Optimization of hospital ward resources with patient relocation using Markov chain modeling

Andersen, Anders Reenberg; Nielsen, Bo Friis; Reinhardt, Line Blander

Published in:
European Journal of Operational Research

Link to article, DOI:
10.1016/j.ejor.2017.01.026

Publication date:
2017

Document Version
Peer reviewed version

Optimization of Hospital Ward Resources with Patient Relocation using Markov Chain Modelling

Technical University of Denmark

Anker Engelunds Vej 1, 2800 Kgs. Lyngby

Andersen, Anders Reenberg * arean@dtu.dk

Nielsen, Bo Friis bfni@dtu.dk

Reinhardt, Line Blander lbre@dtu.dk

Abstract

Overcrowding of hospital wards is a well known and often revisited problem in the literature, yet it appears in many different variations. In this study, we present a mathematical model to solve the problem of ensuring sufficient beds to hospital wards by re-distributing beds that are already available to the hospital. Patient flow is modelled using a homogeneous continuous-time Markov chain and optimization is conducted using a local search heuristic. Our model accounts for patient relocation, which has not been done analytically in literature with similar scope. The study objective is to ensure that patient occupancy is reflected by our Markov chain model, and that a local optimum can be derived within a reasonable runtime.

Using a Danish hospital as our case study, the Markov chain model is statistically found to reflect occupancy of hospital beds by patients as a function of how hospital beds are distributed. Furthermore, our heuristic is found to efficiently derive the optimal solution. Applying our model to the hospital case, we found that relocation of daily arrivals can be reduced by 11.7% by re-distributing beds that are already available to the hospital.

Keywords— OR in health services, Queueing, Markov chain, Stochastic optimization, Heuristics

1 Introduction

Overcrowding of hospital wards is a well known problem in the Danish health care sector. A report from the Ministry of Health [4] indicates that most regions of Denmark experience problems with overcrowding of hospital wards. In addition, the patient organization *Danish Patients* in corporation with *Danish Nurses Organization and the Danish Medical Association* reports that patient admission in hallways and depots is a recurrent necessity for a range of hospitals [2], and in which case both objective and subjective quality of care may suffer a great decrease [1, 21]. Hence, in order to provide patients with the best possible treatment, overcrowding should be reduced as much as possible.

An increasing number of Danish hospitals are developing methods to cope with overcrowding through capacity balancing, where patient relocation, is coordinated using daily capacity meetings, as well as dedicated staff for *patient flow* coordination [3]. Using such methods, some hospitals have succeeded in significantly decreasing the number of patients hospitalized to alternative locations. The hospitals relocate patients from

*Corresponding author: Department of Engineering Management, Building 426, 2800 Kgs. Lyngby. E-mail: arean@dtu.dk
This research was funded and supported by Region Sjælland. URL: www.regionsjaelland.dk

wards with overcrowding to wards where sufficient nursing resources are still available, and thus match resources with demand. However, we conducted interviews with a specific hospital and found that this approach entails costs for both planning, relocation of patients and some decrease in quality of care. In this case, quality of care is decreased due to a mismatch between the optimal type of care and what type of care is alternatively offered to the patient. Hence, a problem arises containing *two* different types of penalty for the hospital management to consider. First, there is the tangible cost of spending man hours on defining and implementing a plan, and secondly, management have to consider the risk of inducing a lower quality of care, either through placing patients in buffer beds or relocating patients to other wards.

The objective of this study, is to provide hospital management with a tactical decision tool, capable of optimizing the match between resources and demand. We focus on a specific case where patients are always relocated whenever ward resources are depleted. The main methodological approach will be mathematical modeling. More specifically, we model bed occupancy using a homogeneous continuous-time Markov chain, and optimize the response using a local search heuristic.
In Section 2 we present the specific problem of this study. In Section 3 our solution approach is presented, divided into two parts. The first part describes the Markov chain, we use to model patient flow behavior. The second part connects this Markov chain model to a local search heuristic. Section 4 demonstrates the usability of our solution approach for a specific hospital case, and tests on a number of different parameter settings are presented. Lastly, we present our conclusion in Section 5.

1.1 Literature Review

Modelling and optimization of hospital bed utilization is a recurrent topic dating back to Newsholme, 1932 [18]. The specific problem structure differs from one study to another, however, all focus on one of three major objectives: (1) Testing scenarios [5, 12], (2) deriving the required number of beds for one or more wards [13–16, 19, 20, 25], or (3) balancing beds with demand [9, 10, 17]. In achieving these, two methodological aspects are considered: (1) The methods used to *model* the system in focus, and (2) the methods used to *study* and *optimize* the system.

Different approaches of modelling the system are known from the literature. These are usually either simulation [12, 16, 25], queueing theoretic approaches [9, 13–15, 17, 19], or a mixture of these [5, 10, 20].
In Goldman et al., 1968 [12], utilization and costs are tested for various bed allocation policies using a simulation model. Harris, 1984 [16], develops a simulation model to assist decision making in the area of operating theatre time tables and the resultant bed requirements. Lastly, Zeraati et al., 2005 [25], use a statistical simulation to estimate the number of required beds for an obstetrics ward.
In the area of queueing theoretic models, two studies by Gorunescu et al., 2002, and Li et al., 2009 [13,14,17], exploit $M/PH/c/N$ and $M/PH/c$ models, to assess a mixture of patient flow. Furthermore, Green, 2002 [15], use an $M/M/s$ model to estimate bed availability in different intensive care and obstetrics units, and Pendergast et al. 1988 [19], use clinical judgment and basic probability theory to derive future hospital bed requirements. Lastly, Cochran et al., 2008 [9], develops a queueing network model that is implemented as a capacity balancing tool between different hospital units.
Exploiting the use of both simulation and queueing theory, Cochran et al., 2006 [10], use queueing networks to assess the flow between units of an obstetrics hospital, and define utilization targets. A Discrete-event-simulation model is then used to maximize the flow. A related approach is used in Akkerman et al., 2009 [5], where Markov chain theory and simulation is used to evaluate a number of different management scenarios. This specific Markov chain model is found to produce useful insight into the theoretical number of required beds, but a simulation model is required to derive the amount of patient rejections.

The second methodological aspect that is considered in most studies, is studying and optimizing the system

in focus. Naturally, scenario testing is more straightforward, whereas bed requirement or capacity balancing would suggest the application of a more elaborate approach. Here, we found only a few studies [17, 20] that exploit advantages of heuristic or mathematical programming elaborately, leaving this area rather unexploited. In Pinto et al., 2014, a simulation-optimization model is developed to analyze dynamic features of the system and find the best configuration of beds. Moreover, Li et al., 2009 [17], applied a $M/PH/c$ model from [13] in a multi-objective goal programming model to reallocate beds.

Two studies in the area of capacity balancing that are rather similar to this paper, are uncovered [9, 10]. However, in case of overcrowding, none of these studies modelled the effect of patients being relocated to alternative locations. In this paper, we present an approach to balance capacity in a system of queues, where patients are relocated when capacity is insufficient. To achieve this, we use a homogeneous continuous-time Markov chain model.
A range of studies, using Markov chains to model patient flow, have already been conducted [6, 8, 11, 24, 26]. As relevant examples, Broyles et al., 2010 [8], predicts distribution and expected admissions, and Bartolomeo et al., 2008 [6], determines the probabilities of readmission for two different patient categories. However, none of these exploit the advantages of Markov chains, to model patient relocation, and subsequently use these models to optimize the system.

2 Problem Description

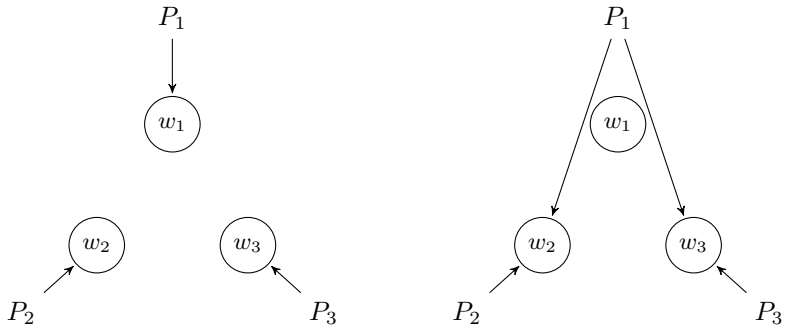
In this study, we consider a Danish hospital where an organizational structure for patient relocation has been fully implemented.
That is, even though minor changes in the distribution of resources might take place on a daily basis, most actions to avoid overcrowding are performed using patient relocation. Any greater changes in the distribution of resources are not practical if they occur too frequently, and are thus considered more as a tactical decision. Deciding on the best allocation of resources, is therefore an important decision, as the result will affect how patients are hospitalized, and the hereto related costs, through a period of several months.
For this reason, the decision this study will focus upon, is how resources should be allocated among the hospital wards. As hospitalizations are usually dependent on a range of different resources, we assume that *one* hospitalization can only take place when *one* "sufficiently" staffed and equipped bed is available. That is, we disregard the possibility that a hospitalization may in some instances take place without sufficient staff or equipment. Thus, if an entrance to a ward is restricted by the lack of resource units, we assume that an alternative ward always exists somewhere else. We have found through interviews with hospital employees that this is a reasonable assumption.

Taking all of the above considerations into account, the overall goal of this study, is to develop a mathematical model that can be used to efficiently minimize the number of rejections at *preferred* wards, by changing the distribution of bed resources.

For the remaining of this paper, a patient hospitalization at a preferred ward, will be denoted as a *primary hospitalization*. A patient relocation to an alternative ward, will be denoted as a *secondary hospitalization*. In the same way, patient blocking at a preferred ward, is denoted a *primary rejection*, as well as patient blocking at the alternative ward is denoted a *secondary rejection*.

2.1 Dynamics of the System

As mentioned above, wards have limited resources, and as a result, arriving patients are relocated whenever resources have been depleted (no staffed and equipped beds are available). During such a relocation, patient



(a) All wards are available. (b) Ward 1 is blocked.

Figure 1: Graphical representation of the hospitalization procedure. Patient types are P_1, P_2 and P_3 for which the preferred hospitalization is at ward w_1, w_2 and w_3, respectively. (a) represents the system under regular load, and (b) the result if ward 1 was to be fully loaded.

characteristics are required to match with the specialization of the alternative ward. Thus, relocating a patient to an arbitrary free ward, is not always a feasible solution.

We interpret these hospitalization operations as a queueing system with N different patient types, arriving at N parallel service stations. The number of servers at each station is equal to the number of staffed and equipped beds at each ward. If all servers are occupied at a station the station is blocked, but a queue is not created. Instead, arrivals will be distributed with a probability to other stations, or disappear from the system entirely. This is illustrated in Figure 1, where a system of $N = 3$ patient types and wards is (a) under regular load, and (b), blocked for ward 1.

Due to these system operations, resources allocated to a ward will not only affect the amount of primary rejections, but also the amount of secondary hospitalizations at that ward. Moreover, notice that treatment time is tied to the patient type, and therefore independent of the ward in which hospitalization takes place. Wards will therefore experience a mixture of different patients with different lengths of stay.

3 Modelling & Solution Approach

To solve the problem of optimizing the distribution of beds, we model the ward occupancy density functions using a homogeneous continuous-time Markov chain. This model approach, is presented in Section 3.1. From the density functions, we derive the specific probabilities of wards blocking, followed by the overall expected number of arriving patients experiencing a primary rejection. This is used as our objective value, as the system is optimized using a local search heuristic. The specific heuristic we use, is presented in Section 3.2.

3.1 A Homogeneous Continuous-Time Markov Chain

As mentioned in Section 2, we consider N patient types, $i \in \{1, 2, \cdots, N\}$, as well as N ward types, $j \in \{1, 2, \cdots, N\}$. To model the ward occupancy density functions for each ward, we introduce a homogeneous continuous-time Markov chain (CTMC) model with state $s = (w_{11}, w_{21}, \cdots, w_{ij}, \cdots, w_{NN})$ and state space S, where w_{ij} is the number of type i patients hospitalized in ward j. Let M_j define the amount of allocated beds to ward j. Hence, M_j is the maximum amount of patients that may be hospitalized in ward j. Further, let f_j be the number of free beds at ward j, so $f_j = M_j - \sum_{i \in I} w_{ij}$. For the purpose of presenting our modelling approach, we include f_j in the state representation to get:

$$s \quad = \quad \left[\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{21} & w_{22} & \cdots & w_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{NN} \end{pmatrix}, (f_1, f_2, \cdots, f_N) \right] \quad \in S$$

Note that f_j is otherwise redundant to the model. Now, λ_i is the arrival rate of patient type i, and μ_i the service rate of patient type i. We assume patients arrive according to a Poisson process and that inter-service time distributions are exponentially distributed. The reader should notice that if the latter does not hold, rejection systems, such as this, are in general robust to the distribution of inter-service times [7, p. 121]. In addition, we statistically test the CTMC model fit for a specific case-hospital in Section 4.1.6.
Let $p(f_1, f_2, \cdots, f_N)_{ij}$ define the fraction of patients of type i that are hospitalized in ward j as function of the number of free beds at all wards in the system, f_1, f_2, ...and f_N. Let Q define the transition rate matrix of the CTMC, with q_{ss^*} the transition rate from a current state $s \in S$ to a new state $s^* \in S$. In the following, $p(f_i = 0, f_j > 0, \cdots, f_N > 0)$ is abbreviated $p(f_i = 0)$, just as $p(f_i = 0, f_k = 0, f_j > 0, \cdots, f_N > 0)$ is abbreviated $p(f_i = 0, f_k = 0)$, and so on. Moreover, we refer to a new state $s^* = (\cdots, w_{ij}+1, \cdots, f_j-1, \cdots)$ to indicate the arrival of a patient i to a ward j, and $s^* = (\cdots, w_{ij} - 1, \cdots, f_j + 1, \cdots)$ for a corresponding discharge. The transition rates are then,

$$q_{ss^*} = \begin{cases} \lambda_i & \text{if } s^* = (\cdots, w_{ii} + 1, \cdots, f_i - 1, \cdots) \text{ and } f_i > 0 \quad \forall i \in I \\ \lambda_i p(f_i = 0)_{ij} & \text{if } s^* = (\cdots, w_{ij} + 1, \cdots, f_j - 1, \cdots) \text{ and } f_i = 0, f_j > 0 \quad \forall i, j \in I, i \neq j \\ \lambda_i p(f_i = 0, f_k = 0)_{ij} & \text{if } s^* = (\cdots, w_{ij} + 1, \cdots, f_j - 1, \cdots) \text{ and } f_i = 0,\ f_k = 0, f_j > 0 \quad \forall i, j, k \in I, i \neq j \neq k \\ \vdots & \qquad \vdots \\ \lambda_i p(f_i = 0, f_k = 0, \cdots, f_l = 0)_{ij} & \text{if } s^* = (\cdots, w_{ij} + 1, \cdots, f_j - 1, \cdots) \text{ and } f_i = 0, f_k = 0, \cdots, f_l = 0, f_j > 0 \quad \forall i, j, k \cdots l \in I, i \neq j \neq k \neq \cdots \neq l \\ \mu_i w_{ij} & \text{if } s^* = (\cdots, w_{ij} - 1, \cdots, f_j + 1, \cdots) \text{ and } w_{ij} > 0 \quad \forall i, j \in I \end{cases}$$

where all other transition rates, q_{ss^*}, are 0.
Notice, as treatment times differ between patient types, the state definition contains an element for every combination of patient type and ward. The variables w_{ii} count primary hospitalizations, whereas the variables w_{ij} count secondary hospitalizations. The model can only jump to a state, where the number of secondary hospitalizations is increased, if capacity is full at the preferred ward. For instance if $N = 3$ and $M_1 = 10$, $M_2 = 15$ and $M_3 = 20$, $s = (w_{11}, w_{21}, w_{31}, w_{12}, w_{22}, w_{32}, w_{13}, w_{23}, w_{33}) = (7, 2, 1, 1, 4, 2, 3, 2, 10) \rightarrow s^* = (7, 2, 1, \mathbf{2}, 4, 2, 3, 2, 10)$ is allowed, because ward 1 is full. However, $s = (7, 2, 1, 1, 4, 2, 3, 2, 10) \rightarrow s^* = (7, 2, 1, 1, 4, 2, 3, \mathbf{3}, 10)$ is not possible, as ward 2 is still open.
The transition rate depends on how many other wards are blocked. $s = (7, 2, 1, 1, 4, 2, 3, 2, 10) \rightarrow s^* = (7, 2, 1, \mathbf{2}, 4, 2, 3, 2, 10)$ has rate $q_{ss^*} = \lambda_1 p(f_1 = 0)_{12}$, as only ward 1 is blocked. Now, $s = (7, 2, 1, 1, 4, 2, 3, 2, 15) \rightarrow s^* = (7, 2, 1, \mathbf{2}, 4, 2, 3, 2, 15)$ has rate $q_{ss^*} = \lambda_1 p(f_1 = 0, f_3 = 0)_{12}$, as both ward 1 and ward 3 are blocked.

The total state space size, $|S|$, of the CTMC is the product of N polynomials of the order N, as shown in (1).

$$|S| = \prod_{j=1}^{N} \left(\frac{1}{N!} \cdot \prod_{i=1}^{N} (M_j + i) \right) \tag{1}$$

Let us consider a case where $N = 3$, and $M_1 = 27$, $M_2 = 23$ and $M_3 = 24$. Then, from (1), the state space, S, has a size of $|S| = 30,876,300,000$ states – which is rather difficult to cope with computationally. Thus, in order for our CTMC to be applicable for even small cases, a rather large fraction of the state space needs to be truncated. To attain this, we use two recursive procedures presented in the following Section 3.1.1.

3.1.1 The Truncation Procedures

Let u_{ij} be an upper bound on the number of patients of type i that is hospitalized in ward j (w_{ij}) for $i \neq j$, so $\sum_{k=u_{ij}+1}^{M_j} Prob\{w_{ij} = k\}$ is *sufficiently* small, where $Prob\{w_{ij} = k\}$ is the probability of attaining a state where $w_{ij} = k$. Taking this idea further, we also let L_j and U_j define the lower and upper bounds on the total amount of patients hospitalized in ward j. In this case, L_j and U_j are chosen so $\sum_{k=L_j}^{U_j} Prob\{\sum_i w_{ij} = k\}$ is *sufficiently* large, but the number of truncated states are maximized. Here, $Prob\{\sum_i w_{ij} = k\}$ is the probability of attaining a state where the sum of patients in ward j is equal to k. Let ϕ_j be the number of free slots at ward j in the truncated system, then $\phi_j = U_j - \sum_i w_{ij}$. Thus as $U_j \leq M_j$, we have that $\phi_j \leq f_j$.
Our goal is then to adjust u_{ij}, L_j and U_j, so reasonable accuracy is maintained, within the practical limits of computing the probability distribution. To attain this, we notice that the hospitalization of patients at each ward is closely related to an $M/M/c/c$ queueing system, cf. Figure 2. That is, a queueing system with capacity equal to the number of beds. The probability that there are n beds occupied in such a system can be derived using (2),

$$p_n = \frac{(\lambda/\mu)^n/n!}{\sum_{i=0}^{c}(\lambda/\mu)^i/i!} \tag{2}$$

where λ is the arrival rate, μ the service rate, and c the number of beds in the system [7, p. 121]. We use (2) to determine bounds on the total amount of occupied beds at each ward, L_j and U_j, as well as for each secondary hospitalization pair, u_{ij}. For the latter, consider that w_{ij} is stochastically smaller or equal to the occupancy in an $M/M/c/c$ system where the arrival rate is the maximum fraction of arriving type i patients to ward j, $\lambda_i \cdot \max\{p(\cdot)_{ij}\}$, and service rate μ_i. The probability mass of such a system, derived using (2), will be at least as shifted in positive direction as the marginal probability mass of w_{ij} in the CTMC. We refer to this $M/M/c/c$ system as the *right-shifted* distribution. Letting τ ($0 \leq \tau \leq 1$) define a truncation tolerance, the upper bound, u_{ij}, is derived using Algorithm 1.

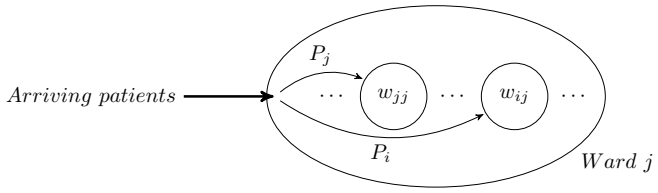


Figure 2: Graphical representation of patients hospitalized in ward j. Patients of different types are hospitalized as long as $\sum_i w_{ij} < M_j$, where M_j corresponds to the parameter c in (2).

Algorithm 1 Procedure for deriving u_{ij} for $i \neq j$

$lamdba \leftarrow \lambda_i \cdot \max\{p(\cdot)_{ij}\}$ ▷ Initialize
$mu \leftarrow \mu_i$
$c \leftarrow M_j$
$right \leftarrow erlangB(c, lambda, mu)$ ▷ Calculate distribution using (2) and save as an array
$st \leftarrow length(right)$
while $sum(right) \geq 1 - \tau$ **do**
    $right[st] \leftarrow 0$
    $st \leftarrow st - 1$
end while
$u \leftarrow st + 1$ ▷ The final bound is the number of non-zero elements plus 1
return u

Here, *right* represents the right-shifted distribution for w_{ij}. The upper bound, u_{ij}, is then determined by recursively truncating *right* starting at the highest occupancy, and lastly adding one to ensure that the least probability mass larger or equal to $1 - \tau$ is left.

In deriving L_j and U_j we consider both a lower and an upper bound of occupancy in ward j. Therefore, in order to ensure sufficient probability mass in the CTMC, and derive the maximum number of states that may be truncated from ward j, we have to consider both a left-shifted *and* right-shifted distribution at the same time. The procedure we use to derive L_j and U_j, is presented below:

1. Determine the minimum service rates in ward j: $\mu_{MIN} \leftarrow \min_{i \in I}\{\mu_i\}$.
2. Determine the maximum arrival rate to ward j: $\lambda_{MAX} \leftarrow \lambda_j + \sum_{i \in I \setminus \{j\}} \lambda_i \cdot \max\{p(\cdot)_{ij}\}$.
3. Calculate left- and right-shifted distributions using (2).
   (a) $right \leftarrow erlangB(M_j, \lambda_{MAX}, \mu_{MIN})$.
   (b) $left \leftarrow erlangB(M_j, \lambda_j, \mu_j)$.
4. Truncate states constrained by τ: $L_j, U_j \leftarrow ipmodel(right, left, \tau)$.

Notice that our procedure lastly takes an Integer Programming model, *ipmodel*(), to maximize the number of truncated states. We propose to formulate this as a *Knapsack Problem* variation, minimizing the number of states in the truncated system constrained by the probability mass tolerance τ.
This concludes the approach we use to derive u_{ij}, L_j and U_j. Notice, how the resulting transition rate matrix, Q, will be dependent on whether $U_j < M_j$ or $U_j = M_j$, leading to different representations of the matrix.

3.1.2 Computing the State Probability Distribution

We have derived a method for reducing the state space as a function of the tolerance τ, and are therefore set to generate the transition rate matrix Q. We assume that most non-acute wards will have long expected length of stay relative to the respective fluctuations in arrival rate. We further assume that most arrivals and discharges occur during the day, causing the system to be "inactive" during the night, so any remaining time-dependency will be negligible in the scope of deriving a long-term allocation of beds for the hospital. Thus, we consider the CTMC as a steady-state process. Now let π define the steady-state probability distribution of the CTMC. Then, we are faced with solving the global balance equations in (3),

$$\pi Q = 0 \tag{3}$$

where $\|\pi\|_1 = 1$. We have found that a solution to (3) can be derived within reasonable runtime using the method of *successive overrelaxation* [22, p. 311]. That is, (3) is written on the form $Ax = b$ by transposing, so we get: $Q^T\pi^T = 0$. Further, $Q^T\pi^T = (D - X - Y)\pi^T = 0$, where D, X and Y are the diagonal, lower- and upper- strictly triangular matrices of Q^T. Let x^k be the $k'th$ iteration solution to π^T. Then we can recursively derive π^T, using (4),

$$x^{k+1} = (1 - \omega)x^k + \omega\{D^{-1}(Xx^{k+1} + Yx^k)\} \tag{4}$$

until convergence. The relaxation parameter, ω, may be adjusted to ensure the fastest rate of convergence. As our case is dependent on different representations of Q, and we want our implementation to be flexible, we chose to conduct a range of tests to search for a fixed relaxation parameter that would result in a reasonable convergence time for "most" cases. We calculated the *iteration matrix* $H_w = (D - \omega X)^{-1}[(1 - \omega)D + \omega Y]$. We then adjusted ω to maximize the distance between the unit dominant and subdominant eigenvalue of H_w, with a view to maximize the convergence rate. It was found that a high distance could be obtained with

a relatively high relaxation parameter – usually around 1.7 to 1.8. Thus for the remaining of this paper, $\omega = 1.75$.
Regarding the question of *when* convergence has occurred, we decided to check this on the largest relative tolerance $\delta = max_i(\left|x_i^k - x_i^{k-m}\right| / \left|x_i^k\right|)$. Where m is set to increase as δ decreases – recall $\lim_{\delta \to 0} x^k = \pi^T$, and thus the rate of convergence is expected to decrease as x^k is closing in on π^T.

To assess our approach, we conducted a series of tests for $N = 3$, $M_1 = 27$, $M_2 = 23$ and $M_3 = 24$, and different settings of the truncation parameter τ. In Table 1, the total runtime of our approach implemented in `Java`, along with state space sizes, are presented for $\tau = 0.05, 0.01$ and 0.001.
Each of these settings were assessed by comparing the respective marginal distributions of π – that is, the distribution of how many beds are expected to be occupied for each ward. Obviously, the tails approach zero as the truncation is relaxed. However, the algorithm only takes 69 seconds to finish for $\tau = 0.05$, against 1,947 seconds for $\tau = 0.001$. Additionally, in case we are only interested in the blocking probabilities, we would be able to make do with the largest truncation value – given that we always end up with a CTMC model representation where $U_j = M_j \quad \forall j \in I$. However, to gain a more generic use of our model, we find it more appropriate to use $\tau = 0.01$.

Table 1: Results from adjustment of τ.

τ	Total Runtime (s)	$\lvert S \rvert$
0.05	69	517,000
0.01	483	1,358,760
0.001	1,947	3,563,520

3.2 A Heuristic Optimization Model

In Section 3.1 we presented an approach to model the ward occupancy for N wards and correspondingly N patient types. We now consider the number of beds, available to ward i, M_i, as a decision variable that may be adjusted to optimize the overall system performance. In general, we consider the following optimization problem:

$$\begin{aligned} \text{min.} \quad & f(\boldsymbol{M}) && \text{(5a)} \\ \text{s.t.} \quad & \\ & \textstyle\sum_{i \in I} M_i = \Theta && \text{(5b)} \\ & M_i \geq 1 \ \forall i \in I && \text{(5c)} \\ & M_i \in \mathbb{N} \end{aligned}$$

Where, as previously defined, I is the set of wards. Here, (5b) ensures that all available resources, Θ, are utilized. Moreover, (5c) ensures that wards contain at least one bed. The objective function (5a) evaluates the system performance as a function of $\boldsymbol{M} = (M_1 \ M_2 \ \cdots \ M_N)^T$, where in this case, a *large* value indicates a *poor* performance. As shown in the following, the objective function can easily be replaced and customized to the specific hospital preferences. In this study, we propose an objective value that increases as more "work" is spent on relocating patients. Consider $f(\boldsymbol{M}) = \sum_{i \in I} \pi_i^B(\boldsymbol{M})$, where $\pi_i^B(\boldsymbol{M})$ is the probability of all beds being occupied in ward i, with beds distributed as in $\boldsymbol{M}$. In this case, we would get some kind of measure for the total amount of work – recall when $\pi_i^B(\boldsymbol{M})$ increases, so does the amount of relocated patients from ward i. However, the expression does not incorporate the weight of patient types arriving with different rates. So we insert λ_i, to get (6), the total expected number of primary rejections.

$$f(\boldsymbol{M}) = \sum_{i \in I} \lambda_i \pi_i^B(\boldsymbol{M}) \tag{6}$$

Returning to the idea that (2) can be used to approximate the occupancy at a single ward, we have a way to estimate $f(\boldsymbol{M})$. Specifically, we use (2) to estimate $\pi_i^B(\boldsymbol{M})$, by calculating the blocking probability $p_c = B(c, \lambda/\mu)$ – known as the *Erlang-B* formula.
Inserting $B(c, \lambda/\mu)$ into (6), we are now able to derive an estimate of the objective value using (7). Doing so, gives us the opportunity to derive an estimate of the optimal solution in a few seconds.

$$\hat{f}(\boldsymbol{M}) = \sum_{i \in I} \lambda_i B(M_i, \lambda_i/\mu_i) \tag{7}$$

Now, from (5b) we have that $\sum_{i \in I} M_i = \Theta$. Therefore, $M_N = \Theta - \sum_{i \in I \setminus \{N\}} M_i$, reducing (7) to a function of $N-1$ variables, $\hat{f}(M_1, M_2, \cdots, M_{N-1})$. Let $\hat{f}_{M_i}(\cdot)$ be the i'th partial derivative of $\hat{f}(M_1, M_2, \cdots, M_{N-1})$, with the derivative of $B(M_i, \lambda_i/\mu_i)$ presented in (10) in Appendix A. The horizontal tangent plane of $\hat{f}(M_1, M_2, \cdots, M_{N-1})$ can then be found from the system of equations: $\hat{f}_{M_1}(\cdot) = 0 \wedge \hat{f}_{M_2}(\cdot) = 0 \wedge \cdots \wedge \hat{f}_{M_{N-1}}(\cdot) = 0$. We solve this, using the *Newton-Raphson method*.
Now, recall that the difference between (6) and (7), is the relocation of patients from fully occupied to free wards. Therefore, as the probabilities $p(\cdot)_{ij}$ from the CTMC model decrease, (6) approaches (7). In other words, an optimal solution derived using (7) is likely close to the optimal solution using (6). To locate the optimal solution to the optimization problem (5a)-(5c), an idea would therefore be to *slowly* change the solution configuration, starting with an initial guess derived from the estimate (7).

Let $N(\boldsymbol{M})$ define the "neighborhood" of the bed distribution $\boldsymbol{M}$, and still consider that $M_N = \Theta - \sum_{i \in I \setminus \{N\}} M_i$, so now $\boldsymbol{M} = (M_1 \; M_2 \; \cdots \; M_{N-1})^T$. Then, $(\boldsymbol{M} + \boldsymbol{\nu}) \in N(\boldsymbol{M})$, where $\|\boldsymbol{\nu}\| = 1$ and the elements $\nu_i \in \{0, -1, 1\}$. This leads to a maximum neighborhood size of $|N(\boldsymbol{M})| = N^2 - 1$ or $O(N^2)$. Hence, in case $N = 3$, $|N(\boldsymbol{M})| = 3^2 - 1 = 8$ solutions.
Consider if $\boldsymbol{M}^*$ is the currently best known solution to (5a)-(5c), then an idea would be to systematically check $N(\boldsymbol{M}^*)$ for an even better solution, and update $\boldsymbol{M}^*$ in case such a solution is found. This leads to the local search heuristic presented in Algorithm 2.

The heuristic progresses by firstly generating an initial solution from the horizontal tangent plane of (7). This is conducted using the function *init*() that, based on the Newton-Raphson method, takes an initial guess M^0. The "raw" output is most likely not integral, so we round to the integer solution yielding the lowest objective value. Next, the currently best known objective value, f^*, is calculated.
Then, *generateneigh*() is used to generate a list of the entire neighborhood. For larger cases, a probabilistic candidate list might be more appropriate, choosing a random fraction of the solutions in $N(\boldsymbol{M})$. Elements of the list should in any case be placed in random order.
Due to the local progression of the heuristic, and a relatively long function evaluation time, we further introduce a list of *banned* solutions, C. As the heuristic can only move one step at a time, there will always be an overlap between the neighborhood of iteration k and $k+1$. For this reason, we add all previously evaluated solutions to a list (line 10, Algorithm 2), to ensure that we do not spend time on evaluating a solution more than once.

4 Implementation & Results

In this section, we directly implement the methods from Section 3 to obtain an improved distribution of beds for a case-hospital. In modelling the system behavior, we have limited our scope to the hospitalization of patients to the medical area of the hospital. More specifically, we focus on patient flow in gastrology,

Algorithm 2 Heuristic to optimize the bed requirements problem in (5).

$M^* \leftarrow init(M^0)$ ▷ Initialize using the horizontal tangent plane of (7)
$f^* \leftarrow objval(M^*)$
$N \leftarrow generateneigh(M^*)$ ▷ Generate list of neighborhood in random order
$C \leftarrow \varnothing$
$j \leftarrow 0$
while $j < length(N)$ and $elapsedtime < timelimit$ **do**
    $j \leftarrow j + 1$
    $M \leftarrow N[j]$
    if $checkbanned(C, M) == false$ **then** ▷ Check banned or constraint violation.
        $C \leftarrow add(C, M)$ ▷ Add M to the list of banned solutions.
        $f \leftarrow objval(M)$
        if $f < f^*$ **then**
            $f^* \leftarrow f$ ▷ Update values
            $M^* \leftarrow M$
            $N \leftarrow generateneigh(M^*)$ ▷ Generate the new neighborhood, again in random order
            $j \leftarrow 0$
        end if
    end if
end while

pneumology, endocrinology and geriatrics, respectively. For the case hospital, these areas make up *three* different wards and correspondingly *three* different patient types.

In Section 4.1, we present the data obtained from the case-hospital and statistically test our homogeneous continuous-time Markov chain (CTMC) model. Next, Section 4.2 presents the implementation of our solution approach. Lastly, we assess the robustness of Algorithm 2, and investigate the solution behavior when the CTMC model parameters are adjusted. This is presented in Section 4.3.

4.1 Case & Data Description

The patient flow was investigated through interviews with hospital staff. Furthermore, we retrieved data from the period of 01-05-2014 to 30-04-2015 on patient arrival and discharge times. From this, we were able to categorize patients on diagnosis and thus also treatment type, giving us the opportunity to determine preferred and alternative wards.

4.1.1 Arrival Rates

Patient hospitalization data was used to derive hourly arrival rates for each of the three patient types, showing clear repetitive patterns on a weekly scale of the hourly arrival rate. In Figure 3, the empirical hourly arrival rates are presented for all patient types. As expected for non-acute wards, most patients are hospitalized during the daytime, whereas an almost negligible fraction of patients arrive during the night. Further, the arrival rates seem to slightly decrease during the weekend, and regain its level starting Monday. The empirical average arrival rates were estimated to $\lambda_1 = 5.42$, $\lambda_2 = 3.96$, $\lambda_3 = 2.52$ patients per day, respectively.

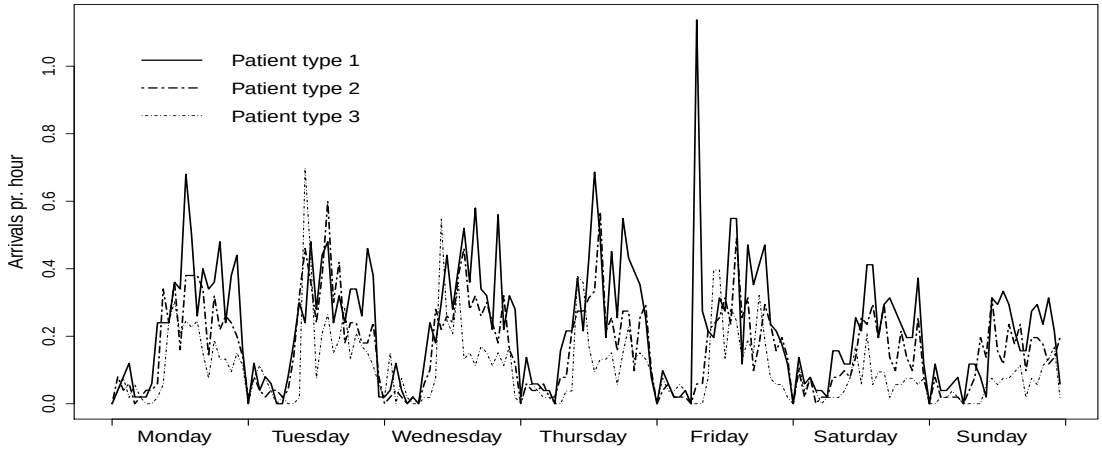


Figure 3: Empirical hourly arrival rate. Presented for each of the three patient types.

4.1.2 Service Rates

To derive service rates for each of the three patient types, we calculated the *Length of Stay* (LOS) using the same patient hospitalization data, as was used to derive the arrival rates. Time-dependency was checked on a daily level by deriving the average hourly LOS from time of arrival. Performing a graphical representation showed no signs of seasonality, neither did an estimate of the autocorrelation function. Regarding load-dependency, we did not obtain sufficient data to confirm nor reject such behavior. For the case-hospital, capacity meetings are often held to ensure that patients are immediately relocated upon blocking. As a result, neither LOS increase nor early-discharge due to overcrowding is rarely the case. The overall distribution of LOS was investigated graphically, where the patient type 1 and 2 distributions show close similarity to an exponential distribution (See Figure 4). With a longer average LOS, the patient type 3 distribution has probability mass that is moved more to the right, quite similar to a gamma distribution.
Due to the similarities that was found between patient type 1 and 2, we tested their difference in mean LOS using a Wilcoxon rank-sum test [23]. With a p-value of 0.2105, we found no significant difference in mean LOS between the two patient types. Figure 4 suggests that there is no difference in statistical distribution either. For patient type $i \in \{1, 2, 3\}$, the resulting service rate ($1/LOS_i = \mu_i$) was estimated based on an empirical average to $\mu_1 = \mu_2 = 0.19$ and $\mu_3 = 0.11$ patients per day, respectively.

4.1.3 Relocations

We investigated the secondary hospitalization options for each of the three patient types. From data, we obtained the 80% most common diagnoses for each patient type, and for each of these diagnoses, hospital staff identified the alternative locations they would usually offer to these patients. This allowed us to draw a picture of how relocated patient are usually distributed. Here, we found that patients have secondary hospitalization options both within and outside the three wards; hence, in case of blocking, a fraction of patient will always be lost from the system. Moreover, we found that patients would usually have a third hospitalization options – however, for this case, we found it reasonable to assume that a third hospitalization options is always situated outside the system. The resulting relocation probabilities are presented in Table 2, showing that a reasonably large fraction of patients has to be relocated elsewhere.

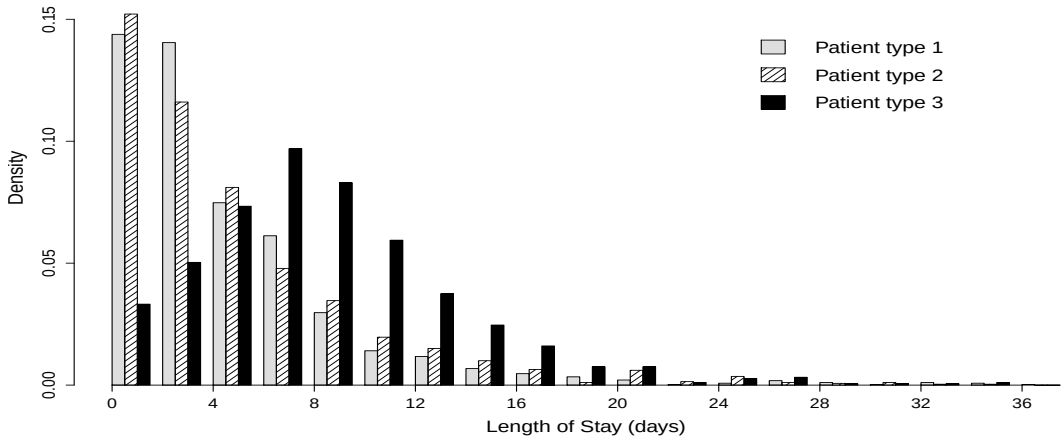


Figure 4: Empirical distribution of *length of stay*, measured in number of days. Presented for each of the three patient types.

Table 2: Probability that patient type $i \in \{1, 2, 3\}$ (P_i), is relocated to ward $j \in \{1, 2, 3\}$ (w_i), in case ward i is blocked.

P_i/w_j	1	2	3	Other
1	-	0.05	0.23	0.72
2	0.10	-	0.27	0.63
3	0.06	0.00	-	0.94

4.1.4 Other Characteristics of the System

From the distinct fluctuating arrival rate, one would naturally expect the level of hospitalized patients to be fluctuating as well. Figure 5 shows the empirical probability of a patient being discharge as function of hour of the week. As expected for non-acute patients, discharges mainly occur on weekdays during the daytime, with a negligible number of patients discharged during the night. Comparing with the arrival rates in Figure 4, we notice that the system is mainly "active" between 07:00 and 23:00.
Observations from 14-09-2015 to 31-10-2015 were obtained to investigate the time-dependent behavior of ward occupancy in the system. Figure 6 shows the average number of occupied beds every 8'th hour during the week. From here, we notice some time-dependent behavior as the occupancy is usually lower during the middle of the day. This behavior repeats on a daily basis, with a small overall decrease during the weekend for ward 2 and 3. Taking the time-dependency of hourly arrival rate and discharge hour into account might be necessary for purposes of accurately predicting the occupancy for each specific hour of the week. However, as our aim is to derive a long-term allocation of beds for the hospital, we consider the observed fluctuations as negligible for this case.

4.1.5 Truncation of the CTMC

For practical reasons we are often required to truncate the CTMC prior to implementation. We start this process by firstly considering the data obtained from the case-hospital. In Section 4.1.2 we found that $\mu_1 = \mu_2$; hence the number of patients of type 1 and 2, can be contained in only two "bins" of the state space. In other words, w_{11} and w_{12} are merged into w_{121}, as well as w_{21} and w_{22} are merged into w_{122}. Moreover, from Table 2, we have that $p(f_1, f_2, f_3)_{32} = 0$ in all cases, so w_{32} can be neglected. This results

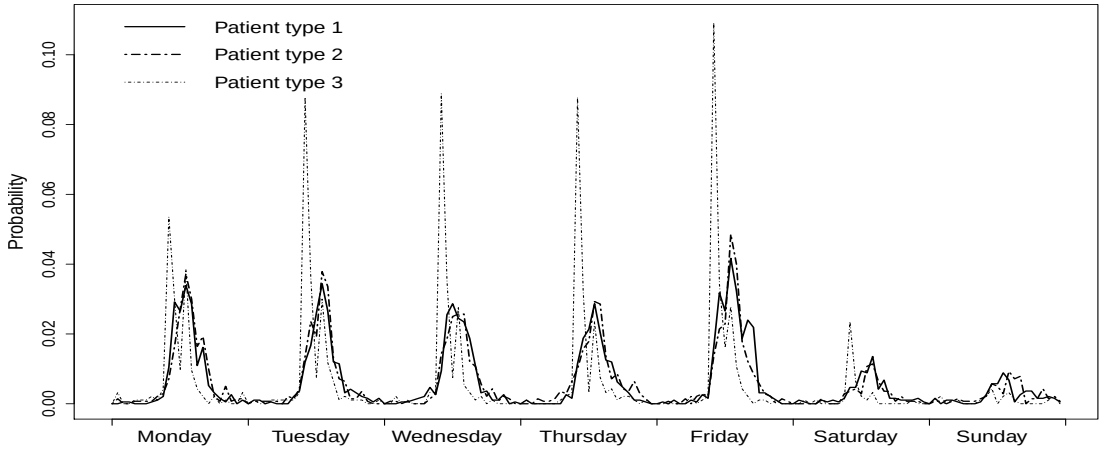


Figure 5: Empirical probability of a patient being discharged as function of hour of the week. Presented for each of the three wards.

in the state representation:

$$s \quad = \quad \left[\begin{pmatrix} w_{121} & - & w_{13} \\ - & w_{122} & w_{23} \\ w_{31} & - & w_{33} \end{pmatrix}, (f_1, f_2, f_3) \right] \quad \in S$$

Now we apply the truncation procedures described in Section 3.1.1. We use the truncation parameter $\tau = 0.01$, as proposed in Section 3.1.2. For the case-hospital, the total number of beds $\Theta = 74$, so we calculate u_{ij}, L_j and U_j for any feasible value of $M_j \in \{0, 1, \cdots, \Theta - (N-1)\} \quad \forall j \in J = \{1, 2, 3\}$.
To illustrate the resulting truncation, the total state space size, $|S|$, for a non-truncated model with $N = 3$ wards, where $M_1 = 27$, $M_2 = 23$ and $M_3 = 24$ has $|S| = 30,876,300,000$ states. The truncated model, with the same settings, has $|S| = (\sum_i^{u_{31}}(U_1 - i - \max\{L_1 - 1; 0\} + 1))(U_2 - L_2 + 1)(\sum_{j=0}^{u_{13}} \sum_{k=0}^{\min\{U_3 - i; u_{23}\}}(U_3 - j - k - \max\{L_3 - i - j; 0\} + 1)) = 1,358,760$ states — a substantial reduction of the state space.

4.1.6 Statistical Testing of the CTMC

We conducted a statistical test to assess the CTMC model fitness with observations on ward occupancy. To our knowledge, there exists no standard technique to test the fitness of a CTMC with a complexity as considered in this study. Thus in this section we present a heuristic approach that combines a simulation of the CTMC behavior and compare this to hospital data on ward occupancy.

To begin with, our null-hypothesis is that the observed values are generated by the CTMC process. If that is the case, we would expect the observed frequency of occupied beds to be quite similar to the marginal distributions of π for each ward. A standard approach would be to test the observed frequencies against the corresponding expected frequencies from the CTMC using a *chi-squared* test. However, such as test would require each of the observed values to be independent, which is not the case here.

Let ω_j be the expected number of occupied beds from the CTMC for ward $j \in J$, where $J = \{1, 2, 3\}$. Then $\omega_j = \sum_{k=0}^{M_j} k \cdot \pi_{kj}$, where π_{kj} is the probability that ward $j \in J$ has k occupied beds. Further, let o_{ij} and e_{ij} be the observed and expected frequency of i occupied beds in ward j. Then, we define our test statistic as (8),

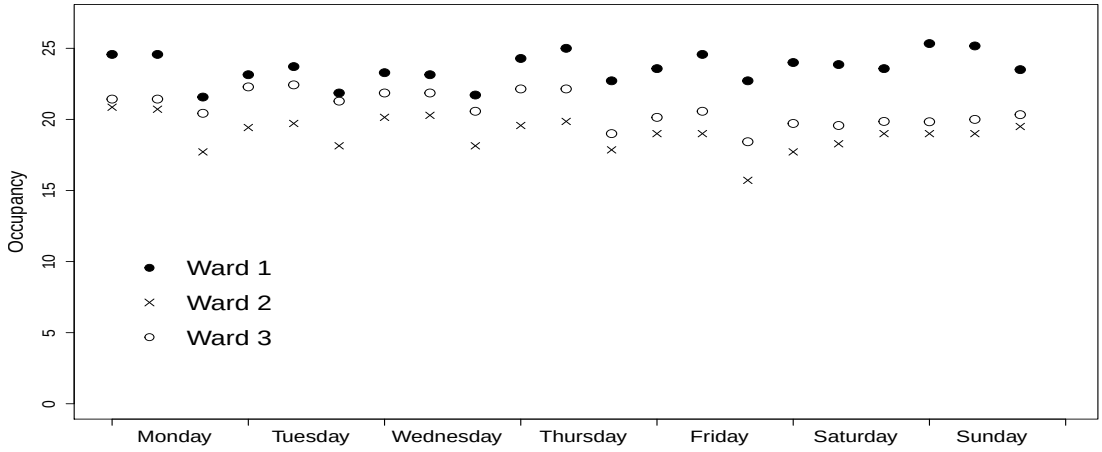


Figure 6: Empirical average number of occupied beds for each of the three wards. Observations were obtained for every 8'th hour of the week.

$$T = \sum_{j \in J} \sum_{i \in I} (o_{ij} - e_{ij})^2 / \omega_j \tag{8}$$

where $I = \{0, 1, 2, \cdots, M_j\}$ is the set of beds that can be occupied, and $J = \{1, 2, 3\}$ the set of wards. In order to quantify the fit of our CTMC model, we require a measure of how (8) relates to the model noise. For this reason, we introduce the simulated model residual (9), where y_{ij} is the simulated frequency of i occupied beds in ward j. Thus, by replicating (9), we determine the distribution of noise that is expected by our CTMC model, and then compare our results from (8) hereto.

$$z = \sum_{j \in J} \sum_{i \in I} (y_{ij} - e_{ij})^2 / \omega_j \tag{9}$$

We implemented the simulation of the CTMC model as a Discrete-Event-Simulation using `Matlab`. Replications of (9) were conducted $n = 30,000$ times (Appendices, Figure 7).
A total of 432 (144 pr. ward) observations were obtained from the period of 14-09-2015 to 31-10-2015. Using these, we calculated $T = 0.45$. A fraction of 32.03% simulated residuals, scoring higher than T, were found. Thus, with a significance level of $\alpha = 0.05$, we *accept* the null-hypothesis.

The power of our test was evaluated by conducting a range of experiments where model parameters were adjusted until under 5% significance would be obtained. Specifically, we adjusted the arrival rates proportionally with 10% increment at the same time.
Results are presented in Table 3, showing a 10% increase (Test 1) and 20% decrease (Test 3), were sufficient adjustments to gain less than 5% significance.

Table 3: Assessment of the power of our test. Conducted by proportionally changing the three arrival rate parameters (λ_i) and evaluating the resulting simulated p-value (p).

#	Change	λ_1	λ_2	λ_3	p-value
0	1.0	5.42	3.96	2.52	0.32
1	1.1	5.96	4.36	2.77	0.01
2	0.9	4.88	3.56	2.27	0.56
3	0.8	4.34	3.17	2.02	0.02

4.2 Optimizing the Case-Hospital

We now consider the heuristic in Algorithm 2 applied to the case-hospital. We initialize the heuristic, by handing $M^0 = (27\ 23)^T$ to *init*(). From the Newton-Raphson method, we get $M_1 = 31.80$ and $M_2 = 23.50$. The rounded integer solutions are then: (1) $f(\lfloor M_1^0 \rfloor, \lfloor M_2^0 \rfloor) = 1.473$, (2) $f(\lceil M_1^0 \rceil, \lceil M_2^0 \rceil) = 1.468$, (3) $f(\lfloor M_1^0 \rfloor, \lceil M_2^0 \rceil) = 1.470$ and (4) $f(\lceil M_1^0 \rceil, \lfloor M_2^0 \rfloor) = 1.467$. (4) returns the lowest value, so we set $M^* = (32\ 23)^T$, and proceed. The initial objective value is then $f^* = 1.603$ patients per day, and initial neighborhood $N = \{(33\ 23), (31\ 23), (32\ 24), (32\ 22), (33\ 24), (31\ 22), (31\ 24), (33\ 22)\}$. As the intention is to use our method as a *tactical* decision tool, a maximum time-limit of 4 hours is considered reasonable. The entire heuristic and its components are implemented in `Java`.
Each iteration is presented below, showing all function evaluations and how the list of banned solutions is updated progressively:

- *Iteration 1* – Checking: $f(31, 22) = 1.641$ and $f(33, 23) = 1.600$. $f(33, 23) = 1.600 < f^*(32, 23) = 1.603$, so we update, $f^* \leftarrow f$, to $f^* = 1.600$. The list of banned solutions is now: $C = \{(32\ 23), (31\ 22)\}$.
- *Iteration 2* – Checking: $f(34, 24) = 1.606$, $f(32, 22) = 1.623$, $f(34, 22) = 1.624$, $f(33, 22) = 1.620$ and $f(32, 24) = 1.592$. $f(32, 24) = 1.592 < f^*(33, 23) = 1.600$, so we update, $f^* \leftarrow f$, to $f^* = 1.592$. The list of banned solutions is now: $C = \{(32\ 23), (31\ 22), (34\ 24), (32\ 22), (34\ 22), (33\ 22), (32\ 24)\}$.
- *Iteration 3* – Checking: $f(31, 25) = 1.607$, $f(33, 24) = 1.596$, $f(31, 23) = 1.617$, $f(33, 25) = 1.603$, $f(32, 25) = 1.600$ and $f(31, 24) = 1.606$. After checking all available solutions in the neighborhood, f^* was not improved, so we **stop** and conclude that $M_1 = 32$, $M_1 = 24$ and $M_3 = \Theta - (M_1 + M_2) = 18$ is at least a local optimum. The total runtime was **1791.12** seconds (≈ 30 min.).

Notice, the final solution, $\boldsymbol{M}^* = (32\ 24)^T$, is actually an element in the initial neighborhood, and could have been selected by chance during the first iteration. Moreover, in generating the initial solution, $(32\ 24)^T$ would have been obtained by simply rounding to nearest integer.
Now, consider that the search space for three wards has a total size of $n = \sum_{i=1}^{\Theta-2}(\Theta - 1 - i) = (1/2)\Theta^2 - (3/2)\Theta + 1$. Thus for $\Theta = 74$, $n = 2628$ solutions. As n is reasonably low for this case, a complete enumeration of the search space is possible. To investigate how results from the heuristic relates to the optimal solution, we conducted a complete enumeration with the result presented in Appendices, Figure 8. Interestingly, the objective function contains only a single extrema – a global minimum in $M_1 = 32$ and $M_2 = 24$. Hence, we can conclude that the solution found from the heuristic, $\boldsymbol{M}^* = (32\ 24)^T$, is in fact the *global* optimal solution to the problem. We ask the reader to notice that the procedure of complete enumeration spend 464,212.02 seconds ($\approx$ 5 days and 9 hours) to complete. Thus, even though the procedure is possible, it is certainly not practical. The heuristic in Algorithm 2, solved the problem in just under 30 min. That is, 99.6% faster.

The optimal solution is compared to the current distribution of beds in Table 4. As for the current distribution, $f(\boldsymbol{M}) = 1.804$ patients per day, the optimal solution yields a 11.77% reduction in number of primary rejections. We notice for the current case, the highest probability of ward blocking, π_i^B, takes place in ward

1. Unfortunately, we find that patient type 1 has the highest arrival rate of 5.42 patients per day as well. Thus, it would be expected, in order to minimize $f(\boldsymbol{M})$, additional resources has to be pushed to ward 1 with a view to decrease π_1^B. Conversely, patient type 3 has the lowest arrival rate of 2.52 patients per day, and with $\pi_3^B = 0.161$, ward 3 is expected to reject 0.41 patients per day – 0.56 patients fewer than for ward 1. Turning to the optimal solution, we find the probability of rejection has been vastly increased for ward 3, but decreased for both ward 1 and 2. Further, we find that $\max_{i\in I}\{\lambda_i\pi_i^B\}$ has been decreased from 0.969 to 0.803 patients per day, and $\min_{i\in I}\{\lambda_i\pi_i^B\}$ decreased from 0.405 to 0.335 patients per day.

Table 4: The optimal solution compared to the current distribution of beds. Presented with objective values, $f(\boldsymbol{M})$, beds M_i and blocking probabilities π_i^B. The product $\lambda_i\pi_i^B$ shows the expected daily amount of primary rejections for each ward.

	Current			**Optimal**		
Ward	M_i	π_i^B	$\lambda_i\pi_i^B$	M_i	π_i^B	$\lambda_i\pi_i^B$
1	27	0.178	0.969	32	0.083	0.454
2	23	0.109	0.430	24	0.084	0.335
3	24	0.161	0.405	18	0.318	0.803
$f(\boldsymbol{M})$	-	-	1.804	-	-	1.592

4.3 Case Testing

With a view to investigate the solution behavior of our heuristic, we conducted a series of tests with various parameter adjustments. The hospital is planning to introduce a number of organizational changes, with the result of increased patient arrival rate, but additional overall bed capacity. Thus in our last test, we demonstrate how our approach may be used as a tool to assess future changes to the organization.

We conducted a total of five different basic tests, where patient flow or available resources were changed. In Table 5 the parameters that were subject to change are presented in bold font, the rest are from the hospital case.
The results for each of Test 1-5 are presented in Table 6, with firstly the initial solution, then the optimal solution, and lastly data on the heuristic progression. The total number of function evaluations that are avoided as a result of the list of banned solutions are presented in the second last column.
Giving the five tests a closer look, we expect for Test 1-3 that an increase in arrival rate results in a corresponding increase in allocated resources. This behavior is found for each of the three tests, where resources are allocated to respond to the increased demand for primary hospitalizations.

In Test 4 nothing was changed but the total amount of available beds Θ. We conducted this test, to assess the potential improvements caused by a relatively small increase in resources. We find that, as more resources are available in the system, additional surplus is created and the fractional distribution of beds between the wards is more balanced. In the original hospital case, the optimal fractional distribution was 43.2%, 32.4% and 24.3% for ward 1, 2 and 3, respectively. In Test 4 with $\Theta = 80$, this distribution changes to 42.5%, 31.3% and 26.3%. More importantly, as all wards receive more bed resources, the objective value is reduced correspondingly. Adding six additional beds yields a 38.9% reduction in the number of primary rejections.

Test 5 was conducted to assess the effect from relocation in the system on the optimal solution. To emphasize, we increased the demand for secondary hospitalizations in ward 3 substantially, by maximizing $p(f_1 = 0)_{13}$ and $p(f_2 = 0)_{23}$, keeping all other parameters fixed. Through these adjustments, we expect to increase the distance from initial to optimal solution. Moreover, we clarify how the optimal solution relates to a large probability of relocation within the system. Interestingly, it might seem natural to allocate beds to the ward

Table 5: Test parameters used to assess Algorithm 2. Parameters subject to change are presented in bold font.

#	λ_1	λ_2	λ_3	Θ	p_{13}	p_{23}
1	**6.775**	3.96	2.52	74	0.23	0.27
2	5.42	**4.95**	2.52	74	0.23	0.27
3	5.42	3.96	**3.15**	74	0.23	0.27
4	5.42	3.96	2.52	**80**	0.23	0.27
5	5.42	3.96	2.52	74	**0.95**	**0.73**

with an increased demand (ward 3), however the optimal solution reveals the objective value is minimized by moving beds to ward 1 and 2, and avoiding relocation in the first place.

Table 6: Results from the five parameter adjustment tests. Both initial and optimal solutions are presented. Information on the heuristic progression is presented in the last four columns.

	Initial				**Optimal**							
#	M_1	M_2	M_3	$f(\boldsymbol{M})$	M_1^*	M_2^*	M_3^*	$f^*(\boldsymbol{M})$	Iter.	Eval.	A. Eval.	Runtime (s)
1	38	22	14	2.376	39	23	12	2.354	4	12	5	368.27
2	31	28	15	2.165	32	29	13	2.158	3	12	5	686.39
3	31	23	20	2.180	32	23	19	2.175	3	11	6	1757.23
4	33	25	22	1.106	34	25	21	1.103	2	11	6	3925.02
5	32	23	19	1.733	33	25	16	1.688	7	21	4	3253.00

4.3.1 Assessment of Expected Hospital Changes

For our last test, we consider a number of organizational changes planned to be introduced in the spring of 2016. Patients of another organizational region are to be rerouted to the case-hospital. As a result, patient arrival rate is expected to increase. Moreover, the case hospital are given additional resources to cope with the increase in demand, and for the area of gastrology, pneumology, endocrinology and geriatrics, available resources will increase from 74 to 93 beds. Patient arrival rate of type 1 and 2 are now expected at 9.84 and 3.44 patients per day, respectively.

Just as previously, we generate the initial solution, starting with $M^0 = (27\ 23)^T$. Rounding to the smallest estimated objective value, we set $M^* = (56\ 20)^T$ and initial objective value $f^* = 1.965$. After 4 iterations we find the new distribution of beds at $M_1 = 56$, $M_2 = 21$ and $M_3 = 16$, with an objective value of $f^* = 1.958$ patients per day. The total number of function evaluations is 9 with an overall runtime of **961.35** seconds.

5 Conclusion & Future Work

With a view to optimizing the distribution of bed resources, we presented a solution approach consisting of two main components. The first was a homogeneous continuous-time Markov chain (CTMC) used to evaluate the patient flow behavior. The second incorporated the Markov chain model in a heuristic to optimize the distribution of bed resources. For a specific hospital case, our approach was used to find a 11.8% reduction in number of primary rejections – that is, the number of patients rejected on first arrival to the hospital. In addition, we found that a relatively small ($\approx$ 8%) increase in bed resources to the medical area has a potential to reduce this rejection rate from the current configuration with 38.9% fewer patients per day. Regarding this, hospital management should consider how the increase in resources relates to the overall

cost, and if a potential increase in cost is compensated by the increased service level.

We collected data for the case-hospital by conducting interviews with hospital staff, and using patient data already registered in the hospital system. During this process we found dependencies in the flow system that stretches toward far more wards than were resources to include in this study. On the other hand, we found it reasonable to assume the medical area as an isolated system with patients going *out*, rather than *in* from other wards. Hourly arrival rate was found to be time-dependent, but with discharges mainly occuring during the day, the time-dependent behavior could be neglected, as was confirmed from observations of ward occupancy. Additionally, we found it reasonable to assume that patient length of stay was independent of the system load. However, for other applications where load-dependency cannot be neglected, such behavior can be implemented by defining service rates of the CTMC as function of the ward occupancy.
We statistically tested the CTMC model by replicating simulations of the CTMC itself. These were compared to hospital observations, and a simulated p-value of $p = 0.32$ was derived. We concluded that the CTMC model is not significantly different from the observed ward occupancy.
The local search heuristic was evaluated using a range of different tests. Firstly, the case-hospital result was checked by conducting a complete enumeration of the search space. Here, we found the heuristic solution was in fact the global optimal solution to the problem. However, as complete enumeration is foreseeable for this problem size, it is certainly not practical as a decision tool. Even though global optimality cannot be proven without, we propose to use our approach, with a 99.6% reduction in runtime. Secondly, we tested our local search heuristic conducting five tests with different parameter adjustments, and one additional test resembling a future organizational change. The local search heuristic performed well in all tests.

5.1 Future Work

For future work, a larger number of wards should be considered. We notice that such an expansion would require a substantial increase in state space, possibly reducing the practical use of our modelling approach. It should be considered how other methods could help to support the CTMC model approach with a view to decrease runtime spend on function evaluations. Moreover, as the problem complexity grows, other local search techniques, such as Tabu Search, might be more suitable approaches.
Lastly, to further support our modelling approach, simulation experiments should be conducted to assess the nature of the system under different parameter settings, as well as the CTMC robustness to different inter-arrival and service time distributions.

Acknowledgements

This research was supported by the Danish governmental organization Region Sjælland. The managing organization of seven public hospitals located on Zealand and Falster. We particularly thank the department of Production, Research and Innovation for providing us with the necessary data to conduct this research. In addition, we would like to thank Senior Project Manager, Pernille Kirkvåg for providing us with essential information on patient flow, as well as Associate Prof. Anders Stockmarr for statistical advice.

References

[1] Flere fejl på afdelinger med overbelægning. URL: http://www.danskepatienter.dk/nyheder/flere-fejl-p-overbelagte-afdelinger, Accessed: 18-03-2016.

[2] Overbelægning er stadig et problem på sygehusene. Danske Patienter. URL: http://www.danskepatienter.dk/nyheder/overbel-gning-er-stadig-problem-p-sygehusene, Accessed: 18-03-2016.

[3] Sikkert patientflow – erfaringer fra et forbedringsprojekt. Dansk Selskab for Patientsikkerhed, 2015.

[4] Status på sundhedsområdet. Sundheds- og Ældreministeriet, 2015.

[5] Renzo Akkerman and Marrig Knip. Reallocation of beds to reduce waiting time for cardiac surgery. *Health Care Management Science*, 2009.

[6] Nicola Bartolomeo, Paolo Trerotoli, Annamaria Moretti, and Gabriella Serio. A markov model to evaluate hospital readmission. *Bmc Medical Research Methodology*, 8(1):23, 2008.

[7] Richard J. Boucherie and Nico M. van Dijk. *Queueing Networks: A Fundamental Approach*. Springer, 2011.

[8] James R. Broyles, Jeffery K. Cochran, and Douglas C. Montgomery. A statistical markov chain approximation of transient hospital inpatient inventory. *European Journal of Operational Research*, 207(3):1645–1657, 2010.

[9] J. K. Cochran and K. Roche. A queuing-based decision support methodology to estimate hospital inpatient bed demand. *Journal of the Operational Research Society*, 59(11):1471–1482, 2008.

[10] Jeffery K. Cochran and Aseem Bharti. Stochastic bed balancing of an obstetrics hospital. *Health Care Management Science*, 9(1):31–45, 2006.

[11] Xinyang Deng, Yong Deng, Xinyang Deng, Qi Liu, and Qi Liu. Newborns prediction based on a belief markov chain model. *Applied Intelligence*, 43(3):473–486, 2015.

[12] J Goldman, H A Knappenberger, and J C Eller. Evaluating bed allocation policy with computer simulation. *Health Services Research*, 3(2):119–29, 119–129, 1968.

[13] F. Gorunescu, S. I. McClean, and P. H. Millard. A queueing model for bed-occupancy management and planning of hospitals. *Journal of the Operational Research Society*, 53(1):19–24, 2002.

[14] F. Gorunescu, Sally I. McClean, and Peter H. Millard. Using a queueing model to help plan bed allocation in a department of geriatric medicine. *Health Care Management Science*, 5(4):307–312, 2002.

[15] LV Green. How many hospital beds? *Inquiry - the Journal of Health Care Organization Provision and Financing*, 39(4):400–412, 2002.

[16] R. A. Harris. Hospital bed requirements planning. *European Journal of Operational Research*, 25(1):121–126, 1984.

[17] Xiaodong Li, Patrick Beullens, Dylan Jones, and Mehrdad Tamiz. Optimal bed allocation in hospitals. *Lecture Notes in Economics and Mathematical Systems*, 618:253–265, 2009.

[18] H.P. Newsholme. Hospital bed accommodation. *Public Health - the Journal of the Society of Medical Officers of Health*, 46:73–77, 1932.

[19] J F Pendergast and W B Vogel. A multistage model of hospital bed requirements. *Health Services Research*, 23(3):381–399, 1988.

[20] Luiz Ricardo Pinto, Francisco Carlos Cardoso de Campos, Ignez Helena Oliva Perpetuo, and Yara Cristina Neves Marques Barbosa Ribeiro. Analysis of hospital bed capacity via queuing theory and simulation. pages 1281–1292, 2014.

[21] D. Sinreich, O. Jabali, and N. P. Dellaert. Reducing emergency department waiting times by adjusting work shifts considering patient visits to multiple care providers. *Iie Transactions Industrial Engineering Research and Development*, 44(3):163–180, 2012.

[22] William J. Stewart. *Probability, Markov Chains, Queues, and Simulation - The Mathematical Basis of Performance Modeling*. Princeton University Press, 1 edition, 2009.

[23] F WILCOXON. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[24] Xiaolei Xie, Jingshan Li, Colleen H. Swartz, and Yue Dong. Modeling and analysis of hospital inpatient rescue process: A markov chain approach. *Ieee International Conference on Automation Science and Engineering*, pages 6653987, 978–983, 2013.

[25] Hojjat Zeraati, Farid Zayeri, Gholamreza Babaee, Navid Khanafshar, and Fatemeh Ramezanzadeh. Required hospital beds estimation: A simulation study. 2005.

[26] Xin Li Zhang, Ting Zhu, Li Luo, Chang Zheng He, Yu Cao, and Ying Kang Shi. Forecasting emergency department patient flow using markov chain. *2013 10th International Conference on Service Systems and Service Management - Proceedings of Icsssm 2013, Int. Conf. Serv. Syst. Serv. Manage.*, pages 6602537, 278–282, 2013.

Appendices

A Equations

Derivative of the Erlang-B formula,

$$\frac{dB}{dM_i} = -\frac{\left(\frac{\lambda_i}{\mu_i}\right)^{M_i+1}\Gamma(M_i+1)\left(\frac{\lambda_i}{\mu_i}\right)^{M_i}\left(G_{2,3}^{3,0}\left(\begin{matrix}1,1\\0,0,M_i+1\end{matrix}\middle|\frac{\lambda_i}{\mu_i}\right)\left(\frac{\lambda_i}{\mu_i}\right)^{M_i}\lambda_i+\Gamma(M_i+1)\left(\Psi(M_i+1)-ln\left(\frac{\lambda_i}{\mu_i}\right)\right)\left(\mu_i\left(\frac{\lambda_i}{\mu_i}\right)^{M_i+1}-\left(\frac{\lambda_i}{\mu_i}\right)^{M_i}\lambda_i\right)\right)\mu_i}{e^{\lambda_i/\mu_i}\ M_i!\left(\left(\frac{\lambda_i}{\mu_i}\right)^{M_i+1}\mu_i\Gamma(M_i+1)+\lambda_i\left(\frac{\lambda_i}{\mu_i}\right)^{M_i}\left(\Gamma\left(M_i+1,\frac{\lambda_i}{\mu_i}\right)-\Gamma(M_i+1)\right)\right)^2} \tag{10}$$

where $G_{p,q}^{m,n}\left(\begin{matrix}a_1,\ldots,a_n,a_{n+1},\ldots,a_p\\b_1,\ldots,b_m,b_{m+1},\ldots,b_q\end{matrix}\middle|\,z\right)$ is the Meijer-G function, $\Gamma(x)$ and $\Gamma(s,x)$ the complete and upper incomplete gamma functions, respectively; and $\Psi(x)$ the digamma function.

B Figures

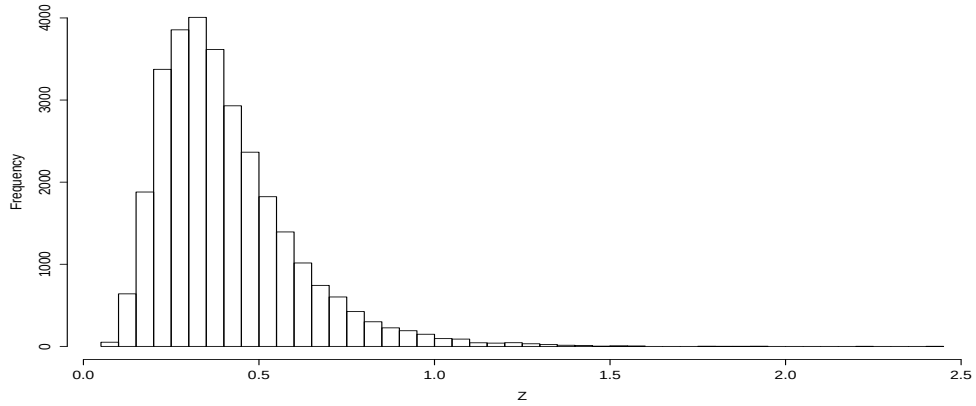


Figure 7: Simulated distribution of (9). Conducted with 30,000 replications.

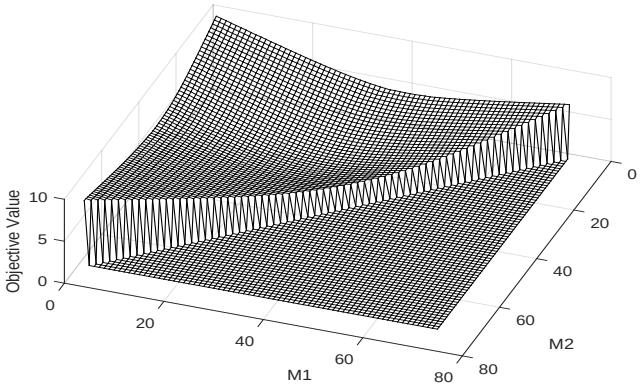


Figure 8: Complete enumeration of the search space for the current distribution of beds.