An approximation of the inpatient distribution in hospitals with patient relocation using Markov chains

Andersen, Anders Reenberg; Nielsen, Bo Friis; Plesner, Andreas Lindhardt

Published in:
Healthcare Analytics

Link to article, DOI:
10.1016/j.health.2023.100145

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Healthcare Analytics 3 (2023) 100145

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

An approximation of the inpatient distribution in hospitals with patient relocation using Markov chains

Anders Reenberg Andersen [a,b,*], Bo Friis Nielsen [a], Andreas Lindhardt Plesner [a]

[a] Technical University of Denmark, Department of Applied Mathematics and Computer Science, Anker Engelunds Vej 1, 2800 Kgs. Lyngby, Denmark
[b] Innovation and Research Centre for Multimorbidity, Ingemannsvej 18, 4200 Slagelse, Denmark

ARTICLE INFO

Keywords:
Bed management
Inpatient flow
Markov chain
Queueing
Stochastic modeling

ABSTRACT

Many hospitals struggle with insufficient capacity for their inpatients. As a result, hospitals may benefit from an approach that evaluates the occupancy of inpatient wards. In this study, we approximate the occupancy distributions of inpatient wards, accounting for the cases where patients relocate due to a shortage of beds. The approximation employs a homogeneous continuous-time Markov chain to evaluate each ward as a queue containing multiple classes of patients. We avoid computational intractability by evaluating each ward separately and accommodating patients arriving from the remaining wards by interrupting the arrival processes, where the interruption times follow hyper-exponential distributions. Numerical experimentation shows that our approach is robust concerning the type of length-of-stay distribution and generally results in a minor loss of accuracy. Further validation indicates that our model reflects the occupancy distributions of inpatient wards in a Danish hospital.

1. Introduction

Inpatient admissions continue to challenge hospitals worldwide. In response to this development, many countries dedicate large fractions of their gross domestic product to healthcare, averaging 8.8% for the countries in the Organisation for Economic Co-operation and Development. Simultaneously, many hospitals aim to gain a higher throughput by reducing the bed capacity and enforcing a shorter length-of-stay for their inpatients [1], thus underlining the need for more efficient ways of utilizing the available capacity.

Several scientific studies seek to provide methods for improving bed capacity planning in hospitals. He et al. [2], Baru et al. [3] and Bhattacharjee and Ray [4] present the most recent reviews of the literature on the topic. All three reviews acknowledge the importance of providing hospitals with an efficient allocation of resources, and Baru et al. note that the problem is a recurrent topic in the literature. As a result, bed capacity planning for hospitals is a well-researched problem. Conversely, He et al. notice a potential for improving the patients' flow by considering multiple wards in the same model. The authors note that collaboration difficulties often arise between wards, but internal coordination and centralized bed allocation strategies can help wards balance their goals and improve the overall delivery of care. Thus, according to He et al., an optimal strategy cannot be obtained by considering wards as separate units.

In this paper, we address the conclusion made by He et al. by providing hospitals with a method for approximating the occupancy of multiple inpatient wards by considering the hospitals that relocate patients. We do not aim to describe the exact behavior of the inpatient flow but rather provide a model with manegable inputs that lead to adequate estimates of the wards' occupancy. Hospitals employing our approximation will be able to evaluate new resource allocation strategies, for instance reallocation of beds, creation of new wards, or rules for relocating patients.

1.1. Literature review

He et al. [2] and Baru et al. [3] divide the literature on bed capacity planning into two overall modeling approaches: Simulation-based models and Markov chain-based models, where the latter includes models from queueing theory. Bhattacharjee and Ray [4] make a similar conclusion for the wider scope of modeling patient flow in hospitals. He et al. find that simulation is the dominant modeling approach, whereas Markov chains are usually better suited for simpler systems. In addition, He et al. and Baru et al. find a few studies approaching bed capacity planning with mathematical programming.

Bekker et al. [5] and Dijkstra et al. [6] provide the most recent studies on bed capacity planning. Both studies accommodate the problem

* Corresponding author at: Technical University of Denmark, Department of Applied Mathematics and Computer Science, Anker Engelunds Vej 1, 2800 Kgs. Lyngby, Denmark.
E-mail address: arean@dtu.dk (A.R. Andersen).

https://doi.org/10.1016/j.health.2023.100145
Received 19 November 2022; Received in revised form 30 January 2023; Accepted 30 January 2023

of ensuring sufficient capacity for COVID-19 patients in the Netherlands. Bekker et al. present methods for predicting the admission rate and bed occupancy, and Dijkstra et al. optimize the balance of patients over multiple hospitals. Lam et al. [7] explore a similar approach, but for patients in Singapore. In general, many recent studies focus on either the allocation of resources [8–13] or patients [14]. Conversely, some studies focus mainly on estimating the demand for beds. Davis and Fard [15] provide theoretical bounds for an improved bed demand forecasting model, and Wu et al. [16] develop two heuristics for estimating the blocking probability in a tandem queueing system. de Bruin et al. [17] and Proudlove [18,19] analyze the demand for beds in a substantial number of hospital wards, but without inter-dependence and with fairly simple models.

Other studies combine an analytical model with simulation. For instance, Kokangul [20] uses simulation to estimate the number of control parameters and their associations in a non-linear optimization model. Wang et al. [21] present a simulation–optimization framework incorporating an analytical surrogate to the objective function. The framework recursively combines information from both a queueing model and simulated samples from the system. Bierlaire et al. [22] present a similar approach, but with an application to urban traffic.

Bekker et al. [23] study a problem in line with the findings of He et al., where the authors assess a range of bed management policies incorporating multiple wards. Bekker et al. find that a so-called *earmark policy* and a *threshold policy* result in close-to-optimal decisions. Andersen et al. [24] and Andersen et al. [25] present similar approaches by considering a set of inter-dependent wards. Both studies provide methods for deriving a close-to-optimal allocation of resources, but only for three wards.

In this paper, we extend the findings by Andersen et al. [24] by providing an approximation of the inpatient occupancy distributions for a large set of hospital wards. We assume that new arrivals can be relocated when all beds in the preferred ward are occupied. Only [23–25] consider analytical models containing multiple wards in the same hospital. However, Bekker et al. assume that blocked patients are lost from the system, and Andersen et al. provide an approach that is intractable for systems exceeding 3 wards.

He et al., Baru et al. and Bhattacharjee and Ray find that Markov chains often restrict the size of the model. We seek to overcome this obstacle with an aggregation method that results in only a minor loss of accuracy. Our approximation resembles aggregation methods such as Norton's theorem, Chandy et al. [26], and the generalization to state-dependent routing by Boucherie and van Dijk [27]. However, our approach differs from these methods by creating a dependence between the wards with relocated patients.

To summarize, our contributions to the literature are

- We provide an approximation of the inpatient occupancy distributions. Our approximation accounts for relocated patients and a substantial number of wards.
- We base the approximation on Markov chain modeling and show that our approach results in only a minor loss of accuracy.
- We validate our approximation by statistically comparing the model to the occupancy in a Danish hospital.

The rest of the paper is organized as follows: In Section 2 we present the details of the problem from both an overall and formal perspective. Next, in Section 3 we present our modeling approach, where the first part contains a non-aggregated approach to modeling the system, and the second part introduces the approximation. Section 4 contains our numerical experiments, and Section 5 validates our approximation using data from a Danish hospital. Finally, Section 6 presents our conclusions.

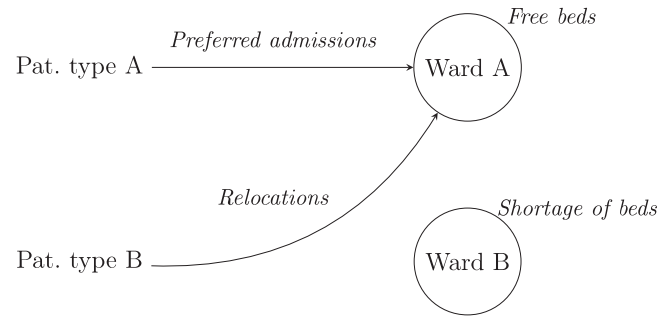

Fig. 1. Patients can be relocated to alternative wards when the beds in their preferred ward are in shortage.

2. Problem description

We study the problem of evaluating the occupancy distributions of inpatient wards in a hospital. We consider the wards that typically follow after the acute part of the patients' pathways, where patients can be characterized by their diagnoses and treatment needs. All patients receive the best care if they are admitted to a certain ward. Thus, if capacity was infinite, each ward would only have to treat a single type of patients. Naturally, wards contain a finite capacity of beds, and, therefore, patients cannot always be admitted to the preferred ward. Thus, in the case of bed shortage, patients will either have to receive care in temporary buffer-beds, leave the hospital, or relocate to an alternative ward.

Certain hospitals monitor their utilization of capacity closely, ensuring that patients become relocated such that only a few admissions occur in buffer-beds (cf. Fig. 1). This type of hospital is the focus of this study since they have become increasingly more common in Denmark. We elaborate on the formal details of our assumptions below. Note that Appendix C, Table C.13 contains an overview of the fundamental symbols in this paper.

2.1. Arrivals and length-of-stay

Let $\mathcal{W} = \{A, B, C, \dots\}$ denote the set of inpatient wards and $\mathcal{P} = \{A, B, C, \dots\}$ the set of patient types. We assume that each patient type in $\mathcal{P}$ prefers a ward in $\mathcal{W}$, and as a result $|\mathcal{W}| = |\mathcal{P}| = n$.

We further assume that patients arrive to the hospital according to a time-homogeneous Poisson process with rate $\lambda_p \in \mathbb{R}^+$, and that an admitted patient occupies a bed with random exponentially distributed length-of-stay with mean $1/\mu_p \in \mathbb{R}^+$, where in both cases $p \in \mathcal{P}$.

Section 5 validates these assumptions and show that they can be used for adequately approximating the occupancy distributions of inpatient wards in a hospital, even when the arrival rates depend on the day of the week.

2.2. Capacity and relocation of patients

Let $M_w \in \mathbb{N}$ denote the bed capacity of ward $w \in \mathcal{W}$, and $k_{wp} \in \mathbb{N}_0$ the number of patients of type $p \in \mathcal{P}$ that are currently admitted to ward $w \in \mathcal{W}$. When $w = p$, the variable k_{wp} accounts for admissions to a preferred ward, whereas when $w \neq p$, the variable k_{wp} accounts for patients that have been relocated from their preferred ward to the alternative ward w. A ward w accepts admissions as long as $\sum_{i \in \mathcal{P}} k_{wi} < M_w$. However, patients of type $p \neq w$ can only be admitted to ward w if $\sum_{i \in \mathcal{P}} k_{ai} = M_a$, where $a \in \mathcal{W}$ and $p = a$, i.e. when all the beds in the ward preferred by the patients of type p are occupied. We assume the relocated patients of type p choose an alternative ward w with a probability of $r_{pw} \in \mathbb{R}^+$. We further allow the sum $\sum_{i \in \mathcal{W}} r_{pi}$ to be less than 1, since patients may be relocated to a different hospital.

The assumptions in Sections 2.1–2.2 form a system consisting of n parallel queues, where patients are redirected whenever the queues are in shortage of beds. During a shortage, the system provides an alternative possibility for the patients to receive treatment in queues with idle capacity. Section 3 delves into this behavior by exploring the system further.

3. Modeling approach

Let the matrix $\boldsymbol{s} \in \mathbb{N}_0^{n\times n}$ denote the current *state* in the hospital, and $\mathcal{S}$ the associated state space. Furthermore, let k_{wp} define the elements of the matrix $\boldsymbol{s}$.

Now, recall the assumptions in Sections 2.1–2.2. The occupancy of beds in the hospital can be described by a homogeneous Continuous-Time Markov Chain (CTMC) with transition rates, $q_{\boldsymbol{s}\boldsymbol{s}^*} \in \mathbb{R}$,

$$q_{\boldsymbol{s}\boldsymbol{s}^*} = \begin{cases} \lambda_p & \text{if } k_{wp}+1 \text{ in } \boldsymbol{s}^*, \text{ where } w=p \text{ and } \sum_{i\in\mathcal{P}} k_{wi} < M_w \text{ in } \boldsymbol{s}. \\ \lambda_p r_{pw} & \text{if } k_{wp}+1 \text{ in } \boldsymbol{s}^*, \text{ where } w\neq p, \ \sum_{i\in\mathcal{P}} k_{ai} = M_a, \\ & p=a \text{ and} \\ & \sum_{i\in\mathcal{P}} k_{wi} < M_w \text{ in } \boldsymbol{s}. \\ \mu_p k_{wp} & \text{if } k_{wp}-1 \text{ in } \boldsymbol{s}^* \text{ and } k_{wp}>0 \text{ in } \boldsymbol{s}. \end{cases}$$

where the diagonal elements are $q_{\boldsymbol{s}\boldsymbol{s}} = -\sum_{\boldsymbol{s}^*\in\mathcal{S}\setminus\boldsymbol{s}} q_{\boldsymbol{s}\boldsymbol{s}^*}$ and all other transition rates $\boldsymbol{s}\neq\boldsymbol{s}^*$ are zero. In the remainder of this paper, we denote the above model *the complete CTMC*.

Note however that $|\mathcal{S}|$ can be intractably large, depending on n and the values of M_w. Andersen et al. [24] describe a similar CTMC, and show that the entire state space has exactly $|\mathcal{S}| = \prod_{i=1}^n \left(1/n! \prod_{j=1}^n (M_i + j)\right)$ states. As a result, a hospital with merely $n=6$ inpatient wards and $M_A = M_B = \cdots = 15$ beds leads to $|\mathcal{S}| \approx 25.5\cdot 10^{27}$ states. The memory requirements for the associated state distribution would be $2.0\cdot 10^{14}$PB. For this reason, Andersen et al.'s CTMC model will often lead to intractable computations for systems of realistic size. In this paper, we provide an approximation of the occupancy distributions for realistically sized systems. Section 3.1 elaborates on the details of our approach.

3.1. An approximation

In this section, we present an approximation of the complete CTMC. Our approximation puts the focus on one specific ward while the role of the other wards is merely to act as additional arrival processes to the ward being evaluated. As defined in Section 2.1, patients primarily arrive to wards according to independent Poisson processes. Thus, wards containing idle beds will not relocate patients to other wards, but when all beds are occupied, wards relocate patients with a certain probability to the ward in focus. As a result, the relocation process from a specific ward is Poisson whenever all beds are occupied. The process is what one could think of as an Interrupted Poisson Process (IPP). However, the term IPP is usually used in a slightly restricted setting.

The $M/M/c/c$ queue is a blocking system, meaning that customers arriving when all servers are busy are denied service and turned away. These so-called blocked customers are also sometimes denoted lost customers, and in a hospital, the lost customers form an overflow process.

The overflow process for the $M/M/1/1$ queue was analyzed by Kuczura [28] and termed IPP. Kuczura showed that the process of blocked customers given by the IPP is a renewal process with hyper-exponentially distributed arrival times.

In a hospital context, the time between two relocated patients is thus a Phase Type (PH) distribution (Neuts [29], p. 41; Bladt and Nielsen [30], p. 125). The PH distribution is a probability distribution defined by the time to absorption in a CTMC. Let $\boldsymbol{\beta} \in \mathbb{R}^o$ denote the probability vector of initializing in the $o \in \mathbb{N}$ phases of a PH distribution, and let $\boldsymbol{\Gamma} \in \mathbb{R}^{o\times o}$ denote the phase-type generator. The transition rate matrix of the CTMC associated with the PH distribution is

$$\begin{pmatrix} \boldsymbol{\Gamma} & \boldsymbol{\gamma} \\ \boldsymbol{0} & 0 \end{pmatrix},$$

where $\boldsymbol{\gamma} = -\boldsymbol{\Gamma}\boldsymbol{e}$ denotes the exit-rate vector, and $(\boldsymbol{\beta}, 0)$ the initial distribution. In this paper, we use the notation $(\boldsymbol{\beta}, \boldsymbol{\Gamma})$ for the representation of the PH distribution.

Kuczura [28] results in the two PH representations

$$(\boldsymbol{\beta}', \boldsymbol{\Gamma}') = \left((1,0), \begin{pmatrix} -\alpha-\gamma_1' & \gamma_1' \\ \gamma_2' & -\gamma_2' \end{pmatrix}\right), \text{and}$$

$$(\boldsymbol{\beta}, \boldsymbol{\Gamma}) = \left((\beta_1, \beta_2), \begin{pmatrix} -\gamma_1 & 0 \\ 0 & -\gamma_2 \end{pmatrix}\right)$$

that define the same distribution. The specific relationships between the two sets of parameters was given by Kuczura. The density $f(x)$ can be expressed as

$$f(x) = \boldsymbol{\beta}' e^{\boldsymbol{\Gamma}' x}\boldsymbol{\gamma}' = \boldsymbol{\beta} e^{\boldsymbol{\Gamma} x}\boldsymbol{\gamma} = \beta_1\gamma_1 e^{-\gamma_1 x} + \beta_2\gamma_2 e^{-\gamma_2 x},$$

where $\boldsymbol{\gamma}'$ and $\boldsymbol{\gamma}$ are the associated exit-rate vectors. Correspondingly, the process of blocked patients from the $M/M/c/c$ queue is also a renewal process with hyper-exponentially distributed time intervals, i.e. the density $f(x)$ can be expressed as

$$f(x) = \sum_{i=1}^{c+1} \beta_i\gamma_i e^{-\gamma_i x}.$$

In our approximation, we take inspiration from this result in the way we model the process of relocated patients to the ward in focus. In the $M/M/c/c$ queue, the time periods, where relocations can occur, are exponentially distributed. This is not true for hospitals, since inpatient wards contain multiple types of patients. These periods will instead have a hyper-exponential distribution. In general, the lengths of the time periods will be dependent as opposed to the case of the $M/M/c/c$ queue. Moreover, the time between successive transitions to the state with all beds occupied in the $M/M/c/c$ queue follows an H_c distribution. This follows from a reformulation of the result mentioned above on the H_{c+1} distribution. The relocation process from wards that only admit Poisson arrivals are modeled exactly by the $M/M/c/c$ queue, while it is not necessarily true in other cases.

3.1.1. Model definition

We exploit the hyper-exponential behavior in our approach by decomposing the complete CTMC into n different models, where each model accounts for the state transitions in a single ward. For the wards not in focus, the relocation process is approximated as sequences of independent hyper-exponentially distributed intervals by two different hyper-exponential distributions. One distribution models periods with relocation of patients corresponding to all beds being occupied, and another distribution corresponding to periods where beds are idle. We denote the wards with idle beds as *open*, and the wards with all beds occupied as being in *shortage*. The relocated patients are generated according to a Poisson process during intervals where the wards are in shortage. Thus, the approximation accounts for a total of $2(n-1)$ hyper-exponential distributions. By using this approach, the resulting state space is substantially reduced and far more computationally tractable than in the complete CTMC (cf. Section 3.2).

Consider the PH representations $(\boldsymbol{\beta}_w^{open}, \boldsymbol{\Gamma}_w^{open})$ and $(\boldsymbol{\beta}_w^{shortage}, \boldsymbol{\Gamma}_w^{shortage})$ of the hyper-exponential distributions accounting for when ward $w \in \mathcal{W}$ is *open* and in *shortage* of beds, respectively. The parameters o_w^{open} and $o_w^{shortage} \in \mathbb{N}$ denote the number of phases, and $\boldsymbol{\beta}_w^{open}$ and $\boldsymbol{\beta}_w^{shortage} \in \mathbb{R}^{o_w}$ denote the initial probability distributions of each hyper-exponential distribution. Additionally, the parameters $-\boldsymbol{\Gamma}_w^{open}\boldsymbol{e} = \boldsymbol{\gamma}_w^{open}$ and $-\boldsymbol{\Gamma}_w^{shortage}\boldsymbol{e} = \boldsymbol{\gamma}_w^{shortage} \in \mathbb{R}^{o_w}$ denote the exit-rate vectors. Let elements γ_{wj}^{open}, where $j \in \{1,2,\dots,o_i^{open}\}$, and $\gamma_{wj}^{shortage}$, where $j \in \{1,2,\dots,o_w^{shortage}\}$, denote the exit-rates associated with phase j in ward

w. Correspondingly, let elements β_{wj}^{open} and $\beta_{wj}^{shortage}$ denote the initial probabilities associated with phase j in ward w.

Now, let $\mathcal{T}_m$ denote the state space of the model for a ward in focus $m \in \mathcal{W}$. Further, let $\mathbf{k} \in \mathbb{N}_0^n$ denote a vector with elements k_p corresponding to the number of patients of type $p \in \mathcal{P}$ that are currently admitted to ward m. Also, let $\mathbf{h} \in \prod_{w \in \mathcal{W}\setminus\{m\}} \bigcup_{j \in \{open, shortage\}} \{1, 2, \dots, o_w^j\}$ denote the states of the $n-1$ wards that are not in focus. When ward w interrupts the relocation process we have $h_w \in \{1, 2, \dots, o_w^{open}\}$, where h_w is the current phase in the hyper-exponential distribution associated with the *open* ward. Correspondingly, when ward w is in *shortage* of beds, the current phase $h_w \in \{1, 2, \dots, o_w^{shortage}\}$. Let $\mathbf{b} \in \prod_{w \in \mathcal{W}\setminus\{m\}} b_w$, where $b_w \in \{open, shortage\}$ indicates whether h_w is associated with an *open* ward or a ward in *shortage* of beds. Components $\mathbf{h}$ and $\mathbf{b}$ therefore indicate the status of all wards in the set $\mathcal{W}\setminus\{m\}$. State $t \in \mathcal{T}_m$ combines both $\mathbf{k}$, $\mathbf{h}$, and $\mathbf{b}$. That is, $t = [\mathbf{k}, \mathbf{h}, \mathbf{b}]$.

Let the element m encompass both the ward and patients preferring the ward in focus. Further, let parameter $q_{tt^*}^m \in \mathbb{R}$ yield the transition rate from a current state $t \in \mathcal{T}_m$ to a new state $t^* \in \mathcal{T}_m$ of a time-homogeneous CTMC. Then,

$$q_{tt^*}^m = \begin{cases} \lambda_m & \text{if } k_m + 1 \text{ in } t^*, \text{ and } \sum_{i \in \mathcal{P}} k_i < M_m \text{ in } t. \\ \lambda_p r_{pm} & \text{if } k_p + 1 \text{ in } t^* \text{ where } p \neq m. \text{ Further} \\ & b_w = shortage, \\ & \text{where } p = w \text{ and } \sum_{i \in \mathcal{P}} k_i < M_m \text{ in } t. \\ \gamma_{wi}^{shortage} \beta_{wj}^{open} & \text{if } h_w = j \text{ and } b_w = open \text{ in } t^*. \text{ Further } h_w = i, \\ & \text{and } b_w = shortage \text{ in } t. \\ \gamma_{wi}^{open} \beta_{wj}^{shortage} & \text{if } h_w = j \text{ and } b_w = shortage \text{ in } t^*. \text{ Further } h_w = i \\ & \text{and } b_w = open \text{ in } t. \\ \mu_p k_p & \text{if } k_p - 1 \text{ in } t^* \text{ and } k_p > 0 \text{ in } t. \end{cases}$$

where the diagonal elements $q_{tt}^m = -\sum_{t^* \in \mathcal{T}_m \setminus t} q_{tt^*}^m$ and all other transition rates are zero.

The cases where $q_{tt^*}^m = \lambda_m$, and $\lambda_p r_{pm}$, account for patients arriving to ward m. Here, the latter is the rate of the relocation process associated with ward $w \neq m$. The case where $q_{tt^*}^m = \gamma_{wi}^{shortage} \beta_{wj}^{open}$, accounts for a transition into a state where ward w is *open*. That is, where the relocation process is interrupted. Note that β_{wj}^{open} is the probability that the PH distribution governing the *open* time will start in state j, and $1/\gamma_{wj}^{open}$ is the expected open time of the ward. Conversely, the case where $q_{tt^*}^m = \gamma_{wi}^{open} \beta_{wj}^{shortage}$ accounts for the transition into a state allowing relocated patients to enter ward m. The last case, $q_{tt^*}^m = \mu_p k_p$, accounts for the patients discharging from ward m.

Fig. 2 illustrates the behavior of a relocation process with $o_w^{shortage} = 2$ and $o_w^{open} = 2$ for ward w. Thus, two H_2 distributions govern the periods in which ward w is respectively *open* and in *shortage* of beds. The number inside each circle reflects the state in the respective distributions, and the arrows between the circles depict a transition from a state with *shortage* to an *open* state, and vice versa. The bold arrows depict an arrival of a relocated patient to ward m.

3.2. Solution approach

Let $\boldsymbol{Q}_m$ denote a matrix of order $|\mathcal{T}_m|$ where the matrix elements are the transition rates $q_{tt^*}^m$. Furthermore, let the row vector $\boldsymbol{\pi}_m \in \mathbb{R}^{|\mathcal{T}_m|}$, where $\|\boldsymbol{\pi}_m\|_1 = 1$, denote the stationary state probabilities of the approximation. The state probabilities are essential for analyzing the characteristics of ward m, such as the expected bed occupancy and shortage probability. In order to derive these measures, we need to solve the following linear system of equations,

$$\boldsymbol{\pi}_m \boldsymbol{Q}_m = \mathbf{0}. \tag{1}$$

The size of $\mathcal{T}_m$ can be intractably large. This especially applies to real-life systems, as we demonstrate later in Section 5. We therefore rely on the numerical approach of Gauss–Seidel (GS) [31, p.301] to solve the system in (1), and store matrix $\boldsymbol{Q}_m$ using a compact format. We note that similar iterative approaches are also applicable, such as successive over-relaxation and the power method, but we found these to converge much slower than GS in our initial tests. The details of how we generate the system in Eq. (1) are described in Section 3.2.1.

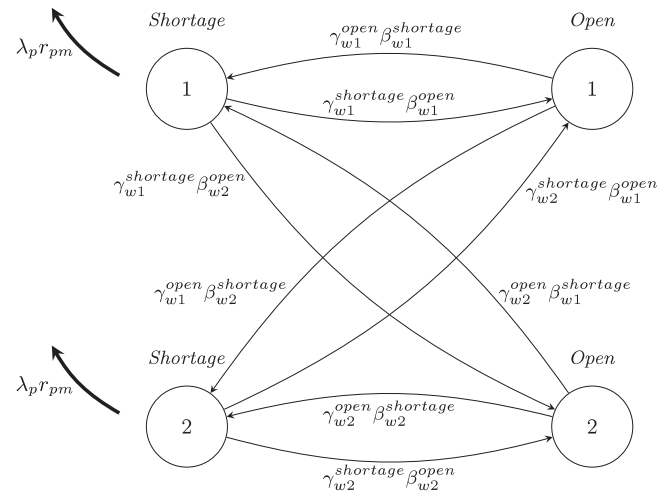

Fig. 2. Example of a relocation process associated with ward w. The circles reflect the states of two H_2 distributions. The arrows between the circles depict the transitions between states in the process, and the bold arrows the relocation of a patient to ward m.

In order to solve the system in Eq. (1), one obviously needs to start by defining the parameters of $\boldsymbol{Q}_m$, including the parameters of the PH distributions $(\boldsymbol{\beta}_w^{open}, \boldsymbol{\Gamma}_w^{open})$ and $(\boldsymbol{\beta}_w^{shortage}, \boldsymbol{\Gamma}_w^{shortage})$. For the cases where hospitals are only interested in the performance of the current system, the parameters can be fitted using patient data. For all other cases, e.g. if the hospital wants to analyze a new configuration of capacity, the parameters can be fitted using samples from a simulation of the periods where the wards are respectively open and in shortage of beds.

Finally, the model needs a decision of how many phases each hyper-exponential distribution should contain. In Sections 4–5, we conduct our computations using distributions with $o_w^{open} = 2$ (corresponding to the H_2 distribution), and distributions with $o_w^{shortage} = 1$ (corresponding to the exponential distribution).

3.2.1. Generating the model in practice

Consider the hyper-exponential distributions of the $n-1$ wards that are not in focus (i.e. the set $\mathcal{W}\setminus\{m\}$). Now, let $u \in \mathbb{N}$ denote the sum of the number of phases of both distributions, and for convenience assume that u is independent of ward w. That is, $o_A^{open} = o_B^{open} = \cdots = o^{open}$, $o_A^{shortage} = o_B^{shortage} = \cdots = o^{shortage}$, and $u = o^{open} + o^{shortage}$. In this case the state space of the approximation has a size of $|\mathcal{T}_m| = u^{n-1}/n! \prod_{j=1}^{n} (M_m + j)$ states.

Let $u^{max} = \max\{o^{open}, o^{shortage}\}$ and $u^{min} = \min\{o^{open}, o^{shortage}\}$. The maximum number of non-zero rates in any row of the transition rate matrix is $z^{max} = 1 + u^{max}(n-1) + 2n$ leading to an upper bound of $z^{max}|\mathcal{T}_m|$ non-zero elements in the entire transition rate matrix. The corresponding lower bound has $z^{min}|\mathcal{T}_m|$ non-zero elements in the entire matrix, where each row has a minimum of $z^{min} = 2 + u^{min}(n-1)$ non-zero rates. These bounds can be useful for determining the memory requirements for the transition rate matrix, $\boldsymbol{Q}_m$, and thus if the problem has a feasible size or not. Consider a hospital with $n = 6$ wards, $M_m = 15$ beds, $o^{open} = 2$ and $o^{shortage} = 1$ phases. The memory required to store $\boldsymbol{Q}_m$ in a compact format is in this case between 1.1 GB and 3.6 GB (assuming 12 bytes per non-zero element). Section 4.2 shows the actual memory usage of systems ranging between 2–6 wards.

In the process of computing the rates, it is possible to allocate the required memory one row at a time. First, one declares a two-dimensional array of rate values and an additional array of column indices (i.e. the indices of t^*). Second, one loops over each of the $|\mathcal{T}_m|$ states, and at each state declares two dummy vectors of size z^{max}. The non-zero transitions (both rates and column indices) of the current state are counted and stored in the dummy vectors. The count is then used to create the required number of slots for the non-zero transitions in the second dimension of the arrays from the first step. The rates and column indices are then finally copied into the newly allocated slots before moving on to the next state.

3.2.2. Truncation

We may evaluate even larger systems by introducing local capacity limits on each patient type with only a small loss of accuracy. Consider for instance a hospital with plenty of beds where relocated patients are rare. This hospital can almost be characterized by a series of parallel $M/M/c/c$ queues, since the states accounting for the relocated patients can be neglected.

In this study, we truncate the state space using local capacity limits, and set the limits such that the probability of exceeding them is negligible. This probability is evaluated using the bound from Chebyshev's inequality with sample mean and variance [32]. That is, for random variable X,

$$Pr\{|X-\hat{\mu}_p| \geq L_p\hat{\sigma}_p\} \leq \frac{g_{N+1}\left(\frac{NL_p^2}{N-1+L_p^2}\right)}{N+1}\left(\frac{N}{N+1}\right)^{1/2}, \tag{2}$$

where N is the sample size, $\hat{\mu}_p$ is the sample mean, and $\hat{\sigma}_p$ the sample standard deviation for patient type $p \in \mathcal{P}$. These parameters can be estimated using either patient data from the hospital or simulation, similar to the parameters for $(\boldsymbol{\beta}_w^{open}, \boldsymbol{\Gamma}_w^{open})$ and $(\boldsymbol{\beta}_w^{shortage}, \boldsymbol{\Gamma}_w^{shortage})$. We evaluate the function g_{N+1} using the approach in [32].

When we derive the local capacity limit $y_p > 0$, where $p \in \mathcal{P}$, we set $L_p = (y_p - \hat{\mu}_p)/\hat{\sigma}_p$ and calculate the bounding probability using Eq. (2). Starting with a value of $y_p = 1$, we increment y_p until Eq. (2) yields a bounding probability below a certain threshold. In this study, we use a threshold of $1 \cdot 10^{-3}$ through all of our numerical tests in Section 4.

3.3. Simulation of the complete CTMC

Discrete Event Simulation (DES) is a robust alternative to evaluating the occupancy distributions of the complete CTMC. In our subsequent experiments, we employ DES to assess the error of our approximation, and to compare the complete CTMC to the occupancy in a real-life hospital.

Our DES implementation use exactly the same features and parameters as described for the complete CTMC. Thus, patients are generated according to time-homogeneous Poisson processes, and they stay at the hospital for an exponentially distributed time. Similarly, the bed capacities and relocation probabilities determine the ward of admission.

We evaluate the occupancy distributions of each ward by observing the current number of admissions at the arrival of a new patient. The resulting frequency distributions lead to an estimate of the occupancy distributions.

In all of our experiments, we start the DES by generating a new arrival to an empty system. For this reason, we let the system stabilize (also denoted *burn-in*) prior to observing the wards' occupancy distributions. In most of our experiments, the simulation stops when an event occurs after a predefined time-limit. We denote this limit the *overall simulation time*. The description of each experiment provides the remaining details.

4. Numerical study

In this section, we present the numerical experiments that validate the error and sensitivity to the length-of-stay distribution of our approximation.

We implemented the approximation and the DES in the *C++* programming language. A single program containing both models is available for download at *GitHub* (see [33]). For the approximation, we used an Expectation–Maximization (EM) algorithm from the *EMpht* program to fit the parameters for the hyper-exponential distributions [34]. Asmussen et al. [35] provide the details of the algorithm. The source code for the *EMpht* program is written in *C* and incorporated directly into our implementation of the approximation.

All experiments were conducted on an HPC-system using an Intel Xeon Processor 2660v3 with ten 2.60 GHz cores (though our implementation only utilized a single core). Each job was allocated a maximum of 128 GB of memory.

4.1. Measures of error

We employed two measures of error to validate the difference between our approximation and the complete CTMC.

Let d_j and $\hat{d}_j$ denote the marginal probability of j occupied beds, where $j \in \{0, 1, \ldots, M_w\}$ in the complete CTMC and approximation, respectively. Furthermore, let $D_j = \sum_{i=0}^{j} d_i$ and $\hat{D}_j = \sum_{i=0}^{j} \hat{d}_i$ denote the cumulative probabilities. Our first measure evaluates the supremum difference between the cumulative probabilities,

$$\epsilon_0 = \sup_{j\in\{0,1,\ldots,M_w\}} \{|D_j - \hat{D}_j|\}. \tag{3}$$

Although typically used for continuous distributions, the measure in Eq. (3) resembles the test statistic from the Kolmogorov–Smirnov test.

Our second measure is the Goodness of Fit (GOF) definition proposed by de Bruin et al. [17], which is further based on Kleijnen et al. [36]. The measure evaluates the similarity of two probability distributions through,

$$\epsilon_1 = 1 - \frac{1}{2}\sum_{j=0}^{M_w} |d_j - \hat{d}_j|. \tag{4}$$

Here, $0 \leq \epsilon_1 \leq 1$, and $\epsilon_1 = 1$ if the distributions are equal.

4.2. Assessment of error

In this section, we assess the memory usage and error of the approximation by comparing the approximated occupancy distributions to the occupancy distributions from the complete CTMC in systems containing 2–6 wards.

The complete CTMC. The systems containing 2–3 wards were evaluated numerically using the power method [31, p. 301], whereas the systems containing 4–6 wards were evaluated using DES. The simulations, used a burn-in time of 365 days and an overall simulation time of 1 825 000 days.

The approximation. We fitted the parameters of the hyper-exponential distributions based on simulations of the system. The simulations used a burn-in time of 365 days and was not stopped until each ward had at least 50 samples of periods where the wards had been open and in shortage of beds.

The approximated occupancy distributions were replicated 50 times for systems containing 2–3 wards, and 3 times for systems containing 4–6 wards. These replications were conducted to account for the variation of the parameter estimates (although we eventually discovered that the variation was close to negligible).

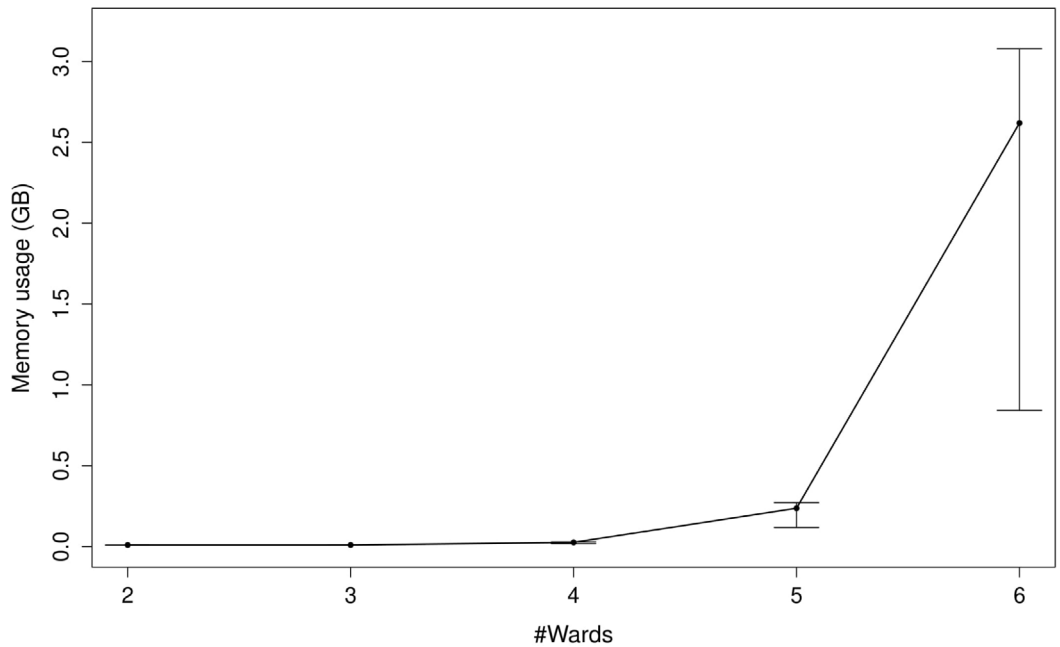

Fig. 3. The average recorded memory usage for systems containing 2–6 wards. The horizontal bars reflect the minimum and maximum usage for each number of wards.

4.2.1. Input parameters

All tests used a fixed mean length-of-stay of $1/\mu_A = 1/\mu_B = \cdots = 1/\mu = 10$ days. The systems containing 2–3 wards used a fixed capacity of $M_A = M_B = \cdots = M = 3$ beds, and the systems containing 4–6 wards used a fixed capacity of 15 beds.

For the arrival rate, we let $\lambda_A = \lambda_B = \cdots = \lambda = \rho M \mu$ with ρ varying from 0.5 to 0.9 with an increment of 0.1. In all of our tests, the patients were relocated with a uniform distribution over the alternative wards, leading to equal occupancy distributions in all wards of the system.

4.2.2. Results

Fig. 3 illustrates the average recorded memory usage of the approximation. Each level of ρ leads to a different state space truncation (cf. Section 3.2.2). The memory requirements, therefore, differ between experiments with the same number of wards. The horizontal bars in Fig. 3 reflect the minimum and maximum memory usage for each number of wards.

Table 1 presents the average and standard deviation shortage probability, supremum error and GOF of the 3–50 replications of the approximation. The confidence intervals of the simulated occupancy distributions had a maximum difference of $3.3 \cdot 10^{-3}$ for a confidence level of 0.05 [37].

All of our tests resulted in occupancy distributions that were close to the distributions of the complete CTMC. The supremum error never exceeded an average value of $2.96 \cdot 10^{-2}$, and the GOF remains above an average value of $9.70 \cdot 10^{-1}$. We found that the error of systems with 4 to 6 wards were quite similar, with the exception of one outlier in the system with 4 wards. The 2 and 3-ward systems separated themselves from the rest by featuring a lower slope (see Fig. 4). This behavior was likely due to the ward capacity increasing from 3 to 15 beds simultaneously to the system's size.

As expected, the error seemed to be an increasing function of ρ and the shortage probability of the complete CTMC, which was a result of the increased influence of the approximated relocation processes. Conversely, the tests did not indicate that the error would increase substantially for systems where the loads are larger than in these experiments.

4.3. Sensitivity to the length-of-stay distribution

The length-of-stay distributions of inpatients can be far from exponentially distributed (see e.g. [24]). We therefore assessed the sensitivity of our approximation by comparing the model to simulations of the complete system, where we replaced the exponential length-of-stay distributions with log-normal distributions. The configurations of the complete system and the approximation were otherwise identical to the tests in Section 4.2.

4.3.1. Input parameters

We tested systems containing 2 to 6 wards with ρ ranging from 0.5 to 0.9. In addition, we used a fixed mean length-of-stay of $1/\mu = 10$ days across all patient types in $\mathcal{P}$. The standard deviation, σ, of the simulated log-normal distribution was tested on five different levels: $1/(2\mu)$, $1/\mu$, $2/\mu$, $4/\mu$ and $6/\mu$.

4.3.2. Results

The resulting occupancy distributions were close to the distributions of the complete system across all tests (see Table 2). Specifically, the average value of ϵ_0 never exceeded $3.02 \cdot 10^{-2}$, and ϵ_1 remained above an average value of $9.81 \cdot 10^{-1}$. Switching to the log-normal distribution and increasing the standard deviation to above the level of the exponential distribution appeared to have a negligible effect. Fig. 5 illustrates this behavior. The largest relative increase in mean error from $1/\mu_i$ to $6/\mu_i$ is merely 6%. This result was likely caused by the close resemblance to the $M/M/c/c$ queueing system. In the $M/M/c/c$ queue, the state distribution is completely insensitive to the type of the service-time distribution [38, p. 122].

5. Application to a hospital case

In this section, we use data from a real hospital to validate our approach. The purpose of the section is to demonstrate that the model's assumptions (cf. Section 2.1–2.2) adequately reflect the wards' occupancy, and that the model can be employed to evaluate changes to the organizational structure in a hospital.

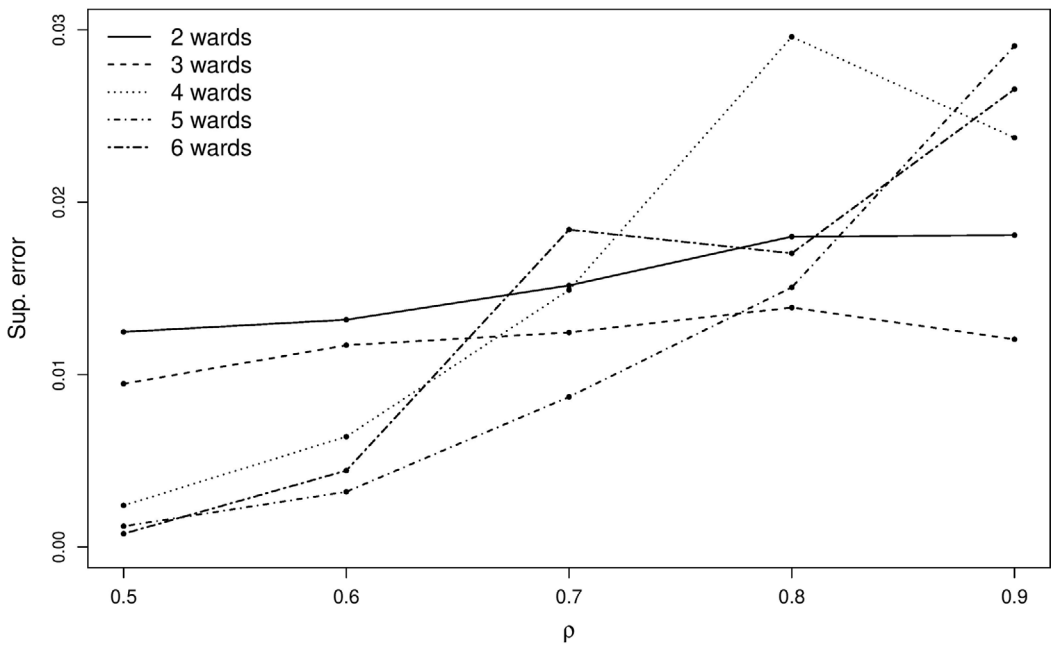

Fig. 4. The supremum error (cf. Eq. (3)) as function of ρ for systems containing 2–6 wards.

Table 1
Assessment of error between the approximation and the complete CTMC. Shows the average and standard deviation of the shortage probability, sup. error and GOF of the 3–50 replications of the approximation.

#Wards	ρ	Shortage prob.	Shortage prob.		Sup. error		GOF	
		Complete CTMC	Average	Std. dev.	Average	Std. dev.	Average	Std. dev.
2	0.5	$1.70\cdot10^{-1}$	$1.73\cdot10^{-1}$	$9.17\cdot10^{-3}$	$1.25\cdot10^{-2}$	$9.70\cdot10^{-3}$	$9.87\cdot10^{-1}$	$9.64\cdot10^{-3}$
2	0.6	$2.35\cdot10^{-1}$	$2.39\cdot10^{-1}$	$1.15\cdot10^{-2}$	$1.32\cdot10^{-2}$	$1.04\cdot10^{-2}$	$9.87\cdot10^{-1}$	$1.03\cdot10^{-2}$
2	0.7	$2.99\cdot10^{-1}$	$3.04\cdot10^{-1}$	$1.42\cdot10^{-2}$	$1.52\cdot10^{-2}$	$1.10\cdot10^{-2}$	$9.85\cdot10^{-1}$	$1.09\cdot10^{-2}$
2	0.8	$3.58\cdot10^{-1}$	$3.69\cdot10^{-1}$	$1.70\cdot10^{-2}$	$1.80\cdot10^{-2}$	$1.25\cdot10^{-2}$	$9.82\cdot10^{-1}$	$1.24\cdot10^{-2}$
2	0.9	$4.12\cdot10^{-1}$	$4.25\cdot10^{-1}$	$1.69\cdot10^{-2}$	$1.81\cdot10^{-2}$	$1.16\cdot10^{-2}$	$9.82\cdot10^{-1}$	$1.15\cdot10^{-2}$
3	0.5	$1.72\cdot10^{-1}$	$1.74\cdot10^{-1}$	$7.03\cdot10^{-3}$	$9.48\cdot10^{-3}$	$7.66\cdot10^{-3}$	$9.90\cdot10^{-1}$	$7.63\cdot10^{-3}$
3	0.6	$2.40\cdot10^{-1}$	$2.44\cdot10^{-1}$	$9.55\cdot10^{-3}$	$1.17\cdot10^{-2}$	$8.70\cdot10^{-3}$	$9.88\cdot10^{-1}$	$8.56\cdot10^{-3}$
3	0.7	$3.07\cdot10^{-1}$	$3.12\cdot10^{-1}$	$1.28\cdot10^{-2}$	$1.24\cdot10^{-2}$	$1.09\cdot10^{-2}$	$9.87\cdot10^{-1}$	$1.08\cdot10^{-2}$
3	0.8	$3.69\cdot10^{-1}$	$3.77\cdot10^{-1}$	$1.23\cdot10^{-2}$	$1.39\cdot10^{-2}$	$8.80\cdot10^{-3}$	$9.86\cdot10^{-1}$	$8.65\cdot10^{-3}$
3	0.9	$4.24\cdot10^{-1}$	$4.33\cdot10^{-1}$	$1.16\cdot10^{-2}$	$1.20\cdot10^{-2}$	$8.84\cdot10^{-3}$	$9.88\cdot10^{-1}$	$8.77\cdot10^{-3}$
4	0.5	$5.86\cdot10^{-3}$	$6.01\cdot10^{-3}$	$1.12\cdot10^{-4}$	$2.41\cdot10^{-3}$	$3.74\cdot10^{-4}$	$9.98\cdot10^{-1}$	$2.66\cdot10^{-4}$
4	0.6	$2.33\cdot10^{-2}$	$2.31\cdot10^{-2}$	$1.15\cdot10^{-3}$	$6.40\cdot10^{-3}$	$3.17\cdot10^{-3}$	$9.94\cdot10^{-1}$	$3.17\cdot10^{-3}$
4	0.7	$6.69\cdot10^{-2}$	$6.53\cdot10^{-2}$	$5.39\cdot10^{-3}$	$1.49\cdot10^{-2}$	$7.62\cdot10^{-3}$	$9.85\cdot10^{-1}$	$6.99\cdot10^{-3}$
4	0.8	$1.42\cdot10^{-1}$	$1.49\cdot10^{-1}$	$1.20\cdot10^{-2}$	$2.96\cdot10^{-2}$	$2.39\cdot10^{-2}$	$9.70\cdot10^{-1}$	$2.30\cdot10^{-2}$
4	0.9	$2.35\cdot10^{-1}$	$2.42\cdot10^{-1}$	$6.92\cdot10^{-3}$	$2.37\cdot10^{-2}$	$9.05\cdot10^{-3}$	$9.76\cdot10^{-1}$	$7.93\cdot10^{-3}$
5	0.5	$6.24\cdot10^{-3}$	$6.00\cdot10^{-3}$	$6.26\cdot10^{-5}$	$1.20\cdot10^{-3}$	$5.74\cdot10^{-4}$	$9.98\cdot10^{-1}$	$2.73\cdot10^{-4}$
5	0.6	$2.40\cdot10^{-2}$	$2.35\cdot10^{-2}$	$3.55\cdot10^{-4}$	$3.20\cdot10^{-3}$	$1.55\cdot10^{-3}$	$9.96\cdot10^{-1}$	$1.17\cdot10^{-3}$
5	0.7	$6.66\cdot10^{-2}$	$6.50\cdot10^{-2}$	$2.50\cdot10^{-3}$	$8.71\cdot10^{-3}$	$3.46\cdot10^{-3}$	$9.91\cdot10^{-1}$	$3.07\cdot10^{-3}$
5	0.8	$1.42\cdot10^{-1}$	$1.41\cdot10^{-1}$	$5.79\cdot10^{-3}$	$1.51\cdot10^{-2}$	$6.19\cdot10^{-3}$	$9.84\cdot10^{-1}$	$5.46\cdot10^{-3}$
5	0.9	$2.35\cdot10^{-1}$	$2.22\cdot10^{-1}$	$2.08\cdot10^{-2}$	$2.91\cdot10^{-2}$	$1.37\cdot10^{-2}$	$9.71\cdot10^{-1}$	$1.36\cdot10^{-2}$
6	0.5	$6.04\cdot10^{-3}$	$5.97\cdot10^{-3}$	$1.08\cdot10^{-5}$	$7.68\cdot10^{-4}$	$2.14\cdot10^{-4}$	$9.99\cdot10^{-1}$	$1.98\cdot10^{-4}$
6	0.6	$2.32\cdot10^{-2}$	$2.35\cdot10^{-2}$	$1.74\cdot10^{-4}$	$4.43\cdot10^{-3}$	$1.10\cdot10^{-3}$	$9.96\cdot10^{-1}$	$1.09\cdot10^{-3}$
6	0.7	$6.51\cdot10^{-2}$	$6.86\cdot10^{-2}$	$2.39\cdot10^{-3}$	$1.84\cdot10^{-2}$	$8.13\cdot10^{-3}$	$9.82\cdot10^{-1}$	$8.13\cdot10^{-3}$
6	0.8	$1.40\cdot10^{-1}$	$1.40\cdot10^{-1}$	$1.12\cdot10^{-2}$	$1.70\cdot10^{-2}$	$1.79\cdot10^{-2}$	$9.82\cdot10^{-1}$	$1.72\cdot10^{-2}$
6	0.9	$2.35\cdot10^{-1}$	$2.40\cdot10^{-1}$	$1.66\cdot10^{-2}$	$2.66\cdot10^{-2}$	$9.04\cdot10^{-3}$	$9.73\cdot10^{-1}$	$9.00\cdot10^{-3}$

We begin by presenting an overview of the data and parameters relating to the inpatient wards in the hospital. Next, we present our data in a statistical validation of the complete CTMC. The DES evaluates the state probabilities of the complete CTMC, since deriving an analytical (as well as numerical) solution was computationally intractable. After validating the model, we show that the difference between the approximation, the simulated complete CTMC and the observed occupancy is minor. Finally, we present an example of evaluating a new organizational structure using our approximation.

Our case is based on 11 inpatient wards in a Danish hospital. Arrivals to these wards often originate as acute admissions, and the majority of arrivals are therefore random non-elective admissions. The hospital uses a centralized coordination unit to control the patient flow. The coordination unit monitors the occupancy in the 11 wards and ensures that patients are relocated according to the system described in Section 2.

5.1. Data and parameters

Table 3 presents an overview of the ward and patient type characteristics. We omit the underlying diagnoses and clinical specializations to ensure anonymity of the hospital. The data for our study covers the

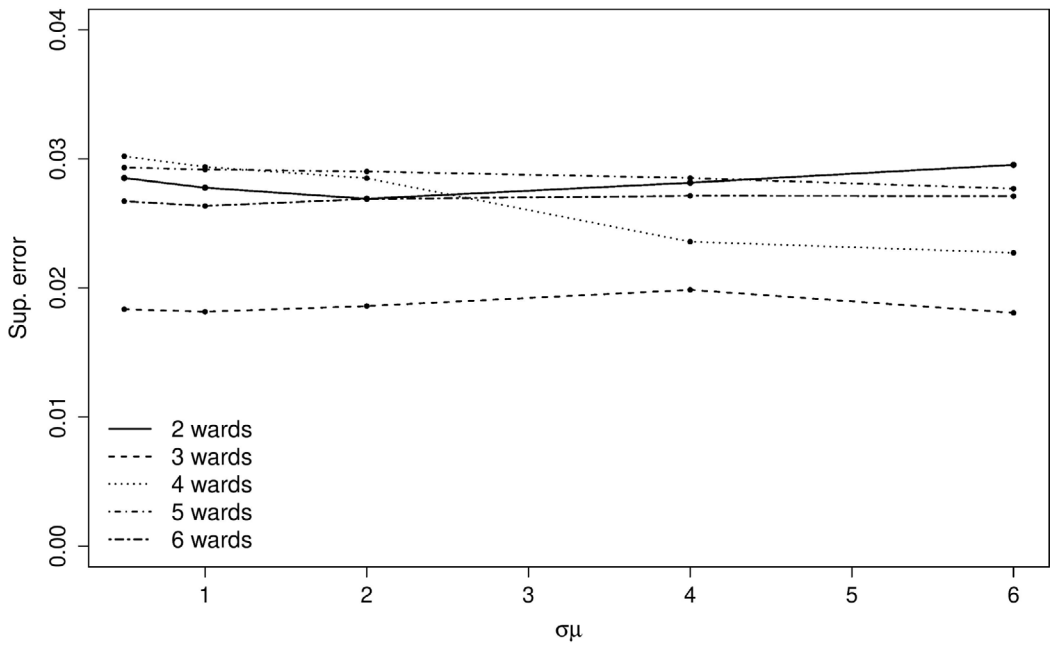

Fig. 5. The approximation compared to the complete systems containing 2–6 wards with log-normal length-of-stay distribution. The abscissa accounts for the standard deviation, and the ordinate accounts for the maximum average sup. error across the five (0.5 to 0.9) varying levels of ρ.

Table 2
The sensitivity of the CTMC approximation to the standard deviation of the length-of-stay distribution. The asterisks indicate that we present the maximum average and standard deviation of the five levels (0.5–0.9) of ρ.

#Wards	$\sigma\mu$	Sup. error		GOF	
		Average*	Std. dev.*	Average*	Std. dev.*
2	1/2	$2.85 \cdot 10^{-2}$	$1.46 \cdot 10^{-2}$	$9.81 \cdot 10^{-1}$	$1.40 \cdot 10^{-2}$
2	1	$2.78 \cdot 10^{-2}$	$1.45 \cdot 10^{-2}$	$9.81 \cdot 10^{-1}$	$1.39 \cdot 10^{-2}$
2	2	$2.69 \cdot 10^{-2}$	$1.48 \cdot 10^{-2}$	$9.81 \cdot 10^{-1}$	$1.43 \cdot 10^{-2}$
2	4	$2.82 \cdot 10^{-2}$	$1.53 \cdot 10^{-2}$	$9.82 \cdot 10^{-1}$	$1.46 \cdot 10^{-2}$
2	6	$2.95 \cdot 10^{-2}$	$1.46 \cdot 10^{-2}$	$9.83 \cdot 10^{-1}$	$1.42 \cdot 10^{-2}$
3	1/2	$1.84 \cdot 10^{-2}$	$9.06 \cdot 10^{-3}$	$9.87 \cdot 10^{-1}$	$8.52 \cdot 10^{-3}$
3	1	$1.82 \cdot 10^{-2}$	$9.17 \cdot 10^{-3}$	$9.86 \cdot 10^{-1}$	$8.60 \cdot 10^{-3}$
3	2	$1.86 \cdot 10^{-2}$	$9.35 \cdot 10^{-3}$	$9.86 \cdot 10^{-1}$	$8.72 \cdot 10^{-3}$
3	4	$1.99 \cdot 10^{-2}$	$9.59 \cdot 10^{-3}$	$9.87 \cdot 10^{-1}$	$9.08 \cdot 10^{-3}$
3	6	$1.81 \cdot 10^{-2}$	$9.45 \cdot 10^{-3}$	$9.86 \cdot 10^{-1}$	$8.95 \cdot 10^{-3}$
4	1/2	$3.02 \cdot 10^{-2}$	$2.35 \cdot 10^{-2}$	$9.98 \cdot 10^{-1}$	$02.23 \cdot 10^{-2}$
4	1	$2.94 \cdot 10^{-2}$	$2.40 \cdot 10^{-2}$	$9.98 \cdot 10^{-1}$	$2.27 \cdot 10^{-2}$
4	2	$2.85 \cdot 10^{-2}$	$2.38 \cdot 10^{-2}$	$9.98 \cdot 10^{-1}$	$2.26 \cdot 10^{-2}$
4	4	$2.36 \cdot 10^{-2}$	$2.37 \cdot 10^{-2}$	$9.94 \cdot 10^{-1}$	$2.21 \cdot 10^{-2}$
4	6	$2.27 \cdot 10^{-2}$	$2.23 \cdot 10^{-2}$	$9.90 \cdot 10^{-1}$	$2.13 \cdot 10^{-2}$
5	1/2	$2.93 \cdot 10^{-2}$	$1.40 \cdot 10^{-2}$	$9.98 \cdot 10^{-1}$	$1.39 \cdot 10^{-2}$
5	1	$2.92 \cdot 10^{-2}$	$1.39 \cdot 10^{-2}$	$9.98 \cdot 10^{-1}$	$1.39 \cdot 10^{-2}$
5	2	$2.90 \cdot 10^{-2}$	$1.36 \cdot 10^{-2}$	$9.98 \cdot 10^{-1}$	$1.35 \cdot 10^{-2}$
5	4	$2.85 \cdot 10^{-2}$	$1.35 \cdot 10^{-2}$	$9.95 \cdot 10^{-1}$	$1.34 \cdot 10^{-2}$
5	6	$2.77 \cdot 10^{-2}$	$1.32 \cdot 10^{-2}$	$9.93 \cdot 10^{-1}$	$1.31 \cdot 10^{-2}$
6	0.5	$2.67 \cdot 10^{-2}$	$1.67 \cdot 10^{-2}$	$9.98 \cdot 10^{-1}$	$1.55 \cdot 10^{-2}$
6	1	$2.64 \cdot 10^{-2}$	$1.65 \cdot 10^{-2}$	$9.98 \cdot 10^{-1}$	$1.57 \cdot 10^{-2}$
6	2	$2.69 \cdot 10^{-2}$	$1.74 \cdot 10^{-2}$	$9.98 \cdot 10^{-1}$	$1.67 \cdot 10^{-2}$
6	4	$2.71 \cdot 10^{-2}$	$1.82 \cdot 10^{-2}$	$9.95 \cdot 10^{-1}$	$1.76 \cdot 10^{-2}$
6	6	$2.71 \cdot 10^{-2}$	$1.84 \cdot 10^{-2}$	$9.94 \cdot 10^{-1}$	$1.73 \cdot 10^{-2}$

period from the 1st of Jan., 2019 to the 29th of Feb., 2020 and was obtained from the hospital's patient register system containing admission and discharge times, wards, diagnoses, and ages of the admitted patients. From these data, we derived the mean number of arrivals per day, the mean length-of-stay of each patient type (cf. second and third column in Table 3), and relative frequencies reflecting the relocation of patients in the system. The latter is presented in Appendix A, Table A.9.

Further analysis indicated that days of the week have an effect on the mean number of arrivals. Specifically, Appendix B, Table B.11 indicates that fewer patients arrive during the weekend, whereas the weekdays coincide with the overall mean. Section 5.3 delves into the impact of this behavior on the model's adequacy.

For the input parameters, we let the arrival rate, λ_p, equal the mean number of arrivals, the rate, μ_p, equal the reciprocal of the mean length-of-stay, and the probabilities r_{pw} equal the relative frequencies in Table A.9.

The ward capacities, M_w, were obtained from the documentation of the capacity coordination meetings in the hospital, where the values in parentheses indicate the allocated number of hallway-beds. In this paper, we include the hallway-beds in the wards' capacity. The observed ward occupancy distributions were obtained from the period 1st of Dec., 2019 to 29th of Feb., 2020.

5.2. Statistical validation

We validated the complete CTMC by comparing the simulated occupancy distributions to the observed distributions from the hospital. Three types of hypothesis tests were employed based on the null-hypothesis that the complete CTMC reflects the occupancy of the real wards. Each test was conducted by estimating the distribution of the test statistic (under the null-hypothesis) through repeated simulations of the complete CTMC. Regardless of the specific test statistic, we stopped each simulation when the sum of frequencies in the ward occupancy distributions reached the same number of observations as in the observed occupancy distributions. The observed test statistic was then compared to the set of the simulated statistics to determine a p-value for the test.

We use this approach, since Pearson's χ^2 test requires that observations are independent, in which case the test statistic would follow the χ^2 distribution. However, in our system, successive observations of the ward occupancy are dependent entailing that the theoretical sampling distribution is unknown.

In our first tests, we used the supremum difference in Eq. (3) as our test statistic. The remaining tests used Pearson's statistic, $\sum_{j \in 0,1,\ldots,M_w} (\omega_j - e_j)^2 / e_j$, and the sum of the squared difference, $\sum_{j \in 0,1,\ldots,M_w} (\omega_j - e_j)^2$, respectively. Here, $\omega_j \in \mathbb{N}$ and $e_j \in \mathbb{R}^+$ are the

Table 3
Estimated arrival rate, Length-Of-Stay (LOS), capacity and the bed load associated with each ward and patient type. The parentheses contain the number of hallway-beds.

Wards & patient types	Mean arrivals per day (λ_p)	Mean LOS in days ($1/\mu_p$)	Capacity (M_w)	Load per bed ($\lambda_p/M_w\mu_p$)
A	14.206	2.917	52 (4)	0.797
B	11.368	3.995	40 (2)	1.135
C	8.068	4.492	26 (2)	1.393
D	6.542	1.361	20 (0)	0.445
E	4.767	3.918	20 (0)	0.933
F	2.898	4.215	18 (2)	0.679
G	4.215	4.227	22 (0)	0.810
H	2.392	7.311	24 (0)	0.729
I	1.915	7.438	22 (2)	0.647
J	0.943	8.249	8 (0)	0.972
K	0.955	1.470	3 (0)	0.468

Table 4
Result of validating the null-hypothesis that the complete CTMC reflects the occupancy in the hospital. Tests are conducted using three different test statistics, where asterisks mark the p-values below 0.05.

Ward	Sup. error		Pearson's statistic		Squared difference	
	Value	*p*-value	Value	*p*-value	Value	*p*-value
A	0.16669	0.09194	780.65457	0.00824*	11 334.31632	0.08146
B	0.12893	0.12068	173.85506	0.04742*	4003.57897	0.05358
C	0.18811	0.00298*	151.99977	0.01706*	4870.66950	0.00148*
D	0.09397	0.28794	87.11908	0.13990	2246.25536	0.15884
E	0.19167	0.02458*	137.06917	0.02798*	3598.64270	0.01596*
F	0.10429	0.45320	37.11216	0.22472	680.16812	0.16426
G	0.05274	0.77682	22.56196	0.58638	604.62651	0.45924
H	0.14010	0.40852	64.49395	0.15550	1100.72300	0.07104
I	0.31426	0.05598	115.78919	0.04728*	1620.49825	0.00482*
J	0.20353	0.28164	12.44600	0.29426	83.68756	0.16846
K	0.33132	0.00005*	49.12755	0.00008*	790.36829	0.00048*

Table 5
Validation of error for the approximation (with modified parameters). Compares the approximation to a simulation of the complete CTMC (Approx. ⇔ Complete), the complete CTMC to a simulation containing the parameter modifications (Complete ⇔ Modified), and lastly the approximation to the observed occupancy distributions (Approx. ⇔ Observed).

Ward	Approx. ⇔ Complete		Modified ⇔ Complete		Approx. ⇔ Observed	
	Sup. error	GOF	Sup. error	GOF	Sup. error	GOF
A	$2.42\cdot10^{-2}$	$9.76\cdot10^{-1}$	$1.49\cdot10^{-3}$	$9.98\cdot10^{-1}$	$1.42\cdot10^{-1}$	$8.47\cdot10^{-1}$
B	$2.36\cdot10^{-3}$	$9.97\cdot10^{-1}$	$2.80\cdot10^{-4}$	$1.00\cdot10^{-0}$	$1.28\cdot10^{-1}$	$8.60\cdot10^{-1}$
C	$5.15\cdot10^{-3}$	$9.95\cdot10^{-1}$	$7.14\cdot10^{-4}$	$9.99\cdot10^{-1}$	$1.83\cdot10^{-1}$	$8.16\cdot10^{-1}$
D	$5.13\cdot10^{-3}$	$9.95\cdot10^{-1}$	$4.67\cdot10^{-4}$	$9.99\cdot10^{-1}$	$9.69\cdot10^{-2}$	$8.64\cdot10^{-1}$
E	$1.61\cdot10^{-2}$	$9.84\cdot10^{-1}$	$1.24\cdot10^{-3}$	$9.99\cdot10^{-1}$	$1.77\cdot10^{-1}$	$8.22\cdot10^{-1}$
F	$5.27\cdot10^{-2}$	$9.47\cdot10^{-1}$	$6.60\cdot10^{-4}$	$9.99\cdot10^{-1}$	$1.43\cdot10^{-1}$	$8.32\cdot10^{-1}$
G	$2.43\cdot10^{-2}$	$9.76\cdot10^{-1}$	$5.28\cdot10^{-4}$	$9.99\cdot10^{-1}$	$6.83\cdot10^{-2}$	$9.09\cdot10^{-1}$
H	$2.79\cdot10^{-2}$	$9.72\cdot10^{-1}$	$4.26\cdot10^{-3}$	$9.96\cdot10^{-1}$	$1.51\cdot10^{-1}$	$7.75\cdot10^{-1}$
I	$1.95\cdot10^{-2}$	$9.81\cdot10^{-1}$	$4.64\cdot10^{-3}$	$9.95\cdot10^{-1}$	$3.30\cdot10^{-1}$	$6.55\cdot10^{-1}$
J	$1.12\cdot10^{-2}$	$9.89\cdot10^{-1}$	$1.15\cdot10^{-3}$	$9.99\cdot10^{-1}$	$2.14\cdot10^{-1}$	$7.39\cdot10^{-1}$
K	$3.95\cdot10^{-2}$	$9.61\cdot10^{-1}$	$1.01\cdot10^{-2}$	$9.90\cdot10^{-1}$	$2.90\cdot10^{-1}$	$7.10\cdot10^{-1}$

Table 6
Shortage probabilities of each ward for the approximation (with modified parameters), a simulation of the complete CTMC, a simulation containing the modified parameters, and lastly the observed distributions.

Ward	Approximation	Complete CTMC	Modified CTMC	Observed
A	0.060	0.068	0.069	0.016
B	0.220	0.222	0.222	0.131
C	0.366	0.372	0.371	0.220
D	0.001	0.001	0.001	0.003
E	0.158	0.167	0.167	0.059
F	0.151	0.187	0.186	0.126
G	0.171	0.189	0.189	0.154
H	0.102	0.119	0.115	0.054
I	0.071	0.082	0.079	0.057
J	0.258	0.268	0.267	0.211
K	0.252	0.289	0.299	0.093

observed and expected frequencies in the ward occupancy distribution with j occupied beds.

Let matrix X_i denote the simulated frequencies of replication i, where element $x_{ij} \in \mathbb{N}_0$ denotes the frequency for $j \in \{0, 1, \dots, M_w\}$ occupied beds. We replicated the simulation a total of 50 000 times; hence $i \in \{1, 2, \dots 50000\}$. Now, let $\tau(\boldsymbol{X}_i, \boldsymbol{\pi}) \in \mathbb{R}^+$ denote the test statistic of replication i, where $\boldsymbol{\pi}$ is the state distribution of the complete CTMC. We derived $\boldsymbol{\pi}$ by simulating the system until convergence, which we found corresponds to a simulation time of 1 825 000 days. For the Pearson's and squared difference statistics, we evaluated e_j by multiplying the ward-marginal distribution from $\boldsymbol{\pi}$ by the number of observations. For the supremum difference statistic, we used the same ward-marginal distribution to estimate D_j (cf. Eq. (3)).

Let $\tau(\boldsymbol{\omega}, \boldsymbol{\pi})$ denote the test statistic for the observed frequencies, $\boldsymbol{\omega}$. The fraction of replications where $\tau(\boldsymbol{X}_i, \boldsymbol{\pi}) > \tau(\boldsymbol{\omega}, \boldsymbol{\pi})$ gives the p-value of the test.

5.2.1. Results

Table 4 presents the test statistic and the resulting p-values for each hospital ward. We found a notable difference in the sensitivity between the three tests. Pearson's statistic generally derived the smallest p-values, whereas Eq. (3) derived the largest. Using a significance level of 0.05, we found that Eq. (3) rejected the null-hypothesis in 3 cases, Pearson's statistic rejected 6 cases, and the squared difference rejected 4 cases. All three tests rejected Ward C, E and K, although the p-value of Ward E was borderline.

On the other hand, the three tests accepted 5 of the 11 wards, and the model seems to fit Ward G particularly well. This is confirmed by the graphical comparison in Fig. 6.

We also found that wards with hallway-beds often result in a low p-value. A descriptive investigation of these wards shows that the poor fit is often the result of an overestimation of the right-tail probabilities. The top graph of Fig. 6 illustrates this behavior. Conversely, our tests almost consistently accept wards *without* hallway-beds. Thus, the deviance between the model and the observed distributions must be largely due to the model not completely reflecting the hospital's policy regarding the use of hallway-beds. Patients might be discharged early to avoid the use of hallway-beds, and the hallway-beds might not be fully available, or only used when other wards are overcrowded.

The following section delves into the specific differences between our approximation, the complete CTMC, and the observed occupancy distributions.

5.3. Validation of the approximation

In this section, we compare the approximated occupancy distributions to the observed distributions from the hospital. Subsequently, we validate the effect of evaluating the system with a day-dependent arrival rate.

Our approximation would be computationally intractable if it was to account for the individual mean length-of-stays of all 11 patient types.

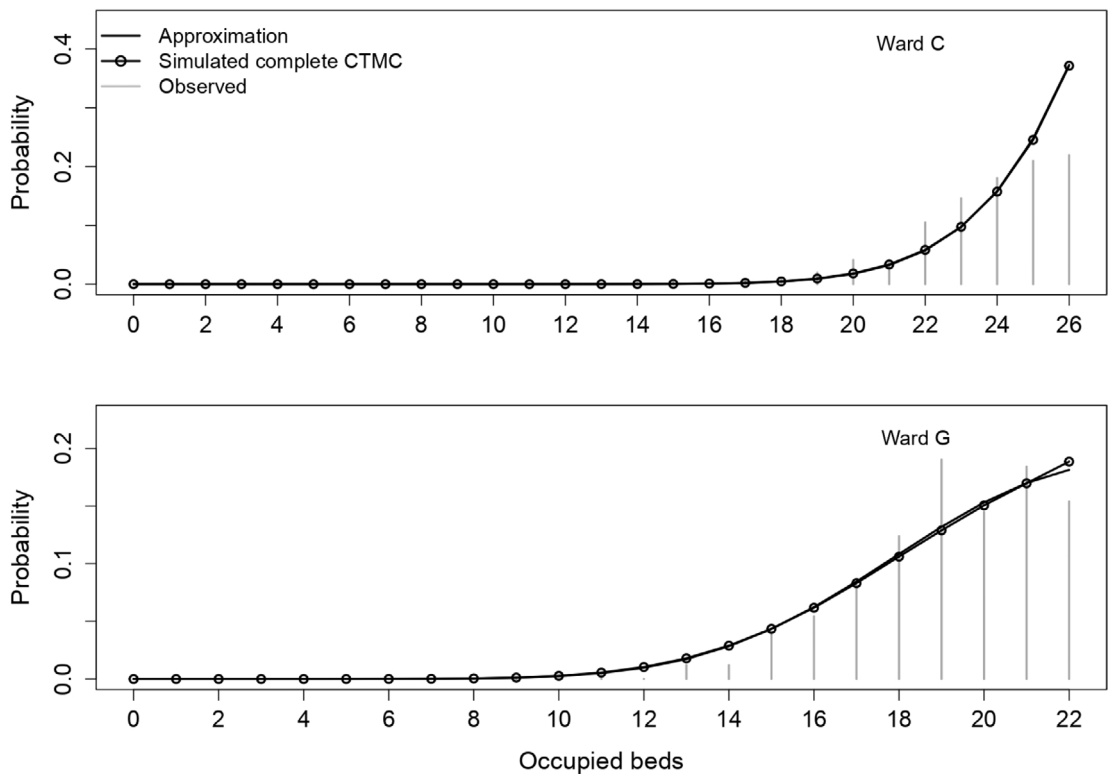

Fig. 6. The approximated, simulated (based on the complete CTMC) and observed ward occupancy distribution for Ward C and G, respectively.

Table 7
Mean occupancy of each ward. Shows the approximated occupancy (column 2) and the day-dependent occupancy (column 3–9). The last column displays the average occupancy of all seven days of the week.

Ward	Approx.	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	*Average*
A	44.77	41.33	43.60	45.65	46.92	46.89	44.96	41.81	44.45
B	37.17	35.31	36.75	37.33	37.68	37.52	36.42	34.83	36.55
C	24.48	23.56	24.13	24.28	24.35	24.41	24.14	23.51	24.05
D	9.11	9.11	9.40	9.53	10.24	10.35	9.57	8.73	9.56
E	16.81	16.08	16.40	16.78	16.91	17.21	17.13	16.35	16.69
F	14.75	13.57	14.34	14.81	15.28	15.46	14.91	13.87	14.61
G	18.89	18.08	18.47	18.85	19.07	19.37	19.12	18.30	18.75
H	19.62	18.67	19.14	19.68	20.07	20.47	20.29	19.25	19.65
I	16.95	16.36	16.53	16.76	17.13	17.63	17.79	17.00	17.03
J	6.24	5.63	5.98	6.16	6.27	6.25	6.41	6.04	6.11
K	1.70	1.54	1.81	1.80	1.85	1.86	1.81	1.70	1.77

Table 8
Shortage probability and expected daily number of preferred admissions of each ward in the current and new organizational structure. Ward L* comprises the consolidation of Ward A, C and G.

Ward	Capacity (M_w)	Current		New	
		Shortage prob.	Daily pref. adm.	Shortage prob.	Daily pref. adm.
A	52	0.06	13.35	–	–
B	40	0.22	8.87	0.21	9.02
C	26	0.37	5.11	–	–
D	20	0.00	6.54	0.00	6.54
E	20	0.16	4.01	0.15	4.07
F	18	0.15	2.46	0.09	2.64
G	22	0.17	3.49	–	–
H	24	0.10	2.15	0.06	2.26
I	22	0.07	1.78	0.04	1.83
J	8	0.26	0.70	0.22	0.73
K	3	0.25	0.71	0.24	0.72
L*	100	–	–	0.08	24.46

The minimum requirements for storing the non-zero transition rates for the largest ward are 4650.6PB. For this reason, we evaluated the system by employing the same mean length-of-stay to all patient types. Specifically, we set the mean length-of-stay to a value of 1, and the arrival rates to λ_i/μ_i to maintain the loads in the system. This allowed us to reduce the vector $\boldsymbol{k}$ to a scalar.

Consequently, the state space reduces to a size of $|\mathcal{T}_m| = u^{n-1}(M_m + 1)$. The resulting complete memory usage for the largest ward is now 1.0 GB. Since we no longer account for the differences between patients, we expect that the error of the approximation is larger than determined in Section 4.2. We evaluate the loss of accuracy by comparing the complete CTMC to a simulation of the system with the same parameter modifications as were used in the approximation.

5.3.1. Results

Table 5 compares the occupancy distributions of all 11 wards, and Fig. 6 visualizes the occupancy distributions of Ward C and G. Table 5 includes the errors, firstly between the modified approximation and a simulation of the complete CTMC, next between a simulation of the system with the same parameter modifications and the complete CTMC, and finally between the modified approximation and the observed occupancy distributions. Table 6 presents the associated shortage probabilities.

Table A.9
Relocation probabilities, r_{pw}, used in the evaluation of the hospital case. The last column shows the probability that a patient of type $p \in \mathcal{P}$ is relocated to another hospital instead of relocated to an alternative ward in the set $\mathcal{W}$.

$\mathcal{P}/\mathcal{W}$	A	B	C	D	E	F	G	H	I	J	K	Other
A	–	0.083	0.158	0.009	0.036	0.143	0.168	0.185	0.114	0.022	0.006	0.076
B	0.150	–	0.077	0.013	0.077	0.277	0.124	0.082	0.034	0.025	0.083	0.058
C	0.263	0.059	–	0.004	0.013	0.131	0.204	0.191	0.098	0.016	0.004	0.017
D	0.005	0.202	0.000	–	0.053	0.000	0.005	0.000	0.000	0.000	0.005	0.730
E	0.128	0.080	0.033	0.000	–	0.045	0.082	0.078	0.219	0.002	0.005	0.328
F	0.137	0.187	0.155	0.008	0.019	–	0.154	0.097	0.054	0.009	0.087	0.093
G	0.154	0.068	0.184	0.010	0.095	0.169	–	0.092	0.113	0.006	0.005	0.104
H	0.189	0.051	0.159	0.000	0.028	0.194	0.160	–	0.155	0.009	0.006	0.049
I	0.195	0.094	0.105	0.000	0.060	0.081	0.107	0.158	–	0.172	0.002	0.026
J	0.061	0.049	0.019	0.000	0.038	0.027	0.064	0.000	0.387	–	0.000	0.355
K	0.012	0.460	0.036	0.000	0.053	0.025	0.022	0.000	0.000	0.000	–	0.392

Table A.10
Relocation probabilities, r_{pw}, used for evaluating the new organizational structure. Ward and patient type L* denotes the merging of the former Ward A, C and G.

$\mathcal{P}/\mathcal{W}$	B	D	E	F	H	I	J	K	L*	Other
B	–	0.013	0.077	0.277	0.082	0.034	0.025	0.083	0.351	0.058
D	0.202	–	0.053	0.000	0.000	0.000	0.000	0.005	0.010	0.730
E	0.080	0.000	–	0.045	0.078	0.219	0.002	0.005	0.243	0.328
F	0.187	0.008	0.019	–	0.097	0.054	0.009	0.087	0.446	0.093
H	0.051	0.000	0.028	0.194	–	0.155	0.009	0.006	0.508	0.050
I	0.094	0.000	0.060	0.081	0.158	–	0.172	0.002	0.406	0.026
J	0.049	0.000	0.038	0.027	0.000	0.387	–	0.000	0.144	0.354
K	0.460	0.000	0.053	0.025	0.000	0.000	0.000	–	0.070	0.391
L*	0.073	0.008	0.038	0.144	0.172	0.109	0.018	0.005	–	0.434

For the observed distributions, our evaluations of GOF (cf. Eq. (4)) were similar to the results that were obtained by de Bruin et al. [17]. The errors between the approximation and the complete CTMC were expectedly larger than encountered in our experiments from Section 4.2. Ward F displays the largest supremum error of $5.27 \cdot 10^{-2}$, exceeding the maximum average error of $2.96 \cdot 10^{-2}$ in Table 1. Conversely, the comparison between the simulation with modified parameters and the complete CTMC shows that the loss of accuracy due to the modification is minor. Ward K displays the most substantial supremum error of $1.01 \cdot 10^{-2}$, but most of the distributions are almost identical to the complete CTMC.

Table 7 presents the effect of using the day-dependent arrival rates from Appendix B, Table B.11 by simulating the mean occupancy on each day of the week. The confidence intervals of the simulated mean occupancy have a maximum difference of 0.116 for a confidence level of 0.05. The table shows that although the estimates differ between the days, they only deviate by a maximum of 12.0% and an average of 3.4% from the approximated occupancy in the second column. Thus, the occupancy do not deviate substantially when taking the day-dependent arrival rate into account.

5.4. Example of an application

In this section, we demonstrate how to evaluate a new organizational structure in a hospital using our approximation. We base our demonstration on the data from Sections 5.1–5.2 and evaluate the effect of merging 3 out of the 11 wards in the hospital. The example is for demonstration purposes only, and does not account for the medical aspects of merging the wards in the hospital.

Other possible applications of our approximation include optimizing the allocation of beds, assessing the creation of new wards in the hospital, assessing the implications of extending the length-of-stay, and deriving new relocation rules.

We validate the solution of our example using the shortage probability and the expected number of preferred admissions per day. The latter is defined by the product between the arrival rate λ_p, and the probability that ward $w \in \mathcal{W}$ can admit a new arrival of type $p \in \mathcal{P}$, where $p = w$.

Appendix B, Table B.12 contains a complete overview of the relocated patients, and shows that Ward A, C and G share a substantial 20.7% of the relocated patients. Thus, we expect that merging the capacities of these wards will have a notable effect on the number of preferred admissions of the affected patient types. Let L* denote the new ward *and* the associated patient type, and let $\mathcal{W}' = \mathcal{P}' = \{A, C, G\}$ denote the set of wards and patient types from the current system that are merged together. Ward L^* has a capacity of $M_{L^*} = \sum_{i \in \mathcal{W}'} M_i$ beds, an arrival rate of $\lambda_{L^*} = \sum_{i \in \mathcal{P}'} \lambda_i$ and mean length-of-stay of $1/\mu_{L^*} = \sum_{i \in \mathcal{P}'} (\lambda_i / \lambda_{L^*})/\mu_i$. For the probability of relocating *to* Ward L*, we let $r_{p,L^*} = \sum_{i \in \mathcal{W}'} r_{pi}$, and for patients that relocate *from* Ward L^*, we let $r_{L^*,w} = \sum_{i \in \mathcal{P}'} r_{iw} \lambda_i / \lambda_{L^*}$. The resulting parameters are $M_{L^*} = 100$ beds, $\lambda_{L^*} = 26.49$ admissions per day, and $1/\mu_{L^*} = 3.61$ days. Appendix A, Table A.10 presents the adjusted relocation probabilities.

5.4.1. Results

Table 8 presents the improved shortage probabilities and expected number of preferred admissions of all wards. We find that merging Ward A, C and G increases the number of preferred admissions of the merged patient types from 21.96 to 24.46 patients per day. This corresponds to a relative improvement of 11.4%. The improvement is 6.3% for the preferred admissions among all wards. The organizational changes result in minor improvements for the wards that were not included in the merge.

6. Conclusion

Efficient resource allocations are essential to any hospital. This particularly applies to hospitals that are subject to high uncertainty, increasing demand, and scarce resources. The scarcity pushes the inpatient wards in hospitals to relocate patients resulting in an additional layer of difficulty for hospitals to assess the occupancy of their inpatient wards.

In this paper, we accommodate the need for creating an overview of the occupancy in a hospital, by providing an approximation of the occupancy distributions. The approximation evaluates the wards in sequence. Each evaluation includes a ward in focus, and the alternative wards by modeling the time where they are open and in shortage of beds, respectively.

We conducted several numerical experiments indicating that our approximation is adequate. Specifically, we found an average supremum difference between the approximated and exact occupancy distributions between $7.68 \cdot 10^{-4}$ and $2.96 \cdot 10^{-2}$. The experiments also indicated that the difference is not notably affected by the shortage probability nor the type of the length-of-stay distribution.

In addition, we conducted a statistical test indicating that our assumptions adequately reflect most inpatient wards in a Danish hospital. The deviations were most apparent in wards containing hallway-beds. A further validation showed that the approximated occupancy distributions resemble the observed distributions (similar to de Bruin et al. [17]), and that they do not deviate substantially when we include time-dependent arrivals.

Table B.11
The mean number of arrivals on each day of the week.

Ward	Overall mean (λ_p)	Mean number of arrivals						
		Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
A	14.21	14.77	16.03	16.74	16.66	14.56	10.62	9.98
B	11.37	13.03	12.61	13.03	13.49	11.84	8.07	7.42
C	8.07	8.62	9.02	8.59	8.72	8.85	7.08	5.57
D	6.54	7.33	6.11	6.95	7.92	6.57	5.40	5.48
E	4.77	4.75	5.10	5.20	4.66	6.08	4.00	3.55
F	2.90	3.75	3.10	3.39	3.34	3.00	2.00	1.68
G	4.22	4.35	4.52	4.54	4.33	4.89	3.30	3.57
H	2.39	2.67	2.74	2.62	2.64	2.80	1.77	1.48
I	1.92	2.12	2.00	1.98	2.15	2.44	1.58	1.12
J	0.94	1.38	1.25	1.21	1.00	1.18	0.33	0.23
K	0.96	1.48	1.05	1.10	1.13	0.93	0.66	0.33

Table B.12
The expected daily number of patients of type $p \in \mathcal{P}$ that are relocated to ward $w \in \mathcal{W}$.

$\mathcal{P}/\mathcal{W}$	A	B	C	D	E	F	G	H	I	J	K	Other
A	–	0.080	0.153	0.009	0.035	0.138	0.162	0.179	0.110	0.021	0.006	0.073
B	0.379	–	0.195	0.033	0.195	0.700	0.313	0.207	0.086	0.063	0.210	0.147
C	0.789	0.177	–	0.012	0.039	0.393	0.612	0.573	0.294	0.048	0.012	0.051
D	0.000	0.001	0.000	–	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004
E	0.102	0.064	0.026	0.000	–	0.036	0.065	0.062	0.174	0.002	0.004	0.261
F	0.074	0.101	0.084	0.004	0.010	–	0.083	0.052	0.029	0.005	0.047	0.050
G	0.123	0.054	0.147	0.008	0.076	0.135	–	0.073	0.090	0.005	0.004	0.083
H	0.054	0.014	0.045	0.000	0.008	0.055	0.045	–	0.044	0.003	0.002	0.014
I	0.030	0.015	0.016	0.000	0.009	0.013	0.017	0.025	–	0.027	0.000	0.004
J	0.015	0.012	0.005	0.000	0.010	0.007	0.016	0.000	0.098	–	0.000	0.090
K	0.003	0.127	0.010	0.000	0.015	0.007	0.006	0.000	0.000	0.000	–	0.108

Table C.13
Overview of the fundamental symbols and definitions.

Symbol	Definition
$\mathcal{W}$, $\mathcal{P}$	The set of inpatient wards and patient types.
w, p, m	An arbitrary ward, patient type, and the ward in focus, respectively.
λ_p, μ_p	The arrival and service rates of patient type p.
r_{pw}	The relocation probability of patient type p and ward w.
M_w	The bed capacity of ward w.
k_{wp}, k_p	The number of currently admitted patients in the complete CTMC and approximation, respectively.
$\boldsymbol{s}$, t	The state definitions of the complete CTMC and approximation, respectively.
$\mathcal{S}$, $\mathcal{T}_m$	The state space of the complete CTMC and approximation, respectively.
$\boldsymbol{Q}_m$, q_{ss^*}, $q^m_{tt^*}$	The transition matrix of the approximation, and rates of the complete CTMC and approximation, respectively.
$\boldsymbol{\pi}_m$	The state distribution of the approximation.
$\boldsymbol{\beta}^{open}_w$, $\boldsymbol{\beta}^{shortage}_w$	Initial probability distributions of the alternative ward w.
$\boldsymbol{\Gamma}^{open}_w$, $\boldsymbol{\Gamma}^{shortage}_w$	Phase-type generators of the alternative ward w.
$\boldsymbol{\gamma}^{open}_w$, $\boldsymbol{\gamma}^{shortage}_w$	Exit-rates of the alternative ward w.
o^{open}_w, $o^{shortage}_w$	The number of phases of the alternative ward w.
$\boldsymbol{h}$, h_w	The current phase of the hyper-exponential dist. related to an alternative ward.
$\boldsymbol{b}$, b_w	Indicates if alternative ward w is *open* or in *shortage*.
d_j, $\hat{d}_j$, D_j, $\hat{D}_j$	Exact and estimated marginal and cumulative probabilities for j beds.
ρ	The ward load (per bed).
i, j	Generic index variables.

To propagate the use of models for bed capacity planning, we propose that future work investigates the sensitivity to load-dependent behavior, such as premature discharges to avoid overcrowding.

Furthermore, the simplicity of the hyper-exponential distribution makes the model ideal for computational implementation as well as parameter fitting, but our approach does not impose any restrictions on the type of PH distribution. Future work should therefore investigate the potential benefits of employing general PH distributions in the model.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

We thank Region Zealand and the Innovation and Research Centre for Multimorbidity for providing the necessary support for this study. Funding was provided by both The Innovation and Research Centre for Multimorbidity and the Technical University of Denmark.

Appendix A. Additional parameters

See Tables A.9 and A.10.

Appendix B. Additional results

See Tables B.11 and B.12.

Appendix C. Symbols

See Table C.13.

References

[1] OECD, OECD health statistics, 2021, URL https://www.oecd.org/health/health-statistics.htm.

[2] L. He, S. Chalil Madathil, A. Oberoi, G. Servis, M.T. Khasawneh, A systematic review of research design and modeling techniques in inpatient bed management, Comput. Ind. Eng. (ISSN: 18790550) 127 (2019) 451–466, http://dx.doi.org/10.1016/j.cie.2018.10.033, 03608352.

[3] R.A. Baru, E.A. Cudney, I.G. Guardiola, D.L. Warner, R.E. Phillips, Systematic review of operations research and simulation methods for bed management, in: Iie Annual Conference and Expo 2015, 2015, pp. 298–306.

[4] P. Bhattacharjee, P.K. Ray, Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: A review and reflections, Comput. Ind. Eng. (ISSN: 18790550) 78 (2014) 299–312, http://dx.doi.org/10.1016/j.cie.2014.04.016, 03608352.

[5] R. Bekker, M. uit het Broek, G. Koole, Modeling COVID-19 hospital admissions and occupancy in the Netherlands, European J. Oper. Res. (ISSN: 18726860) 304 (1) (2023) 207–218, http://dx.doi.org/10.1016/j.ejor.2021.12.044, 03772217.

[6] S. Dijkstra, S. Baas, A. Braaksma, R.J. Boucherie, Dynamic fair balancing of COVID-19 patients over hospitals based on forecasts of bed occupancy, Omega (U. K.) (ISSN: 18735274) 116 (2023) 102801, http://dx.doi.org/10.1016/j.omega.2022.102801, 03050483.

[7] S.S.W. Lam, A.R. Pourghaderi, H.R. Abdullah, F.N.H.L. Nguyen, F.J. Siddiqui, J.P. Ansah, J.G. Low, D.B. Matchar, M.E.H. Ong, An agile systems modeling framework for bed resource planning during COVID-19 pandemic in Singapore, Front. Public Health (ISSN: 22962565) 10 (2022) 714092, http://dx.doi.org/10.3389/fpubh.2022.714092.

[8] M. Barbato, A. Ceselli, M. Premoli, On the impact of resource relocation in facing health emergencies, European J. Oper. Res. (2022).

[9] D.J. Breuer, S. Kapadia, N. Lahrichi, J.C. Benneyan, Joint robust optimization of bed capacity, nurse staffing, and care access under uncertainty, Ann. Oper. Res. (ISSN: 15729338) 312 (2) (2022) 673–689, http://dx.doi.org/10.1007/s10479-022-04559-w, 02545330.

[10] E.J. Delgado, X. Cabezas, C. Martin-Barreiro, V. Leiva, F. Rojas, An equity-based optimization model to solve the location problem for healthcare centers applied to hospital beds and COVID-19 vaccination, Mathematics (ISSN: 22277390) 10 (11) (2022) 1825, http://dx.doi.org/10.3390/math10111825.

[11] X. Gong, X. Wang, L. Zhou, N. Geng, Managing hospital inpatient beds under clustered overflow configuration, Comput. Oper. Res. 148 (2022) 106021.

[12] T. Latruwe, M. Van der Wee, P. Vanleenhove, J. Devriese, S. Verbrugge, D. Colle, A long-term forecasting and simulation model for strategic planning of hospital bed capacity, Oper. Res. Health Care (2022) 100375.

[13] S. Priyan, S. Banerjee, An interactive optimization model for sustainable production scheduling in healthcare, Healthc. Anal. (ISSN: 27724425) (2022) http://dx.doi.org/10.1016/j.health.2022.100124.

[14] Y. Jiang, F. Yang, Z. Tang, Q.L. Li, Admission control of hospitalization with patient gender by using Markov decision process, Int. Trans. Oper. Res. (ISSN: 14753995) 30 (1) (2023) 70–98, http://dx.doi.org/10.1111/itor.12931, 09696016.

[15] S. Davis, N. Fard, Theoretical bounds and approximation of the probability mass function of future hospital bed demand, Health Care Manag. Sci. (ISSN: 15729389) 23 (1) (2020) 20–33, http://dx.doi.org/10.1007/s10729-018-9461-7, 13869620.

[16] X. Wu, J. Li, C.H. Chu, Modeling multi-stage healthcare systems with service interactions under blocking for bed allocation, European J. Oper. Res. (ISSN: 18726860) 278 (3) (2019) 927–941, http://dx.doi.org/10.1016/j.ejor.2019.05.004, 03772217.

[17] A.M. de Bruin, R. Bekker, L. van Zanten, G.M. Koole, Dimensioning hospital wards using the Erlang loss model, Ann. Oper. Res. (ISSN: 15729338) 178 (1) (2010) 23–43, http://dx.doi.org/10.1007/s10479-009-0647-8, 02545330.

[18] N.C. Proudlove, The 85% bed occupancy fallacy: The use, misuse and insights of queuing theory, Health Serv. Manag. Res. (ISSN: 17581044) 33 (3) (2020) 110–121, http://dx.doi.org/10.1177/0951484819870936, 09514848.

[19] N. Proudlove, Use and misuse of queueing theory for hospital capacity decisions, in: Operations Management for Healthcare, Routledge, 2022, pp. 197–216.

[20] A. Kokangul, A combination of deterministic and stochastic approaches to optimize bed capacity in a hospital unit, Comput. Methods Programs Biomed. (ISSN: 18727565) 90 (1) (2008) 56–65, http://dx.doi.org/10.1016/j.cmpb.2008.01.001, 01692607.

[21] X. Wang, X. Gong, N. Geng, Z. Jiang, L. Zhou, Metamodel-based simulation optimisation for bed allocation, Int. J. Prod. Res. (2019) 1–21, http://dx.doi.org/10.1080/00207543.2019.1677962, ISSN: 1366588x, 00207543.

[22] M. Bierlaire, C. Osorio Pizano, M.I. of Technology. Department of Civil, C. Osorio Pizano, A simulation-based optimization framework for urban transportation problems, 2013.

[23] R. Bekker, G. Koole, D. Roubos, Flexible bed allocations for hospital wards, Health Care Manag. Sci. (ISSN: 15729389) 20 (4) (2017) 453–466, http://dx.doi.org/10.1007/s10729-016-9364-4, 13869620.

[24] A.R. Andersen, B.F. Nielsen, L.B. Reinhardt, Optimization of hospital ward resources with patient relocation using Markov chain modeling, European J. Oper. Res. (ISSN: 18726860) 260 (1) (2017) 1152–1163, http://dx.doi.org/10.1016/j.ejor.2017.01.026, 03772217.

[25] A.R. Andersen, W. Vancroonenburg, G. Vanden Berghe, Strategic room type allocation for nursing wards through Markov chain modeling, Artif. Intell. Med. (ISSN: 18732860) 99 (2019) 101705, http://dx.doi.org/10.1016/j.artmed.2019.101705, 09333657.

[26] K.M. Chandy, U. Herzog, L. Woo, Parametric analysis of queueing networks, Ibm J. Res. Dev. (ISSN: 21518556) 19 (1) (1975) 36–42, 00188646.

[27] R.J. Boucherie, N.M. van Dijk, A generalization of Norton's theorem for queueing networks, Queueing Syst. Theory Appl. (ISSN: 15729443) 13 (1–3) (1993) 251–259, http://dx.doi.org/10.1007/BF01158934, 02570130.

[28] A. Kuczura, The interrupted Poisson process as an overflow process, Bell Syst. Tech. J. 52 (3) (1973) 437–448.

[29] M. Neuts, Matrix-Geometric Solutions in Stochastic Models, The Johns Hopkins University Press, 1981.

[30] M. Bladt, B.F. Nielsen, Matrix-Exponential Distributions in Applied Probability, Springer, 2017, http://dx.doi.org/10.1007/978-1-4939-7049-0, ISBN: 149397047x, 1493970496, 9781493970490, 9781493970476.

[31] W.J. Stewart, Probability, Markov Chains, Queues, and Simulation: The Mathematical Basis of Performance Modeling, Princeton University Press, ISBN: 9780691140629, 2009.

[32] J.G. Saw, M.C.K. Yang, T.C. Mo, Chebyshev inequality with estimated mean and variance, Am. Stat. (ISSN: 15372731) 38 (2) (1984) 130, http://dx.doi.org/10.2307/2683249, 00031305.

[33] A.R. Andersen, Areenberg/RelSys: RelSys - first release, 2022, URL https://doi.org/10.5281/zenodo.6037435.

[34] S. Asmussen, O. Nerman, M. Olsson, The EMpht-programme, 1998, URL https://web.archive.org/web/20180617130551/http://home.math.au.dk/asmus/pspapers.html.

[35] S. Asmussen, O. Nerman, M. Olsson, Fitting phase-type distributions via the EM algorithm, Scand. J. Stat. (ISSN: 14679469) 23 (4) (1996) 419–441, 03036898.

[36] J.P.C. Kleijnen, R.C.H. Cheng, B. Bettonvil, Validation of trace-driven simulation models: More on bootstrap tests, in: Proceedings of the 2000 Winter Simulation Conference, Vols 1 and 2, 2000, pp. 882–892.

[37] E. Wilson, Probable inference, the law of succession, and statistical inference, J. Amer. Statist. Assoc. 22 (158) (1927) 209–212, http://dx.doi.org/10.1080/01621459.1927.10502953, ISSN: 1537274x, 01621459.

[38] R.J. Boucherie, N.M. Dijk, Queueing Networks: A Fundamental Approach, Springer, ISBN: 978-1-4419-6471-7, 2011, http://dx.doi.org/10.1007/978-1-4419-6472-4.