J Sched (2007) 10:5–23
DOI 10.1007/s10951-006-0323-7

Cyclic preference scheduling of nurses using a Lagrangian-based heuristic

Jonathan F. Bard · Hadi W. Purnomo

Abstract This paper addresses the problem of developing cyclic schedules for nurses while taking into account the quality of individual rosters. In this context, quality is gauged by the absence of certain undesirable shift patterns. The problem is formulated as an integer program (IP) and then decomposed using Lagrangian relaxation. Two approaches were explored, the first based on the relaxation of the preference constraints and the second based on the relaxation of the demand constraints. A theoretical examination of the first approach indicated that it was not likely to yield good bounds. The second approach showed more promise and was subsequently used to develop a solution methodology that combined subgradient optimization, the bundle method, heuristics, and variable fixing. After the Lagrangian dual problem was solved, though, there was no obvious way to perform branch and bound when a duality gap existed between the lower bound and the best objective function value provided by an IP-based feasibility heuristic. This led to the introduction of a variable fixing scheme to speed convergence. The full algorithm was tested on data provided by a medium-size U.S. hospital. Computational results showed that in most cases, problem instances with up to 100 nurses and 20 rotational profiles could be solved to near-optimality in less than 20 min.

J. F. Bard (✉)
Graduate Program in Operations Research & Industrial Engineering, 1 University Station C2200,
The University of Texas, Austin, TX 78712-0292, USA
e-mail: jbard@mail.utexas.edu

H. W. Purnomo
American Airlines, AMR Corp Headquarter HDQ1, Mail Drop 5358, Fort Worth, TX 76155, USA
e-mail: hadiwaskito@yahoo.com

Keywords Cyclic scheduling · Preference scheduling · Nurse rostering · Lagrangian relaxation · Bundle method

1 Introduction

The demand for nurses in the United States has been outpacing the supply for more than a decade. The situation is now at the point where the rules for good practice are being stretched to the limit and patient care is being jeopardized (Spratley et al., 2000). The majority of researchers in the field argue that the work environment must make the career more attractive. Nurses have consistently identified dissatisfaction with schedules, inadequate investment in information technology, and a lack of opportunity to deploy their skills and competencies to improve patient care as their principal grievances (Kimball and O'Neil, 2002).

While technology alone cannot be expected to improve the quality of the work environment, a big step in that direction involves better human resources planning. As part of the effort to cope with shortages, many hospitals have adopted scheduling policies that give increased weight to the preferences and requests of their nursing staff, often at a considerable cost. The rationale for this accommodation is that more individual control over schedules will lead to higher morale, a more attractive work environment, increased flexibility to deal with personal matters, and ultimately, higher retention rates.

For a planning horizon that may extend up to 6 weeks, the primary goal of midterm scheduling is to generate a set of rosters for the nursing staff that specifies their work assignments. In the more traditional case, fixed patterns of days on and days off are established and the staff is rotated continuously through them. This is known as *cyclic scheduling* (Emmons, 1985; Howell, 1998; Millar and Kiragu, 1998).

At the other extreme is *self-scheduling* which uses a sign-up procedure (e.g., see Griesmer, 1993). In light of their eligibility and contractual obligations, nurses are asked to sign up for those shifts that they wish to work over the planning horizon. When violations or conflicts occur, the nurse manager tries to resolve them through consensus—often a difficult process.

A third approach, *preference scheduling*, applies a common set of rules and a cost measure that together are designed to achieve a balance between staff satisfaction and the use of outside resources (e.g., see Berrada et al., 1996; Burke et al., 1999; De Causmaecker and Vanden Berghe, 2003; Bard and Purnomo, 2005a). The rules (constraints) are hospital-dependent but generally may be categorized as either hard—must be satisfied, or soft—can be violated at a cost. A common approach is to rebuild the rosters from scratch at the beginning of the planning horizon starting from a template or a sign-up schedule. In its original conception, preference scheduling dealt mostly with individual requests such as specific days off (e.g., see Warner, 1976), but now includes issues related to the quality of the schedule as judged by the presence of undesirable work patterns.

The purpose of this paper is to offer management greater flexibility in constructing rosters by combining the principal components of cyclic and preference scheduling in a single model. To find solutions to the corresponding large-scale integer program (IP), it was necessary to develop a hybrid algorithm comprising both heuristic and exact procedures. The initial approach centered on a Lagrangian relaxation and the use of standard subgradient optimization to solve the Lagrangian dual. Slow convergence and several unusual properties of the relaxation led to the use of a bundle method coupled with a novel branching scheme. When integrated with an IP-based heuristic to find feasible solutions, the overall methodology proved to be effective in finding near-optimal solutions to problem instances with up to 100 nurses, often in less than 20 min.

In the next section, we give a brief review of the staff scheduling literature with an emphasis on nurse rostering. In Section 3, we define what is meant by a rotational profile and then formally state the cyclic preference scheduling problem. The IP models are presented in Section 4 followed by the details of the proposed solution methodology and implementation issues in Section 5. Computational results are highlighted in Section 6. We close with some remarks on the effectiveness of the approach. Several theoretical aspects of the relaxed model are addressed in Appendix A.

2 Literature review

A high proportion of hospital staffing costs are associated with nursing resources, so generating schedules that better match supply with demand can have a significant impact on the operating budget (Pierskalla and Brailer, 1994). In the last two decades, most of the research on nurse scheduling has concentrated on rostering with the goal of accommodating individual preferences such as requests for specific shifts or days off. Preference scheduling models embody work rules and some method of quantifying the violation of preference requests. A common way to do this is to assign penalties based on the severity of a violation (Warner, 1976; Jaumard et al., 1998).

One disadvantage of preference scheduling is its inherent inconsistency. Due to the implicit assumption of independence of consecutive planning horizons, nurses may have noticeably different shift assignments from week to week and month to month. This can be unsettling from a personal point of view. An alternative approach that has not yet been widely adopted due to the inherent difficulty in finding solutions is to include a cyclic feature into the problem along with the preferences. The cyclic period is typically half the length of the general planning horizon used in preference scheduling.

Early research on nurse scheduling was primarily aimed at developing efficient heuristics. Miller et al. (1976) were the first to formally address the preference scheduling problem. Starting with an initial solution, they developed a greedy neighborhood search procedure to find local optima. Howell (1998) solved the cyclic scheduling problem by combining intuitive information of what constitutes a good schedule with greedy exchange procedures. More recently, metaheuristics, such as tabu search, simulated annealing, and genetic algorithms, have been used to solve various midterm scheduling problems (Brusco and Jacobs, 1995; Dowsland, 1998; Nonobe and Ibaraki, 1998; Burke et al., 1999, 2004a; Aickelin and Dowsland, 2000, 2004; Kawanaka et al., 2001). A memetic approach (i.e., a genetic algorithm hybridized with a steepest descent heuristic) and a memetic/tabu search hybrid are discussed by Burke et al. (2001).

Nevertheless, it is often difficult for heuristics to cope with conflicting hard and soft constraints in a computationally efficient manner. Motivated by the need to balance solution quality with computational effort, De Causmaecker and Vanden Berghe (2003) showed how to combine metaheuristics and coverage relaxation algorithms to address practical concerns in a real scheduling environment.

Considering exact methods, there are two principal ways of formulating staff scheduling problems as IPs. The first is the pattern-view formulation and leads to a set-covering-type problem with a large number of columns. In these models, each column represents a particular scheduling pattern or roster consisting of a sequence of shifts and days-off assignments that span the planning horizon. The second formulation is based on the shift view of the problem and often contains a large number of rows, most associated with individual employee constraints. Each formulation has its

own advantages, but the underlying problem remains NP-hard (Lau, 1996).

Exact algorithms typically involve some form of decomposition or the use of cutting planes derived from polyhedral theory. In some cases, though, commercial software is sufficient to find good solutions; e.g., see Isken (2004) and Randhawa and Sitompul (1993). The column generation approach, an example of decomposition, uses the pattern-view formulation as a master problem and either heuristics or the shift-view formulation as subproblems to generate candidate rosters. In a call center application, Caprara et al. (2003) simplified the subproblems into network flow problems that were easily solved.

An easy way to overcome the formidable size of the set-covering formulation is to generate only a subset of columns at a time. Warner (1976) used 50 columns generated by a greedy algorithm to set up a problem with 20 nurses and a block pivoting strategy to find feasible solutions. A similar idea was pursued by Bard and Purnomo (2005a). Taking this a step further, Jaumard et al. (1998) developed a branch-and-price (B&P) algorithm for the preference scheduling problem. In B&P, a master problem is created that contains the demand constraints only. The hard and soft constraints are contained in a series of subproblems, one for each nurse, which are solved iteratively to generate columns for the master problem. Branch and bound is then used to achieve integrality. Their preliminary testing showed that instances with up to 41 nurses could be solved for 2-week blocks in about 16.5 h on a Sparc Sun 5 workstation.

Beginning with the shift-view formulation, Valouxis and Housos (2000) developed a hybrid approach that solved a simplified version of the original preference scheduling problem that was constructed by ignoring several difficult-to-model constraints. After finding a solution to the reduced IP, a local search heuristic based on tabu search was used to achieve feasibility. Alternatively, goal programming is a common methodology for dealing with soft constraints. In this approach, rules are prioritized and treated as goals or objectives to be satisfied. The optimization is carried out sequentially so to ensure that goal achievement is preserved; e.g., see Berrada et al. (1996) and Ferland et al. (2001). Topaloglu and Ozkarahan (2004) also solved a midterm scheduling problem that considered both preferences and cyclical factors.

Other methods, such as constraint programming (CP), have also been devised to solve nurse scheduling problems. CP is an artificial intelligent technique that, unlike integer programming, does not make use of an explicit mathematical representation of the hard and soft constraints but applies logic rules instead. Cheng et al. (1997) used 20 different constraint-type rules to construct rosters for up to 30 nurses. For more general staff scheduling problems that make use of CP, see Meyer auf'm Hofe (1997, 2001).

Nevertheless, capturing rostering knowledge in the form of logic rules as required by CP, is not without its challenges. Inadequate or inflexible constructs may produce questionable results. Case-based reasoning is a different artificial intelligence technique that aims to imitate human-style decision making through analogy. Previous problems and solutions are stored in a *case-base* and accessed by processes associated with identification, retrieval, adaptation, and storage. An example related to nurse rostering is given by Petrovic et al. (2002). For a survey of the nurse rostering literature, see Burke et al., 2004b; for a general bibliography on staff scheduling, see Ernst et al. (2004).

3 Problem statement

When the nursing staff is fixed, the objective of cyclic scheduling is to generate a set of rosters that minimizes the number of uncovered shifts over the planning horizon. To ensure fairness, the nurses are sequentially assigned to the optimal rosters on a 2-week rotating basis, although in some cases, a subset of the nurses may be given invariant assignments. When the demand changes sufficiently, the entire process is repeated. Hospitals in the United States typically employ nurses to work either 8- or 12-h shifts, giving rise to five standard shift types: three 8-h shifts called Day or D (7 a.m.–3 p.m.), Evening or E (3 p.m.–11 p.m.), Night or N (11:00 p.m.–7 a.m.) and two 12-h shifts called AM (7 a.m.–7 p.m.) and PM (7 p.m.–7 a.m.). An AM shift starts at the same time as a D shift and ends midway into an E shift. A similar interpretation exists for a PM shift. European hospitals use a similar shift structure but their lengths and overlaps may vary.

Over a 2-week planning period, a nurse may generally work up to two different shifts, such as D and E. Fundamental to the idea of cyclic scheduling in this context is the rotational profile.

Definition 1. A *rotational profile* for a nurse is defined by the triplet (eligible shifts, ratio, total hours).

The *ratio*, often expressed as a percentage, indicates the minimum number of eligible shifts of given type that must be assigned, while the *total hours* specifies the total number of working hours that must be assigned to the nurse every 2 weeks. For example, an E/N nurse with a 30% ratio whose contract calls for 80 h of work in 2 weeks (E/N, 30%, 80) must be assigned 10 shifts over this period. At least three of those shifts must be E and at least three N. A D/AM nurse with a 25% ratio who is contracted for 72 h of work in 2 weeks (D/AM, 25%, 72) must be assigned at least two D and two AM shifts. The possibilities are (3D, 4AM) and (6D,

2AM). A ratio of 100% simply means that the nurse does not rotate.

Problem inputs include the demand per shift, the number of nurses contracted for every rotational profile, work rules, preference violation penalties, and limits on the use of outside resources, such as float pool and agency nurses. Demand is specified as a lower and upper bound on the number of nurses needed per shift. Because of nationwide staff shortages, it would be unusual to be able to cover all demand. Therefore, it is assumed that gaps in the schedule will be filled with outside resources. Short-term scheduling addresses this issue.

Work rules are a function of contractual agreements, labor laws, hospital policies, and preference considerations. They can be classified as either hard or soft constraints. The hard constraints included here are:

a. All full-time nurses must be assigned either 72 or 80 h within a 2-week planning period, depending on their contract. When a nurse is assigned fewer hours than specified in the contract, she is still paid her full weekly salary. Cancellations and overtime are taken into account on a daily basis and are not part of our model (see Bard and Purnomo (2005b) for a discussion of the daily adjustment problem).
b. A nurse can only be assigned to the shifts that define her rotational profile.
c. The number of consecutive working days, also commonly called the *workstretch*, cannot exceed some value, call it D^{maxon}, which is always ≥ 2. For a nurse working 8-h shifts, this parameter is typically 5 or 6.
d. A nurse can work for at most 12 h in a day and can be assigned at most one shift per day. Also, there needs to be at least an 8-h break between consecutive assignments. Compliance is generally automatic but additional restrictions are required for those profiles in which back-to-back shifts (i.e., no break between two shifts on consecutive days) are possible. In particular, the sequences N/D, PM/D, N/AM, and PM/AM must be excluded from consideration.
e. Nurses must work two weekend shifts in the same weekend every 2 weeks. For our purposes, the first weekend shift starts at 7:00 p.m. on Friday and the last weekend shifts starts at 3 p.m. on Sunday. N and PM on Sunday are not considered weekend shifts.

The soft constraints included in the problem are:

f. Days-on and days-off patterns. There are two undesirable working patterns. The first is evidenced by 1 day off between 2 working days and is denoted by on-off-on or 1-0-1. The second is evidenced by a day on between 2 days off and is denoted by off-on-off or 0-1-0. It is more desirable to have at least 2 consecutive days off. In our implementation, nurses who only work 12-h shifts (AM, PM, or both) are not subject to the 1-0-1 and 0-1-0 soft constraints because they would lead to too many violations and hence most hospitals view them as too restrictive.
g. Different shift assignments on consecutive working days. This situation may occur when a rotational nurse is assigned to work a sequence such as D/E/D without an intervening day off. This type of pattern is highly undesirable because it disrupts the body's circadian rhythm.

Definition 2. An *optimal solution* to the cyclic preference scheduling problem is one that minimizes a weighted combination of preference violations and the number of outside resources subject to the hard constraints (a)–(e).

Definition 3. Let $V_{\max}$ be the maximum number of violations of the soft constraints permitted and let r_a be the *penalty coefficient* associated with $a \in [1, V_{\max}]$ violations. Then $r_a = 2^{a-1}$.

The rationale for the use of an exponential function in Definition 3 can be found in Bard and Purnomo (2005a).

4 Model formulation

Our model for cyclic preference scheduling takes the shift view and is written to include both the hard and soft constraints. Specific expressions are similar to those used by Valouxis and Houxos (2000) and others. The following notation is used in the remainder of the paper.

Indices and sets

i	index for nurses; $i \in N$
d	index for days; $d \in D$
a	index for the number of preference violations; $a = 1, \ldots, V_{\max}$
m	index for weekends in the 14-day planning period; $m \in W$
t	index for shifts; $t \in T$
$t_1(t_2)$	first (second) shift type for a rotational nurse
T_i	set of shift types that nurse i is hired to work
T	set of all possible shift types considered, $T = \bigcup_{i \in N} T_i = \{\text{D, E, N, AM, PM}\}$
D_{W}	set of weekend days in a 2-week period
W	set of weekends under consideration
N	set of nurses to be scheduled
N_{R}	set of nurses with nondegenrate rotational profiles (i.e., two possible shift types); $N_{\text{R}} \subseteq N$
N_{BB}	set of nurses with back-to-back rotational profiles (N/D, PM/D, N/AM, PM/AM); $N_{\text{BB}} \subseteq N_{\text{R}} \subseteq N$
D	set of days for which the model is to be solved; $\lvert D\rvert = 14$

Parameters

$V_{\max}$	maximum number of violations allowed for each nurse (=5 in the computations)
r_a	penalty assigned to a midterm schedule that has a violations (maximum value of $a = 5$ and $r_a = 2^{\alpha-1}$, so the maximum value of r_a in the computations is $2^{5-1} = 16$)
M	large number representing the cost of an outside nurse (undercoverage) in a period (=50 in the computations, which is approximately $3 \times$ (max value of r_a) $= 48$)
H_i	number of hours nurse i is contracted to work every 2 weeks (=72 or 80)
h_t	length of shift t (hours)
$LD_{dt}(UD_{dt})$	lower (upper) demand requirement for shift t on day d
${D_i}^{\max on}$	maximum number of consecutive days (workstretch) that nurse i is permitted to work
P_{it}	minimum number of shifts of type t that nurse i must work every 2 weeks
${W_i}^{\max}$	number of weekend shifts nurse i must work every 2 weeks
$TR_{\max}$	maximum number of shift transitions allowed on consecutive days during the 14-day planning horizon (=3 in the computations)
$O_{dt}^{\max}$	maximum number of outside nurses that can be assigned to shift t on day d

Decision variables

x_{idt}	(binary) 1 if nurse i works shift t on day d, 0 otherwise
w_{im}	(binary) 1 if nurse i works on weekend m, 0 otherwise
v_{ia}	(binary) 1 if nurse i has a violations in his or her midterm schedule, 0 otherwise
b_{id}	(accounting) 1 if nurse $i \in N_{\mathrm{R}}$ works shift t_1 on day d and shift t_2 on day $d+1$, 0 otherwise; $t_1 \neq t_2$
p_{id}	(accounting) 1 when nurse i has a 0-1-0 pattern that starts on day d, 0 otherwise
q_{id}	(accounting) 1 when nurse i has a 1-0-1 pattern that starts on day d; 0 otherwise
y_{dt}	number of outside nurses assigned to shift t on day d
s_{dt}	excess number of nurses assigned to shift t on day d

$$\theta_{\mathrm{IP}} = \text{Minimize} \sum_{i \in N} \sum_{a=1}^{V_{\max}} r_a v_{ia} + M \sum_{d \in D} \sum_{t \in T} y_{dt} \tag{1a}$$

$$\text{subject to} \sum_{i \in N} x_{idt} - s_{dt} + y_{dt} = \mathrm{LD}_{dt}, \quad d \in D, t \in T \tag{1b}$$

$$\sum_{d \in D} x_{idt} \geq P_{it}, \quad i \in N_R, t \in T_i \tag{1c}$$

$$\sum_{d \in D} \sum_{t \in T_i} h_i x_{idt} = H_i, \quad i \in N \tag{1d}$$

$$\sum_{t \in T_i} x_{idt} \leq 1, \quad i \in N, d \in D \tag{1e}$$

$$x_{idt_2} + x_{i,d+1,t_1} \leq 1, \quad i \in N_{\mathrm{BB}}, d \in D \tag{1f}$$

$$\sum_{l=d}^{d+D_i^{\max on}} \sum_{t \in T_i} x_{ilt} \leq D_i^{\max on}, \quad i \in N, d \in D \tag{1g}$$

$$\sum_{d \in D_W} \sum_{t \in T_i} x_{idt} = W_i^{\max} w_{im}, \quad i \in N, m \in W \tag{1h}$$

$$\sum_{m \in W} w_{im} = 1, \quad i \in N \tag{1i}$$

$$\sum_{t \in T_i} x_{idt} + \left(1 - \sum_{t \in T_i} x_{i,d+1,t}\right) + \sum_{t \in T_i} x_{i,d+2,t} + p_{id} \geq 1, \quad i \in N, d \in D \tag{1j}$$

$$\left(1 - \sum_{t \in T_i} x_{idt}\right) + \sum_{t \in T_i} x_{i,d+1,t} + \left(1 - \sum_{t \in T_i} x_{i,d+2,t}\right) + q_{id} \geq 1, \quad i \in N, d \in D \tag{1k}$$

$$1 - x_{idt_\alpha} + 1 - x_{i,d+1,t_\beta} + b_{id} \geq 1, \quad i \in N_{\mathrm{R}}, d \in D, \alpha \neq \beta \in \{1, 2\} \tag{1l}$$

$$\sum_{d \in D} b_{id} \leq TR_{\max}, \quad i \in N_{\mathrm{R}} \tag{1m}$$

$$\sum_{d \in D} (p_{id} + q_{id} + b_{id}) = \sum_{a=1}^{V_{\max}} a v_{ia}, \quad i \in N \tag{1n}$$

$$\sum_{a=1}^{V_{\max}} v_{ia} \leq 1, \quad i \in N \tag{1o}$$

$$0 \leq s_{dt} \leq UD_{dt} - LD_{dt}, \quad 0 \leq y_{dt} \leq O_{dt}^{\max}, \forall t, d \tag{1p}$$

$$b_{id}, p_{id}, q_{id} \in [0, 1], \quad \forall i, d; v_{ia} \in \{0, 1\}, \forall i, a; w_{im} \in \{0, 1\}, \forall i, m \tag{1q}$$

$$x_{idt} \in \{0, 1\}, \quad \forall i, t, d, \text{ where } x_{i,14+l,t} \equiv x_{ilt}, \; l = 1, \ldots, D_i^{\max on} \tag{1r}$$

The objective function, represented by Eq. (1a), is the weighted sum of preference violations and the cost of covering gaps with outside nurses. The choice of the parameter M implicitly defines the tradeoff between satisfying the collective preferences of the nurses and incurring additional costs by allowing for shortages. In general, $M \gg$ the penalty coefficient r_a. Equation (1b) corresponds to the demand

requirement for each shift t on day d and represents a transformation from a two-sided inequality into a single equality constraint with an upper bound on the slack variable s_{dt}, as indicated in Eq. (1p). Because an optimal solution will always exist with s_{dt} and y_{dt} integral, they can be treated as continuous variables. Note that some authors, such as Jaumard et al. (1998), express demand in terms of periods rather than shifts. The conversion of one to the other is straightforward.

The remaining constraints are written for each nurse i. The constraint, represented by Eq. (1c), guarantees that at least P_{it} shifts of type t are assigned every 2 weeks for $i \in N_R$, where P_{it} is determined from the ratio percentage. For nurses with a single shift profile; i.e., $i \in N \backslash N_R$, Eq. (1c) can be removed. Equation (1d) states that the total number of hours assigned to nurse i must be equal to the number of hours H_i, that she is contractually obligated to work every 2 weeks.

The constraint, represented by Eq. (1e), restricts a nurse to at most one shift assignment within 24 h. Because the length of a shift is at most 12 h, constraints (1c)–(1e) automatically ensure an 8-h break between shifts for nurses with rotational profiles except for the back-to-back cases mentioned in the description of hard constraint (d). These cases are handled by the constraint, represented by Eq. (1f), which permits only one assignment of either an N or PM shift (t_2) on day d, or a D or AM shift (t_1) on day $d + 1$.

The constraint, represented by Eq. (1g), limits the work-stretch of nurse i to no more than D_i^{maxon} days in any time window of $D_i^{\text{maxon}} + 1$ consecutive days. This corresponds to rule (c). In the implementation, the parameter D_i^{maxon} was set to 5 for nurses who work for 8-h shifts only and 4 for nurses who work both 8- and 12-h shifts. Because the problem is cyclic, day 14 is followed by day 1. This is indicated in Eq. (1r). The weekend rule (e) is modeled by constraints (1h)–(1i). Weekends are defined by N and PM shifts for Friday, D, E, N, AM, and PM shifts for Saturday, and D, E, and AM shifts for Sunday. Together, these constraints require that nurse i work exactly $W_i^{\max}$ weekend days every 2 weeks. Although the days must fall on the same weekend, it is an easy matter to allow split weekends. Note that the value of $W_i^{\max}$ is a function of the rotational profile. In our implementation, if nurse i works only 12-h shifts, then she will be assigned just one weekend day ($W_i^{\max} = 1$) every 2 weeks; otherwise, $W_i^{\max} = 2$.

The constraints, represented by Eqs. (1j)–(1o), determine the quality of the rosters. The undesirable patterns are counted in the model by the variables p_{id}, q_{id}, and b_{id}. A 0-1-0 pattern starting on day d implies that $\sum_{t \in T_i} x_{idt} = 0$, $\sum_{t \in T_i} x_{i,d+1,t} = 1$, and $\sum_{t \in T_i} x_{i,d+2,t} = 0$. The constraint, represented by Eq. (1j), sets $p_{id} = 1$ when such a pattern exists. Because all of the other variables in the constraint are binary and all of the data are integral, p_{id} will always be integral in an optimal solution so it can be treated as a continuous variable. The constraint, represented by Eq. (1k), is the corresponding constraint for 1-0-1 patterns, which detect the existence of $\sum_{t \in T_i} x_{idt} = 1$, $\sum_{t \in T_i} x_{i,d+1,t} = 0$, and $\sum_{t \in T_i} x_{i,d+2,t} = 1$ starting on day d. The total number of these patterns for nurse i is given by the summation $\sum_{d \in D}(p_{id} + q_{id})$. Implicit in the formulation is that a roster is a circulation so that in Eqs. (1j) and (1k), day $14 + d =$ day d.

The constraint, represented by Eq. (1l), detects a shift transition during consecutive days, and must be included for every possible combination of shift transitions that nurse i may have. The maximum number permitted is given by the parameter $TR_{\max}$, as indicated in Eq. (1m). The constraint, represented by Eq. (1n), counts the number of preference violations and Eq. (1o) determine which penalty coefficient r_a will be in effect. If it were desirable to account for the severity of each violation, we could do this by multiplying each variable in Eq. (1n) by the appropriate weight.

Although shifts shorter than 8 h are not included in the model given by Eqs. (1a)–(1r), Definition 1 is broad enough to allow shifts of any length and total working hours that would reflect the use of part-timers. Also not included are grade considerations and the use of higher skilled workers to fill in for lower skilled workers when there is idle time in their schedules. Not accounting for seniority and individual skills opens up the possibility that the model might produce schedules in which some shifts are staffed by inexperienced nurses only, an undesirable situation. Our approach to dealing with this imbalance is called *downgrading* and was addressed in earlier work (see Bard and Purnomo, 2005c). The downgrading component of the model was omitted to avoid unnecessary notation.

The final point that requires clarification is the way annual leave, training days, requests, and other departures from a fixed cycle can be accommodated by the model. In each of these cases, the first step, say for nurse i, is to reset both the ratio parameter P_{it} in Eq. (1c) and the total working hours parameter H_i in Eq. (1d) to reflect the new scheduling constraints. This is equivalent to adjusting nurse i's rotational profile for the upcoming planning period. The next step is to remove the variables x_{idt} from the model that correspond to shifts no longer permitted in a solution due to the reduced work period. It may also be necessary to modify the weekend constraints (1h)–(1i), depending on the particular situation.

4.1 Problem difficulty and LP relaxation

The size of model (1a)–(1r) is determined largely by the accounting constraints, represented by Eqs. (1j)–(1l). While the other constraints only grow linearly with the number of nurses in a unit, these constraints grow at a rate proportional to $O(|N| \cdot |D|)$. The number of variables grows at a rate proportional to $O(|N| \cdot |D| \cdot |T|)$. A small problem with 20 nurses and a 2-week planning horizon contains roughly 1500

variables and 700 constraints. A medium-size problem for the same planning horizon with 50 nurses requires about 5000 variables and 3500 constraints.

Attempts to solve several instances of (1) with CPLEX 7.5 proved frustrating. The best results obtained within a 4-h time limit had a 3.3% optimality gap. Starting with values as high as 50%, the optimality gap decreased sharply at first, but then failed to show much improvement as the search tree grew. In fact, the best lower bound provided by CPLEX never differed from the LP solution obtained at the root node even after hours of computations and the generation of numerous cuts along the way. As more nodes were explored, solutions with fewer violations of the soft constraints, represented by Eqs. (1j)–(1l), were found, rather than solutions with fewer outside nurses.

The nature of these results was not unexpected due to the relative weights of the coefficients in Eq. (1a). In all instances, the LP solution attained the minimum number of outside nurses possible. Somewhat surprisingly, though, the best IP solutions found by CPLEX also attained the minimum number of outside nurses, but this was likely due to the characteristics of the data rather than a general principle. Also, all LP solutions at the root node had zero values for the days on and days off accounting variables p_{id} and q_{id} and few nonzero values for the switching variables b_{id}. At subsequent nodes, the unchanging lower bounds were a consequence of the infinite possibilities of generating fractional solutions over the full set of binary variables.

In general, when artificially weighted objective function terms are present in a problem, large optimality gap reductions may result from incremental reductions in the more heavily weighted terms. In our case, eliminating a single outside nurse yielded a sharp decrease in the percentage gap. For problems with single shift lengths, we have the result given below, which suggests that the LP relaxation of model (1a)–(1r) cannot be relied upon to provide a tight lower bound on the optimum.

Proposition 1. *At optimality, the second term in the objective function* (1a)*,* $\sum_{d\in D}\sum_{t\in T} y_{dt}$*,*

1. *always achieves its minimum value, and*
2. *is always integral in the relaxed LP solution to the model, represented by Eq.* (1)*, when all shifts are of the same length.*

Proof: Part 1 follows from the fact that the objective function coefficient M is arbitrarily large. For part 2, we note that when a nurse can only be assigned to shifts of the same length, Eq. (1d) can be written as $\sum_{d\in D}\sum_{t\in T_i} x_{idt} = H_i/h_t, i \in N$, where h_t is constant. The right-hand side of this equation, H_i/h_t, is integral by definition or else there would be no feasible solution. Summing over i gives $\sum_{i\in N}\sum_{d\in D}\sum_{t\in T_i} x_{idt} = \frac{1}{h_t}\sum_{i\in N} H_i$, which is still integral.

Next, we sum the equalities in Eq. (1b), over d and t to get

$$\sum_{d\in D}\sum_{t\in T}\left(-s_{dt} + \sum_{i\in N} x_{idt} + y_{dt}\right) = \sum_{d\in D}\sum_{t\in T} LD_{dt}$$

or

$$\sum_{d\in D}\sum_{t\in T}(-s_{dt} + y_{dt}) + \sum_{i\in N}\sum_{d\in D}\sum_{t\in T} x_{idt} = \sum_{d\in D}\sum_{t\in T} LD_{dt}$$

The second term on the left-hand side is integral because the summation over $t \in T$ can be replaced by the summation over $t \in T_i$, which was shown to be integral. Because demand data LD_{dt}, are integral, the first term on the left-hand side, $\sum_{d\in D}\sum_{t\in T}(-s_{dt} + y_{dt})$, is also integral.

The fact that we wish to minimize the number of outside nurses in objective function, represented by Eq. (1a), coupled with the demand constraint, represented by Eq. (1b), implies that $S_{dt} \times y_{dt} = 0$ for all $d \in D, t \in T$. Therefore, if $\sum_{d\in D}\sum_{t\in T} y_{dt}$ is not integral, there is at least one day d and shift t for which y_{dt} is fractional and at least one nurse i for which $x_{idt} \in (0, 1)$. The latter assertion follows from the integrality of LD_{dt} in Eq. (1b) and the fact that $S_{dt} = 0$. Now, let the specific instance be $d = d_1, t = t_1$, and $i = i_1$ and let $f_{d_1t_1} = y_{d_1t_1} = \lfloor y_{d_1t_1} \rfloor$ and $x_{i_1d_1t_1} = 1 - f_{d_1t_1}$, where $\lfloor\varphi\rfloor$ is the largest integer less than or equal to φ. If we make the current values of $y_{d_1t_1}$ and $x_{i_1d_1t_1}$ integral by putting $y_{d_1t_1} \leftarrow \lfloor y_{d_1t_1} \rfloor$ and $x_{i_1d_1t_1} \leftarrow 1$, and adjust the values of $(x_{idt}, w_{im}, v_{ia}, p_{id}, q_{id}, b_{id})$ as necessary in constraints (1c)–(1o) to obtain a new feasible solution, we then get a smaller objective function value. This follows because M is arbitrarily large. As such, it doesn't matter whether the first term in Eq. (1a) increases as a result of the adjustment.

A check of constraints (1c)–(1o) indicates that the proposed marginal adjustment in the decision variables will produce a feasible solution with no increase in the other y_{dt} variables. If more than one nurse has a fractional value in the term $\sum_{i\in N} x_{idt}$ in Eq. (1b), similar arguments can be used to find a new feasible solution with an improved objective function value. As a consequence, $\sum_{d\in D}\sum_{t\in T} y_{dt}$, must be integral or we cannot claim to have found the optimal LP solution to model (1). □

Empirically, we observed that Proposition 1 held for all instances regardless of shift lengths, and that the value of $\sum_{d\in D}\sum_{t\in T} y_{dt}$ in the best IP solutions obtained always matched the value in the corresponding LP solutions.

4.2 Lagrangian relaxation and dual bounds

Further investigation of model (1a)–(1r) revealed when either the demand constraints Eq. (1b) or the pattern constraints Eqs. (1j) and (1k) were removed, CPLEX could easily find the optimum. In the first case, the remaining constraints define the feasible rosters for each nurse i. In the second case, the remaining constraints define rosters that more closely reflect pure cyclic scheduling without preference considerations.

When a tractable IP results after some constraints are removed from a problem, the use of Lagrangian relaxation (LR) to find bounds is indicated. For the two promising cases identified above, it is easier to construct feasible solutions to the full problem when the pattern constraints, represented by Eqs. (1j) and (1k), are relaxed and the corresponding IP is solved to get $\hat{x}_{idt}$. Once $\hat{x}_{idt}$ is found, it is straightforward to calculate the values of p_{id} in Eq. (1j) and q_{id} in Eq. (1k) to obtain a feasible solution. All that remains is to update the first objective function term, $\sum_{i\in N}\sum_{a=1}^{V_{\max}} r_a v_{ia}$, to account for the additional violations.

To formulate the relaxed problem, let $\boldsymbol{\lambda} \in \Re_+^{\rho}$ and $\boldsymbol{\mu} \in \Re_+^{\rho}$ be the multipliers associated with the constraints, represented by Eqs. (1j) and (1k), respectively, where $\rho = |N| \times |D|$, and augment Eq. (1a) as follows:

$$\theta_{\mathrm{LR}}(\boldsymbol{\lambda}, \boldsymbol{\mu})$$

$$= \text{Minimize} \quad \sum_{i\in N}\sum_{a=1}^{V_{\max}} r_a v_{ia} + M \sum_{d\in D}\sum_{t\in T} y_{dt} \tag{2a}$$

$$-\sum_{i\in N}\sum_{d\in D}\lambda_{id}\left(\sum_{t\in T_i} x_{idt} - \sum_{t\in T_i} x_{i,d+1,t} + \sum_{t\in T_i} x_{i,d+2,t} + p_{id}\right) \tag{2b}$$

$$-\sum_{i\in N}\sum_{d\in D}\mu_{id}\left(-\sum_{t\in T_i} x_{idt} + \sum_{t\in T_i} x_{i,d+1,t} - \sum_{t\in T_i} x_{i,d+2,t} + q_{id} + 1\right) \tag{2c}$$

$$\text{subject to Eqs. (1b)–(1i), Eqs. (1l)–(1r)} \tag{2d}$$

For fixed values of $\boldsymbol{\lambda} \equiv (\lambda_{id})$ and $\boldsymbol{\mu} \equiv (\mu_{id})$, it is well known that $\theta_{\mathrm{LR}}(\boldsymbol{\lambda}, \boldsymbol{\mu}) \le \theta_{\mathrm{IP}}$, so the goal is to find the values of $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ that maximize the objective function in Eq. (2). This leads to the Lagrangian dual (LD) problem that can be stated as follows:

$$\theta_{\mathrm{LD}} = \max_{\boldsymbol{\lambda},\boldsymbol{\mu}\ge 0} \theta_{\mathrm{LR}}(\boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{3}$$

It is also well known that $\theta_{\mathrm{LD}} \ge \theta_{\mathrm{LP}}$, so for a particular relaxation we would like to determine whether a strict inequality holds; in other words, whether the LD bound is better than the LP bound.

In our initial testing, we found that the optimal multipliers $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^*$ were always zero, a result that is established in Appendix A. Unfortunately, the lower bound on the original objective function provided by the corresponding solution was not very tight because two of the three preference constraints do not play a role in the problem. As an alternative, we propose to relax the demand constraints, represented by Eq. (1b), only. This gives two advantages: first it allows the remaining constraints to be decomposed by nurse i, and second, because all nurses with identical rotational profiles are subject to the same constraints, aggregation is possible. This means that the relaxed problem can be stated in terms of profiles rather than nurses.

If we now let $\boldsymbol{\mu} \equiv (\mu_{dt})$ be the Lagrange multipliers for the demand constraints, represented by Eq. (1b), $j \in N^p$ be the index for rotational profiles, n_j^R be the number of nurses with profile j, then the corresponding model is

$$\theta_{\mathrm{LR}}(\boldsymbol{\mu}) = \text{Minimize} \quad \sum_{j\in N^p} n_j^R \theta_j^{SP} + \sum_{d\in D}\sum_{t\in T}\mu_{dt} S_{dt} + \sum_{d\in D}\sum_{t\in T}(M - \mu_{dt}) y_{dt} + \sum_{d\in D}\sum_{t\in T}\mu_{dt} LD_{dt} \tag{4a}$$

$$\text{subject to} \quad 0 \le s_{dt} \le UD_{dt} - LD_{dt}, \quad 0 \le y_{dt} \le O_{dt}^{\max}, \forall d, t \tag{4b}$$

Subproblem j

$$\theta_j^{SP}(\boldsymbol{\mu}) = \text{Minimize} \quad \sum_{a=1}^{V_{\max}} r_a v_{ja} - \sum_{d\in D}\sum_{t\in T}\mu_{dt} x_{jdt} \tag{4c}$$

$$\text{subject to} \quad \text{Eqs. (1c)–(1o), Eqs. (1q)–(1r)} \tag{4d}$$

where the index i in the model, represented by Eq. (1), is replaced by the index j here. In the next section, we present our solution approach to Eq. (4).

5 Solution methodology

Given a multiplier vector $\boldsymbol{\mu}$, the value of $\theta_{\mathrm{LR}}(\boldsymbol{\mu})$ in Eq. (4a) can be easily computed by solving the $|N^p|$ subproblems, represented by Eqs. (4c)–(4d) whose objective function values $\theta_j^{\mathrm{SP}}(\boldsymbol{\mu})$ define the first term in Eq. (4a). The remaining terms are a function of the slack and gap variables, s_{dt}, and y_{dt}, respectively. Because these variables have bound constraints only as indicated in Eq. (4b), their optimal values can be determined by inspection. In particular,

- when $\mu_{dt} > 0$, set $s_{dt} = 0$; otherwise, set $s_{dt} = UD_{dt} - LD_{dt}$
- when $M - \mu_{dt} \ge 0$, set $y_{dt} = 0$; otherwise, set $y_{dt} = O_{dt}^{\max}$

To find the best bound on the original IP, represented by model (1), the Lagrangian dual, $\theta_{\mathrm{LD}} = \max_{\mu} \theta_{\mathrm{LR}}(\boldsymbol{\mu})$, must be solved. Our LD solution strategy is to first run a standard subgradient dual ascent algorithm (Nemhauser and Wolsey, 1988) and then switch to a bundle method (Lemarechal, 1989) once a sufficient number of subgradients have been identified. With some frustration, we discovered that once the Lagrangian dual was solved, there was no obvious way to perform branch and bound because there were no fractional variables. Branching on violations of the relaxed demand constraints was not possible either because undercoverage is allowed. As a consequence, we developed a heuristic branching strategy to improve the lower bound and speed convergence.

In our general experience, solving the Lagrangian dual rarely if ever yields an optimal solution, or even a feasible solution, of the original problem. For the LR approach to be effective then, a separate heuristic is needed to convert relaxed solutions obtained from model (4) to feasible solutions of model (1). These solutions provide an upper bound on θ_{IP} and help to reduce the size of the search tree once branching begins. In the following subsections, we describe the bundle method, the upper bound heuristic, and the variable fixing procedure that substitutes for branch and bound.

5.1 Bundle method

A basic criticism of the subgradient algorithm is that it fails to make use of all but the most recent information. A second criticism is that the subgradients obtained from the relaxed constraints in Eq. (4a) may not provide improving directions. As a consequence, $\theta_{\mathrm{LR}}(\boldsymbol{\mu}^k)$ is not monotone increasing.

An alternative strategy for updating the multipliers is to use what is called a *bundle* of past subgradients denoted by B. Theoretically, there are likely to be an infinite number of subgradients in the subdifferential of $\theta_{\mathrm{LR}}(\boldsymbol{\mu}^k)$ of which only a subset provide an improving direction. The idea is to use the bundle to construct an approximation of the subdifferential to obtain a more promising direction. In this approach, the new subgradient is defined as a convex combination of all subgradients in the current bundle; i.e., $\{\mathbf{g}^i : i \in B\}$. To find the convex multipliers, which we call λ_i, we need to solve the following quadratic program (QP) at iteration k:

$$\theta^k_{\mathrm{BD}} = \text{Minimize} \quad 0.5\tau_k \left\| \sum_{i\in B} \mathbf{g}^i \lambda_i \right\|^2 + \sum_{i\in B} \alpha^k_i \lambda_i \tag{5a}$$

$$\text{subject to} \quad \sum_{i\in B} \lambda_i = 1 \tag{5b}$$

$$\lambda_i \geq 0, \quad \forall i \in B \tag{5c}$$

where τ_k is the step size and α^k_i is a linearization error factor associated with subgradient i. The calculation of α^k_i is discussed below.

The quadratic term in Eq. (5a) is derived from the formulation that gives the steepest descent direction for the subgradients in the bundle (see Lemarechal, 1989). The second term in Eq. (5a) is a linearization error factor. Early applications had a constraint of the form $\sum_{i\in B} \alpha^k_i \lambda_i \leq \Delta_k$ instead, but this required a dynamic modification of the upper bound error Δ_k, which proved too unwieldy. After solving model (5) to get $\boldsymbol{\lambda}^*$, the new subgradient is $\mathbf{g}^{\mathrm{bundle}} = \sum_{i\in B} \mathbf{g}^i \lambda^*_i$.

The step-size parameter τ_k is initially set to 1 in our implementation. Although it can be held fixed throughout the algorithm, a more common approach is to update it based on a trust region strategy which is tied to the occurrence of taking either a null-step (NS) or a serious step (SS) at the current iteration. An SS is performed when the new subgradient gives a significant improvement, as determined by the following inequality:

$$\theta_{\mathrm{LR}}(\boldsymbol{\mu}^{k+1}) - \theta_{\mathrm{LR}}(\boldsymbol{\mu}^k) \geq m_1 \theta^k_{\mathrm{BD}} \tag{6}$$

where m_1 is the trust-region parameter whose value is typically set to 0.1 (Crainic et al., 2001).

The left-hand side of Eq. (6) indicates the change in the relaxed solution between successive iterations; the right-hand side is the current threshold value.

In the bundle method, the standard multiplier updating formula is only used when an SS is taken. When Eq. (6) is not satisfied, an NS is taken, and although the multipliers are not updated, the subgradient obtained from solving QP, represented by model (5), is stored as part of the bundle.

Testing in general has shown that too large a step-size τ_k will cause too many consecutive null steps to be taken between serious steps. This is equivalent to a long drought of nonimproving steps. In contrast, when τ_k is too small, many serious steps will be taken but each will yield only minimal improvement. With this in mind, we use the following formulas to either increase or decrease the step size (Frangioni and Gallo, 1999).

Increase step:

$$\tau_{k+1} = \max\left\{\tau_k, \min\left\{t_M, M\tau_k, \right.\right.$$
$$\left.\left. 2\tau_k\theta^k_{\mathrm{BD}}\left(\theta^k_{\mathrm{BD}} - (\theta_{\mathrm{LR}}(\boldsymbol{\mu}^k) - \theta_{\mathrm{LR}}(\boldsymbol{\mu}^{k-1}))\right)\right\}\right\} \tag{7a}$$

Decrease step:

$$\tau_{k+1} = \min\left\{\tau_k, \max\left\{t_m, m\tau_k, \right.\right.$$
$$\left.\left.\left(\sum_{i\in B} \alpha_i + \theta_{\mathrm{LR}}(\boldsymbol{\mu}^k) - \theta_{\mathrm{LR}}(\boldsymbol{\mu}^{k-1})\right) \Big/ 2\sum_{i\in B} \alpha^k_i \right\}\right\} \tag{7b}$$

In (7), t_m and t_M, m and M are parameters that are determined empirically. In our implementation, we used $t_m = 0.01$, $t_M = 100$, $m = 0.4$ and $M = 3.5$.

The presence of the linear term in Eq. (5a) is to ensure that the less accurate subgradients play a lesser role in determining the search direction. At iteration k, α_i^k is computed for all $i \in B$ as follows:

$$\alpha_i^k = \theta_{\mathrm{LR}}(\boldsymbol{\mu}^k) - \theta_{\mathrm{LR}}(\boldsymbol{\mu}^i) - (\boldsymbol{\mu}^k - \boldsymbol{\mu}^i)\mathbf{g}^i \tag{8}$$

The right-hand side of Eq. (8) is similar to a first-order Taylor series expansion of $\theta_{\mathrm{LR}}(\boldsymbol{\mu})$ around $\boldsymbol{\mu}^i$. Frangioni and Gallo (1999) suggested that the aggregate error $\sum_{i\in B} \alpha_i^k$ can be used to help in the determination of SS and NS. For a parameter $m_2 > 0$, when $\alpha_i^k \le m_2 \sum_{i\in B} \alpha_i^k$ is not satisfied, he recommends that the step size be decreased using the formula in Eq. (7b). Typically, m_2 is set to 0.9. In our case, testing showed that this rule was too restrictive and hence omitted from our algorithm.

Bundle_Algorithm

Input:	Current multipliers $\boldsymbol{\mu}^k = \{\mu_{dt}^k, d \in D, t \in T\}$, bundle B, maximum size of bundle $B_{\max}$, subgradient parameter $\{m_1\}$, step-size parameters $\{t_m, t_M, m, M\}$
Output:	New multiplier $\boldsymbol{\mu}^{k+1}$
Step 1:	Solving QP (5a)–(5c) to get $(\boldsymbol{\lambda}^*, \theta_{\mathrm{BD}})$ and a new subgradient $\mathbf{g}^{\mathrm{bundle}}$.
	If $(\lvert B\rvert = B_{\max})$ then \ check size of bundle
	$\omega = \arg\min\{\lambda_i : i \in B\}$
	$B \leftarrow B \setminus \{\mathbf{g}^\omega\}$
	Let $\mu^{\mathrm{temp}} = \mu^k + \tau_k \mathbf{g}^{\mathrm{bundle}}$ \ find temporary multipliers
	Solve (4) with $\boldsymbol{\mu}^{\mathrm{temp}}$ to get $(x_{idt}^{\mathrm{temp}}, s_{dt}^{\mathrm{temp}}, y_{dt}^{\mathrm{temp}})$ and $\theta_{\mathrm{LR}}(\boldsymbol{\mu}^{\mathrm{temp}})$
	Compute subgradient $\mathbf{g}^k = \left(LD_{dt} - \sum_{i\in N} x_{idt}^{\mathrm{temp}} + s_{dt}^{\mathrm{temp}} - y_{dt}^{\mathrm{temp}}\right)$
Step 2:	Insert $\mathbf{g}^k$ into bundle: $B \leftarrow B \cup \{\mathbf{g}^k\}$.
Step 3:	If the condition given by Eq. (6) is satisfied, then
	serious step (SS): update multipliers, $\boldsymbol{\mu}^{k+1} = \boldsymbol{\mu}^{\mathrm{temp}}$, and
	increase step size τ based on Eq. (7a).
Else	
	null step (NS): keep current multipliers as base, $\boldsymbol{\mu}^{k+1} = \boldsymbol{\mu}^k$, and
	decrease step-size τ based on (7b).

Both the subgradient and bundle algorithms are known to exhibit slow convergence in the tail so they are usually terminated when no improvement in the objective function value is observed in some predetermined number of iterations. At that point, branching is initiated. We take a similar approach based empirically on the performance of these algorithms on the problem, represented by model (4). The specific logic is discussed at the end of the section.

5.2 Feasibility IP heuristic

Because Lagrangian relaxation algorithms only provide lower bounds, an efficient heuristic is needed to construct feasible solutions to the original problem. Early testing indicated that 5–6 s were required on average to perform one iteration of the subgradient algorithm for 100-nurse instances, and that LD required about 120 iterations or 30 min to converge. As expected, none of the solutions to model (4) was feasible to model (1) so upon termination, only a lower bound on θ_{IP} was available.

To construct feasible solutions, we developed a second IP model that makes use of intermediate solutions of the Lagrangian relaxation problem, represented by model (4). As mentioned, the IP solution to each subproblem $j \in N^P$ for a given multiplier value $\boldsymbol{\mu}$ is a roster that satisfies all the hard constraints. From this observation, we formulated a set-covering-type IP to represent the demand constraints with rosters as columns. The gap and slack variable bound constraints were also included along with the requirement that each nurse be given a schedule. This formulation corresponds to the pattern view of the nurse scheduling problem. To ensure quick solutions, the number of columns per nurse was limited to 20.

To translate the demand constraint, represented by Eq. (1b), from the constraint-based, shift-view formulation to a set-covering-type formulation, we need to introduce some additional notation. Let $K(j)$ be a subset of feasible rosters,

for rotational profile j, $\zeta_{j\kappa}$ a nonnegative decision variable indicating the number of nurses with rotational profile j who are assigned to roster κ in the IP heuristic, $c_{j\kappa}$ be the penalty cost of rotational profile j when roster κ is assigned, X^{κ}_{jdt} be a parameter derived from the solution of the decomposed LR subproblems that is equal to 1 if roster κ for rotational profile j covers shift t on day d and 0 otherwise. The model used to find feasible solutions is

$$\theta_{\mathrm{HR}} = \text{Minimize} \quad \sum_{j\in N^P}\sum_{\kappa\in K(j)} c_{j\kappa}\zeta_{j\kappa} + M\sum_{d\in D}\sum_{t\in T} y_{dt} \tag{9a}$$

$$\text{subject to} \quad \sum_{j\in N^P}\sum_{\kappa\in K(j)} X^{\kappa}_{jdt}\zeta_{j\kappa} - s_{dt} + y_{dt} = LD_{dt}, \quad d\in D, t\in T \tag{9b}$$

$$\sum_{\kappa\in K(j)} \zeta_{j\kappa} = n^R_j, \quad j\in N^P \tag{9c}$$

$$0\le s_{dt}\le UD_{dt}-LD_{dt}, \quad 0\le y_{dt}\le O^{\max}_{dt}, \quad \forall d,t;\zeta_{jk}\ge 0 \text{ and integer}, \forall j,k \tag{9d}$$

The objective function in Eq. (9a) represents the cost of a schedule. The coefficients $c_{j\kappa}$ can be computed for each rotational profile j once roster κ is specified. The constraint, represented by Eq. (9b), is the equivalent of Eq. (lb) and Eq. (9c) is a generalization of the assignment constraint. To complete the formulation, bounds on the slack and gap variables are introduced in Eq. (9d), as in Eq. (1p).

It is interesting to note that the LP relaxation of model (9) produces dual variables, call it vector $\boldsymbol{\pi}$, for the demand constraint Eq. (9b) that are closely related to the multipliers $\boldsymbol{\mu}$. When the $\boldsymbol{\pi}$ values are intermittently substituted into Eq. (4a) in place of the current $\boldsymbol{\mu}$ values, Caprara et al. (1999) showed that $\theta_{\mathrm{LR}}(\boldsymbol{\mu})$ may converge more rapidly. Although there is no theoretical backing for this idea, we use it in our algorithm whenever the solution to model (9) produces an improved bound.

5.3 Variable fixing heuristic

In our solution approach, the subproblems, represented by Eqs. (4b) and (4c) are always solved optimally so at termination of LD, there are no fractional values of the decision variables on which to perform branch and bound. Short of enumerating all feasible rosters, there is no clear way to iterate toward the optimal solution of model (1). Instead, we propose a partial enumeration scheme based on the solution obtained from the IP heuristic, represented by model (9).

The idea is to sequentially fix the rosters for each rotational profile one at a time during the LR iterations after a threshold number of iterations is reached. In other words, rather than trying to maximize $\theta_{\mathrm{LR}}(\boldsymbol{\mu})$ and get the best lower bound possible on θ_{IP}, we start fixing variables once a good feasible solution has been obtained with the IP heuristic. Although such fixing alters the nature of the LR problem and so is not likely to produce the optimal value of θ_{LD}, it hastens the overall computations and, for our problem, provided very good feasible solutions.

Two rules were used to determine the order in which the rotational profiles are fixed. The first is based on the number of nurses in a profile, which are ordered from smallest to largest with ties broken arbitrarily. This sorting scheme is referred to as *smallest number ordering*, and is implemented by storing the profiles in the set $F = \{j_1, j_2, \ldots, j_{|N^P|}\}$, where

$$n^R_{j_1} \le n^R_{j_2} \le \cdots \le n^R_{j_{|N^P-1|}} \le n^R_{j_{|N^P|}}$$

The second rule is based on the opportunity cost of removing a rotational profile from a feasible schedule. After the first IP heuristic solution is obtained from model (9), call it $(\hat{\zeta}_{j\kappa}, \hat{s}_{dt}, \hat{y}_{dt})$, we compute the level of coverage associated with each profile and then sort them from the largest to the smallest. Letting cov_j be the total, nonredundant demand covered by rotational profile j, the elements of F are ordered such that

$$\mathrm{cov}_{j_1} \ge \mathrm{cov}_{j_2} \ge \cdots \ge \mathrm{cov}_{j_{|N^P-1|}} \ge \mathrm{cov}_{j_{|N^P|}}$$

where $\mathrm{cov}_j = \sum_{d\in D}\sum_{t\in T}\hat{x}_{jdt} - \sum_{d\in D}\sum_{t\in T}\hat{s}_{dt}$ and $\hat{x}_{jdt}$ is determined from $\hat{\zeta}_{j\kappa}$. The rationale for this ordering is that the greater the coverage, the more important the profile.

As profiles are fixed, both Eqs. (4) and (9) must be updated before being solved. Let,

$\bar{F}$ = set of rotational profiles that have been fixed
$\hat{\mathbf{X}}^{\kappa}_j$ = m-dimensional column vector associated with roster κ and rotational profile j in the best feasible solution found to date; i.e., the incumbent
S_j = set of fixed rosters for rotational profile j; $S_j = \{(\hat{\mathbf{X}}^{\kappa}_j, \hat{\zeta}_{j\kappa}) : \forall \hat{\zeta}_{j\kappa} \ge 1, \kappa \in K(j)\}$
θ_{fixed} = total contribution of all $j\in\bar{F}$ to the LR objective function

Using this notation, Eq. (4a) can be rewritten as

$$\theta_{\mathrm{LR}}(\boldsymbol{\mu}) = \text{Minimize} \sum_{j\in N^P\setminus\bar{F}} n^R_j\theta^{\mathrm{SP}}_j + \mu_{dt}s_{dt} + \sum_{d\in D}\sum_{t\in T}(M-\mu_{dt})y_{dt} + \sum_{d\in D}\sum_{t\in T}\mu_{dt}LD_{dt} + \theta_{\mathrm{fixed}} \tag{10a}$$

where

$$\theta_{\text{fixed}} = \sum_{j \in \bar{F}} \sum_{\kappa \in S_j} \hat{\zeta}_{j\kappa} \hat{\theta}^{\text{SP}}_{j\kappa} \tag{10b}$$

and

$$\hat{\theta}^{\text{SP}}_{j\kappa} = c_{j\kappa} - \sum_{d \in D} \sum_{t \in T} \mu_{dt} \hat{x}^{\kappa}_{jdt} \tag{10c}$$

The contribution of the fixed profiles $j \in \bar{F}$ to $\theta_{\text{LR}}(\boldsymbol{\mu})$ in Eq. (10a) can be computed directly because the values of the decision variables x_{jdt} are known. The calculations are given in Eqs. (10b) and (10c). Note that in Eq. (10b), the term $\hat{\zeta}_{j\kappa}\hat{\theta}^{\text{SP}}_{j\kappa}$ appears rather than $n^R_j \theta^{\text{SP}}_j$ and is now summed over all columns $\kappa \in S_j$ since the incumbent solution for profile j may have $|S_j|$ different rosters, each with $\hat{\zeta}_{jk}$ nurses. Recall that only a single roster results when the subproblem, represented by Eqs. (4c)–(4d), is solved.

The final point about Eq. (10a) is that the contribution θ_{fixed} of the fixed profiles is updated dynamically when a new incumbent is found by the IP heuristic. In other words, we do not necessarily keep the original values $(\hat{x}^{\kappa}_{jdt}, \hat{\zeta}_{jk})$ obtained from the first run of model (9) but replace them with the values associated with the incumbent.

5.4 Full Lagrangian relaxation algorithm

We now describe how the various algorithms presented in the previous subsections are combined to solve the original problem (1a)–(1r). The procedure starts with an initial set of multipliers μ^1 whose component values are randomly selected from the set $\{1, 5, 10\}$; that is, $\mu^1_{dt} = \text{RND}(1, 5, 10), \forall d \in D, t \in T$. In addition, one subproblem defined by Eqs. (4c) and (4d) is set up for each rotational profile $j \in N^P$. To keep new notation to a minimum, we do not parameterize all options in the algorithm.

First (F)	operator that takes the first element in the set F
k_{bundle}	parameter for starting Bundle_Algorithm ($k_{\text{bundle}} = 20$)
k_{heur}	frequency parameter for calling the IP heuristic ($k_{\text{heur}} = 10$)
k_{rfix}	frequency parameter for fixing rotational profiles ($k_{\text{rfix}} = 20$)
fix	Boolean variable indicating whether variable fixing has started; fix $= \langle$false$\rangle$ if $k < k_{\text{fix}}$ and $\langle$true$\rangle$ otherwise
S	set of solutions associated with rotational profiles fixed in (10); $S = \{S_j, \forall j \in \bar{F}\}$
$\mathbf{X}_{\text{BEST}}$	incumbent solution
θ_{BEST}	incumbent objective function value

Cyclic_scheduling_algorithm

Input: Initial multiplier values $\boldsymbol{\mu}^1 = \left\{\mu^1_{dt} : \forall d \in D, t \in T\right\}$, subgradient optimization parameters, bundle parameters

Output: $\theta_{\text{BEST}}, \mathbf{X}_{\text{BEST}}$

Step 0: (Initialization) Set $k = 1$, *fix* $= \langle$false$\rangle$, $\theta_{\text{BEST}} = \infty$, $S = \emptyset$, $\bar{F} = \emptyset$, $\mathbf{X}_{\text{BEST}} = \emptyset$ and set up subproblems, represented by Eqs. (4c) and (4d) for each rotational profile $j \in N^P$.

Step 1: (Solve Lagrangian relaxation problem) Solve the following problem
$\theta_{\text{LR}}(\boldsymbol{\mu}) = \text{Minimize} \quad \sum_{j \in N^P \setminus \bar{F}} n^R_j \theta^{\text{SP}}_j + \mu^k_{dt} s_{dt} + \sum_{d \in D} \sum_{t \in T} \left(M - \mu^k_{dt}\right) y_{dt}$
$+ \sum_{d \in D} \sum_{t \in T} \mu^k_{dt} LD_{dt} + \theta_{\text{fixed}}$
subject to Eqs. (4b)–(4d)
to get $\hat{\mathbf{X}}_j = \left(\hat{x}^k_{jdt}, \hat{s}^k_{dt}, \hat{y}^k_{dt}, \forall d \in D, \forall t \in T\right), j \in N^P \setminus \bar{F}$ and add corresponding column to model (9).

Step 2: (Multiplier calculation) Obtain new multipliers μ^{k+1} as follows:
If $(k \le k_{\text{bundle}})$, run Subgradient_Algorithm; put $B \leftarrow B \cup \{\mathbf{g}^k\}$
If $(k > k_{\text{bundle}})$, run Bundle_Algorithm; put $B \leftarrow B \cup \{\mathbf{g}^{\text{bundle}}\}$

Step 3: (Find feasible solution) If $(k - 1 \bmod k_{\text{heur}} = 0$ and $k > 0)$ then
3a. Solve the IP heuristic given by model (9) to get $\left(\hat{x}^k_{jdt}, \hat{s}^k_{dt}, \hat{y}^k_{dt}, \forall d \in D, t \in T; \zeta_{j\kappa}, j \in N^P, \kappa \in K(j)\right)$ and θ_{HR}.
3b. If $(\theta_{\text{HR}} < \theta_{\text{BEST}})$ then
Put $\theta_{\text{BEST}} \leftarrow \theta_{\text{HR}}$
$\mathbf{X}_{\text{BEST}} = \left\{\left(\left(\hat{s}^k_{dt}, \hat{y}^k_{dt}\right), \forall d \in D, \forall t \in T\right), \left(\left(\hat{\mathbf{X}}^k_j, \hat{\zeta}^k_{j\kappa}\right), \forall \kappa \in K(j), j \in N^P\right)\right\}$
$S = \bigcup_{p \in \bar{F}} \left\{\left(\hat{X}_p, \hat{\zeta}_{p\kappa}\right) \in \mathbf{X}_{\text{BEST}}\right\}$; recalculate θ_{fixed} in (10b) and (10c)

Solve the LP relaxation of model (9) and use the dual variables π associated with Eq. (9b) in place of the multipliers $\boldsymbol{\mu}^{k+1}$ derived in Step 2; i.e., put $\boldsymbol{\mu}^{k+1} \leftarrow \boldsymbol{\pi}$

Step 4: (Begin variable fixing) If ($k > 30$ and $fix = \langle \text{false} \rangle$) then
Sort profiles in set F by either coverage or smallest nurse order. Set $fix = \langle \text{true} \rangle$.

Step 5: (Heuristic fixing rule) If ($k - 1 \bmod k_{\text{rfix}} = 0$ and $fix = \langle true \rangle$) then
$p = \text{First}(F)$, $F \leftarrow F \setminus \{p\}$ and $\bar{F} \leftarrow \bar{F} \cup \{p\}$
$S_p = \left\{ \left(\hat{X}_p, \hat{\zeta}_{p\kappa} \right) \in \mathbf{X}_{\text{BEST}} \right\}$ \ fix rotational profile p in LR problem (10)
$S \leftarrow S \cup \{S_p\}$

Step 6: (Termination test) If ($[\theta_{\text{BEST}} - \theta_{\text{LR}}(\boldsymbol{\mu}^k)]/\theta_{\text{LR}}(\boldsymbol{\mu}^k) \leq 0.005$ or $F = \emptyset$) then stop; else, put $k \leftarrow k + 1$ and go to Step 1.

At Step 1, the Lagrangian relaxation problem is solved to get $\theta_{\text{LR}}(\mu^k)$. In the process, each subproblem whose profiles are not fixed is solved separately and the corresponding rosters are used to populate the columns of the heuristic IP given by model (9). Concurrently, the values of s_{dt} and y_{dt} are trivially determined by examining the corresponding objective function coefficients in Eq. (10a). At Step 2, new multiplier values are found by either subgradient optimization if $k \leq k_{\text{bundle}} = 20$ or the bundle method otherwise. The corresponding subgradient is stored in the set B regardless of the approach, and if $|B| > B_{\max} = 60$, an element of B is removed. We found that it was best to use the subgradient algorithm in the early iterations for two reasons: (1) it is computational inexpensive and (2) it provided steady improvement in $\theta_{\text{LR}}(\mu)$ after a few iterations. We switched strategies at iteration 21 when the tailing off effect became noticeable.

At Step 3, a new feasible solution is found every $k_{\text{heur}} = 10$ iterations starting at iteration 11 by solving model (9). CPLEX is used at this step with a termination criterion of either 60 s or 0.1% optimality gap. If $\theta_{\text{HR}} < \theta_{\text{BEST}}$, then a better solution has been found triggering the following adjustments: the incumbent is updated, the set S is updated, the fixed term in model (10) is recomputed, and the current values of the multipliers $\{\mu_{dt}^{k+1}, \forall d \in D, t \in T\}$ derived in Step 2 are replaced with the dual variables $\{\pi_{dt}, \forall d \in D, t \in T\}$ associated with the LP solution to Eq. (9). Because μ and π are closely related, the expectation is that a "big jump" in the Lagrangian objective function will be realized when $\theta_{\text{LR}}(\boldsymbol{\pi})$ is solved at the next iteration rather than $\theta_{\text{LR}}(\boldsymbol{\mu})$. To ensure that the set-covering problem solves quickly, a maximum of 20 columns is allowed for each nurse. When this number is reached, all columns not in the solution to (9) are discarded.

The purpose of Step 4 is to determine when the variable fixing component of the algorithm starts. At the appropriate iteration, one of the two ordering schemes is selected for the set F. In the next section, we provide computational results for both options. When Step 5 is reached, we begin fixing rotational profiles, one at a time every $\tau_{\text{rfix}} = 20$ iterations starting at iteration 31. This requires an updating of the sets F, $\bar{F}$, and S. The more profiles that are fixed, the fewer subproblems that have to be solved.

The final step checks to see if either of the termination criteria is satisfied. The first is a bounds test based on a 0.5% optimality gap. Although we cannot be assured that a better solution does not exist even when $\theta_{\text{LR}}(\boldsymbol{\mu}^k) > \theta_{\text{BEST}}$ due to the nonmonotonicity of the Lagrangian function, none was ever found when the test was omitted. The second stopping criterion comes into play when the set F is empty, implying that there are no more free variables.

6 Computational results

The cyclic scheduling algorithm was implemented in Visual C++ and linked to CPLEX 7.5, which was used to solve the rostering subproblems, represented by Eqs. (4c) and (4d) and the IP heuristic given in (9a)–(9d). Its performance was measured on 15 problem instances of various sizes that were generated from data obtained from a 400-bed U.S. hospital. The experimental design was aimed at determining the effectiveness of the two approaches used to solve the Lagrangian dual and the quality of the solutions provided by the IP heuristic. All computations were performed on a Dell PC with a 1.1-GHz processor and 256 MHz of memory.

6.1 Problem sets

The characteristics of the data sets used in the testing are summarized in Table 1. For each instance, column 2 indicates the total number of nurses considered simultaneously. Recall that schedules are developed independently by each unit in a hospital. The total number of subproblems given in column 3 is equivalent to the total number of rotational profiles after nurse aggregation. The larger this number, the more difficult the problem is to solve. When a profile is limited to one or two shift types, as is the case here, 15 is the maximum number of subproblems that are possible without considering ratios and total working hours (5 single-shift profiles $+ \binom{5}{2}$ two-shift profiles). The eligible shifts for the subproblems, represented by Eqs. (4c) and (4d) were randomly generated

Table 1 Input characteristics of problem instances

Problem No.	No. of nurses	No. of sub-problems	Total demand (h)	Total supply (h)	Total slack allowed (h)	Total gap allowed (h)
1	20	5	1344	1600	464	336
2	20	8	1504	1536	672	672
3	30	5	2206	2400	368	336
4	30	8	2240	2312	712	672
5	50	8	3276	3792	1024	672
6	50	12	3504	3840	896	672
7	50	15	4556	3840	1144	1344
8	80	10	6152	6112	1192	1344
9	80	12	6264	6144	1192	1344
10	80	15	6248	6128	1192	672
11	80	20	6248	6128	1192	672
12	100	12	7484	7672	912	1344
13	100	15	7588	7680	672	672
14	100	18	7588	7696	784	672
15	100	20	7572	7672	882	672

from these 15. Ratios and total hours (either 72 or 80) were then assigned. The actual data sets can be downloaded from http://www.cs.nott.ac.uk/∼tec/NRP/.

The demand, supply, allowed slack (s_{dt}), and allowed gap (y_{dt}) are measured in hours and determine the tightness of a problem's feasible region. The total demand for all shifts over the 2-week planning horizon is given in column 4. The total supply is the sum of working hours for all nurses to be scheduled and is given in column 5. The allowed slack is the total surplus hours that can be assigned to a shift. In the demand constraint, represented by Eq. (4b), this is controlled by the upper bound on s_{dt}, which is $UD_{dt} - LD_{dt}$. The allowed gap is computed by summing the upper bound on the gap variables, $O_{dt}^{\max}$, over all shifts and days, and converting the result to hours. Depending on the problem set, a gap of either one or two nurses per shift was permitted per day. For $O_{dt}^{\max} = 1$, the total gap hours were either 336 or 672 over 14 days, depending on whether the rotational profiles included the three 8-h shift types only (24 h/day) or all five shift types (48 h/day). Those scenarios with 1344 total gap hours had $O_{dt}^{\max} = 2$ and always included all five shift types (96 h/day). No problem sets had shift types other than {D, E, N} or {D, E, N, AM, PM}.

For a fixed number of profiles, the problem instances in Table 1 are generally the most difficult that we could construct. As the difference between supply and demand decreases, and as the allowed slack and allowed gap decrease, the difficulty of a problem increases. When the LP relaxation of model (1) was solved for instances with looser feasible regions, their objective function values, θ_{LP}, were almost always within a small percentage of the best IP solution found, θ_{BEST}. Hence, those results are not reported here.

6.2 Output

Table 2 summarizes the computational results. The quality of a schedule is measured by the average number of violations per nurse (column 2), the total surplus hours (column 3), and the total gap hours (column 4). Under the "general results" columns, the values reported are the better of the two values obtained using the coverage ordering and the smallest number ordering schemes for constructing F. Violations include the undesirable days-on and days-off patterns, and the number of switches in shift assignments on consecutive days. Surplus hours occur when the number of nurses assigned to a shift exceeds the demand, giving rise to idle time. Similarly, the "gap hours" indicate the total amount of undercoverage that exists in the schedule for the upcoming 2 weeks. The number of surplus and gap hours should be no greater than the maximum hours shown in the last two columns of Table 1. Because of the complementary condition $s_{dt} \times y_{dt} = 0, \forall d \in D, t \in T$, it is impossible to have a schedule in which the surplus and gap hours are both at their maximum values.

The next two columns in Table 2 give the lower bounds obtained from two different relaxations of model (1). The "LP soln" in column 5 is the objective function value found by relaxing the integrality requirements in the original model. Column 6 gives the highest value of Lagrangian objective function $\theta_{\text{LR}}(\boldsymbol{\mu})$ obtained before variable fixing was invoked at iteration 30. This is the best lower bound on θ_{IP} that was achieved using the two subgradient strategies. Of course, once variable fixing begins, this bound may and did, in fact, increase in some cases.

The last six columns in Table 2 highlight the performance of the two ordering schemes used to fix variables.

Table 2 Computational results

Problem no.	General results					Coverage ordering			Smallest number ordering		
	Violations/ nurse	Surplus hours	Gap hours	LP soln, θ_{LP}	LD soln, θ_{LD}	IP soln, θ_{BEST}	Time (s)	Total iterations	IP soln, θ_{BEST}	Time (s)	Total iterations
1	2	256	0	0	11	36	177	90	36	328	130
2	1.5	136	120	400	418	424	229	150	426	656	350
3	2	268	64	360	403	412	373	180	413	209	110
4	1.3	104	56	280	236	316	389	90	313	753	150
5	1.3	584	80	320	363	448	432	100	448	700	190
6	1.18	228	40	160	111	213	506	130	214	590	150
7	1.12	0	724	2467	2549	2550	296	30	2550	296	30
8	1.06	192	232	960	1016	1017	211	30	1017	211	30
9	1.11	132	288	1280	1319	1339	1338	130	1339	2117	210
10	0.95	64	176	560	474	770	1318	290	775	1783	230
11	1.2	152	320	480	594	595	210	30	595	210	30
12	1.15	452	252	920	1169	1180	530	50	1235	1406	170
13	1.15	364	56	41	282	406	1795	170	406	2207	210
14	1.10	276	168	481	658	767	2375	250	767	3185	290
15	1.08	292	192	681	876	880	315	30	880	315	30

Performance is measured by (i) the quality of the feasible solution found with the IP heuristic, (ii) the total computation time, and (iii) the total number of iterations. Computation times are measured in seconds and include the initialization process, the time for subgradient optimization, and the time required to solve the IP heuristic. Recall that the first 20 iterations use the subgradient algorithm to find the multiplier values, while the remainder use the bundle method. Variable fixing starts at iteration 31.

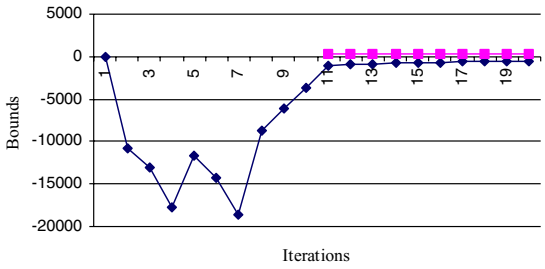
With respect to schedule quality, we see from Table 2 that the average number of violations per nurse is no more than two, and exhibits a slight downward trend as problem size increases. The total slack and gap hours are significantly less than the maximum allowed. The average total slack is approximately 29% of the maximum allowed, while the average total gap hours is about 25%, which is a good sign.

With respect to solution quality, the first measure of interest is the lower bound. As expected, the best Lagrangian solution prior to variable fixing, θ_{LD}, was generally greater than the LP solution, θ_{LP}, obtained from problem (1). Only instances 4, 6, and 10 yielded a result with $\theta_{LP} > \theta_{LD}$, which indicates that the Lagrangian dual problem was not always solved to optimality. In fact, when we applied our preliminary column generation algorithm to problem (1), we found that it produced a lower bound that was 19.2% on average greater than θ_{LD}. Theory tells us that these two bounds should be equal. The magnitude of their difference, however, was not surprising because our primary goal was to find good feasible solutions quickly and not to solve the Lagrangian dual to optimality. This was the main reason for starting the variable fixing scheme at iteration 31.

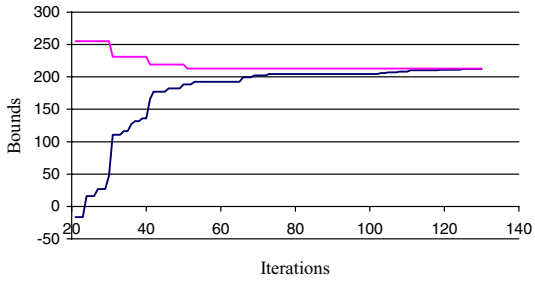
With this said, the relatively high quality of the lower bound θ_{LD} can be attributed to the effectiveness of the subgradient algorithm in finding good multipliers on the first 20 iterations, enhanced with the inclusion of the dual variables obtained from the IP heuristic. In four of the larger instances, θ_{LD} was within 0.5% to the best integer solution found by the heuristic after 30 iterations. Two of these were the 100-nurse instances with 20 rotational profiles each.

The second measure of computational performance is the quality of the IP solution reported at termination. By examining the optimality gap between θ_{BEST} and θ_{LD}, we can determine how far we might be in the worst case from the true (unknown) value of θ_{IP}. Using the θ_{BEST} values in column 7, the largest gap is 227% for problem 1 and the smallest is virtually 0% for problems 7 and 8. The average gap is 29.6% (15.5% when problem 1 is excluded) but the variance is high so few generalizations are possible.

To get a better idea of overall algorithmic performance we have included a plot of the upper and lower bounds for problem no. 6 as a function of the number of iterations. Figure 1(a) contains the first 20 iterations and Fig. 1(b) contains the remainder. The first few iterations are erratic and illustrate the nonmonotonic property of the standard subgradient method. At iteration 10, a feasible solution ($\theta_{HR} = 317$) is obtained from the IP heuristic, and at iteration 21, the bundle method is initiated. From that point on, the lower bound is guaranteed to be nondecreasing due to the serious-step and null-step logic. Nevertheless, convergence is relatively slow with the exception of the steep jumps at iterations 41 and 61 that result from the use of the dual variables, $\boldsymbol{\pi}$, rather than the multipliers, $\boldsymbol{\mu}$, in the solution to model (4). In contrast, the upper bound behaves as a step function because it is only updated



(a) Upper and lower bounds for first 20 iterations



(b) Upper and lower bounds for iteration 21 to iteration 130

Fig. 1 Algorithmic performance for problem no. 6.

every 10 iterations. In this problem, the best solution is found at iteration 51 ($\theta_{\text{BEST}} = 213$).

Comparing the results of the two variable fixing schemes in the last six columns of Table 2, we see that coverage ordering was superior to smallest number ordering in almost all instances on all metrics. The former produced feasible solutions that were at least as good as the latter in all but instance 4; however, the differences were negligible with the exception of instance 12. With respect to computational efficiency (time and number of iterations), the coverage ordering scheme did better in 10 instances and worse in only one. Overall run-time averages were 699 and 997 s, respectively, and overall iteration averages were 116 and 154. In general, the better the quality of the lower bound $\theta_{\text{LR}}(\boldsymbol{\mu})$ prior to variable fixing, the fewer iterations required by the algorithm.

7 Summary and conclusions

The purpose of this paper has been to describe a dual-ascent Lagrangian heuristic for solving the cyclic preference scheduling problem for nurses. We began by examining two different relaxations, the first involving the preference constraints and the second involving the demand constraints. A theoretical examination of the IP that resulted when the preference constraints were placed in the objective function as a penalty term, led to the conclusion that is was not likely to provide good bounds.

Relaxing the demand constraints, however, allowed us to decompose the original problem into a set of subproblems, one for each rotational profile, and in the process, achieve both a measure of computational efficiency as well as the potential for strong bounds. To find the best lower bound we used a standard subgradient algorithm in the early stages of the Lagrangian iterations and a bundle method in the later stages. High quality upper bounds were obtained by periodically solving an IP heuristic whose columns were derived from the subproblem solutions. The computations for this component of the algorithm never took more than a few minutes.

To increase the speed of convergence, a variable fixing strategy was developed in place of traditional branch and bound. The idea was to sequentially fix rotational profiles to increase the lower bound. Two approaches were investigated, one based on the criterion of minimum number of nurses in a profile and the other on the opportunity cost of removing a profile from a solution. Empirically, the latter was more effective and converged more rapidly. The performance of the full algorithm was demonstrated on problem instances with up to 100 nurses and 20 rotational profiles.

To increase the usefulness of the model, we would like to extend the types of preference violations that it can handle, which are now limited to coverage patterns. An interesting area of future research would be to develop a procedure that would be able to incorporate personalized restrictions for each nurse. A related problem centers on the augmentation of the staff. If it were possible to hire one or more nurses for the upcoming planning period, we would like to be able to determine the best profiles to assign them. Although this problem sounds simple, its complexity is the same as the original.

Appendix A: analysis of alternative Lagragrian relaxation model

In this appendix, we show that the relaxation given by the model (2), is not likely to yield good lower bounds to the original problem because the multipliers are always zero. The proof is based on enumeration of all feasible points to model (1), which are then used to set up an explicit form of model (3). First, though, we confirm that the optimal solution to the LR problem, represented by model (2), contains the minimum possible number of outside nurses.

Lemma 1. *In an optimal solution to model* (2)*, the term in the objective function associated with the outside nurses,* $M \sum_{d\in D} \sum_{t\in T} y_{dt}$*, always achieves its lowest feasible value.*

Proof: To begin, we need to show that the multipliers λ_{id} and μ_{id} are bounded in (3) for all i and d. To do this, it is convenient to view (3) as a 2-player max–min game in which the first player picks the multipliers $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ and the second player picks the variables $(\mathbf{x}, \mathbf{p}, \mathbf{q}, \mathbf{s}, \mathbf{y})$. To

simplify the notation, let $u_{id} = \sum_{t\in T_i} x_{idt} - \sum_{t\in T_i} x_{i,d+1,t} + \sum_{t\in T_i} x_{i,d+2,t}$ and let $w_{id} = -\sum_{t\in T_i} x_{idt} + \sum_{t\in T_i} x_{i,d+1,t} - \sum_{t\in T_i} x_{i,d+2,t} + 1$ where $u_{id}, w_{id} \in \{-1, 0, 1, 2\}$ for all feasible rosters. Now, looking at Eq. (2b), if $\lambda_{id} > 0$, then the second player would like to make $u_{id} + P_{id} > 0$ as long as $\lambda_{id}(u_{id} + P_{id}) \geq \sum_{a=1}^{V_{\max}} r_a v_{ia}$. Thus, the first player will always pick λ_{id} so that it satisfies the following condition:

$$\lambda_{id} \leq \min\left\{\sum_{a=1}^{V_{\max}} r_a v_{ia}^k \big/ \left(u_{id}^k + p_{id}^k\right) : k \text{ is a feasible roster}\right\}$$

Because the same arguments are valid for μ_{id}, q_{id}, and w_{id} as well, we conclude that λ_{id} and μ_{id} are bounded. The statement of the lemma follows because M is arbitrarily large and the summation $\sum_{d\in D}\sum_{t\in T} y_{dt}$ is independent of the multipliers $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. □

Proposition 2. *The optimal solution of the Lagrangian dual problem* (3) *is* $\lambda_{id} = 0$ *and* $\mu_{id} = 0$ *for all* $i \in N$ *and* $d \in D$.

Proof: In light of Lemma 1, let y_{dt}^* be the optimal values of the outside nurse variables in (2) and let us introduce a modified version of the LR problem without the demand constraint, represented by Eq. (1b). □

$$\begin{aligned}
\hat{\theta}_{\mathrm{LR}}(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= \theta_{\mathrm{LR}}(\boldsymbol{\lambda}, \mu) - M\sum_{d\in D}\sum_{t\in T} y_{dt}^* \\
&= \text{Minimize} \quad \sum_{i\in N}\sum_{a=1}^{V_{\max}} r_a v_{ia} + \sum_{i\in N}\sum_{d\in D}(-\boldsymbol{\lambda}_{id}\boldsymbol{p}_{id} - \boldsymbol{\mu}_{id}\boldsymbol{q}_{id}) \\
&\quad - \sum_{i\in N}\sum_{d\in D}\lambda_{id}\left(\sum_{t\in T_i} x_{idt} - \sum_{t\in T_i} x_{i,d+1,t} + \sum_{t\in T_i} x_{i,d+2,t}\right) \\
&\quad - \sum_{i\in N}\sum_{d\in D}\mu_{id}\left(-\sum_{t\in T_i} x_{idt} + \sum_{t\in T_i} x_{i,d+1,t} - \sum_{t\in T_i} x_{i,d+2,t} + 1\right) \\
&\quad \text{subject to Eqs. (1c)}-\text{(1i)}, \ \text{Eqs.(1l)}-\text{(1r)} \qquad (11)
\end{aligned}$$

The corresponding LD problem is

$$\hat{\theta}_{\mathrm{LD}} = \max_{\boldsymbol{\lambda},\boldsymbol{\mu}\geq 0} \hat{\theta}_{\mathrm{LR}}(\boldsymbol{\lambda}, \boldsymbol{\mu}) \qquad (12)$$

which can be viewed as a maximization problem in $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ over the set of discrete points. Because the demand constraint, represented by Eq. (1b), has been omitted, each point represents a roster for a nurse determined by the remaining constraints, represented by Eqs. (1c)–(1i), Eqs. (1l)–(1r). For nurse i, let n_i be the number of feasible rosters, $\mathbf{X}_i^k = (x_{idt}^k, P_{id}^k, q_{id}^k, s_{dt}^k, y_{im}^k)$ the kth feasible roster, and $Q_i = \{\mathbf{X}_i^k : k = 1, \ldots, n_i\}$ the set of all feasible rosters. Thus, for each nurse i, we have

$$\hat{\theta}_{\mathrm{LR}}(\boldsymbol{\lambda}_i, \boldsymbol{\mu}_i) = \max_{\mathbf{X}_i^k \in Q_i} \hat{\theta}_{\mathrm{LR}}\left(\boldsymbol{\lambda}_i, \boldsymbol{\mu}_i; \mathbf{X}_i^k\right) \qquad (13)$$

where $\boldsymbol{\lambda}_i = (\lambda_{id})$, $\boldsymbol{\mu}_i = (\mu_{id})$, and $\hat{\theta}_{\mathrm{LR}}(\boldsymbol{\lambda}_i, \boldsymbol{\mu}_i; \mathbf{X}_i^k)$ is the objective function value for problem (4) at $\mathbf{X}_i^k$. The Lagrangian function $\hat{\theta}_{\mathrm{LR}}(\boldsymbol{\lambda}_i, \boldsymbol{\mu}_i)$ to be maximized is piecewise linear and concave.

In general, a roster is defined as a 14-day sequence of days on and days off for a particular rotational profile. For example, for nurse i who works only evening shifts, the kth roster in simplified form might be E-E-E-E-Off-E-Off-Off-E-E-E-E-E-Off or $\mathbf{x}_i^k = (1, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0)$. Based on this definition and the constraints in (11), the size of set Q_i is 150 for nonrotational nurses. It grows exponentially as more shift types are allowed.

Now, for every feasible roster for nurse i, we can find the associated values of $u_{id} = u_{id}^k = \sum_{t\in T_i} x_{idt}^k - \sum_{t\in T_i} x_{i,d+1,t}^k + \sum_{t\in T_i} x_{i,d+2,t}^k$ and $w_{id}^k = -\sum_{t\in T_i} x_{idt}^k + \sum_{t\in T_i} x_{i,d+1,t}^k - \sum_{t\in T_i} x_{i,d+2,t}^k + 1$ for $d = 1, \ldots, |D|$, where u_{id}^k, $w_{id}^k \in \{-1, 0, 1, 2\}$ for all k. Using this notation, the Lagrangian dual problem, represented by (13), can be written equivalently as

$$\begin{aligned}
&\text{Maximize} \quad \eta_i \\
&\text{subject to} \sum_{a=1}^{V_{\max}} r_a v_{ia}^k + \sum_{d\in D}\left(-p_{id}^k\lambda_{id} - q_{id}^k\mu_{id}\right) - \sum_{d\in D} u_{id}^k\lambda_{id} \\
&\qquad - \sum_{d\in D} w_{id}^k\mu_{id} \geq \eta_i, \forall \mathbf{X}_i^k \in Q_i \quad \boldsymbol{\lambda}_i \geq \mathbf{0}, \boldsymbol{\mu}_i \geq \mathbf{0}
\end{aligned} \qquad (14)$$

which is a linear program in the η_i, λ_i, and μ_i variables, where η_i, has been introduced to transform the piecewise linear function $\theta_{\mathrm{LR}}^i(\boldsymbol{\lambda}_i, \boldsymbol{\mu}_i)$ in Eq. (13) into a linear function subject to linear constraints. The size of (14) depends on the number of feasible schedules for nurse i. The largest instance arises for a rotating nurse and has 23 variables and 644 constraints.

Because the objective in model (14) is maximization, the only case in which a solution will contain values of λ_{id} or μ_{id} greater than 0 is when the corresponding values of u_{ia}^k or w_{id}^k are negative. In general terms then, a necessary condition for $\boldsymbol{\lambda}_i = \boldsymbol{\mu}_i = \mathbf{0}$ in an optimal solution to (14) is that there does not exist a column in the constraint matrix associated with one of the λ_{id} or variables that has a -1 in each row. It can be verified that no such column exists by examining the appropriate constraints in the model, represented by model (2), and recognizing that each row in (14) is derived from a feasible roster for nurse i.

To complete the proof, note that model (14) is a relaxation of the Lagrangian dual problem given in (2) with the demand constraint removed. Therefore, solving (14) jointly for all nurses with objective function $\sum_{i \in N} \eta i$, subject to the additional constraint, represented by Eq. (1b), cannot produce a large objective function value. This follows, in part, because the set Q_i is not dependent on Eq. (1b).

The implication of Proposition 2 is that Lagrangian dual, represented by Eq. (3), can be solved in one step by fixing $\boldsymbol{\lambda} = \boldsymbol{\mu} = \mathbf{0}$ and solving (2). It is interesting to note that the Lagrangian function $\theta^i_{\mathrm{LR}}(\boldsymbol{\lambda}_i, \boldsymbol{\mu}_i)$ in Eq. (13) is nonincreasing at $\boldsymbol{\lambda}_i = \boldsymbol{\mu}_i = \mathbf{0}$. This can be seen from (16) and the fact that $P^k_{id}, q^k_{id} = 0$ or 1 and that $\min\{u^k_{ia}, w^k_{id} : \forall k\} = -1$. If, say, $u^{\hat{k}}_{ia} = -1$, $P^{\hat{k}}_{id} = 0$ and all other values of $\boldsymbol{\lambda}_i$ and $\boldsymbol{\mu}_i = \mathbf{0}$, then increasing λ_{id} slightly from 0 causes the left-hand side of constraint $\hat{k}$ in (14) to decrease by $|u^{\hat{k}}_{ia}\lambda_{id}|$. This, in turn, allows η_i on the right-hand side to potentially increase by the same amount. Nevertheless, because there always exists a $k \neq \hat{k}$ such that $u^k_{ia} \geq 0$, η_i cannot be increased. In light of the fact that $\theta^i_{\mathrm{LR}}(\boldsymbol{\lambda}_i, \boldsymbol{\mu}_i)$ is concave, this means that it is maximized at the origin.

References

Aickelin, U. and K. Dowsland, "Exploiting problem structure in a genetic algorithm approach to a nurse rostering problem," *Journal of Scheduling*, **3**(3), 139–153 (2000).

Aickelin, U. and K. Dowsland, "An indirect algorithm for a nurse-scheduling problem," *Computers & Operations Research*, **31**(5), 761–778 (2004).

Bard, J. F. and H. W. Purnomo, "Preference scheduling for nurses using column generation," *European Journal of Operational Research*, **164**(2), 510–534 (2005a).

Bard, J. F. and H. W. Purnomo, "Hospital-wide reactive scheduling of nurses with preference considerations," *IIE Transactions on Operations Engineering*, **37**(7), 589–608 (2005b).

Bard, J. F. and H. W. Purnomo, "A column generation-based approach to solve the preference scheduling problem for nurses with downgrading," *Socio-Economic Planning Sciences*, **39**(3), 193–213 (2005c).

Berrada, I., J. A. Ferland, and P. Michelon, "A multi-objective approach to nurse scheduling with both hard and soft constraints," *Socio-Economic Planning Sciences*, **30**(3), 183–193 (1996).

Brusco, M. J. and L. W. Jacobs, "Cost analysis of alternative formulations for personnel scheduling in continuously operating organisations," *European Journal of Operational Research*, **86**(2), 249–261 (1995).

Burke, E. K., P. De Causmaecker, and G. Vanden Berghe, "A hybrid tabu search algorithm for the nurse rostering problem," in: B. McKay et al. (Eds.), *Simulated Evolution and Learning, Lecture Notes in Artificial Intelligence*. Springer, Berlin (1999), Vol. 1585, pp. 187–194.

Burke, E. K., P. I. Cowling, P. De Causmaecker, and G. Vanden Berghe, "A memetic approach to the nurse rostering problem," *Applied Intelligence*, **15**(3), 199–214 (2001).

Burke, E. K., P. De Causmaecker, and G. Vanden Berghe, "Novel meta-heuristic approaches to nurse rostering problems in Belgian hospitals, Chap. 44, in J. Leung (Ed.), *Handbook of Scheduling: Algorithms, Models, and Performance Analysis*, CRC Press, Boca Raton, FL (2004a), pp. 44:1–44:18.

Burke, E. K., P. De Causmaecker, G. Vanden Berghe, and H. Van Landeghem, "The state of the art of nurse rostering," *Journal of Scheduling*, **7**(6), 441–499 (2004b).

Caprara, A., M. Fischeti, and P. Toth. "A heuristic method for the set covering problem," *Operations Research*, **47**(5), 730–743 (1999).

Caprara, A., M. Monaci, and P. Toth, "Models and algorithms for a staff scheduling problem," *Mathematical Programming, Series B*, **98**, 445–476 (2003).

Cheng, B. M. W., J. H. M. Lee, and J. C. K. Wu, "A nurse rostering system using constraint programming and redundant modeling," *IEEE Transactions in Information Technology in Biomedicine*, **1**(1), 44–54 (1997).

Crainic, T. G., A. Frangioni, and B. Gendron, "Bundle-based relaxation methods for multicommodity capacitated fixed charge network design," *Discrete Applied Mathematics*, **112**, 73–99 (2001).

De Causmaecker, P. and G. Vanden Berghe, "Relaxation of coverage constraints in hospital personnel rostering," in: E. K. Burke and P. De Causmaecker (Eds.), *Practice and Theory of Automated Timetabling, Vol. IV, 4th International Conference*, PATAT 2002, Gent, Belgium, LNCS, Springer, Berlin (2003), Vol. 2740, pp. 129–147.

Dowsland, K. A., "Nurse scheduling with tabu search and strategic oscillation," *European Journal of Operational Research*, **106**(2–3), 393–407 (1998).

Emmons, H., "Work-force scheduling with cyclic requirements and constraints on days off, weekends off, and work stretch," *IIE Transactions*, **17**(1), 8–15 (1985).

Ernst, A.T., H. Jiang, M. Krishnamoorthy, and D. Sier, "Staff scheduling and rostering: a review of applications, methods and models," *European Journal of Operational Research*, **153**, 3–27 (2004).

Ferland, J.A., I. Berrada, I. Nabli, B. Ahiod, P. Michelon, V. Gascon, and E. Gagné, "Generalized assignment type goal programming problem: application to nurse scheduling," *Journal of Heuristics*, **7**, 391–413 (2001).

Frangioni, A. and G. Gallo, "A bundle dual-ascent approach to linear multicommodity min-cost flow problems," *INFORMS Journal on Computing*, **11**, 370–393 (1999).

Griesmer, H., "Self-scheduling turned us into a winning team," *Management Decisions*, **56**(12), 21–23 (1993).

Howell, J. P., "Cyclical scheduling of nursing personnel," *Hospital J.A.H.A.*, 40, 77–85 (1998).

Isken, M., "An implicit tour scheduling problem with application in healthcare," *Annals of Operations Research*, **128**, 91–109 (2004).

Jaumard, B., F. Semet, and T. Vovor, "A generalized linear programming model for nurse scheduling," *European Journal of Operational Research*, **107**, 1–18 (1998).

Kawanaka, H., K. Yamamoto, T. Yoshikawa, T. Shinogi, and S. Tsuruoka, "Genetic algorithm with constraints for the nurse scheduling problem," in: Proceedings of Congress on Evolutionary Computation, *IEEE Press*, Seoul, South Korea (2001), Vol. 2, pp. 1123–1130.

Kimball, B. and E. O'Neil, "The American nursing shortage," The Robert Wood Johnson Foundation, Princeton, NJ (2002).

Lau, H.C., "On the complexity of manpower shift scheduling," *Computers & Operations Research*, **23**(1), 93–102 (1996).

Lemarechal, C., "Nondifferentiable optimization," in: G. L. Nemhauser, A. H. G. Rinnooy Kan, and M. J. Todd (Eds.), *Handbooks in Operations Research and Management Science*, Vol. 1: *Optimization*, North-Holland, Amsterdam, The Netherlands (1989), pp. 529–572.

Meyer auf' m Hofe, H. "ComPlan/SIEDAPlan: personnel assignment as a problem of hierarchical constraint satisfaction," in: *Proceedings of the 3rd International Conference on the Practical Application of Constraint Technology*, London, (1997), pp. 257–271.

Meyer auf' m Hofe, H. "Solving rostering tasks as constraint optimisation, in E.K Burke and W. Erben (Eds.), *Practice and Theory of Automated Timetabling, Vol. III*, 3rd International Conference, PATAT 2000, Konstanz, Germany, LNCS, Vol. 2079, pp. 191–212, Springer, Berlin (2001).

Millar, H.H. and M. Kiragu, "Cyclic and non-cyclic scheduling of 12-hour shift nurses by network programming," *European Journal of Operational Research*, **104**, 582–592 (1998).

Miller, H.E., W.P. Pierskalla, and G. J. Rath, "Nurse scheduling using mathematical programming," *Operations Research*, **24**(5), 857–870 (1976).

Nemhauser, G.L. and L.A. Wolsey, *Integer and Combinatorial Optimization*, Wiley, New York (1988).

Nonobe, K. and T. Ibaraki, "A tabu search approach to the constraint satisfaction problem as a general problem solver," *European Journal of Operational Research*, **106**, 599–623 (1998).

Petrovic, S., G. Beddoe, and G. Vanden Berghe, "Storing and adapting repair experiences in employee rostering," in: E. K. Burke and P. De Causmaecker (Eds.), *Practice and Theory of Automated Timetabling, Vol. IV, 4th International Conference*, PATAT 2002, Gent, Belgium, LNCS, (2003), Vol. 2740, pp. 148–165.

Pierskalla, W.P. and D.J. Brailer, "Applications of operations research in health care delivery," *Handbooks in Operations Research and Management Science*, North Holland, Amsterdam, The Netherlands (1994), Vol. 6, pp. 469–505.

Randhawa, S.U. and D. Sitompul, "A heuristic-based computerized nurse scheduling system," *Computer & Operations Research*, **20**(8), 837–844 (1993).

Spratley, E., A. Johnson, J. Sochalski, M. Fritz, and W. Spencer, "The registered nurse population," Findings from the National Sample Survey of Registered Nurses," U.S. Department of Health and Human Services (2000).

Topaloglu, S. and I. Ozkarahan, "An implicit goal programming model for the tour scheduling problem considering the employee work preferences," *Annals of Operations Research*, **128**, 135–158 (2004).

Valouxis, C. and E. Housos, "Hybrid optimization techniques for the workshift and rest assignment of nursing personnel," *Artificial Intelligence in Medicine*, **20**, 155–175 (2000).

Warner, D.M., "Scheduling nursing personnel according to nursing preference: a mathematical programming approach," *Operations Research*, **24**(5), 842–856 (1976).