Ann Oper Res (2010) 178: 5–21
DOI 10.1007/s10479-009-0547-y

Stochastic dynamic nursing service budgeting

Gergely Mincsovics · Nico Dellaert

Published online: 20 May 2009

Abstract We address the nursing service budgeting problem from the department manager's point of view. The model allocates the budget dynamically to three types of nursing care capacities: 1) permanent nurses, 2) temporary nurses, and 3) overtime. The quarterly tactical decisions are the aggregate weekly shift pattern of permanent nurses and the policy for hiring temporary nurses and using overtime. The decisions are optimized with respect to nursing care shortage and a soft-constraint on the annual budget. For the aggregate weekly shift pattern, permanent nurses require a notification lead-time of one quarter to prepare the personal rosters. Our model offers a solution to the nursing service budgeting problem that extends the existing literature by using a Markovian demand model, resolving the anticipation of the operational decisions, and applying general budget as well as shortage penalty functions.

Keywords Manpower planning · Non-linear stochastic dynamic programming · Health service · Optimization · Stochastic processes

1 Introduction

For many hospitals, the costs and availability of nurses are of great concern. How to best allocate the nursing budget is a complicated problem that the operations research (OR) literature can help address. Today, the existing OR-literature on nurse planning is mainly concentrated on nurse scheduling models (see e.g. Burke et al. 2004). Interestingly, the nursing service budgeting problem has been utterly ignored in the last two decades.

The nurse workforce management process comprises decisions that are situated at different levels of the decision-making hierarchy. In general three stages are distinguished:

G. Mincsovics (✉) · N. Dellaert
Department of Technology Management, Eindhoven University of Technology, P.O. Box 513,
5600MB Eindhoven, The Netherlands
e-mail: G.Z.Mincsovics@tue.nl

N. Dellaert
e-mail: N.P.Dellaert@tue.nl

budgeting, scheduling, and daily staffing (Brusco and Showalter 1993). It is a complex task to model the entire hierarchical structure; even the simplest sensible budgeting calculations are difficult. Cavouras and McKinley (1997) warn for the necessity of feedback reporting among the decision-makers at different levels. Since the levels of the decision-making hierarchy are all interrelated, isolated optimization of the different levels can easily yield poor performance (Lowerre 1979). Easton et al. (1992) also point out the importance of integrating staffing and scheduling decisions over a year-long horizon. Following Abernathy et al. (1973), very recently, Li et al. (2007) recall the importance of integrating the decision levels in the workforce planning. These two papers suggest iterative methods for solving the hierarchical workforce planning models, they propose.

The more practice-oriented nursing service budgeting literature takes an accounting point-of-view to be applied by managers. Kirby and Wiczai (1985) recommend starting with a basic budgeting system. The simple calculations illustrated in their paper have two steps: (1) NHPPD (nursing hour per patient day) times annual patient days gives the annual workload, (2) annual workload times productivity factor per full-time equivalent hours (contractual FTE) gives the number of full-time nurses to be hired. Arthur and James (1994) review the major workload measurement practices. The exact determination of the productivity factor is an important element of this line of research (Lowerre 1979).

To the best of our knowledge, only two papers contain annual nursing service budgeting models in the traditional OR-literature. The primary concern of Trivedi (1981) is to ensure balanced staffing, to meet union demands, and to satisfy cost control and containment regulations, such that the number of part-time and the number of full-time nurses satisfy integrity. This complex problem is modeled as mixed-integer goal programming, which allows employing only deterministic demand. Kao and Queyranne (1985) study a set of models along three dimensions: having multi- or single-period, being disaggregate or aggregate for skill classes, and modeling probabilistic or deterministic demand pattern. We refer the reader to Venkataraman and Brusco (1996) for references to simulation based approaches on integrated budgeting and scheduling in services.

The traditional OR-literature does not address important concerns of other, health care originated papers. For example, Jeang (1996) alloys the work of Trivedi (1981) and Kao and Queyranne (1985) in order to build a stochastic model that provides the weekly pattern of the permanent nurses. The main decision is on the weekly pattern of the permanent nurses as in Trivedi (1981), nevertheless the model accounts for uncertainty of demand, as Kao and Queyranne (1985).

In our paper, similarly to Trivedi (1981), Kao and Queyranne (1985) and Jeang (1996), we introduce a new annual nursing service budgeting model with some additional complexities, and illustrate the model's performance via calculations with real-life data. We also follow the mentioned three budgeting papers in not comparing with any previously established model, but claiming a better representation of real-life. It is the model's different outputs and input needs that impede any of such comparison.

The two major additional complexities that we add are the non-stationary stochastic evolving demand and forecast updates. These aspects are not present in any of the above mentioned papers. Our non-stationary stochastic evolving demand representation allows calculation with multiple future demand scenarios at the same time as well as an autocorrelated demand process. Without forecast updates, capacity decisions are assumed to be made such that they do not react to the actual demand realization. In contrast, we use a Markovian demand model that allows us to represent forecast updates and our dynamic budget allocation to become responsive to these updates.

We aim at building a model that satisfies the main concerns of health care specialists and nursing managers and meets the modeling requirements of the state of the art in OR. To achieve this goal, we use the principles of OR literature in health care for conceptual modeling, the guidelines of health care specialists for modeling capacity decisions, and nursing managers' accounting suggestions for determination of important cost aspects. This paper thus helps the communication between OR specialists and health care managers as well as serves as a good basis for developing practical nursing service budgeting models.

In the following two sections, we introduce a set of descriptive models. These are the shortage penalty cost, the budget penalty, the demand and the productivity models. In the end of Sect. 3, we propose a decision structure and build a conceptual optimization model for the stochastic dynamic nurse budgeting problem. Simplifications of this conceptual model result in the final computational model, which is introduced in Sect. 4. We demonstrate the usefulness of the computational model via performing numerical experiments in Sect. 5. Conclusions are drawn in Sect. 6.

2 Longitudinal service budget allocation

This section presents a simple optimization model, which introduces the general notion of the longitudinal service budget allocation. The model represents the budget allocation process focusing on the trade-off between service capacity shortage and budget deviation. The severity of different levels of service capacity shortages and budget deviations are described by penalty functions, which we discuss in more details.

2.1 Capacity shortage penalty function

The concept of shortage penalty cost (a virtual cost) shows a general way of modeling loss of quality due to capacity shortage. This concept was developed for the case of nursing services by Warner and Prawda (1972). It is more general than the service level concept, see e.g. Jeang (1996), who restricts capacity allocation.

According to the shortage penalty concept, predefined penalty costs are assigned to capacity shortages. The following axiomatic statements characterize the shortage penalty function: (1) the penalty is positive in case of shortage and zero if there is no shortage; and (2) the cost is convexly (and non-linearly) increasing function of the shortage.

In our model, we use a generalized shortage penalty cost model. We assume that the shortage penalty cost is an arbitrary time-dependent function of the demand for nursing care and the available nursing capacity in the period.

2.2 Budget deviation penalty function

Although some existing models use penalty functions for budget deviations, the concept of such penalty has not been axiomatically described yet. We mention that the penalty function in the nursing service budgeting model of Trivedi (1981) is assumed to be linear, and that the more abstract service budgeting model of Zimmerman (1976) uses a general function form. In our model, we also use a general form of budget penalty function, which can depend on the annual budget and the annual capacity costs.

2.3 Conceptual optimization model for longitudinal budget allocation

We present a conceptual optimization model, (OM1), to demonstrate the use of shortage and budget penalty costs. The given budget W is distributed over periods with the objective to find a balance between shortages and budget deficit.

Model 1—Conceptual static optimization model for the budgeting problem (OM1)

$$\min_{capacity} \left\{ \left(\sum_t ShortagePenalty(demand_t, capacity_t) \right) \right.$$
$$\left. + BudgetPenalty(W, Cost(capacity)) \right\}.$$

The budget W is thus a resource, which is to be optimally allocated for employing capacity to minimize the shortages and budget deviations. Naturally, the shortage and budget penalty functions need to have the same unit of utility, so that we can calculate a trade-off between them. Our further models preserve the trade-off idea of this conceptual model, and extend it by a detailed decision structure and demand dynamics.

3 Stochastic dynamic nursing service budgeting

Many details are missing from OM1 that are specific to the budgeting of nursing service. Since our goal is to build a nursing service budgeting model, we need to gain understanding how the demand patterns look, what the productive part of the service capacity is, as well as the capacity sources and capacity decisions are.

3.1 Demand pattern characteristics

In order to find the most appropriate demand model, we need to gain understanding of its characteristics. Warner and Prawda (1972) find that demand prediction for nursing care has 5–10% error for a few days ahead. However, after the first few days, error increased to 20–30%. The relatively small error in the short-term allow them to use a deterministic demand model with a twice a week rolling schedule. Kao and Queyranne (1985) compare time-dependent stochastic demand models having independent, monthly periods with their deterministic counterparts. Their results indicate that ignoring demand uncertainty can lead to an underestimate of budget needs. By the computations, Kao and Queyranne (1985) use an ARIMA model to generate a stochastic demand model with independent periods. Kao and Tung (1980) show that the number of monthly patient days at different departments fit different ARIMA models. The type of their best fit ARIMA models suggest that the assumption of independent demand periods is not appropriate. Generally, the assumption of independent periods is not justified and unwanted.

In our model, we use a general stochastic demand process model with a general dependency structure. The only assumption is that the demand process is exogenous, i.e., our decisions do not influence the demand. Similarly to Kao and Queyranne (1985), we simplify the demand model for our computations. However, next to the time-dependency of demand, we allow dependencies between demands of different periods.

3.2 Productivity

The calculation of productivity gives a substantial part of the nursing capacity cost accounting. Keeling (1999) explains that we can expect to have 1,477 hours productive work (productive FTE) from the 2,080 hours contracted (contractual FTE). Lowerre (1979) lists holidays, vacations, personal days, and sick days that constitute the non-productive fraction of the contractual FTE, and shows a sequential procedure for accurate calculation.

Table 1 Some descriptive information on nursing from Siferd and Benton (1992)

Starting time	(on weekdays) three 49%, four to five 29%
Shift pattern	have most or all staff with a permanent shift assignment 49% have most or all staff who work a set pattern of days on and off 33% have most or all staff who work the same days each week 21%
Overtime	authorize nursing overtime 100 to 400 times per year 42% always use voluntary overtime for a shortage of nursing staff 54%
Temporary nurses	authorize use of temporary nursing staff 30 or more times per year 43% sometimes use hospital pool nurses for a shortage of nursing staff 62%
Hiring (and leaves)	hire new nursing staff 2 to 9 times per year 61%

One can find productivity constants in models as either the ratio of productive and contractual FTE as p in Kao and Queyranne (1985), or its reciprocal, as γ in Trivedi (1981). We employ the former, 'p' productivity definition in our model. Contrary to the deterministic productivity formulations in the literature, we take sample paths of productivity to describe its random behavior shift by shift (P_{ti}). To evaluate the available (productive) workforce in a shift, which is supposed to be an integer, we use controlled stochastic rounding ($R(.)$).

3.3 Capacity decisions

In this section, we make our modeling choices for the decisions, their timing and structure, and extend the optimization model OM1 to a more detailed one, OM2. After a short review of the axiomatic models, we recall some results of an empirical study that can serve as a good basis for the model extension.

Practical computational models on budgeting are concerned with calculating the number of full-time, part-time nurses on pay-roll, the overtime to be utilized, and a usually week long pattern that repeats throughout the budgeted horizon (Trivedi 1981; Jeang 1996), or make decisions only at an aggregate, monthly level (Kao and Queyranne 1985). All these models fail to represent the dynamics, if the capacity decisions are taken responsively to the recent demand realizations and the actual remaining budget. Consequently, these models do not anticipate future capacity response (to e.g. a sustaining low level of demand), and do not benefit from the fact that most of the capacity decisions need not to be made in the beginning of the horizon. In Sect. 3.1, we mentioned that the demand forecast error can increase from 5–10% for a few days ahead to 20–30% on a longer term. This observation suggests that delaying the decision making must have some added value.

A better source of input for modeling the decision structure is the empirical literature. Siferd and Benton (1992) surveys hospital nursing units providing useful statistics on nursing service capacity decisions. These statistics are summarized in Table 1 outlining how often certain capacity options are used among their respondents. Afterwards, we group the capacity decisions by their frequency, creating a new table, Table 2. Here, we include the decisions' lead-time in brackets, additionally. We deduce the hierarchical structure of decisions based on this table to build the model, OM2.

The *starting time* abbreviates the number of shifts and the daily time slots that the shifts cover. This is a decision at the strategic level: the feasible set of shift timings remains unchanged for years. In our model, we assume three shifts each day with some exogenous, fixed starting times.

Table 2 Decisions (and their lead times) grouped by their frequency

years	starting time (0)
yearly/quarterly	shift pattern (1), hiring and leaves (1)
per shift	temporary nursing (0 or more), overtime (0 or more)

The *shift pattern* stands for the aggregate weekly pattern of permanent nurses. That is, 21 values describing the regular number of permanent nurses in each shift of the week. Alternatively, the shift pattern can be for example two-week long. Our model calculates the aggregate shift patterns as well as the *hiring* decisions, which may be necessary, (approximately) quarterly. Naturally, it is more practical to shorten or prolong the quarters so that those start and end at the time of *leaves* or in times of better hiring opportunities.

We note that the determination of the number of permanent nurses from their aggregate shift pattern is simple: the total of the aggregate shift pattern needs to be multiplied by the ratio of the single shift workload per year, $365 \times 8\text{ h} = 2920\text{ h}$, and the productive FTE.

The operational level consists of *overtime* and *temporary nurse* hiring decisions. We assume a limited overtime, enough supply to reach the overtime limit, and an infinite supply of temporary nurses that we can unlimitedly use, as in Kao and Queyranne (1985). In our model, we assume to limit the overtime according to a predefined policy, which declares a set of feasible overtime values (Z). Temporary nurse hiring is mostly an ad-hoc operational decision. It is a separated short-term decision even in the best practical models (Bard and Purnomo 2006), not reckoning with the budget constraints on the long term. In the next subsection, we propose a model that overcomes the coordination problem of the operational and higher level decisions.

3.4 Conceptual optimization model for nursing service budgeting

We formulate a conceptual stochastic dynamic programming model for the budgeting problem, OM2, which uses the concept of OM1 and the descriptive models. The model, OM2 includes the demand and productivity aspects as explained in Sects. 3.1 and 3.2, and uses the decision structure of Table 2. In the model, the tactical and operational decisions create an embedded structure of (penalty) cost-to-go functions: the $f(.)$ functions correspond to the tactical decisions, and the $g(.)$ functions to the operational ones. The quarterly cost-to-go $f(.)$ has t index, while the shift index of $g(.)$ is i. There are $T = 5$ quarters since we include the first quarter's permanent shift pattern decision made one quarter in advance before the budgeted year starts (OM2.B); quarter t has I_t shifts.

In line with OM1, the only costs are the budget penalty and some shortage penalties (see $s_{ti}(.\ ,.)$ in (OM2.D) and $B(.\ ,.)$ in (OM2.G)). The model's state space consists of the remaining budget, the demand state, and the shift pattern for the next quarter. By the operational decisions, we need the current quarter's shift pattern, additionally. $E_{.}[.]$ stands for the expected value operator.

Model 2—Conceptual stochastic dynamic nurse budgeting optimization model (OM2)

$$f_0 = f_1(W, m_1), \tag{OM2.A}$$

$$f_1(r_1, m_1) = \min_{\bar{u}_2}\{E_{\mathcal{D}_1(m_1)}[f_2(r_2, \bar{u}_2, D(\mathcal{D}_1(m_1)))]\} \quad \text{with } r_2 = r_1, \tag{OM2.B}$$

$$f_t(r_t, \bar{u}_t, m_t) = \min_{\bar{u}_{t+1}}\{E_{\mathcal{D}_t(m_t),\mathcal{P}_t}[g_{t,1}(r_t - c^u\bar{u}_t, \bar{u}_t, \bar{u}_{t+1}, \mathcal{D}_t(m_t))]\}, \tag{OM2.C}$$

$$g_{t,i}(r_{t,1}, \bar{u}_t, \bar{u}_{t+1}, \mathcal{D}_t(m_t)) = \min_{v_{ti}, o_{ti}} \{s_{ti}(\mathcal{D}_{ti}(m_t), R(u_{ti}, \mathcal{P}_{ti}) + v_{ti} + o_{ti}) + g_{t,i+1}(r_{t,1} - v_{ti} c^v_j - o_{ti} c^o_j, \bar{u}_t, \bar{u}_{t+1}, \mathcal{D}_t(m_t))\}, \tag{OM2.D}$$

$$g_{t,I_t+1}(r_{t,I_t+1}, \bar{u}_t, \bar{u}_{t+1}, \mathcal{D}_t(m_t)) = f_{t+1}(r_{t,I_t+1}, \bar{u}_{t+1}, D(\mathcal{D}_t(m_t))), \tag{OM2.E}$$

$$f_T(r_T, \bar{u}_T, m_T) = E_{\mathcal{D}_T(m_T), \mathcal{P}_T}[g_{T,1}(r_T - c^u \bar{u}_T, \bar{u}_T, 0, \mathcal{D}_t(m_t))], \tag{OM2.F}$$

$$f_{T+1}(W, r_{T+1}) = B(W, W - r_{T+1}). \tag{OM2.G}$$

The tactical decisions can be seen in (OM2.B, OM2.C, OM2.F, OM2.G), which give the quarterly dynamics of the model, whereas the operational decisions are modeled in (OM2.D, OM2.E), which give the dynamics at the shift level. Equation (OM2.A) defines the minimal expected annual quality loss for a budget, W, and an initial demand state, m_1. Equation (OM2.B) determines the first nurse shift pattern of the budgeted year, $\bar{u}_2$. Equation (OM2.C) decides on the permanent shift pattern of a quarter ahead, where the feasible shift pattern, $\bar{u}_t$, is a 21 element long non-negative integer vector. Equation (OM2.D) describes the decisions per shift, where we minimize penalty costs for a given demand by making the best choice for overtime and temporary nurses. The feasible overtime values, v_{ti} are from the set, Z. The number of temporary nurses can be any non-negative integer value. Equation (OM2.E) calculates the transition to the new demand state, based upon the demand pattern realizations that correspond to the previous demand state, m_t. Equation (OM2.F) is similar to (OM2.C), but as this is the last quarter no shift pattern decision is made any more. Finally, in (OM2.F), takes into consideration the consequences of end-of-year budget deviations, how much the starting budget has been depleted by the costs of permanent nurses (OM2.C, OM2.F), and the costs of temporary help (OM2.D).

We represent the evolving demand by a set of year-long sample paths. These sample paths are categorized into groups in each quarter (e.g., low/medium/high total demand in the quarter). The groups are associated with demand states (m_t). The quarterly demand state transition probabilities can be calculated as the number of paths changing demand state accordingly.

The demand state serves as the tactical level information source that let model the dynamics of forecast updates. The tactical level decision, $\bar{u}_t$, is based on the knowledge of the remaining budget and the present demand state, from which the set of relevant path continuations can be extracted. For each of the relevant paths, a deterministic problem is solved resulting in the operational level decisions, v_{ti} and o_{ti}.

Table 3 summarizes the notations of variables used. The second column declares the variable category: input (I), output (O), and auxiliary (–). In the latter category, we classified outputs of no particular importance.

4 Computational model

Since OM2 has more than thousand decision epochs and a large state space, it is not attractive computationally. We build another model, which is a simplified version of OM2, and as such, its minimizations can be evaluated. We take two simplification steps.

4.1 Computational optimization model for the stochastic dynamic nursing service budgeting problem

In the first simplifying step, we reduce the number of decision epochs to the number of periods plus one by creating a hierarchical optimization structure. The numerous decision

Table 3 Notations (variable, variable category (input I, output O, and auxiliary –), description)

indices		
t	–	period index ($t = 1, \ldots, 5$, quarters; quarters $2, \ldots, 5$ are budgeted)
i	–	subperiod index within a period ($i = 1, \ldots, I_t$, 8h shifts, where $I_t = 270, 273$ or 276)
$j(t,i)$	–	(we abbreviate it to j) subperiod index within a midperiod ($j = 1, \ldots, 21$, 8h shifts of the week)
capacity decisions		
$\bar{u}_t$	O	$= (\bar{u}_1, \ldots, u_{21})$ aggregate weekly shift pattern of permanent nurses, where u_j is the number of permanent nurses in each weekly subperiod j of the actual period, and similarly, u_1 is the number of permanent nurses in subperiod i of the actual period
V_t	O	temporary help policy; it gives our preferred exchange rate between the penalty costs and the real money
v_i	O	the number of temporary pool nurses hired in subperiod i of the actual period ($v_i = 0, 1, 2, \ldots$)
o_i	O	the amount of overtime utilized in subperiod i of the actual period ($o_i \in \mathbf{Z}$)
capacity restrictions		
Z	I	the set of feasible amount of overtime per shift
cost functions		
$f_t(.)$	O	expected minimal penalty-cost-to-go from the beginning of period t onwards
$g_{t,i}(.)$	O	expected minimal penalty-cost-to-go from the beginning of subperiod i of period t onwards
cost coefficients		
c_j^u	I	unit shiftly cost of a permanent nurse in the subperiod j of a week
c_j^v	I	unit shiftly cost of a temporary pool nurse in the subperiod j of a week
c_j^o	I	unit shiftly cost of overtime in the subperiod j of a week
budget and cost		
W	I	annual budget
r_t	–	remaining budget in the beginning of period t
$C(\bar{d}, \bar{u}, V)$	O	capacity cost subtotal in the actual period for demand sample path $\bar{d}$, aggregate permanent midperiodly shift pattern ii, and temporary help policy V
penalty cost functions		
$St(d, \bar{u}, V)$	O	nursing care shortage penalty cost subtotal in period t for demand sample path d, aggregate permanent weekly shift pattern $\bar{u}$, and temporary help policy V
$s_j(d, c)$	I	nursing care shortage penalty cost in the subperiod j of a week, when demand for nursing care is d, and the nursing capacity is c
$B(W, C)$	I	budget penalty cost for annual budget W and for annual capacity costs C
demand		
m_t	–	demand state that captures information about the demand process' of period t in the beginning of the period
$\mathcal{D}_t(m_t)$	I	Markovian demand process in period t
d_i	–	demand realization in subperiod i of the actual period
$D(\bar{d})$	–	demand state transition function, e.g. if the demand realization in period t was $\bar{d}$, it gives the next demand state m_{t+1}; it can be interpreted as forecast information for period $t+1$

Table 3 (*Continued*)

productivity		
$\mathcal{P}_{ti}$	I	a random process that gives the fraction of the productive permanent nursing capacity
$R(x)$	–	$= \begin{cases} \lceil x \rceil & \text{with probability } x - \lfloor x \rfloor \\ \lfloor x \rfloor & \text{with probability } \lceil x \rceil - x \end{cases}$, a random variable that helps generate randomized integers from the real number x

epochs of OM2 are a result of the operational decisions. Our goal is to find tactical decisions that define the operational decisions. Therefore, we assume that the operational decisions follow a policy, which is decided at the tactical level. We call this policy the temporary help policy V_t, renewed each quarter. By this way the operational decisions form a consistent part of the budget allocation. For example, if we expect to end up with budget deficit, we will be less willing to use overtime or hire temporary nurses.

In the second simplifying step, we reduce the state space via a mapping. It is the aggregate permanent shift pattern that makes the state space large. We carry 21 dimensions of the shift pattern next to few other dimensions. We propose an approximation that helps in resolving the curse of dimensionality for this situation. Namely, we create a mapping $\bar{u}_t(m_{t-1}, B_t^u)$, which calculates the aggregate shift pattern from only two dimensions: the demand state at the time of the decision m_{t-1}, and the budget part allocated to cover the shift pattern of permanent nurses B_t^u. This way the multiple dimensions of $\bar{u}_t$ are translated to the two dimensions m_{t-1} and B_t^u. The necessary mapping we calculate via a greedy algorithm.

The apparent simplicity compared to OM2 is a result of the disappearing optimizations (OM2.D) and (OM2.E) and the reduced state space dimensionality. Equation (SDNBOM.A) represents the minimal expected annual penalty costs for a budget W and an initial demand state (m_0, m_1). Equation (SDNBOM.B) determines the first quarter's permanent nurse budget, while (SDNBOM.C) also decides on the temporary help policy parameter, V_t, minimizing the expected future penalty costs for the remaining budget.

Model 3—Stochastic dynamic nurse budgeting optimization model (SDNBOM)

$$f_0 = f_1(W, m_0, m_1) \tag{SDNBOM.A}$$

$$f_1(r_1, m_0, m_1) = \min_{B_2^u}\{E_{\mathcal{D}_1}[f_2(r_2, m_1, D(\mathcal{D}_1), B_2^u)]\} \quad \text{with } r_2 = r_1 \tag{SDNBOM.B}$$

$$\begin{aligned} f_t(r_t, m_{t-1}, m_t, B_t^u) = \min_{V_t, B_{t+1}^u} \{&E_{\mathcal{D}_t, \mathcal{P}_t}[S_t(\mathcal{D}_t, \bar{u}_t, V_t) \\ &+ f_{t+1}(r_t - c^u\bar{u}_t - C(\mathcal{D}_t, \bar{u}_t, V_t), m_t, D(\mathcal{D}_t), B_{t+1}^u)]\} \end{aligned} \tag{SDNBOM.C}$$

$$\begin{aligned} f_T(r_T, m_{tT-1}, m_T, B_T^u) = \min_{V_T}\{&E_{\mathcal{D}_T, \mathcal{P}_T}[S_T(\mathcal{D}_T, \bar{u}_T, V_T) \\ &+ f_{Tt+1}(W, r_T - c^u\bar{u}_T - C(\mathcal{D}_T, \bar{u}_T, V_T)))]\} \end{aligned} \tag{SDNBOM.D}$$

$$f_{T+1}(W, r_{T+1}) = B(W, W - r_{T+1}) \tag{SDNBOM.E}$$

where $\bar{u}_t$ and $\mathcal{D}_t$ abbreviates $\bar{u}_t(m_{t-1}, B_t^u)$ and $\mathcal{D}_t(m_{t-1}, m_t)$, respectively.

A fortunate advantage of SDNBOM is that the state space transformation with the $\bar{u}_t(m_{t-1}, B_t^u)$ mapping imported a new dimension, m_{t-1}, to the state space of period t. This implies that we can afford to use a second-order Markovian demand model. We remark that some extra calculations are necessary: in (SDNBOM.C) and (SDNBOM.D) we need to evaluate the $S_t(.)$, the $C(.)$ functions, and $\bar{u}_t$.

4.2 Algorithms for the simplifying calculations

In this section, we discuss implementation issues of algorithms that can evaluate the $S_t(.)$, and $C(.)$ functions, and the $\bar{u}_t$vectors. The algorithms, we present are examples and not intended to provide the optimal operational decisions. Instead, the provided operational decisions are reasonable and, importantly, coordinated with the tactical decisions.

Algorithm 1 calculates the shortage penalty and capacity cost $S_t(.)$ and $C(.)$ for a given temporary help policy and sample path, optimizing the temporary help and overtime. The temporary help policy class has a single parameter, V_t, which gives our preferred exchange rate between the quality-related penalty costs and the money allocated from the budget for temporary help, in quarter t. We invest into overtime and/or temporary nurses up to the capacity level where the shortage penalty/capacity cost ratio in the shift get closest to V_t *while not exceeding it*. The policy parameter V_t, we optimize in the beginning of each quarter. For the calculations, we take a number of realizations of D_t as a function of m_{t-1}, which are sample paths (vectors) with elements d_i, the demand in shift i.

Algorithm 2 is a greedy algorithm, which calculates the aggregate permanent shift pattern, $\bar{u}_t$. Here, $Q_j(k) = E_{D_t}[\sum_{\substack{\text{the weekly index of}\\ \text{shift } i \text{ inquarter } t \text{ is } j}} s_i(D_{ti}, k)]$ is defined as the expected sum of shortage penalties in the actual period incurred if k nursing capacity is used for the weekly index j (e.g., j is 'Monday night', then $Q_j(5)$ is the sum of shortage penalties of Monday night shifts throughout the period if 5 permanent nurses are hired). $(Q_j(0))_j$ is the vector of $Q_j(0)$'s having elements for all j's. B^u is a given upper bound for B^u_t. ΔQ is the vector of the actual Q_j gradients. The algorithm provides the table (B^u_i, u^i) for all shift i, for some given quarter and demand state.

Algorithm 1 Calculation of shortage penalty and capacity costs $S_t(.)$ and $C(.)$

$$q_i(v, o) = \frac{s_i(d_i, R(u_i\mathcal{P}_{ti}) + v + o)}{u_i c^u_i + v c^v_i + o c^o_i}$$

$$(v_i, o_i) = \operatorname*{arg\,max}_{\substack{(v,o)\in\mathbb{N}_0\times Z\\ q_i(v,o)\le V_t}} q_i(v, o)$$

$$C(d, u_t, V_t) = \sum_i v c^v_i + o c^o_i$$

$$S_t(d, u_t, V_t) = \sum_i s_i(d_i, R(u_i\mathcal{P}_{ti}) + v_i + o_i)$$

4.3 Justification of the simplifying steps

The policy class that Algorithm 1 represents is an approximation, which ignores the demand- and budget-responsiveness of the temporary help decisions within the quarter, but it remains quarterly responsive. Under the assumption that the quarters are static (the demand

Algorithm 2 Calculation of the permanent shift pattern $\bar{u}_t(m_{t-1}, B_t^u)$

```
$(i, B_i^u, u^i, Q) := (0, 0, 0, (Q_j(0))_j)$
$\Delta Q := \left(\frac{Q_j - Q_j(1)}{c_j^u}\right)_j$
while $B_{i+1}^u \leq B^u$
    $j* := \arg\max_j \{\Delta Q_j\}$
    $u^{i+1} := u^i$
    $(B_{i+1}^u, u_{j*}^{i+1}) := (B_i^u + c_{j*}^u, u_{j*}^{i+1} + 1)$
    $Q^{\text{new}} := Q_{j*}(u_{j*}^{i+1})$
    $(\Delta Q_{j*}, Q_{j*}) := (Q_{j*} - Q^{\text{new}}, Q^{\text{new}})$
    $i := i + 1$
end while
```

and budget circumstances do not change), Algorithm 1 can provide optimal temporary help decisions. Namely, if $s_i(d_i, .)$ is convex for all i and all possible demand value d_i, then Algorithm 1 becomes equivalent with a greedy algorithm yielding an optimal behavior (Fox 1966). In the greedy algorithm, V_t becomes the terminating gradient value. Since V_t is optimized, the budget spent for temporary help in the quarter is also optimized. Note that if the year is split into more periods, then the approximation improves.

The optimality of Algorithm 2's greedy mechanism may be damaged by poor anticipation of the future use of temporary help. We may assume that either no temporary help is used or we use temporary help only as replacement in case of absenteeism of the permanent workers (Warner and Prawda 1972). Under any of these assumptions, the greedy Algorithm 2 gives an optimal solution (Fox 1966).

4.4 Discussion of the assumptions

Since the nursing service budgeting is a composite problem, many assumptions need to be taken while modeling. We create an explicit list of our modeling assumptions so that one can judge the models applicability. The assumptions are ordered by their time span decreasingly.

Assumption 1 Three shifts a day meaning three fixed starting times.

Siferd and Benton (1992) reports that 49% of the surveyed hospitals use three starting times. In the cases, where more starting times are in use, appropriate alternative algorithms to Algorithm 1 and 2 are difficult to find. Provided these algorithms, however, our model is still applicable without taking this assumption.

Assumption 2 Demand for nursing care is purely exogenous, independent from the capacity level, the quality of care provided or other controllable variables.

Demand for care is not independent on the capacity levels, in general. On the one hand, when staffing at adequate levels patients' stays are shorter than by consistently short staffing, because then adequate nursing care is received (Flood and Diers 1988). On the other hand, by a continuously unsatisfactory level of nursing capacity, patients may tend to select another hospital because of the low quality or the long waiting times. Our assumption can, however, get justification since our model optimizes capacity decisions for a given budget. The budget limits the long-term level of nursing capacity that can be provided, so both the capacity and the service level can be, approximately, regarded as constant.

Assumption 3 Years are independent; no long-term effects are taken into account.

On the strategic long-term, the surrounding population of patients and nurses can increase or decrease. Costs of capacities can change; new regulations may come into effect. We do not model these aspects, although the annual change in the patient population can be incorporated into the demand model.

Assumption 4 In the end of each quarter nurses can be hired or fired in unlimited amount at no cost.

This assumption can be restrictive in times of a nursing shortage (Brusco and Showalter 1993). For the situations, where hiring cannot be solved easily, we suggest using a modified, constrained version of SDNBOM, where we constrain the search space of the aggregate shift pattern of permanent nurses (e.g., $|\sum_j u_{t+1,j} - \sum_j u_{t,j}| \leq 5$ for some t).

Assumption 5 Once we set up an aggregate weekly shift schedule, new permanent nurses are hired, and the permanent nurses establish a set of personal rosters to meet the aggregate schedule in one quarter.

We can use period lengths different from a quarter, which are suitable to describe the lead-time of hiring and personal roster negotiations. Note that our model allows variable period lengths as well.

Assumption 6 Budget is known a quarter in advance, before the budgeted year starts.

For the case, if the budget is not known by the time, we decide on the coming year's first shift pattern, we can say, it is known in stochastic terms. Again, we suggest using an altered version of SDNBOM: an expectation on the budget may follow the first stage's minimization.

Assumption 7 Nursing care shortage depends only on demand, capacity and time.

Although this type of shortage penalty function is a broad generalization of that in Warner and Prawda (1972), it can still carry restrictions to particular situations. For example, a further generalization to dependency on demand, permanent capacity, overtime, temporary capacity and time may be preferable. The SDNBOM model allows this generalization without any additional calculational complexity.

Assumption 8 Budget penalty depends only on the given annual budget and the annual costs.

We generalized the linear soft budget constraint of Trivedi (1981) to an arbitrary penalty function. We are not aware of any sensible broader generalization.

Assumption 9 No difference in efficiency between nurses.

This assumption relates to Assumption 7: we can translate the possible efficiency differences to differences in the generalized shortage penalty function.

Assumption 10 Demand forecast error is negligible in the short-term: demand during a shift is assumed to be known at the beginning of the shift.

Naturally, this assumption can be seriously restrictive, when a large fraction of incoming patients is emergency type, and the claimed closely deterministic short-term demand structure is not valid. Otherwise, it is a reasonable assumption (Warner and Prawda 1972).

Assumption 11 Temporary nurses can be hired only for whole shifts.

Assumption 12 Overtime for less than one shift.

If we needed more nursing care that the permanent nurses can provide, we will use temporary nurses or overtime. As long as we can calculate the best feasible overtime—temporary nurse combination from the desired temporary help capacity, we can modify the SDNBOM model to cover any temporary capacity policy till the shifts are independent. I.e., the SDNBOM model cannot treat policies that have constraints on a set of shifts, e.g., if one had overtime last weekend, she would not be allowed to have overtime this weekend.

Assumption 13 Temporary nurses and overtime volunteers have infinite supply; we can hire them in the beginning of the shift.

Temporary nurses and overtime volunteers are not generally always available (Brusco and Showalter 1993). Constant upper limits on their number can be included in our model. Alternatively, by gradually increasing the hourly cost of temporary nurse hiring, we can also lower the use of temporary nurses to some given limit.

Assumption 14 No carry-over of workload from shift to shift (lost service).

In the call-center staffing literature, Atlason et al. (2005) point out that service of consecutive (short) time periods are interrelated, demand is partially lost, and partially carried over to the next period. We can expect the same interrelation to hold for demand for nursing care.

Assumption 15 Single nurse class and substitution between classes.

If there are fixed ratios between nursing classes, the permanent capacity cost will be approximately linear function of the permanent capacity. For the better application of the greedy algorithm, we need to have this permanent capacity cost function being convexly increasing (Fox 1966). We do not handle substitution between classes.

Table 4 The pure and the relative quadratic penalty cost functions

pure quadratic	relative quadratic
$ShortagePenalty = (demand - capacity)^2$	$ShortagePenalty = \left(\frac{demand-capacity}{demand}\right)^2$

5 Numerical experiments

In this section, we first demonstrate what kind of solution SDNBOM can provide. Afterwards, we show experiments investigating the shortage penalty cost model selection.

To illustrate the functioning of SDNBOM, we used workload data of a mental health inpatient ward (Ridley 2007). The maximum of nursing care demand per shift was eight nurses; the daily total demand was about 11, with around 6, 3, and 1 in the day, evening, and night shifts, respectively. The number of demand sample paths was twelve, which we gained by simulating the ARMA demand process fit to the workload data. The more sample paths, the better the demand process is represented, and the more memory and calculation time is needed. The number of productivity sample paths was three, with an average productivity around 70%. The number of demand states per period was three. Under this setting, one evaluation of the SDNBOM took around 5 minutes (main computer parameters: 2.8 GHz CPU, 1 Gb RAM).

We modeled the capacity shortage costs as being dependent both on the demand and the available capacity, resulting in the relative quadratic penalty cost function (see Table 4), which we considered more realistic than the pure quadratic shortage penalty cost function of Warner and Prawda (1972). Namely, as opposed to the relative quadratic penalty, the pure quadratic penalty has a shortcoming in that it regards the situation with one demand and no nurses as severe as having ten demand and nine nurses.

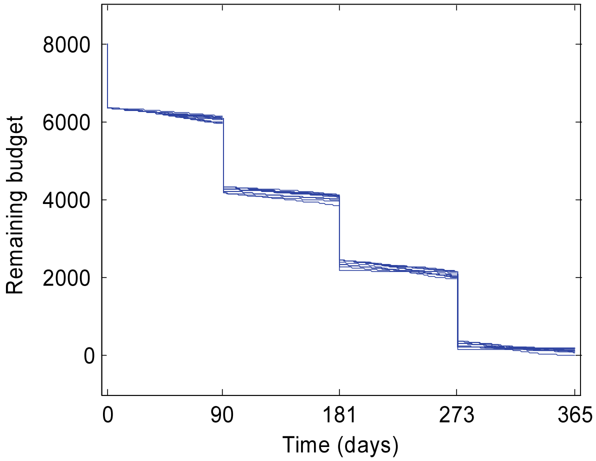
While solving SDNBOM with the setting described above, we archived the state-dependent decisions throughout the year. This archive allowed us to show for the different demand sample paths how the budget is allocated longitudinally (see Fig. 1), and how far capacity is matched to demand (Fig. 2).

In Fig. 1, we can see how the budget is allocated to permanent capacity and temporary capacities (including overtime) along the year for 12 demand scenarios. Each quarter starts with a step downwards, which corresponds to the quarterly permanent capacity expenditures. Within each quarter, some part of the budget is consumed by the temporary capacity expenses. By the end of the year, the remaining budget finishes around zero. For some scenarios, this could only be reached by allowing considerable capacity shortage costs in the last quarter.

Figure 2 depicts the match between demand and capacity for each shift and each scenario, quarterly grouped. Although the majority of the points are on the shortage side because of the limited budget, SDNBOM provides a good match between demand and capacity.

In our further experiments, we evaluated the SDNBOM with pure and relative quadratic penalty cost functions and tested the impact of modeling updated forecasts. We compare the outcomes in Table 5. The table demonstrates the solution of the first quarter's permanent shift pattern for an annual budget of 8,000 and forecast updates. Because of the randomization in the SDNBOM, we can get different results for the same parameter setting. The repeated experiments under the same setting showed that using the pure quadratic penalty the SDNBOM sometimes assigns zero nurses to night shifts (still overtime and temporary labor can be used) and more fluctuations in the number of permanent nurses in a shift of the week. To the contrary, using the relative quadratic penalty seems to result in overstaffing night shifts. Although the SDNBOM does not enable us to justify the use of any penalty

Fig. 1 Illustration of the remaining budget in the course of the year for 12 demand scenarios



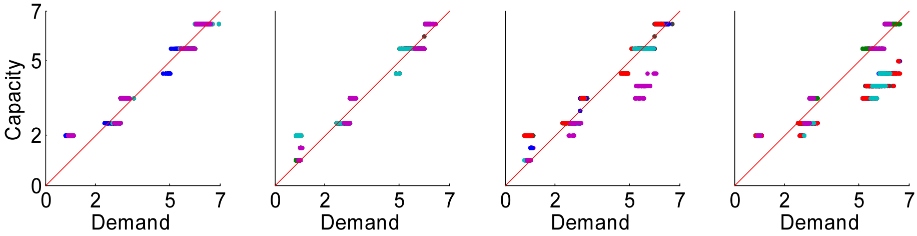


Fig. 2 Demand vs. the available capacity in the four quarters for the same 12 demand scenarios as in Fig. 1. The *diagonal line* indicates from where shortage penalty is to be paid (below the line)

Table 5 Outcomes and outcomes ranges under the same parameter settings for the pure and the relative quadratic penalty cost functions based on 25 runs

Pure quadratic		Permanent shift pattern for the first quarter ($\bar{u}_t$)																					Total FTE
	B_2^u	D	E	N	D	E	N	D	E	N	D	E	N	D	E	N	D	E	N	D	E	N	
median	2,038.3	9	5	2	9	5	2	9	5	2	9	5	2	9	5	2	9	5	2	9	5	2	22.0
minimum	1,428.3	7	2	0	7	3	0	7	3	0	7	3	0	7	3	0	7	2	0	7	3	0	15.0
maximum	2,074.7	10	5	2	9	5	2	10	5	2	10	5	2	9	5	2	10	5	2	9	5	2	22.4

Relative quadratic		Permanent shift pattern for the first quarter ($\bar{u}_t$)																					Total FTE
	B_2^u	D	E	N	D	E	N	D	E	N	D	E	N	D	E	N	D	E	N	D	E	N	
median	1,561.7	5	4	3	5	4	3	6	4	3	5	4	3	6	4	3	5	4	3	5	4	3	17.4
minimum	1,433.9	4	4	3	4	4	3	5	4	3	4	4	3	5	4	3	5	4	3	4	4	3	16.2
maximum	1,861.5	7	5	3	7	5	3	7	5	3	7	5	3	7	4	3	7	5	3	7	4	3	20.6

cost functions, we can conclude that the results are sensitive on the selection of the shortage penalty cost model, and that the preferences of the hospital management could play a large role in deciding on the penalty cost function.

In our further comparisons, we evaluated the SDNBOM with and without modeling forecast updates. In this example, we found that the penalty cost reduction gained from using quarterly forecast updates is 91%. However, for some stable demand process, we generated, the cost reduction becomes marginal, between 0–2%. We note that the response to the quarterly forecast updates is sometimes a big change in the permanent nursing capacity, which is not necessarily wanted.

6 Conclusions

We built a stochastic dynamic optimization model for the nursing service budgeting problem based on existing concepts. We generalized some of these concepts, and in few cases we gave suggestions for further generalizations. First, we built a conceptual optimization model, which consisted of exact descriptive models of the generalized concepts. As far as the underlying concepts and the data are reliable, the OM2 conceptual model provided optimal decisions. Due to its high complexity, we could not evaluate the conceptual model. Therefore, we proposed some simplifying steps, which led to our final optimization model, SDNBOM. We verified the results the SDNBOM provide and illustrated how far it makes demand and capacity match. The SDNBOM model, we could evaluate in some minutes on a single personal computer, for multiple periods.

By building our model, we put emphasis on the precise modeling of reality. Although we cannot expect the results to be optimal, we used our model to test assumptions on the shortage penalty cost models and on the demand forecast updates. We found that different shortage penalty cost functions can lead to quite different staffing decisions. Therefore, the management should carefully select a shortage penalty cost function that appropriately represents their preferences. All in all, our overview and our findings can supplement the development of future practical computational models on nursing service budgeting.

Future research, may address empirical modeling of the shortage and budget penalty functions. Using empirical penalty functions would make further numerical experiments with the SDNBOM interesting. Additionally, studying a set of real-life demand processes would be necessary to draw appropriate conclusions on the value of using forecast updates. Furthermore, it would be interesting to formulate a mixed integer program instead of the simple greedy approximation of Algorithm 2.

Acknowledgements Herewith we thank to Colin Ridley for sending the data that he collected in Ridley (2007) and for providing us with many additional details. We also thank Marion Rauner, and the anonym reviewers for their suggestions towards the improvement of this paper.

References

Abernathy, W. J., Baloff, N., Hershey, J. C., & Wandel, S. (1973). A three-stage manpower planning and scheduling model – a service-sector example. *Operations Research*, *21*(3), 693–711.

Arthur, T., & James, N. (1994). Determining nurse staffing levels, a critical review of the literature. *Journal of Advanced Nursing*, *19*, 558–565.

Atlason, J., Epelman, M. A., & Henderson, S. G. (2005). Optimizing call centers staffing using simulation and analytic center cutting plane methods. Tech. rept. 04-09. IOE Dept., University of Michigan.

Bard, J. F., & Purnomo, H. W. (2006). Incremental changes in the workforce to accommodate changes in demand. *Health Care Management Science*, *9*, 71–85.

Brusco, M. J., & Showalter, M. J. (1993). Constrained nurse staffing analysis. *Omega*, *21*, 175–186.

Burke, E. K., De Causmaecker, P., Vanden Berghe, G., & Van Landeghem, H. (2004). The state of the art of nurse rostering. *Journal of Scheduling*, *7*, 441–499.

Easton, F., Rossin, D., & Borders, W. (1992). Analysis of alternative scheduling policies for hospital nurses. *Production and Operations Management*, *1*(2), 159–174.

Flood, S. D., & Diers, D. (1988). Nurse staffing, patient outcome and cost. *Nursing Management*, *19*(5), 34–43.

Fox, B. (1966). Discrete optimization via marginal analysis. *Management Science, A*, *13*(3), 210–216.

Jeang, A. (1996). Flexible nursing staff planning with adjustable patient demands. *Journal of Medical Systems*, *20*(4), 173–182.

Kao, E. P. C., & Queyranne, M. (1985). Budgeting costs of nursing in a hospital. *Management Science*, *31*(5), 608–621.

Kao, E. P. C., & Tung, G. G. (1980). Forecasting demands for inpatient services in a large public health care delivery system. *Socio-Economic Planning Science*, *14*(5), 97–106.

Keeling, B. (1999). How to allocate the right staff mix across shifts. *Nursing Management*, *30*(9), 16.

Kirby, K. K., & Wiczai, L. J. (1985). Budgeting for variable staffing. *Nursing economic*, *3*(May–June), 160–166.

Li, Y., Chen, J., & Cai, X. (2007). An integrated staff-sizing approach considering feasibility of scheduling decision. *Annals of Operations Research*, *155*, 361–390.

Lowerre, J. M. (1979). On personnel budgeting for continuous operations (with emphasis on hospitals). *Decision Sciences*, *10*, 126–135.

Ridley, C. (2007). Relating nursing workload to quality of care in child and adolescent mental health inpatient services. *International Journal of Health Care*, *20*(5), 429–440.

Siferd, S. P., & Benton, W. C. (1992). Workforce staffing and scheduling: hospital nursing specific models. *European Journal of Operational Research*, *60*, 233–246.

Trivedi, V. M. (1981). A mixed-integer goal programming model for nursing service budgeting. *Operations Research*, *29*(5), 1019–1034.

Venkataraman, R., & Brusco, M. (1996). An integrated analysis of nurse staffing and scheduling policies. *Omega*, *24*(1), 57–71.

Warner, D. M., & Prawda, J. (1972). A mathematical programming model for scheduling nursing personnel in a hospital. *Management Science, Application Series, Part 1*, *19*(4), 411–422.

Zimmerman, J. L. (1976). Budget uncertainty and the allocation decision in a nonprofit organization. *Journal of Accounting Research*, *14*(2), 301–319.