

Turner-Stokes, L (2007), 'Cost-efficiency of longer-stay rehabilitation programmes: Can they provide value for money?', *Brain Injury,* 21 (10), 1015-21.

Turner-Stokes, L and Nyein, K (1999), 'The northwick park care needs assessment (NPCNA): a directly costable outcome measure in rehabilitation', *Clinical Rehabilitation,* 13 (3), 253.

Turner-Stokes, L, Paul, S, and Williams, H (2006), 'Efficiency of specialist rehabilitation in reducing dependency and costs of continuing care for adults with complex acquired brain injuries', *Journal of Neurology, Neurosurgery & Psychiatry,* 77 (5), 634.

Turner-Stokes, L, et al. (1998), 'The northwick park dependency score (NPDS): a measure of nursing dependency in rehabilitation', *Clinical Rehabilitation,* 12 (4), 304.

Turner-Strokes, L (2003), *Rehabilitation following acquired brain injury: national clinical guidelines* (Royal College of Physicians).

Van Der Heijden, MC (1988), 'On the three-moment approximation of a general distribution by a Coxian distribution', *Probability in the Engineering and Informational Sciences,* 2 (02), 257-61.

Van der Putten, J, et al. (1999), 'Measuring change in disability after inpatient rehabilitation: comparison of the responsiveness of the Barthel index and the functional independence measure', *Journal of Neurology, Neurosurgery & Psychiatry,* 66 (4), 480.

Victoria-Feser, MP (1997), 'A robust test for non-nested hypotheses', *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 59 (3), 715-27.

Vreeman, D.J., et al. (2006), 'Evidence for electronic health record systems in physical therapy', *Physical Therapy,* 86 (3), 434.

Wade, DT (1987), 'Neurological rehabilitation', *Disability & Rehabilitation,* 9 (1), 45-47.

Wagner, U and Geyer, ALJ (1995), 'A maximum entropy method for inverting Laplace transforms of probability density functions', *Biometrika,* 82 (4), 887.

Wang, Z (2008), 'Income distribution, market size and the evolution of industry', *Review of Economic Dynamics,* 11 (3), 542-65.

Whitlock, JA (1995), 'Functional outcome after rehabilitation for severe traumatic brain injury', *Archives of Physical Medicine and Rehabilitation,* 76 (12), 1103-12.

Whitt, W (1982), 'Approximating a point process by a renewal process, I: Two basic methods', *Operations Research,* 30 (1), 125-47.

Whitt, W (1984), 'On approximations for queues, III: Mixtures of exponential distributions', *AT&T Bell Laboratories technical journal: a journal of the AT&T companies,* 163.

Wilson, JTL, Pettigrew, LEL, and Teasdale, GM (1998), 'Structured interviews for the Glasgow outcome scale and the extended Glasgow outcome scale: guidelines for their use', *Journal of Neurotrauma,* 15 (8), 573-85.

Woeginger, G (2003), 'Exact algorithms for NP-hard problems: A survey', *Combinatorial Optimization—Eureka, You Shrink!,* 185-207.

Wren, A (1996), 'Scheduling, timetabling and rostering – a special relationship?', *Practice and Theory of Automated Timetabling: Lecture Notes in Computer Science,* 1153, 46-75.

Wright, MH, et al. (1998), 'Convergence properties of the Nelder—Mead simplex method in low dimensions', *SIAM Journal of Optimisation,* 9 (1), 112–47.

Wu, MY, Shu, W, and Gu, J (1997), 'Local search for DAG scheduling and task assignment', *Proceedings of the 1997 International Conference on Paralell Processing,* 174-81.

Xie, H, Chaussalet, TJ, and Millard, PH (2005), 'A continuous time Markov model for the length of stay of elderly people in institutional long-term care', *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 168 (1), 51-61.

Yen, JW and Birge JR (2006), 'A stochastic programming approach to the airline crew scheduling problem', *Transportation Science,* 40 (1), 3-14.

Young, JP (1965), 'Stabilization of inpatient bed occupancy through control of admissions', *Hospitals,* 39 (19), 41.

Zasler, ND (1997), 'Prognostic indicators in medical rehabilitation of traumatic brain injury: a commentary and review', *Archives of Physical Medicine and Rehabilitation,* 78 (8S4), 12-16.

Zhu, XL, et al. (2007), 'Does intensive rehabilitation improve the functional outcome of patients with traumatic brain injury (TBI)? A randomised controlled trial', *Brain Injury,* 21 (7), 681-90.