The Impact of non-Poisson Arrival Processes in Health Care Facilities with Finite Capacity

Christian Rohrbeck

Supervisor: Dave Worthington

Abstract

Queueing theory is widely used in health care applications to find the optimal level of staff and capacity. In some applications, e.g. in a hospital, the performance is measured by the blocking probability, i.e. the percentage of patients lost due to balking. In such systems, the arrivals are usually considered to have exponentially distributed inter-arrival times. However, studies indicate that this assumption is not always reasonable, as the variance is underestimated or overestimated by using Poisson arrival processes. This work considers approaches to approximate or simulate the blocking probability for a non-Poisson arrival process. In particular, the approximation method by Bekker and Koeleman (2011) is evaluated by a simulation study based on Nelson and Gerhardt (2011). The results evidence a high sensitivity of the blocking probability to variability, dependency and non-stationarity of the arrival process.

1 Introduction

In many countries, the health care system is operated or at least supported by national institutions. The National Health Service (NHS) in the United Kingdom, for instance, is primarily funded by general taxation rather than by private insurance payments and provides comprehensive health services. Currently, the costs increase due to the demographic change (Seshamani and Gray, 2004) and technical innovations (Bodenheimer, 2005). Consequently, the budget of the NHS has to increase. However, due to recession and national debts in default, the government cannot subsidise the system with much money. Hence, the government decided to reform the NHS in 2012, in order to provide more efficient and effective care. In other words, costs caused by ineffectiveness in the health care system should be reduced. Queueing theory is a method to detect such ineffectivenesses and to minimise the total costs. The total costs can be divided into waiting costs, i.e. costs associated with patients having to wait for service and capacity costs, i.e. costs of providing the service (Singh, 2006). Thus, queueing theory is used to find the best compromise between the patients who want fast and good service and the health care providers who only have a finite budget they want to spend effectively.

Several health care facilities can be formulated as a queueing system, e.g. flu vaccination or the collection of medicine. Flu vaccination in its simplest form can be modelled as a system with one queue and one server. The patients arrive at the health center and go to the nurse surgery to get the vaccine. If the nurse is already busy, the patients have to queue, i.e. to sit in the waiting room. The pharmacy store has, in contrast to the flu vaccination, usually more than one server to handle the demand for service. In both examples, patients only need one service. Kendal (1953) proposes a standard notation for queueing systems where patients only require one service before leaving; the notation is $(A/B/S/d/e)$. Here, A and B denote the probability distributions of the inter-arrival times and service times respectively. Following Kendal (1953), exponentially distributed inter-arrival times are denoted by M and general distributions by G. The parameter S denotes the number of servers, d the maximal number of patients allowed in the system and e the queuing discipline.

In a lot of applications, for instance flu vaccination or collection of medicine, the queuing discipline is first in first out (FIFO). Furthermore, the arrivals are usually considered to be independent with exponentially distributed inter-arrival times with parameter λ. The number of arrivals at time point t is thus Poisson with parameter λt. Additionally, it is often assumed that the service times are identically and independent distributed with mean $1/\mu$ and independent of the length of the queue.

One important property of a queueing system is the expected number of patients $\mathbb{E}(n)$. Under the assumptions above and for $d = \infty$, $\mathbb{E}(n)$ can be calculated for specific settings of B and S. First, if the service times are also exponentially distributed and $\lambda < \mu \cdot S$, $\mathbb{E}(n)$ and the steady-state behaviour of the system can be calculated directly. In particular, $\mathbb{E}(n)$ and the steady-state behaviour are determined by λ, μ and S; see, for instance, Worthington (2009) for details. Second, if $S = 1$ and $\lambda < \mu$, $\mathbb{E}(n)$ is determined by the Pollaczek– Khinchine formula; for a proof see Gross and Harris (1985). It states that the expected number of patients in the system only depends on λ, μ and the variance of the service times, σ^2, but not on the service time distribution itself. Formally,

$$\mathbb{E}(n) = \frac{\lambda}{\mu} + \frac{\lambda^2(1/\mu^2 + \sigma^2)}{2(1 - \lambda/\mu)}. \tag{1}$$

Nevertheless, the assumption of constant arrival rates seems not reasonable in a lot of applications like accident and emergency (A&E) units, see for example Lane et al. (2000). If $S = \infty$, the expected number of patients in the system at time point t is determined by

$$\mathbb{E}(n(t)) = \int_{-\infty}^{t} \lambda(u)F(t-u)du, \tag{2}$$

where $F(t-u) = \mathbb{P}(\text{service time} > t-u)$. Aside from $\mathbb{E}(n)$, the expected waiting time $\mathbb{E}(W)$ is a further important value to evaluate the performance of a queueing system. According to Little's formula (Little, 1961), the expected time a patient spends in the system, $\mathbb{E}(W)$, determines to the quotient of the expected number in the system, $\mathbb{E}(n)$ and the mean arrival rate $\mathbb{E}(\lambda)$. Formally,

$$\mathbb{E}(W) = \frac{\mathbb{E}(n)}{\mathbb{E}(\lambda)}. \tag{3}$$

Despite its flexibility, the notation by Kendal (1953) cannot capture any system where patients need more than one service. For example, the appointment practice in health care centres is a network of queues, i.e. patients need more than one service. First, patients queue at the reception, get an appointment and then come back some days later to see the doctor. Consequently, each patient requires service at two servers, reception and doctor. Creemers and Lambrecht (2009) show that the inter-arrival times can be considered to be exponentially distributed. If also the service times are exponentially distributed and independent from each other, the system can be formulated as a Jackson Network; see (Jackson, 1957) for details.

A more complicated instance is the modelling of ambulance and emergency units (A&E) in hosipi-

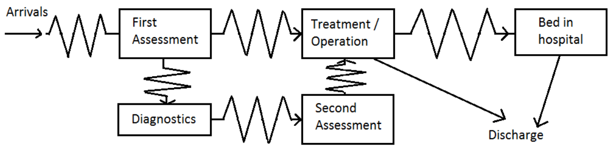

Figure 1: Modelling of an ambulance and emergency department

tals, see Figure 1. Patients arrive at the hospital and the doctor assesses the severity of emergency. If the doctor is unsure about the cause of emergency, the patient is sent to the diagnostics centre and then assessed again. As soon as the cause is known, the patient is treated. After the treatment, the patient either has to stay in hospital to regenerate or is discharged directly. In simulations, it is necessary to consider that both, assessment and treatment, use the same medical staff. In a wide range of publications, such as Deo and Itai (2011) and Izady and Worthington (2012), the inter-arrival times for the emergency department are considered to be exponentially distributed.

But is the assumption of independent exponentially distributed inter-arrival times always reasonable? The disadvantage of this assumption is that the variability of the arrival process is determined by the mean. Current studies indicate that the variance of some arrival processes in health care applications is not well fitted by Poisson. One example is the arrival process of patients at their appointment dates. The reception usually anticipates a mean service time of about 10 to 15 minutes and therefore schedule the appointment dates in this way. Nevertheless, the work by Fontantesi et al. (2002) shows that this is not reasonable as patients often arrive in clusters due to bus schedules, parking space availability, etc. Fontantesi et al. (2002) propose to perform group scheduling and variable sized-blocks of appointment times. Another example is the study by McManus et al. (2003) on surgical caseload that indicates that the variability in scheduled admissions is very high and exceeds that of emergencies. The work by Litvak et al. (2005) implies that reducing the variability has a direct effect on the patients, as the stress of the nursing staff decreases with decreasing variability. Therefore, there is a high demand for methods and results for non-exponentially distributed inter-arrival times in order to provide the optimal level of care.

Recent publications aim to approximate and simulate the impact of non-Poisson arrival processes. Bekker and Koeleman (2011) consider the problem described by McManus et al. (2003) and aim to schedule admissions in the corresponding system which has finite capacity. In particular, Bekker and Koeleman (2011) focus on stationary as well as non-stationary arrival rates and explore existing approximation methods. The results evidence a sensibility of the system to non-stationarity and variability. However, a possible dependency in the arrival process is ignored in the approximations.

Nelson and Gerhardt (2011) provide an algorithm for simulating non-stationary, non-Poisson arrival processes with possible dependency between the arrivals. The algorithm extends the work by Gerhardt and Nelson (2009) in that the methods are also applicable to dependent arrivals. This work aims to apply this simulation method to evaluate the results by Bekker and Koeleman (2011) and to investigate the sensitivity of the system to dependency in the arrival process.

The remainder of this work is organised as follows: In Section 2, the approach by Bekker and Koeleman (2011) is described. First, stationary and then non-stationary arrival rates are considered. In Section 3, the approach by Nelson and Gerhardt (2011) is presented that allows to simulate dependent arrivals with any choice of the first two moments. In Section 4, this simulation method is applied to the framework of Bekker and Koeleman (2011). The work concludes with Section 5, where the contribution of this report is discussed.

2 Bekker and Koelemann (2011)

One important task for a hospital manager is to know the bed demand for a week. The bed demand itself depends on scheduled admissions and emergency admissions. According to Bekker and Koeleman (2011), the arrivals of emergency admissions are well modelled by a Poisson process. However, the studies by McManus et al. (2003) and de Bruin et al. (2010) evidence a higher variability in scheduled admissions than in emergency admissions. This variability is highly affected by the schedule of operations - scheduled operations are usually done on weekdays and their number can vary significantly. Therefore, the arrival process of scheduled admissions is non-Poisson. Additionally, the number of elective patients is very small on weekends. As mentioned in Section 1, reducing the variability would result in less stressed nurses and a higher patient safety. Currently, the number of staffed beds in hospitals is higher on weekdays than on the weekend, partly due to higher staffing costs.

Bekker and Koeleman (2011) use existing approximations for the corresponding queueing system with a non-Poisson arrival process with independent arrivals. Furthermore, the services are assumed to be independent and identically distributed. The approach by Bekker and Koeleman (2011) considers the usage of different admission quota per day of the week. Further, patients are divided into different types, taking differences in the length of stay (LOS) into account. However, this section focuses on introducing the approximations used by Bekker and Koeleman (2011). First, the stationary and then the non-stationary case with time-dependent arrivals is considered. In particular, the arrival rates are assumed to be constant during a day and do not change weekly. Both approaches are explained in the following and numerical examples are given. In Section 4, these numerical examples are then compared to the results of the simulation study.

2.1 Stationary arrivals

In a hospital, the capacity is equal to the number of beds s. If all beds are occupied, an arriving patient is rejected and sent to another hospital, which means the patient does not queue. As each patient only requires one service, i.e. to stay at the ward, the system can be notated based on Kendal (1953). As a patient is either served directly or rejected, the parameter d in the notation is equal to s. Patients arrive with rate λ and L denotes the length of stay at the ward with mean $1/\mu$. The performance of the hospital is evaluated by the blocking probability, i.e. the probability that an arriving patient is rejected. A too high blocking probability indicates the necessity to increase the bed capacity s. If the arrival process would be Poisson with rate λ, the blocking probability could be calculated using the Erlang loss formula

$$B(s, \lambda/\mu) = \frac{(\lambda/\mu)^s/s!}{\sum_{k=0}^{s} (\lambda/\mu)^k/k!}. \tag{4}$$

The Erlang loss formula is applied in several health care applications, as de Bruin et al. (2007), Restrepo et al. (2009) and de Bruin et al. (2010). However, as the assumption of a Poisson arrival process is not reasonable, Equation 4 is not applicable. Therefore, the queueing system has to be modelled as $(G/G/s/s/-)$.

Bekker and Koeleman (2011) approximate the $(G/G/s/s/-)$ model by its infinite-server counterpart, i.e. a $(G/G/\infty)$ model. By considering this infinite-server system, some theoretical results are applicable to approximate the blocking probability. Let c_a denote the coefficient of variation of the inter-arrival time, i.e. the coefficient of the standard deviation and the mean of the inter-arrival time distribution. For instance, if the arrival process is Poisson, $c_a = 1$. For the considered model, the heavy-traffic approximation by Borovkov (1967) is applicable. In general, a heavy-traffic approximation is a stochastic process which has a similar behaviour to a scaled version of the considered queueing model when utilisation of the system is very high. Here, following Borovkov (1967), the number of busy servers $X(\lambda, \mu)$ approaches a normal distribution if λ/μ tends to infinity. Formally,

$$\frac{X(\lambda,\mu) - \lambda/\mu}{\sqrt{z \cdot \lambda/\mu}} \to \mathcal{N}(0,1), \text{ as } \lambda/\mu \to \infty,$$

where

$$z = 1 + (c_a^2 - 1)\frac{1}{\mathbb{E}L}\int_0^\infty \mathbb{P}(L > y)^2 dy. \qquad (5)$$

Here, the term

$$\int_0^\infty \mathbb{P}(L > y)^2 dy$$

is a measurement for the inequality in the length of stay L. The parameter z measures the peakedness of the arrival rate and service times; Whitt (1984) discusses this parameter in particular.

From Equation 5, it is concluded that the peakedness and thus the variance of busy servers increases with c_a. On the other hand, whether the manager of the hospital should consider to reduce the variability in the length of stay depends on the sign of $(c_a^2 - 1)$. In other words, considering the variability in the length of stay is only beneficial if the arrival process is quite stable. Thus, managers should focus on stabilising the arrival process before stabilising the length of stay distribution (Bekker and Koeleman, 2011). Following the square root staffing rule by Whitt (1992), the required number of beds is the sum of the mean of occupied beds λ/μ and a constant times the standard deviation of occupied beds $\sqrt{z \cdot \lambda/\mu}$.

The blocking probability is then approximated by the Hayward approximation by Whitt (1984) and depends on z. Formally,

$$B_c = B_c(s, \lambda/\mu, z) \approx B\left(\frac{s}{z}, \frac{\lambda}{\mu z}\right). \qquad (6)$$

In other words, the Erlang loss formula is applied to the number of servers s and the load of the target λ/μ but both are divided by the peakedness z. As the first component is non-integer in the original definition, the continuous version of the Erlang formula proposed by Jagers and Van Doorn (1986) is applied. The continuous Erlang loss formula implies that with increasing peakedness, the blocking probability increases. Hence, the blocking probability increases in the variation of the arrival processes, determined by c_a.

2.2 Numerical examples

In order to illustrate that c_a and the service time distribution have a high impact on the blocking probability, some numerical examples are run. In particular, the approximations are calculated for some values of c_a and different service time distributions. The setting is the same as in Bekker and Koeleman (2011). Specifically, let $s = 28$, $\lambda = 41/7$ and $1/\mu = 4$, i.e. $\lambda < s\mu$. Bekker and Koeleman (2011) argue that the length of stay at the ward is well modelled by a exponential or a hyper-exponential distribution. The hyper-exponential has higher variance than the exponential distribution with equal mean. These two and the deterministic distribution are the three considered service time distributions. By using Equation 5, Equation 6 and the continuous Erlang loss formula by Jagers and Van Doorn (1986) given by

$$B\left(\frac{s}{z}, \frac{\lambda}{\mu z}\right) = \left[\frac{\lambda}{\mu z}\int_0^\infty \exp\left(\frac{\lambda}{\mu z}t\right)(1+t)^{s/z} dt\right]^{-1},$$

the blocking probability is calculated for the considered choices of c_a and L .

Table 1 provides the approximated blocking probability and standard deviation of the offered load for some choices of c_a^2 and the considered service time distributions. For the hyper-exponential distribution, the parameters are set to $p_1 = 0.5$, $1/\mu_1 = 3$ and $1/\mu_2 = 5$. The results evidence an increasing blocking probability with increasing c_a. Further-

Table 1: Approximated blocking probability and estimated standard deviation of the offered load for different c_a.

c_a^2	Length of stay	z	B_c (%)	$\sqrt{z \cdot \lambda/\mu}$
1	Deterministic	1	5.8	4.84
	Exponential	1	5.8	4.84
	$H_2(p_1 = 0.5)$	1	5.8	4.84
2	Deterministic	2	10.1	6.85
	Exponential	1.5	8.5	5.93
	$H_2(p_1 = 0.5)$	1.48	8.4	5.90
0.5	Deterministic	0.5	2.5	3.42
	Exponential	0.75	4.2	4.19
	$H_2(p_1 = 0.5)$	0.76	4.3	4.21

more, for $c_a^2 < 1$, the deterministic service time distribution leads to a lower blocking probability than the other two. For $c_a > 1$, this relation is the other way round.

2.3 Time-dependent arrivals

In the following, it is assumed that the arrival rate for each hour of the day is constant but can differ from day to day. Furthermore, the arrival rate per day of the week does not change weekly. Consequently, the arrival process can take seven different arrival rates, one for each day. Formally, let $\lambda_1, \cdots, \lambda_7$ denote the arrival rates for the seven days of the week with λ_1 being the arrival rate at Mondays. Further, $\overline{\lambda}$ denotes the average arrival rate.

Bekker and Koeleman (2011) apply the approach by Holtzmann and Jagerman (1979) and split up the peakedness into a random part, z_{rand}, and a predictable part, z_{pred}, both depending on the arrival process. Specifically, the predictable part comes from the changing arrival rates whereas the random part depends on c_a. The random part of the peakedness is determined by Equation 5. Formally,

$$\begin{aligned} z &= z_{rand} + z_{pred} \\ &= 1 + (c_a^2 - 1)\frac{1}{\mathbb{E}L}\int_0^\infty \mathbb{P}(L > y)^2 dy + z_{pred} \end{aligned} \tag{7}$$

Applying the Hayward approximation, see Equation 6, and the continuous Erlang loss formula, it is noted that the blocking probability for a non-stationary arrival process is at least as high as the blocking probability of the stationary arrival process with arrival rate $\overline{\lambda}$.

In order to estimate the predictable peakedness, Bekker and Koeleman (2011) apply the fluid approximation by Massey and Whitt (1993). Therefore, the parameter z_{pred} determines as the coefficient of the variance and the mean of the expected number of occupied beds in the corresponding $(G/G/\infty)$ queueing system. Formally,

$$z_{pred} = \frac{Var[m(t)]}{\mathbb{E}[m(t)]}, \tag{8}$$

where $m(t)$ is the mean number of occupied beds in the system at time point t.

As the arrival process is non-Poisson, Equation 2 cannot be applied directly. However, due to the independence of the arrivals, Equation 2 can be generalised to this specific case of non-Poisson arrivals. In particular, for a discrete time process, the expected number of patients in service at the end of day d, $d \in \mathbb{N}$, is determined by

$$\begin{aligned} m(d) &= \mathbb{E}[\text{number in the system at day } d] \\ &= \sum_{i=-\infty}^{d} \mathbb{E}[\text{arrivals day } i] \cdot \mathbb{P}[\text{still in service}] \\ &= \sum_{i=0}^{\infty} \lambda(d-i)\mathbb{P}(L > i). \end{aligned}$$

However, as the patients can arrive and be discharged at each time of the day, the formula for $m(d)$ has to be formulated in continuous time. In the continuous case, $m(t)$ determines to

$$m(t) = \int_0^\infty \lambda(t-y)\mathbb{P}(L > y)dy \tag{9}$$

According to the arrival rates considered, $\lambda(t)$ is defined as $\lambda(t) = \lambda_i$, if $i = \lceil t \rceil \mod 7 + 1$. Consequently, the interval $(d-1, d)$ corresponds to the day d.

As mentioned in Section 2.2, Bekker and Koeleman (2011) model the service times at the ward by a exponential or a hyper-exponential distribution. For the case of exponential service times and $d \in \mathbb{N}$, a recursive relation is obtained for the expected number of patients in the system, denoted by $m^e(d)$, at time point d. Formally,

$$\begin{aligned} m^e(d) &= \int_0^\infty \lambda(d-s)e^{-s\mu}ds \\ &= \int_0^1 \lambda(d)e^{-s\mu}ds + \int_1^\infty \lambda(d-s)e^{-s\mu}ds \\ &= \frac{\lambda(d)}{\mu}\left(1-e^{-\mu}\right) + e^{-\mu}\int_0^\infty \lambda(d-s-1)e^{-s\mu}ds \\ &= \frac{\lambda(d)}{\mu}\left(1-e^{-\mu}\right) + e^{-\mu}m^e(d-1) \end{aligned}$$

Iteratively applying this recursive relation yields

$$m^e(d) = \frac{(1-e^{-\mu})}{\mu} + \sum_{i=0}^{n-1} \lambda(d-i)e^{-\mu i} + e^{-\mu n}m^e(d-n)$$

By using the periodicity of the daily arrival rates, as $\lambda(d+T) = \lambda(d)$ with $T = 7$, this relation simplifies to

$$m^e(d) = \frac{1}{\mu}\left(\frac{1-e^{-\mu}}{1-e^{-\mu T}}\right)\sum_{i=0}^{T-1} \lambda(d-i)e^{-\mu i} \qquad (10)$$

In case of a hyper-exponential distribution, $m(t)$ determines to

$$\begin{aligned} m(t) &= \int_0^\infty \lambda(t-s)\left(\sum_{j=1}^{J} p_j e^{-\mu_j s}\right) ds \\ &= \sum_{j=1}^{J} p_j \int_0^\infty \lambda(t-s)e^{-\mu_j s}ds \\ &= \sum_{j=1}^{J} p_j m_j^e(t), \end{aligned}$$

where $m_j^e(t)$ is the expected number of patients at time point t for the exponential service time distribution with mean μ_i.

Table 2: Fraction of refused admissions for stationary arrivals

c_a^2	Length of stay	z_{rand}	z_{pred}	B_c (%)
1	Deterministic	1	0.553	8.7
	Exponential	1	0.152	6.7
	$H_2(p_1 = 0.5)$	1	0.150	6.7
2	Deterministic	2	0.553	13
	Exponential	1.5	0.152	9.2
	$H_2(p_1 = 0.5)$	1.48	0.150	9.1
0.5	Deterministic	0.5	0.553	6.1
	Exponential	0.75	0.152	5.2
	$H_2(p_1 = 0.5)$	0.76	0.150	5.3

The coefficient of variance and mean of the expected number in the system is then determined by

$$z_{pred} = \frac{1}{T-1}\sum_{d=1}^{T} \frac{(m(d)-\overline{m})^2}{\overline{m}}, \qquad (11)$$

where $\overline{m} = \sum_{d=1}^{T} m(d)/T = \lambda/\mu$.

2.4 Numerical examples

As for the stationary case, numerical examples for the non-stationary case are run. Bekker and Koeleman (2011) set the arrival rate on working days to 7 and 3 otherwise. This implies a mean arrival rate of 41/7 and an average workload of 23.43. Consequently, z_{rand} is equal to the parameter z in Table 1. The mean service time is set to $1/\mu = 4$ as in Section 2.2 and the same service time distributions are considered. The approximation results in Table 2 illustrate that the influence of the time-dependency on the blocking-probability depends immensely on the service time distribution. In particular, if the service time is deterministically distributed, the predictable peakedness is very high with 0.553. Exponentially and hyper-exponentially distributed service times does not differ much. In difference to Table 1, the approximated blocking probability is for any of the c_a^2 higher for the deterministic than for the other two service time distributions.

3 Nelson and Gerhardt (2011)

Given the numerical examples in Section 2.2 and 2.4, the results should be verified by a simulation study. Therefore, it is necessary to simulate inter-arrival times from a non-stationary, non-Poisson arrival process with the desired properties. For this purpose, the simulation method by Nelson and Gerhardt (2011) is introduced in this section and applied in Section 4. Nelson and Gerhardt (2011) address the simulation of non-stationary, non-Poisson processes with integrable arrival rates and a dependence structure between the arrivals. Specifically, the dependency is modelled by a geometric autocorrelation function.

The worksheet made available by Nelson and Gerhardt (2011) handles the case of piece-wise constant arrival rates. As the numerical examples in Section 2 consider the case of a constant arrival rate during a day, the approach by Nelson and Gerhardt (2011) is applicable. However, Nelson and Gerhardt (2011) argue that the simulated arrival process only captures or approximates some important characteristics of the desired arrival process; in particular, mean, variance and autocorrelation. Nevertheless, these are the characteristics used for the approximation by Bekker and Koeleman (2011).

The simulation method proposed by Nelson and Gerhardt (2011) generates a non-stationary, non-Poisson arrival process with dependent arrivals in two steps. First, a stationary non-Poisson arrival process, the base process, is generated. Second, the inversion method is applied to transform the stationary process into a non-stationary process. The inversion method, for instance described by Çinlar (1975) and Rolski and Szekli (1991) is well-known in order to transform a stationary Poisson process into a non-stationary Poisson process. However, Nelson and Gerhardt (2011) prove that it is applicable in this more general framework of non-Poisson processes with dependent arrivals. A second popular method used to transform stationary Poisson processes into non-stationary Poisson processes is the thinning method by Lewis and Shedler (1979). Nevertheless, the work by Gerhardt and Nelson (2009) implies that this method can also be used for renewal processes, i.e. non-Poisson processes with independent arrivals. This section is organised as follows: In Section 3.1, the notation for the following sections is given. Section 3.2 describes the simulation of the base process. Finally, inversion and thinning methods are described in Section 3.3 and 3.4, respectively.

3.1 Notation

In the first step, inter-arrival times $\{X_n : n \geq 1\}$ of a stationary process are generated. This process captures important features like the autocorrelation of the desired process. The resulting arrival count process, i.e. the number of arrivals occurred until time point t, is denoted by

$$N(t) = \max\left\{ m \geq 0 : \sum_{i=1}^{m} X_i \leq t \right\}.$$

It is assumed that $N(t)$ is initialised in equilibrium with $X_2, X_3, \dots$ identically distributed whereas X_1 has the associated equilibrium distribution. Therefore, $\mathbb{E}[N(t)] = rt$ for some fixed rate $r > 0$. Nelson and Gerhardt (2011) fix the parameter r to 1. The second step of the algorithm then transforms the sequence of inter-arrival times of the stationary process into the inter-arrival times $\{W_n : n \geq 1\}$ of the desired process with integrable arrival rates. Similarly to the stationary process, the arrival count process $I(t)$ is defined by

$$I(t) = \max\left\{ m \geq 0 : \sum_{i=1}^{m} W_i \leq t \right\}.$$

3.2 Base process

In the first step, a stationary process with the desired characteristics is simulated. In order to model a non-Poisson arrival process with dependency of the arrivals, the Markovian Arrival Process (MAP) introduced by Lucantoni et al. (1990) is a common example and used in several publications, e.g. Alfa and Frigui (1996) or Ramirez-Cobo and Carrizosa (2012). For instance, Wang et al. (2010) or Bause and Horvath (2010) consider the problem of fitting such models. In general Markovian Arrival Processes can be seen as a generalisation of the phase-type inter-arrival times.

Nelson and Gerhardt (2011) use the Markov-MECO (Mixture of Erlangs of Common Order) approach proposed by Johnson (1998) for simulating a stationary non-Poisson arrival process with dependence structure. The Markov-MECO is a particular case of a Markovian Arrival Process. Gerhardt and Nelson (2009) argue that by using a MAP base process, the process can be initialised in equilibrium, requiring to compute the distribution of the current state of the continuous time Markov chain in equilibrium (Nelson and Gerhardt, 2011).

The Markov-MECO approach extends the MECO process by Johnson and Taaffe (1989) which can capture the first three moments of the inter-arrival time distribution. Johnson (1998) additionally provides a control over the dependency between the arrivals. In particular, the dependency is modelled by a Markov Chain and the autocorrelation function is geometric; for more details on the Markov-MECO approach, see Johnson (1998). An example of a Markov-MECO process is illustrated in Figure 2. The current inter-arrival time has either an $E_k(\lambda_1)$ or an $E_k(\lambda_2)$ distribution and a Markov chain governs from which distribution the next inter-arrival time is sampled.

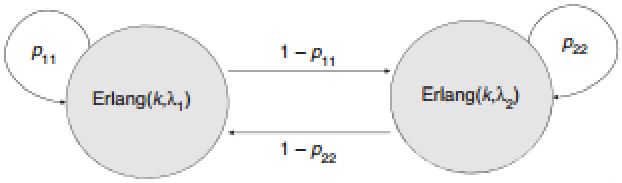

Figure 2: The Markov chain that describes Markov-MECO inter-arrival times. Source: Nelson and Gerhardt (2011), page 6

Although the Markov-MECO approach can handle any first three moments, in the worksheet facilitated by Nelson and Gerhardt (2011), some parameters cannot be chosen by the user. The first parameter is the mean of the base process because it is set to 1. Nevertheless, as the process is transformed to the desired mean arrival rates in the next step by the inversion method, this restriction does not limit the set of arrival processes contained in the Markov-MECO approach.

Second, the user cannot set the skewness, as this parameter is hard to select just by intuition. Nelson and Gerhardt (2011) fix this parameter as the third moment of a Markovian distribution which is fully determined by the mean and the variance. Specifically, if $c_a < 1$, the MECon distribution by Tijms (1994) is used and a balanced hyper-exponential distribution, for example by Sauer and Chandy (1975), otherwise (Nelson and Gerhardt, 2011). Consequently, the user can choose the parameter for the variance c_a^2 and the dependence structure. As the autocorrelation function is geometric, it is sufficient to give the correlation between two consecutive arrivals, ρ_1. The worksheet is available for download at http://users.iems.northwestern.edu/∼nelsonb/NSNR.xls.

3.3 Inversion method

The inversion method is widely used in simulations to transform a stationary into a non-stationary Poisson process. However, Gerhardt and Nelson (2009) and Nelson and Gerhardt (2011) apply this method to transform a non-Poisson arrival process. Specifically, suppose that $r(t)$, $t \geq 0$, is the desired non-negative, integrable arrival rate for $I(t)$. Further, set

$$R(t) = \int_0^t r(s)ds$$

and define

$$R^{-1}(s) = \inf\{t : R(t) \geq s\}$$

for all $s > 0$. By applying this inverse function to the inter-arrival times sampled from the base process, the stationary inter-arrival times $\{X_n : n \geq 1\}$ are transformed to the inter-arrival times $\{W_n : n \geq 1\}$ and arrival times $\{V_n : n \geq 1\}$. Formalising this procedure leads to Algorithm 1.

Algorithm 1 Inversion Method for arrival processes

```
Require: Inter-arrival times $\{X_n : n \geq 0\}$
Require: Integrated arrival rate $R(t)$
Set $W_0 = 0$, $m = 0$, $V_0 = 0$ and $S_0 = 0$
while Further arrivals do
    Set $m = m + 1$
    Set $S_m = X_m + S_{m-1}$
    Set $V_m = R^{-1}(S_m)$
    Set $W_m = V_m - V_{m-1}$
end while
return Inter-arrival times $\{W_n : n \geq 1\}$
return Arrival times $\{V_n : n \geq 1\}$
```

Figure 3 illustrates the algorithm for $r(t) = 2t$, i.e. the mean arrival rate increases with constant rate over time. The circles on the left vertical axis are the simulated arrival times of the base process. Then the inversion method is run on these and the crosses on the horizontal bottom axis are the transformed arrivals.

However, the inversion method does not automatically imply that the transformed arrival counting process $I(t)$ is non-stationary with the desired characteristics. In particular, it is necessary to exclude that the transformed process is Poisson. For this reason, Nelson and Gerhardt (2011) considers the index of dispersion for counts (IDC), e.g. used in Sriram and Whitt (1986) or Kim (2011). The IDC of the base process $N(t)$ is given by

$$\text{IDC} = \lim_{t \to \infty} \frac{Var[N(t)]}{\mathbb{E}[N(t)]}. \tag{12}$$

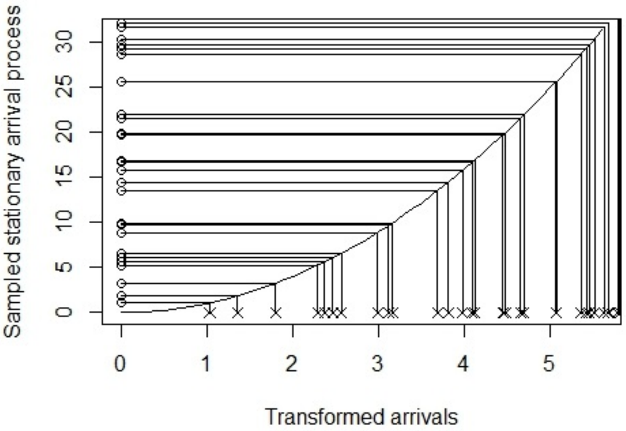

Figure 3: Illustration of the inversion method with $r(t) = 2t$.

For example, if $N(t)$ is a Poisson process, the IDC is 1. Therefore, it can be concluded that the transformed process $I(t)$ is non-Poisson if the corresponding index of dispersion for counts is not 1. Nelson and Gerhardt (2011) prove that this property is given for the inversion method.

Theorem 1. *$\mathbb{E}[I(t)] = R(t)$, for all $t \geq 0$, and* $Var[I(t)] \approx$ IDC$\cdot R(t)$, for large t.

Proof by Nelson and Gerhardt (2011).
The assumption that N(t) is initialised in equilibrium with rate $r = 1$ implies $\mathbb{E}[N(t)] = t$, for all $t \geq 0$. Further, because of equation 12, $Var[N(t)] \approx$ IDC $\cdot\, t$, for large t. Thus

$$\begin{aligned}\mathbb{E}[I(t)] &= \mathbb{E}[\mathbb{E}\{I(t)|N(R(t))\}] \\ &= \mathbb{E}[N(R(t))] \\ &= R(t)\end{aligned}$$

for all $t \geq 0$, while

$$\begin{aligned}Var[I(t)] &= \mathbb{E}[Var\{I(t)|N(R(t))\}] + \\ &\qquad Var[\mathbb{E}\{I(t)|N(R(t))\}] \\ &= 0 + Var[N(R(t))] \\ &\approx \text{IDC} \cdot R(t)\end{aligned}$$

for large t. □

Thus Theorem 1 implies that the transformed process $I(t)$ has approximately the same IDC as the base process $N(t)$. Consequently, if the simulated, stationary arrival process has an IDC which is significantly different from 1, the transformed non-stationary process preserves this property and is thus non-Poisson. In the same way, the result implies that the non-stationary process generated by applying the inversion method to a stationary Poisson process is also Poisson. In summary, applying the inversion method leads to the desired arrival rate under preserving the IDC. The worksheet facilitated by Nelson and Gerhardt (2011) works with piecewise-constant arrival rates which have to be specified by the user. Finally, the user has to set the length of the simulation and the number of replications to run. The simulated inter-arrivals times are then wrote in a Excel-spreadsheet with one replication per column and can be used for the simulation.

Nelson and Gerhardt (2011) mention that the IDC is not a quite intuitive measure of variability and dependence. Therefore, the authors also consider the index of dispersion for intervals (IDI) by Gusella (1991). Formally, the IDI is given by

$$\begin{aligned}\text{IDI} = &\lim_{n\to\infty} \frac{Var[\sum_{i=1}^{n} X_i]}{n\mathbb{E}^2[X_2]} \\ =&c_a^2 \left(1 + 2\sum_{j=1}^{\infty} \rho_j\right),\end{aligned}$$

where ρ_j is the lag-j autocorrelation of the stationary inter-arrival times $X_2, X_3, \ldots$. The work by Whitt (2002) proves that both indexes are identical under some conditions. Specifically, the IDC exists and is equal to the IDI if $\sum_{i=1}^{n} X_i$ of the base process $N(t)$ satisfy a Central Limit Theorem of the form

$$\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} X_n - n\mu\right) \overset{\mathcal{D}}{\to} \mathcal{N}(0,\tau^2).$$

Consequently, the IDC captures both, dependency and variability of the arrivals.

3.4 Thinning method

Another approach used in simulations to transform a stationary to a non-stationary Poisson process is the thinning method. The work by Gerhardt and Nelson (2009) proves that the thinning method is also applicable to transform a stationary, non-Poisson process to a non-stationary non-Poisson under the condition that the arrivals are independent. In particular, let $r^* = \max\{r(t) : t \geq 0\}$ be the maximum arrival rate occurring and assume that r^* is finite.

The first step is again to simulate the inter-arrival times $\{X_n : n \geq 1\}$ of a stationary non-Poisson process with arrival rate r^* and variance $c_a^2/(r^*)^2$, i.e. the inter-arrival distribution has mean $1/r^*$ and variance $c_a^2/(r^*)^2$ (Gerhardt and Nelson, 2009). For instance, the MECO process by Johnson and Taaffe (1989) could be used again. After simulating this stationary process, a simulated arrival at time point t is accepted with probability $r(t)/r^*$, i.e. if $r(t) = r^*$, all sampled arrivals are accepted. Otherwise, the simulated arrival time is rejected and does not appear in the transformed arrival process. This approach leads to Algorithm 2.

Algorithm 2 Thinning Method for non-stationary arrival processes

```
Require: Inter-arrival times $\{X_n : n \geq 1\}$
Require: Arrival rate $r(t)$
Set $r^* = \max\{r(t) : t \geq 0\}$
Set $m = 0$, $p = 0$, $V_0 = 0$, $W_0 = 0$ and $S_0 = 0$
while Further arrivals do
    Set $m = m + 1$
    Set $S_m = X_m + S_{m-1}$
    Generate $U_m \sim Uniform[0,1]$
    if $U_m < r(S_m)/r^*$ then
        Set $p = p + 1$
        Set $V_p = S_m$
        Set $W_p = V_p - V_{p-1}$
    end if
end while
return Inter-arrival times $\{W_n : t \geq 1\}$
return Arrival times $\{V_n : n \geq 1\}$
```

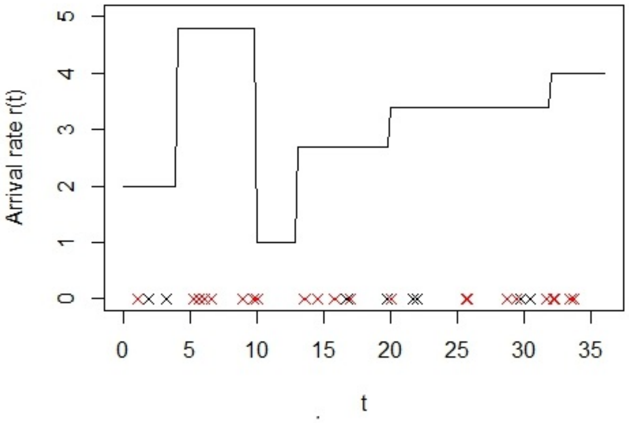

Figure 4: Illustration of the thinning method with piecewise-linear arrival rates with accepted arrival (red) and rejected arrivals (black).

Figure 4 illustrates a resulting arrival process with piecewise-constant arrival rates and $r^* = 4.8$. The red crosses are the generated and accepted arrivals and the black crosses are generated by the non-stationary arrival process but rejected. Gerhardt and Nelson (2009) prove that $\mathbb{E}[I(t)] = R(t)$ for all $t \geq 0$. Further, Gerhardt and Nelson (2009) prove that the index of dispersion for counts of the transformed process $I(t)$ at time point t can be approximated; see Result 2.3 of Gerhardt and Nelson (2009). Gerhardt and Nelson (2009) use Phase-type distributions to simulate the stationary non-Poisson process. Phase-type distributions are widely used to fit renewal processes, see for instance Harper et al. (2011) or Payne et al. (2012). However, as the simulation study in Section 4 is based on the worksheet by Nelson and Gerhardt (2011), the thinning method is not considered in more detail.

This section concludes by mentioning that the processes generated by thinning and inversion are not equivalent in general as an example by Gerhardt and Nelson (2009) evidences. Consider $R(t) = t/2$, i.e. $r(t) = 1/2$, and the base process has rate 1. Then the expectation of the process generated by inversion is $2n + c_a^2 - 1$. Otherwise, the expectation of the process generated by thinning is $2n + \frac{c_a^2-1}{2}$.

4 Simulation Study

The numerical examples in Section 2 evidence two results. First, the blocking probability increases with the variance of the arrival process. Second, a non-stationary arrival process leads to a higher blocking probability than a stationary arrival process with the same mean arrival rate. In order to evaluate the approximations obtained in Section 2.2 and 2.4, several simulations are run. Nevertheless, the approximations do not consider the case of dependent arrivals. Therefore, the sensitivity of the blocking probability on positive and negative correlation is investigated. Another not considered aspect is the possibility of batched arrivals. For instance, in case of a car accident with several casualties, family members are normally taken to the same hospital. This case is as well investigated by simulation.

Following the setting in Section 2, the bed capacity is fixed at $s = 28$ beds and the average length of stay is four days. In particular, the service time distributions considered are the deterministic, the exponential and the hyper-exponential distribution. Here, the hyper-exponential distribution is the mixture of two exponential distributions with parameter $\mu_1 = 3$, $\mu_2 = 5$ and $p_1 = 0.5$.

The samples of the arrival processes are generated by the worksheet facilitated by Nelson and Gerhardt (2011). Each arrival process is the union of ten arrival processes of run length 365 days, i.e. the total run length is about 10 years. Due to overflow problems, the worksheet cannot handle a time period of more than 3000 days. All simulations are run in Simul8 and the generated samples are imported as csv-files. To avoid incorrectness, a warm-up period of 10 days is set. In the simulation, a patient queues at maximum 15 minutes, although the setting in Section 2 states that the patient leaves automatically. However, the results of the simulated queue indicate that the number of patients which get a bed

Table 3: Approximated and simulated blocking probability for stationary independent arrivals and different choices of c_a^2.

c_a^2	Length of stay	approx.(%)	simul. (%)
1	Deterministic	5.8	5.7
	Exponential	5.8	5.7
	$H_2(p_1 = 0.5)$	5.8	5.7
2	Deterministic	10.1	10.5
	Exponential	8.5	8.7
	$H_2(p_1 = 0.5)$	8.4	8.8
0.5	Deterministic	2.5	3.0
	Exponential	4.2	4.4
	$H_2(p_1 = 0.5)$	4.3	4.4

because they wait 15 minutes is negligible. In order to achieve valid results, for each arrival process and each length of stay distribution, ten repetitions are run and the blocking probability is taken as the average number of rejected patients and divided by the number of arrived patients. The Simul8 worksheet and all generated arrival processes are available for download at http://www.lancs.ac.uk/pg/rohrbeck/simulation-project.zip.

This section is organised as follows: First in Section 4.1, the blocking probability of the stationary case is simulated and compared to the approximated results of Section 2.2. Then the case of slightly positive or negative correlated arrivals is considered. Section 4.2 considers the case of non-stationary arrivals in the same way as Section 4.1. Finally, Section 4.3 investigates whether it is important to consider batched arrivals or not in this setting.

4.1 Stationary arrivals

According to the numerical examples in Section 2.2, the arrival rate per day is set to $\lambda = 41/7$. Additionally, the arrivals are independent. Table 3 provides the approximated and simulated blocking probabilities for this setting. The results evidence that the approximation fits the simulated blocking probability very well. In particular, the largest difference of 0.5 occurs for $c_a^2 = 0.4$ and deterministically distributed service times.

All the properties concluded by the approximations are approved by the simulations. First, the blocking probability increases with the coefficient of variation c_a. Second, the deterministic service time distribution is more sensitive to changes in variance than the other two considered distributions. Sensitivity in this context refers to the relative change of the blocking probability with changing c_a. However, also the absolute change is higher. Third, the blocking probability for $c_a^2 = 1$ is equal for each of the three considered service time distributions. Property two and three imply that for $c_a^2 < 1$, the deterministic service time distribution implies a lower blocking probability than the other two distributions. For $c_a^2 > 1$, the relation is the opposite way round. Forth,the difference between exponential and hyper-exponential distributed service times is small.

Next, the sensitivity of the blocking probability on possible dependency in the arrival process, positive as well as negative, is investigated. In order to compare this setting to the one in Section 2.2, the same service time distributions and values for c_a^2 as in Table 3 are considered. In difference to the previous simulations, the parameter value ρ_1 of the geometric autocorrelation function is not equal 0. In particular, in order to model positive correlation, ρ_1 is set to 0.2. Otherwise, ρ_1 is set to -0.2 for negative correlation. For $c_a^2 = 2$, $\rho_1 = -0.2$ is not contained in the feasible region of the worksheet by Nelson and Gerhardt (2011). In this case, ρ_1 is set to the smallest possible value, $\rho_1 = -0.085$. Consequently, six different arrival processes are generated, two for each value of c_a^2.

Table 4 provides the blocking probabilities for the considered parameter settings. Several conclusions are drawn from this results. First, the higher ρ_1, the higher the blocking probability. In particular, if the arrivals are negatively correlated, the block-

Table 4: Simulated blocking probability for positive and negative correlated, stationary arrivals and different choices of c_a^2.

c_a^2	ρ_1	Length of stay	simulated $B_c(\%)$
1	-0.2	Deterministic	4.5
		Exponential	5.1
		$H_2(p_1 = 0.5)$	5.2
	0.2	Deterministic	9.9
		Exponential	8.3
		$H_2(p_1 = 0.5)$	8.4
2	-0.085	Deterministic	9.7
		Exponential	8.2
		$H_2(p_1 = 0.5)$	8.1
	0.2	Deterministic	17.6
		Exponential	13.7
		$H_2(p_1 = 0.5)$	13.8
0.5	-0.2	Deterministic	2.0
		Exponential	3.9
		$H_2(p_1 = 0.5)$	3.9
	0.2	Deterministic	7.5
		Exponential	6.9
		$H_2(p_1 = 0.5)$	7.0

ing probability is lower than in the case of independent arrivals. Contrary, positive correlated arrivals lead to a higher blocking probability. Especially if $c_a^2 = 2$, the absolute change in per cent is large. In other words, the more the arrivals are correlated, the higher is the number of rejected patients.

Second, the absolute difference in the blocking probability between the negative correlated and the independent arrivals is lower than the absolute difference between the positive correlated and independent arrivals. For instance, for $c_a^2 = 1$ and the deterministic distribution, the absolute difference between negative correlated and independent arrivals is 1.2%. On the other hand, the absolute difference between positive correlated and independent arrivals is 4.2%. Thus, the blocking probability is more sensitive to positive than to negative correlation. Third, independent of the correlation of the arrivals, a higher c_a^2 leads to a higher blocking probability. Therefore, the property concluded by the approximations is valid for correlated arrival processes too. Forth, the blocking probabilities for exponentially and hyper-exponentially distributed service times are similar though the blocking probability for the exponential service time distribution is often smaller.

Last, the deterministic service time distribution is more sensitive to correlation than the other two service time distributions; e.g. for $c_a^2 = 1$ the relative difference between positive correlated and independent arrivals is 74% whereas it is 46% for the exponential distribution. Another example is that the blocking probability of the determisitc distribution for $c_a^2 = 0.5$ and $\rho_1 = 0.2$ is higher than for the exponential distribution although it is the other way around for $\rho_1 = 0$. On the other hand, for $c_a^2 = 1$ and $\rho_1 = -0.2$, the deterministic distribution leads to a lower blocking probability than the exponential distribution although the blocking probabilities were equal for $\rho_1 = 0$. In summary, correlation has a high influence on the blocking probability. However, the degree of influence depends on the service time distribution. The determinsitic service time distribution in particular is highly sensitive to changes in the dependence structure of the arrival process.

4.2 Non-stationary arrivals

The theoretical results of Section 2.3 imply an higher blocking probability for non-stationary arrivals than for stationary arrivals with the same average arrival rate. The impact of time-dependent arrivals has been observed in various studies, for instance by Melamed et al. (1992) on compressed video frame bits and by Ware et al. (1998) on computer networks. For example, Biller and Nelson (2005) introduce an approach to fit stochastic models to dependent time-series processes inputs.

According to Section 2.4, the mean arrival rate is set to 7 patients per day on working days and 3 per day on the weekend. This implies an average arrival rate of $\overline{\lambda} = 41/7$. First, the case of inde-

Table 5: Approximated and simulated blocking probability for non-stationary independent arrivals and different choices of c_a^2.

c_a^2	Length of stay	approx.(%)	simul. (%)
1	Deterministic	8.7	9.3
	Exponential	6.7	7.0
	$H_2(p_1 = 0.5)$	6.7	7.0
2	Deterministic	13.0	13.5
	Exponential	9.2	10.0
	$H_2(p_1 = 0.5)$	9.1	10.1
0.5	Deterministic	6.1	7.0
	Exponential	5.2	5.5
	$H_2(p_1 = 0.5)$	5.3	5.5

pendent arrivals is considered. Table 5 provides the simulated and approximated blocking probabilities for the considered arrival processes with different values for c_a^2. Compared to Table 3, the simulated blocking probability is higher for the non-stationary than for the stationary case. Consequently, the second conclusion of Section 2 is approved by the simulation.

As for the stationary case, the blocking probability increases with c_a^2. Further, the results confirm that the deterministic service time distribution leads to higher blocking probabilities than the other two service time distributions for any of the considered values for c_a^2 and is thus more sensitive to non-stationarity. As in Table 3, the difference between the exponential and hyper-exponential service time distributions is very small. However, in difference to Table 3, the difference between simulated and approximated blocking probability is higher. This effect may be due to the higher number of approximations used in the non-stationary case, e.g. the fluid approximation for calculating the predictable peakedness. Nevertheless, the approximations fit the simulated blocking probabilities quite well.

Table 6: Simulated blocking probability for positive and negative correlated, non-stationary arrivals and different choices of c_a^2.

c_a^2	ρ_1	Length of stay	simulated B_c(%)
1	-0.2	Deterministic	8.8
		Exponential	6.6
		$H_2(p_1 = 0.5)$	6.7
	0.2	Deterministic	12.4
		Exponential	9.4
		$H_2(p_1 = 0.5)$	9.4
2	-0.085	Deterministic	12.7
		Exponential	9.5
		$H_2(p_1 = 0.5)$	9.6
	0.2	Deterministic	19.1
		Exponential	14.2
		$H_2(p_1 = 0.5)$	14.3
0.5	-0.2	Deterministic	6.4
		Exponential	5.1
		$H_2(p_1 = 0.5)$	5.0
	0.2	Deterministic	10.0
		Exponential	8.0
		$H_2(p_1 = 0.5)$	8.0

After investigating the case of independent arrivals, the case of dependent arrivals is considered in the following. The parameter values for ρ_1 are set to the same values as in Section 4.1, i.e. negative and positive correlated arrivals are simulated. Although the influence of dependency of the arrivals was already studied in Section 4.1, it is not clear how large the combined effect of non-stationarity and dependency is. In particular, the correlation between dependency of the arrivals and a non-stationary arrival process is investigated. Table 6 provides the simulated blocking probabilities for non-stationary arrival processes with dependent arrivals for the three considered service time distributions.

Several results obtained in Section 4.1 are also valid for this case. First, the blocking probability for a non-stationary arrival process increases with ρ_1 and c_a^2. Second, the blocking probability is more sensitive to positive than to negative autocorrelation. For instance, for the case of $c_a^2 = 0.5$ and deterministic service time distribution, the simulated blocking probability is 43% higher if ρ_1 is 0.2 than for the independent case in Table 5. On the other hand, the blocking probability decreases by 9% if the parameter ρ_1 is -0.2. Third, the deterministic distribution is more sensitive to correlation than the exponential or hyper-exponential distribution. Last, as in all simulations before, the difference between exponential and hyper-expoential service time distribution is not high.

Additionally to these results, further conclusions can be found about the influence of the non-stationarity. First, the relative change caused by the dependency in the arrival process in the non-stationary case is less than in the stationary case; e.g. for $c_a^2 = 1$ and the exponential distribution, the relative difference between independent and positive correlated arrivals is 46% in the stationary case whereas it is 34% in the non-stationary case. In other words, if the arrival process is non-stationary, the blocking probability is less sensitive to dependency.

Second, a higher value of ρ_1 decreases the relative difference between the stationary and non-stationary arrival process. For instance, for the deterministic service time distribution, the relative difference between the stationary and the non-stationary case is 63% for $c_a^2 = 1$ and $\rho_1 = 0$, but 25% for $c_a^2 = 1$ and $\rho_1 = 0.2$. The same result holds for c_a^2. These observations can be formulated in the following two points for the considered service time distributions:

1. The higher c_a^2, the less sensitive is the blocking probability to non-stationarity of the arrival process.
2. The higher ρ_1, the less sensitive is the blocking probability to non-stationarity of the arrival process.

In summary, the non-stationarity of the arrival process increases the blocking probability compared to the stationary case. On the other hand, the blocking probability is less sensitive to changes in the variability or dependency of the arrival process.

Table 7: Approximated and simulated blocking probability for stationary independent arrivals and different choices of c_a^2 for the case of batched arrivals and not batched arrivals.

c_a^2	Length of stay	no batches	batches
1	Deterministic	5.7	5.5
	Exponential	5.7	5.8
	$H_2(p_1 = 0.5)$	5.7	5.7
2	Deterministic	10.5	10.2
	Exponential	8.7	8.9
	$H_2(p_1 = 0.5)$	8.8	8.7
0.5	Deterministic	3.0	2.7
	Exponential	4.4	4.3
	$H_2(p_1 = 0.5)$	4.4	4.0

4.3 Batches of arrivals

All the approximations considered so far assume that patients arrive single. The simulations done in Section 4.1 and Section 4.2 evidence an high importance to know whether the arrivals are truly independent. In the previous sections, dependency was modelled by a geometric autocorrelation function. Here, the dependency is considered to be in the form that some patients arrive in batches. As the worksheet does not give the opportunity to simulate such arrival processes, they are generated by rounding up the arrival times generated for Table 3. In particular, the day is split up into intervals of 2.5 hours length and arrivals in one interval arrive at the same time. This leads to batches of size two or three at some time points.

Table 7 provides the simulated blocking probabilities for the case of batched arrivals. The results evidence a small difference between the case of batched arrivals and the case of independent arrivals. However, the difference is not significant. Therefore, it is concluded that the possibility of small batches of arrivals has not to be considered in this application. All other properties concluded from Table 3 can be also approved for the case of batched arrivals.

5 Summary and Discussion

The motivation for this work was to investigate the impact of non-Poisson arrival processes on the blocking probability of a $(G/G/s/s/-)$ model. In particular, the impact of variance, non-stationarity and dependency was considered. Studies in health care applications motivate the necessity to consider such systems with non-Poisson arrival processes. McManus et al. (2003), for example, evidence that the arrivals of scheduled admissions at a clinical ward is not well modelled by a Poisson process.

Bekker and Koeleman (2011) consider independent, non-Poisson arrivals and approximate the corresponding queueing system by a $(G/G/\infty)$ model. However, possible dependency in the arrival process is not investigated, The blocking probability is then estimated by applying the Hayward approximation and the heavy-traffic approximation by Borovkov (1967). For the non-stationarity arrival process, the fluid approximation is used additionally. In order to measure the impact of variation and non-stationarity, the approximations were applied to several service time distributions and arrival processes with equal mean but different variability. The theoretical results implied that the blocking probability is higher for non-stationary arrival processes and increases with the variation of the arrival process. Furthermore, the sensitivity to the variance depends on the service time distribution. Bekker and Koeleman (2011) also considers shortly the case of health chains, i.e. systems with more than one ward. In this case, heavy-traffic approximations exist too, e.g. see Glynn and Whitt (1991). However, these have a more complicated form than the approximation considered here.

In the simulation studies, the worksheet by Nelson and Gerhardt (2011) was then used to simulate non-Poisson arrival processes which captured the features considered by Bekker and Koeleman (2011); mean, variance and non-stationarity. The approach by Nelson and Gerhardt (2011) samples inter-arrival times from a Markov-MECO process which can capture the first three moments of the desired inter-arrival time distribution. Furthermore, Markov-MECO allows to model dependency between the arrivals by a geometric autocorrelation function. By using the inversion method, the simulated inter-arrival times are then transformed to the desired non-stationary arrival process. The possibility of dependency of the arrivals was also used, in order to investigate the sensitivity of the blocking probability to dependency between the arrivals.

However, whether the Markov-MECO approach is always appropriate to model the arrival process is unsure. The knowledge which descriptors are most important to approximate the arrival process sufficiently, is still limited (Johnson and Taaffe, 1989). The work by Anderson et al. (2004) evidence that an Markovian Arrival process and its reverse cannot be distinguished properly by classical statistical descriptors like mean, variance, etc. Additionally, several authors propose new descriptors for arrival processes. Johnson and Narayana (1996), for example, proposes new descriptors of burstiness of a Markovian arrival processes. Therefore, further research is needed to specify which descriptors are required under which conditions to model the arrival process sufficiently in order to approximate the desired properties of the queueing system.

The simulated blocking probabilities approved the approximations for all considered cases. Further, the influence of possible dependency between the arrivals on the blocking probability was investigated. The results implied that the blocking probability increases with the dependency in the arrival process. This means, a positive correlation between the arrivals led to a higher blocking probability compared to the case of independent arrivals. In the same way, negative correlated arrivals led to a lower blocking probability. Last, the case of batched arrivals was considered and the results evidenced that small batches do not have to be modelled in the considered setting.

However, further research is still necessary. First, as the simulation study evidences that a dependency between the arrivals has a high impact on the blocking probability, it is necessary to find a closed form approximation for the case of dependent arrivals. Following the Hayward approximation, one approach may be to extend the formula of the peakedness in the heavy-traffic approximation by a factor for the autocorrelation between the arrivals. Second, Bekker and Koeleman (2011) use the fluid approximation to model the non-stationary case. Nevertheless, the simulation results indicate that there may exist a better approach to capture the non-stationarity.

Third, Gerhardt and Nelson (2009) also propose the thinning method to transform independent non-Poisson arrivals. However, it is not clear whether this method can also be applied to dependent arrivals in the way that an index measure, such as the IDC, is preserved. Forth, as the number of staffed beds varies during a week, it is necessary to investigate whether the variation of the capacity has a high impact on the blocking probability or whether the blocking probability is possibly reduced due to a higher capacity during days of high admission.

This work considered the problem of modelling arrival processes in health care applications with respect to finite capacity. The consideration of non-Poisson arrival processes in health care applications is of great importance as shown by several studies. For the considered setting, methods to approximate or simulate the performance of the system exist. Specifically, the approaches by Bekker and Koeleman (2011) and Nelson and Gerhardt (2011) lead to good results. In the framework of time-dependent arrival rates, both approaches are applicable and imply the same conclusions, an increasing blocking probability with increasing variance and non-stationarity of the arrival process. However, the theoretical approximations by Bekker and Koeleman (2011) ignore a possible dependency of the arrivals which has a significant influence on the blocking probability, as evidenced by the simulation study performed in this work.

References

Alfa, A. S. and Frigui, I. (1996). Discrete NT-policy single server queue with markovian arrival process and phase type service. *European Journal of Operational Research*, 88(3):599–613.

Anderson, A. T., Neuts, M. F., and Nielsen, B. F. (2004). On the time reversal of Markovian arrival processes. *Stochastic Models*, 20(2):237–260.

Bause, F. and Horvath, G. (2010). Fitting Markovian Arrival Processes by incorporating correlation into phase type renewal processes. In *7th International Conference on Quantitative Evaluation of SysTems (QEST) 2010*, pages 97–106.

Bekker, R. and Koeleman, P. M. (2011). Scheduling admissions and reducing variability in bed demand. *Health Care Management Science*, 14(3):237–249.

Biller, B. and Nelson, B. L. (2005). Fitting time-series input processes for simulation. *Operations Research*, 53(3):549–559.

Bodenheimer, T. (2005). High and rising health care costs. Part 2: technologic innovation. *Annals of internal medicine*, 142(11):932–937.

Borovkov, A. (1967). On limit laws for service processes in multi-channel systems. *Siberian Mathematical Journal*, 8(5):746–763.

Çinlar, E. (1975). *Introduction to stochastic processes.* Prentice-Hall, Englewood Cliffs, NJ.

Creemers, S. and Lambrecht, M. (2009). Queueing models for appointment-driven systems. *Annals of Operation Research*, 178(1):155–172.

de Bruin, A., Bekker, R., van Zanten, L., and Koole, G. (2010). Dimensioning clinical wards using the Erlang loss model. *Operations Research*, 178:23–43.

de Bruin, A. M., van Rossum, A. C., m C, V., and Koole, G. M. (2007). Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science*, 10(2):125–137.

Deo, S. and Itai, G. (2011). Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science*, 57(7):1300–1319.

Fontantesi, J., Alexopoulos, C., Goldsman, D., DeGuire, M., Kopald, D., Holcomb, K., and Sawyer, M. (2002). Non-punctual patients: planning for variability in appointment arrival times. *Journal of Medical Practice Management*, 18(1):14–18.

Gerhardt, I. and Nelson, B. (2009). Transforming renewal process for simulation of nonstationary arrival processes. *Journal on Computing*, 21(4):630–640.

Glynn, P. W. and Whitt, W. (1991). A new view of the heavy-traffic limit theorem for infinite-server queues. *Advances in Applied Probability*, 23(1):188–209.

Gross, D. and Harris, C. M. (1985). *Fundamentals of Queueing Theory*. Wiley, New York, 2 edition.

Gusella, R. (1991). Characterizing the variability of arrival processes with indexes of dispersion. *IEEE Journal On Selected Areas In Communications*, 9(2):203–211.

Harper, P. R., Knight, V. A., and Marshall, A. H. (2011). Discrete conditional phase-type models utilising classification trees: Application to modelling health service capacities. *European Journal of Operational Research*, 219(3):522–530.

Holtzmann, J. and Jagerman, D. (1979). Estimating peakedness from arrival counts. In *Proceedings of ITC-9. Torremolinos, Spain.*

Izady, N. and Worthington, D. (2012). Setting staffing requirements for the time dependent queueing networks: The case of accident and emergency departments. *European Journal of Operations Research*, 219(3):531–540.

Jackson, J. (1957). Networks of waiting lines. *Operations Research*, 5(4):518–521.

Jagers, A. A. and Van Doorn, E. A. (1986). On the continued Erlang loss function. *Operations Research Letters*, 5(1):43–46.

Johnson, M. A. (1998). Markov MECO: a simple markovian model for approximating nonrenewal arrival processes. *Stochastic Models*, 14(1-2):419–442.

Johnson, M. A. and Narayana, S. (1996). Descriptors of arrival-process burstiness with application to the discrete Markovian arrival process. *Queueing Systems*, 23(1):107–130.

Johnson, M. A. and Taaffe, M. R. (1989). Matching moments to phase distributions: Mixture of Erlangs of common order. *Stochastic Models*, 5(4):711–743.

Kendal, D. (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):338–354.

Kim, S. (2011). The two-moment three-parameter decomposition approximation of queueing networks with exponential residual renewal processes. *Queueing Systems*, 68(2):193–216.

Lane, D., Monefeldt, C., and Rosenhead, J. (2000). Looking in the wrong place for healthcare improvements: A system dynamics study of an accident and emergency department. *Journal of the Operational Research Society*, 51(5):518–531.

Lewis, P. A. W. and Shedler, G. S. (1979). Simulation of nonhomogeneous Poisson processes by thinning. *Naval Research Logistics Quarterly*, 26.

Little, J. D. C. (1961). A proof for the queueing formula L= λW. *Operations Research*, 9:383–387.

Litvak, E., Buerhaus, P., Davidoff, F., and Long, M. (2005). Managing unnecessary variability in patient demand to reduce nursing stress and improve patient safety. *Joint Commission on quality and patient safety*, 31(6):330–338.

Lucantoni, D. M., Meier-Hellstern, K. S., and Neuts, M. F. (1990). A single-server queue with server vacations and a class of non-renewal arrival processes. *Advances in Applied Probability*, 22(3):676–705.

Massey, W. and Whitt, W. (1993). Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems*, 13(1):183–250.

McManus, M., Long, M., Copper, A., Mandell, J., Berwick, D., Pagano, M., and Litvak, E. (2003). Variability in surgical caseload and access to intensive care services. *Anesthesiology*, 98(6):1491–1496.

Melamed, B., Hill, J. R., and Goldsman, D. (1992). The TES methodology: Modeling empirical stationary time series. In Swain, J. J., Goldsman, D., Crain, R. C., and Wilson, J. R., editors, *Proc. 1992 Winter Simulation Conference*, pages 135–144, Institute of Electrical and Electronics Engineers, Piscataway, NJ.

Nelson, B. and Gerhardt, I. (2011). Modelling and simulating non-stationary arrival processes to facilitate analysis. *Journal of Simulation*, 5:3–8.

Payne, K., Marshall, A. H., and Cairns, K. J. (2012). Investigating the efficiency of fitting Coxian phase-type distributions to health care data. *IMA Journal of Management Mathematics*, 23(2):133–145.

Ramirez-Cobo, P. and Carrizosa, E. (2012). A note on the dependence structure of the two-state Markovian arrival process. *Journal of Applied Probability*, 49(1):295–302.

Restrepo, M., Henderson, S. G., and Topaloglu, H. (2009). Erlang loss models for the static deployment of ambulances. *Health Care Management Science*, 12(1):67–79.

Rolski, T. and Szekli, R. (1991). Stochastic ordering and thinning. *Stochastic Processes and their Applications*, 37(2):299–312.

Sauer, C. and Chandy, K. (1975). Approximate analysis of central server models. *IBM Journal of Research and Development*, 19(3):301–313.

Seshamani, M. and Gray, A. (2004). Time to death and health expenditure: an improved model for the impact of demographic change on the health care costs. *Age and Ageing*, 33(6):556–561.

Singh, V. (2006). Use of queuing models in health care. http://works.bepress.com/vikas_singh/4/.

Sriram, K. and Whitt, W. (1986). Characterizing superposition arrivals processes in packet multiplexers for voice and data. *IEEE Journal on Selected Areas in Communications*, 4(6):833–846.

Tijms, H. C. (1994). *Stochastic Models: An Algorithmic Approach.* Wiley, New York.

Wang, X., Qu, H., Xu, L., Han, X., and Zhang, J. (2010). A MAP fitting approach with joint approximation oriented to the dynamic resource provisioning in shared data centres. In *2010 Fifth International Conference on Networking, Architecture, and Storage*, pages 100–108.

Ware, P. P., Page, T. W., and Nelson, B. L. (1998). Automatic modeling of file system workloads using two-level arrival processes. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(3):305–330.

Whitt, W. (1984). Heavy-traffic approximations for service systems with blocking. *AT&T Bell Laboratories Technical Journal*, 63(5):689–708.

Whitt, W. (1992). Understanding the efficiency of multi-server service systems. *Management Science*, 38(5):708–723.

Whitt, W. (2002). *Stochastic-Process Limits: An Introduction to stochastic process limits and their application to queues.* Springer, New York.

Worthington, D. (2009). Reflections on queue modelling from the last 50 years. *Journal of the Operational Research Society*, 60:S82–S93.