Stochastic Programming for Nurse Assignment

PRATTANA PUNNAKITIKASHEM, pxp1742@exchange.uta.edu

Department of Industrial and Manufacturing Systems Engineering, The University of Texas at Arlington, Arlington, Texas 76019-0017,USA

JAY M. ROSENBERGER, jrosenbe@uta.edu

Department of Industrial and Manufacturing Systems Engineering, The University of Texas at Arlington, Arlington, Texas 76019-0407,USA

DEBORAH BUCKLEY BEHAN, dbehan@uta.edu

School of Nursing, The University of Texas at Arlington, Arlington, Texas 76019-0017,USA

Abstract. We present a brief overview of four stages of nurse planning. For the last stage, which assigns nurses to patients, a stochastic integer programming model is developed. A Benders' decomposition approach is proposed to solve this problem, and a greedy algorithm is employed to solve the recourse subproblem. Patient-to-nurse ratio constraints are introduced to balance the workload of nurses as well as improve the overall performance of the algorithm. Computational results are provided based upon data from Baylor Regional Medical Center in Grapevine, Texas. Finally, areas of future research are discussed.

Keywords: nurse assignment, stochastic programming, Benders' decomposition,

1 Introduction

One of the greatest problems in health care today is a shortage of nurses. The demand for nurses is growing, while fewer young nurses are available to provide care. A survey by the American Hospital Association found that 75% of vacant hospital staffing positions are for registered nurses [19]. The number of nurses per capita declined by 2% from 1996 to 2000, while the attrition rate of hospital nursing staff grew from 11.7% in 1998 to 26.2% in 2000 [14]. From 1993 to 2001, enrollment in registered nurse degree programs declined by 50,000 nurses [19]. With fewer new nurses entering the profession, the average age of the working registered nurse is increasing [7]. From 1983 to 1998, the number of nurses under 30 years of age decreased by 41% [19]. The elderly population is growing in numbers, and they need substantially more health care services [30]. Moreover, the number of citizens over 65 years old is expected to be 70 million in 2030, more than twice that of 1999 [1]. Consequently, the shortage will become more severe. By 2020, the United States will face a 20% shortage in the number of nurses needed for the nation's health care system [7].

The nursing shortage affects patient care. The National Survey on Consumers' Experiences with Patient Safety and

Quality Information Consumers showed that the most important factors causing medical error are workload, stress, and fatigue of health professionals (74%); not enough time spent with patients (70%); and not enough nurses in health care systems (69%) [28]. A study by the Agency for Healthcare Research and Quality reported that nurses spend insufficient time with patients in hospitals with low staffing levels [26]. Powers [23] observed that excessive workload enhances poor quality of patient care. Given that nursing resources are so scarce, intelligent planning methods are needed to reduce the burden of the shortage. One method to reduce excessive workload on nurses is balancing patient assignments.

Stochastic programming has successfully optimized many industry decisions including budgets for nursing [16], but no one has used it to assign nurses to patients. Optimization research on nurse assignment only includes a few deterministic integer programming models. In this paper, we develop a stochastic programming model for nurse assignment with a recourse penalty function to minimize excess workload for nurses. In the stochastic model, the amount of required care of a patient is random, so it considers several scenarios for a patient's required care. It provides the initial assignment of nurses to patients for a nursing shift in a hospital unit.

In the remainder of Section 1, we describe four stages of nurse planning—*nurse budgeting*, *nurse scheduling*, *nurse rescheduling*, and *nurse assignment*. In Section 2, we present a two-stage stochastic programming model for nurse assignment that minimizes excess workload, and Section 3 presents algorithmic approaches to solve it. Because of the special structure of the recourse function, we solve the second-stage subproblem with a greedy algorithm. Anecdotal accounts from nurses suggest that nurse assignment is usually performed within 30 minutes before each shift. Consequently, the focus of this research is to find a good solution with the time limitation. A patient-to-nurse ratio is introduced to balance the number of patients assigned to each nurse. In Section 4, we compare the performance from our model with those from other methods based upon data from a medical surgery unit at Baylor Medical Center in Grapevine, Texas. Finally, we discuss conclusions and topics of future research in Section 5.

1.1 *Overview of Nurse Planning*

Nurses work in a variety of environments including hospitals, clinics, private doctor's offices, nursing homes, and individual homes. Although our research is on nurses at one hospital, it may be applied to other hospitals and environments with multiple nurses and patients. Hospitals in the United States employ two types of nurses—*registered nurses (RNs)* and *licensed vocational nurses (LVNs)*. We describe the four stages of nurse planning in the Sections 1.1.1 through 1.1.4.

1.1.1 *Nurse Budgeting*

Financial planners create budgets and determine how many nurses they will hire as permanent staff and how many they will hire from an agency. Warner [32] implemented a Markovian analysis to forecast nursing personnel for general wards of a hospital. Kao and Tung [17] predicted patient demands over a year by an autoregressive integrated moving

average forecasting method. Dieck [13] compared the Box-Jenkins modeling and the Winters' heuristic approach for forecasting patients admission to public health facilities. Trivedi [29] developed a mixed integer goal programming model, while Kao and Queyranne [16] applied a stochastic programming approach to optimize a budget for nurses.

1.1.2 *Nurse Scheduling*

In the second stage, a nurse manager forecasts the number of patients that will enter a hospital unit over four to six weeks. Based upon the forecasted number of patients, the manager uses a census matrix to determine the number and level of nurses needed. When the number of nurses of each type is known, a schedule is created that partitions a day into shifts that are typically 8 or 12-hours in length. Typically, the manager posts a schedule two weeks before the beginning of the time horizon. Most academic literature on nurse planning is on scheduling [34, 20, 33, 8, 9, 3, 2, 4, 15, 18, 12, 22]. Because these algorithms only consider the nurse budgeting and scheduling stages, they ignore changes in staff and patient forecasts and assume the schedule will be followed as planned. Anecdotal evidence suggests that changes to the schedule are frequent, so intelligent planning models to reschedule nurses will dramatically improve nurse planning.

1.1.3 *Nurse Rescheduling*

The rescheduling process occurs 90 minutes before each shift. A nurse supervisor reviews the scheduled nurses based upon the activities of the previous shift, activities of other units, and either a census matrix or a patient classification system. If there is a shortage of nurses for the upcoming shift, the supervisor tries to recruit additional nurses who work as needed–*PRN nurses*, nurses who work part time–*part-time nurses*, and nurses who are not scheduled for the upcoming shift–*off-duty nurses*. If an insufficient set of nurses agrees to work the shift, the supervisor, upon approval from a nurse manager, hires temporary agency nurses to satisfy the remaining shortage. If there are too many scheduled nurses for the shift than needed, then the supervisor has surplus PRN nurses and part-time nurses take the day off without pay.

Patient classification systems are the most sophisticated technology for nurse rescheduling. These systems group patients into one of several categories. They estimate how many times certain tasks will be performed in caring for a patient in each category. Using these estimates and the expected time required to perform each task, the systems determine the amount of time to care for a typical patient. As patients are admitted into the unit, the system classifies these patients, and nurse supervisors use the estimated patient care to determine how many nurses are needed for the shift in nurse rescheduling. As a patient's condition changes, he may be given a new patient classification. Although patient classifications systems provide benchmarks for nurse planning, they have several drawbacks as described in Section 2.2. Siferd and Benton [25] developed a stochastic model based upon the patients in a unit to determine how many nurses are required for the shift. Bard and Purnomo [5] presented an integer programming model for daily nurse rescheduling and implemented a branch-and-price algorithm to solve the problem.

1.1.4 *Nurse Assignment*

In the final stage of nurse planning, *nurse assignment*, a charge nurse assigns each patient to a nurse at the beginning of a shift. Typically, the nurse assignment has to be performed within 30 minutes before a shift. Although the charge nurse may update an assignment, in many hospital units, such as medical-surgical units, revised assignments only include assigning a nurse to a new admission; rarely is a patient reassigned a new nurse during the middle of a shift. Consequently, the initial assignment can determine the amount of workload given to each nurse during the shift. A nurse's *workload* is the amount of time required to care for her patients over a time period, and *excess workload* is the difference between the workload and the time available for care. In reality, excess workload results in other nurses assisting overworked nurses. One important consideration in nurse assignment is workload balance.

Developing balanced workloads for nurses is difficult because of the variation of patients' conditions [21]. In practice, most nurse assignments are based upon either an intuitive judgment or the caseload method, in which each nurse is assigned the same number of patients [24]. Modern patient classification systems partition the set of patients into groups, and each group is assigned to a nurse (Overfelt 2004). Walts and Kapadia [31] presented a patient classification system and optimization model to determine the level of staffing to meet the required workload level, but they did not use a detailed nurse assignment model. Mullinax and Lawley [21] developed an integer linear programming model that assigns patients to nurses in a neonatal intensive care unit. The nurseries are divided into a number of physical zones. They used a zone-based heuristic that assigns nurses to zones and computes patient assignments within each zone. Unlike the stochastic programming model in this paper, these approaches and patient classification systems ignore uncertainty, which is a major drawback considering the enormous variance in patient care. We are unaware of any previous research on stochastic nurse assignment.

1.2 *Contribution*

The contribution of this paper includes a stochastic programming model for the nurse assignment problem. The model addresses several important issues that are ignored in academic literature and patient classification systems.

- *Patient Uncertainty.* Traditional nurse assignment models ignore uncertainty. Because of the enormous variance in patient care, the stochastic programming model that considers uncertainty provides more robust solutions.
- *Fluctuations in Patient Care.* Traditional models ignore fluctuations in patient care during the shift. Some patients, such as expectant mothers, require minimal care for part of a shift but require significant care at other times during the shift. The stochastic programming model considers when patients require care.
- *Differences in Nurses.* Traditional models ignore the different skills of the nurses. Many of them use a targeted amount of time to perform certain tasks instead of an average time to complete the task. Some targets may be realistic for some nurses but unrealistic for others. The stochastic programming model considers the skills of each nurse individually.

In addition to the formulation of a new nurse assignment, this paper contributes a Benders' decomposition approach to solve it. We develop an optimal greedy algorithm to solve the recourse subproblem. We demonstrate the effectiveness of the model and algorithms with a computational study based upon data from a medical surgery unit at Baylor Regional Medical Center in Grapevine, Texas, that compares our methods with current approaches.

2 Model Description

In Section 1, we described four stages of nurse planning, and in this section we present a stochastic programming model for the final stage, nurse assignment.

2.1 *Model Assumptions*

Prior to the beginning of a shift, a charge nurse assigns each patient to an RN or an LVN. Although patients can usually be nursed by either type of nurse, state regulations can preclude them from performing certain patient care. Furthermore, some states, such as Texas, require that every patient be assessed by an RN within any 24-hour time period. Consequently, a charge nurse will assign RNs to patients who were assigned LVNs in the previous shift. We assume:

Assumption A1. A charge nurse determines which nurses can be assigned to which patients before optimizing nurse assignment.

Because patients enter and leave the hospital unit throughout a shift, nurse assignments are updated dynamically. However, revised nurse assignments often only include assigning a nurse to a new admission. Rarely is a patient reassigned a new nurse during the middle of a shift due to concerns for continuity of care. Hence, we make the following assumption:

Assumption A2. Nurse assignments are not changed, except when there are newly admitted patients.

Nurses distinguish between two types of patient care. *Direct care* is the amount of time nurses spend with patients, while *indirect care* is time spent on other tasks for patients, such as documentation of a patient's condition. In our stochastic programming model, we divide a nurse shift into several smaller *time periods*. The amount of direct and indirect care the patients require in each time period are given as parameters to the model. Nurses often provide indirect care throughout the shift, but direct care is often determined by a patient's condition, which is usually more urgent. Consequently, we make the following assumption:

Assumption A3. Direct care needs to be performed within the given time period, while indirect care can be performed in any time period from the given period until the end of the shift.

In addition to assumption A3, we assume nurses optimally allocate their indirect care to minimize excess workload. In some assignments, a nurse's patients will require more care than the nurse can provide. In such cases, a charge nurse, a nurse aide, or another nurse may assist the overworked nurse. However, an assignment requiring such assistance is undesirable. Implicitly, we presume that nurses receive assistance when absolutely necessary. The penalty of an assignment will be determined by a nondecreasing piecewise-linear function as shown in Figure 1. Because the function penalizes assignments with overworked nurses, an assignment will not include overworked nurses if such a solution exists. We describe the details of this function in Section 2.2.

During a shift patients may enter the hospital unit by admission from an emergency room, direct admission from a doctor, transferring from another unit, or birth. Patients may leave by discharge, transferring to another unit, or death. After a patient is discharged and his room has been cleaned and sterilized, a charge nurse may assign a newly admitted patient to the original patient's room. The charge nurse will often assign the nurse who cared for the recently discharged patient to the newly admitted patient. She can anticipate some of the patients that will be admitted because they are currently in another hospital unit. However, an *unanticipated patient* may enter a hospital unit during a shift without any warning prior to the shift. Unanticipated patients must be assigned a nurse, so we include them in the set of patients. We can represent an unknown number of patients by increasing the number of patients and randomly allowing their required care to be zero. Similarly, we can model random times for admissions and discharges. In this paper, we assume:

Assumption A4. The set of patients to be assigned includes potential unanticipated patients, so the number of patients is fixed.

2.2 *Stochastic Model for Nurse Assignment*

Let P and N be the sets of patients and nurses for a shift, respectively. We assume that a charge nurse determines which nurses can be assigned to which patients before optimizing patient assignment. For each patient $p \in P$, let $N(p)$ be the set of nurses which can be assigned to patient p. For each nurse $n \in N$, let $P(n)$ be the set of patients that can be assigned to nurse n; that is, $P(n) = \{p \in P | n \in N(p)\}$. For each patient $p \in P$, and nurse $n \in N(p)$, let *assignment variable*

$$X_{pn} = \begin{cases} 1 & \text{if patient } p \in P \text{ is assigned to nurse } n \in N(p), \\ 0 & \text{otherwise.} \end{cases}$$

Let Ξ be a set of random scenarios, and for each $\xi \in \Xi$, let ϕ^{ξ} be the probability that scenario ξ occurs.

A shift is divided into a set of time periods T. As the workload of a nurse increases in a time period $\tau \in T$, her patients receive less care, which is unsafe. We model the penalty for assigning workload to nurses as a monotonically nondecreasing piecewise linear function with k pieces, as shown in Figure 1. For each time period $\tau \in T$ and each nurse $n \in N$, let $A^{\xi}_{\tau ni}$ be the amount of workload assigned to nurse n between $m_{\tau ni}$ and $m_{\tau n(i+1)}$ in scenario $\xi \in \Xi$.

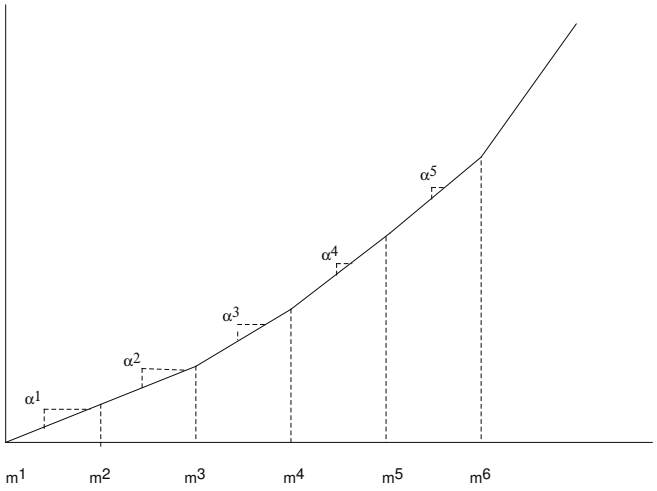

Figure 1: *The penalty for assigning workload to nurses is a monotonically nondecreasing piecewise linear function.*

Let $\alpha_{\tau ni}$ be the marginal penalty of $A_{\tau ni}$ for $1 \leq i \leq k$. Because the penalty is monotonically nondecreasing, $0 = m_{\tau n1} < \ldots < m_{\tau nk}$ and $0 \leq \alpha_{\tau n1} < \ldots < \alpha_{\tau nk}$. For notation, let $m_{\tau n(k+1)}$ be ∞. This penalty function is nondecreasing and piecewise linear, so the marginal penalty for assigning more patient care to an overworked nurse is greater than that of a nurse with less workload. Consequently, the function naturally balances the workload and allows nurses to provide better care. One special case of the penalty function has $k = 2$, $\alpha_{\tau n1} = 0$, $\alpha_{\tau n2} = 1$, and $m_{\tau n2}$ equal to the duration of the time period τ for each $\tau \in T$ and each $n \in N$. We refer to the value of variable $A^{\xi}_{\tau n2}$ as the *excess workload* on nurse n in time period τ and scenario ξ.

For each patient $p \in P$, each scenario $\xi \in \Xi$, and each $t \in T$, let d^{ξ}_{tp} be the amount of direct care required by patient p in time period t. Because patient p may be admitted or discharged during a shift, the patient care may vary dramatically throughout the shift. For each patient $p \in P$, each scenario $\xi \in \Xi$, and each time period $t \in T$, let e^{ξ}_{tp} be the amount of indirect care required by patient p at the beginning of time period t until the end of the shift. For each pair of time periods $(t, \tau) \in T \times T$, where $t \leq \tau$, and each nurse $n \in N$, let *indirect workload variable* $E^{\xi}_{t\tau n}$ be the total indirect care that can be performed during or after time period t and is performed in time period τ by nurse n. The amount of direct and indirect care the patients require in each time period are given as parameters to the model.

The stochastic programming model for patient assignment (SPA) is formulated as

$$\min \sum_{\xi\in\Xi}\sum_{n\in N}\sum_{\tau\in T}\sum_{i=1}^{k}\phi^{\xi}\alpha_{\tau ni}A^{\xi}_{\tau ni} \tag{1}$$

$$\sum_{n\in N(p)} X_{pn} = 1 \qquad \forall p\in P, \tag{2}$$

$$\sum_{p\in P(n)} e^{\xi}_{tpn}X_{pn} = \sum_{\tau=t}^{|T|} E^{\xi}_{t\tau n} \qquad \forall t\in T, n\in N, \xi\in\Xi, \tag{3}$$

$$\sum_{p\in P(n)} d^{\xi}_{\tau pn}X_{pn} + \sum_{t=1}^{\tau} E^{\xi}_{t\tau n} = \sum_{i=1}^{k} A^{\xi}_{\tau ni} \qquad \forall \tau\in T, n\in N, \xi\in\Xi, \tag{4}$$

$$X_{pn}\in\{0,1\} \qquad \forall p\in P(n), n\in N, \tag{5}$$

$$E^{\xi}_{t\tau n}\geq 0 \qquad \forall t,\tau\in T, t\leq\tau, n\in N, \xi\in\Xi, \tag{6}$$

$$m_{\tau n(i+1)} - m_{\tau ni} \geq A^{\xi}_{\tau ni}\geq 0 \qquad \forall \tau\in T, 1\leq i\leq k, n\in N, \xi\in\Xi. \tag{7}$$

Objective (1) minimizes the workload penalty on nurses. The first constraint set—the *nurse assignment constraints* (2) ensure that every patient is assigned to a nurse. The *indirect care constraints* in set (3) determine the total indirect care performed by nurse n from the beginning of time period t until the end of the shift. For each time period $\tau\in T$, the workload of nurse $n\in N$ consisting of direct care and indirect care is defined by a *workload constraint* in set (4). Constraint set (5) requires that the assignment variables be binary, and set (6) ensures the indirect care variables are nonnegative. Constraints (7) give the upper and lower bounds on the marginal workload variables. Observe that for each $\tau\in T, n\in N$, $A_{\tau nk}$ has no upper bound since $m_{\tau n(k+1)}=\infty$.

The following proposition is obvious

Proposition 1. *Let (X^*, A^*, E^*) be an optimal solution to SPA. Then for each $\xi\in\Xi$, $\tau\in T$, $n\in N$, there exists a positive integer $l\leq k$ such that*

$$A^{\xi^*}_{\tau ni} = \begin{cases} m_{\tau n(i+1)} - m_{\tau ni} & 1\leq i<l, \\ \sum_{j=1}^{i} A^{\xi^*}_{\tau nj} - m_{\tau ni} & i=l, \\ 0 & l<i\leq k. \end{cases} \tag{8}$$

Given an assignment $\overline{X}$, the constraints in (3), (4), (6), and (7) can be decomposed by nurse and scenario resulting in $|N|\times|\Xi|$ recourse subproblems. In the next section, we implement a Benders' decomposition to solve SPA. Although typical real-world problems cannot be solved to optimality within 30 minutes, the remainder of this paper focuses on finding a good solution within the time limit.

3 Algorithmic Approach

In this section, we present a Benders' decomposition approach to solve SPA. Moreover, we develop an optimal greedy algorithm for solving the recourse subproblems, and then we discuss patient-to-nurse ratio constraints to improve computational efficiency.

3.1 *Benders Decomposition*

Solving SPA with many scenarios and many time periods using branch and bound may be time consuming. However, two-stage stochastic programming models, like SPA, have a block angular structure that is appropriate for mathematical decomposition. The standard L-shaped method, based upon Benders' decomposition, is the most common solution approach for two-stage stochastic programming problems [6, 11]. Applying Benders' decomposition to SPA, the master problem assigns nurses to patients, and each recourse problem penalizes the assigned workload. Not only does SPA decompose by scenario like the standard L-shaped method, but it also decomposes by nurse into $|N| \times |\Xi|$ linear programming subproblems. Therefore, the subproblems are even more manageable than the standard L-shaped method, which only decomposes by scenario. Let $\overline{X}$ be a given assignment. For each $t \in T$, let $\overline{e}^{\xi}_{tn} = \sum_{p \in P(n)} e^{\xi}_{tpn} \overline{X}_{pn}$, and let $\overline{d}^{\xi}_{tn} = \sum_{p \in P(n)} d^{\xi}_{tpn} \overline{X}_{pn}$. The *primal subproblem* (PS^{ξ}_n) for each nurse $n \in N$ and each scenario $\xi \in \Xi$ is given by

$$\min \sum_{\tau \in T} \sum_{i=1}^{k} \alpha_{\tau ni} A^{\xi}_{\tau ni} \tag{9}$$

$$\sum_{\tau=t}^{|T|} E^{\xi}_{t\tau n} = \overline{e}^{\xi}_{tn} \qquad \forall t \in T, \tag{10}$$

$$\sum_{i=1}^{k} A^{\xi}_{\tau ni} - \sum_{t=1}^{\tau} E^{\xi}_{t\tau n} = \overline{d}^{\xi}_{\tau n} \qquad \forall \tau \in T, \tag{11}$$

$$(A^{\xi}_n, E^{\xi}_n) \text{ satisfy (6) and (7).}$$

In the primal subproblem, the workload variables $A^{\xi}_{\tau ni}$ are obtained, and the indirect care variables $E^{\xi}_{t\tau n}$ determine the time periods in which indirect care is performed. Observe that the solution $(\tilde{A}, \tilde{E})$ is feasible when $\tilde{E}^{\xi}_{ttn} = \overline{e}^{\xi}_{tn}$ and $\tilde{A}^{\xi}_{tnk} = \tilde{E}^{\xi}_{ttn} + \overline{d}^{\xi}_{tn}$ for all $t \in T$, and all other variables are zero.

Each primal subproblem PS^{ξ}_n can be formulated as a network flow problem, as depicted in Figure 2. Consider a directed network $G = (\mathcal{N}, \mathcal{A})$ with node set $\mathcal{N}$ and arc set $\mathcal{A}$, in which $|\mathcal{N}| = (2+k)|T| + 1$ and $|\mathcal{A}| = |T|(|T|+1)/2 + 2k|T|$. The network includes four types of nodes—t nodes (the left nodes in Figure 2), t' nodes (the middle-left nodes), $t'i$ nodes (the middle-right nodes), and a sink node (the right node labeled s). For each time period $t \in T$, a t node with supply $\overline{e}^{\xi}_{tn}$ and a t' node with supply $\overline{d}^{\xi}_{tn}$ are in $\mathcal{N}$. An arc between t and t' nodes is in $\mathcal{A}$ whenever $t \leq t'$, and the flow on this arc represents the value of variable $E^{\xi}_{tt'n}$ in the primal subproblem PS^{ξ}_n. For each $t \in T$ and each $i = 1, \ldots, k$, a $t'i$ node is added to $\mathcal{N}$, and an arc from the t' node to the $t'i$ node is included in $\mathcal{A}$. The flow on the arc

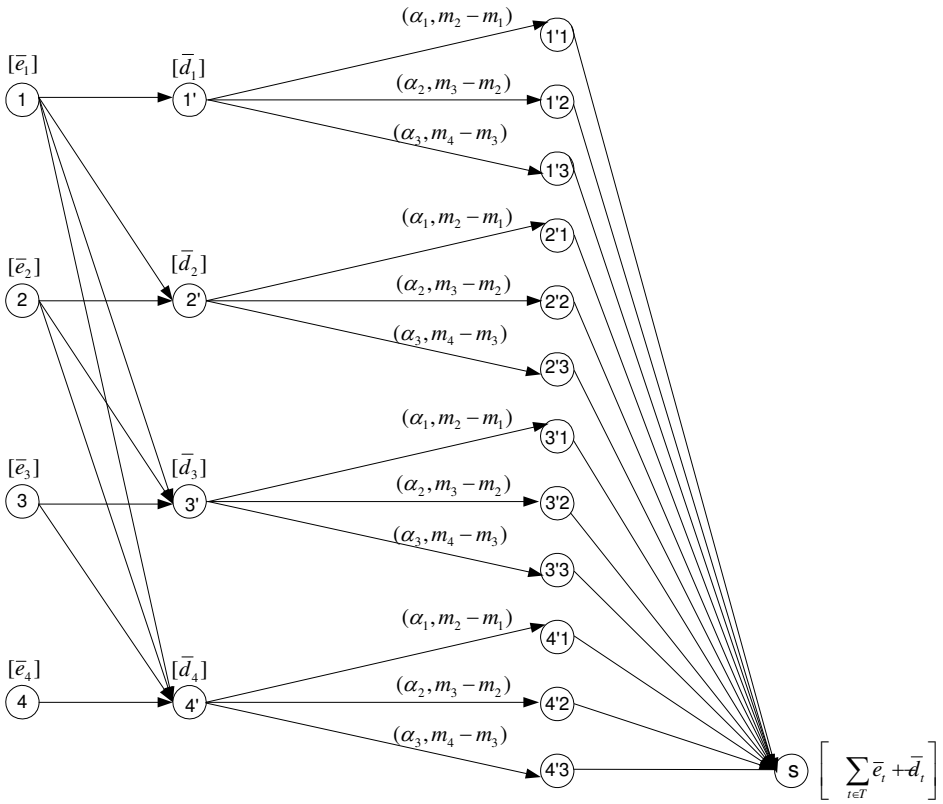

Figure 2: *The network flow primal subproblem*

from the t' node to the $t'i$ node is the value of the variable $A^{\xi}_{t'in}$, so it has a per unit cost of $\alpha_{t'ni}$ and an upper bound of $m_{t'n(i+1)} - m_{t'ni}$. A sink node with a demand of $\sum_{t\in T} \overline{d}^{\xi}_{tn} + \overline{e}^{\xi}_{tn}$ is used, and arcs from the $t'i$ nodes to the sink node are in $\mathcal{A}$.

Let π^{ξ}_{tn}, $Y^{\xi}_{\tau n}$, and $\rho^{\xi}_{\tau ni}$ be the dual variables associated with constraint sets (10) and (11) and the upper bounds in set (7), respectively. The *dual subproblem* (DS^{ξ}_n) is

$$\max \sum_{t\in T} \left[\sum_{i=1}^{k} (m_{ti} - m_{t(i+1)}) \rho^{\xi}_{tni} \right] + \overline{e}_t \pi^{\xi}_t + \overline{d}_t Y^{\xi}_{tn} \tag{12}$$

$$Y^{\xi}_{\tau n} - \rho^{\xi}_{\tau ni} \le \alpha_{\tau i} \qquad \forall \tau \in T, 1 \le i \le k, \tag{13}$$

$$\pi^{\xi}_{tn} \le Y^{\xi}_{\tau n} \qquad \forall t, \tau \in T, t \le \tau, \tag{14}$$

$$\rho^{\xi}_{\tau ni} \ge 0 \qquad \forall \tau \in T, 1 \le i \le k, \tag{15}$$

$$\pi^{\xi}_{tn}, Y^{\xi}_{\tau n} \text{ free} \qquad \forall t, \tau \in T. \tag{16}$$

The solution $(\tilde{\pi}^{\xi}_n, \tilde{Y}^{\xi}_n, \tilde{\rho}^{\xi}_n) = 0$ is always feasible, so both the primal and dual subproblems have optimal solutions.

Let DS be the combination of all dual subproblems DS^{ξ}_n over all nurses and scenarios. Let Δ be the set of extreme points for the dual subproblem DS. The original SPA problem is reformulated as follows:

$$\min \eta \tag{17}$$

$$\eta \ge \sum_{n\in N} \sum_{\xi\in\Xi} \sum_{t\in T} \phi^{\xi} \left[\sum_{p\in P(n)} \left(\tilde{\pi}^{\xi}_{tn} e_{tpn} + \tilde{Y}^{\xi}_{tn} d_{tpn} \right) X_{pn} + \sum_{i=1}^{k} (m_{tni} - m_{tn(i+1)}) \tilde{\rho}^{\xi}_{\tau ni} \right] \quad \forall (\tilde{\pi}, \tilde{Y}, \tilde{\rho}) \in \Delta, \tag{18}$$

where X_{pn} satisfy (2) and (5).

Dual extreme rays are ignored because the dual subproblem is always feasible.

The L-shaped method is described as Algorithm 1. On each iteration, we consider a subset of dual extreme points $\overline{\Delta} \subseteq \Delta$, and let constraint set (18') be the subset of (18) over $\overline{\Delta}$. We solve a restricted master problem (2), (5), (17), and (18') to find an assignment $\overline{X}$ and an anticipated objective value $\overline{\eta}$. Using the assignment $\overline{X}$, we solve the dual subproblem over all of the nurses and scenarios to obtain $(\tilde{\pi}, \tilde{Y}, \tilde{\rho})$. If the anticipated objective value $\overline{\eta}$ is less than the objective value of the dual solution $(\tilde{\pi}, \tilde{Y}, \tilde{\rho})$, then we add a Benders' optimality cut to (18'). Otherwise, the algorithm terminates and the assignment $\overline{X}$ is optimal. In the next section, we solve the subproblems.

3.2 *Greedy Algorithm*

In this section, we present a greedy algorithm to evaluate the recourse function PS^{ξ}_n. Properties of solutions by the greedy algorithm are stated, and we prove that the greedy algorithm is optimal. Finally, we describe how to find a complimentary optimal dual solution. To simplify notation, we ignore the superscript ξ and the subscript n.

The greedy algorithm works under the following reasonable assumption:

Algorithm 1 Nurse Assignment Benders Decomposition Algorithm (NABDA).

$\overline{\Delta} \leftarrow \emptyset$, $STOP \leftarrow FALSE$.

while $STOP = FALSE$ **do**

Solve the restricted master problem (2), (5), (17), and (18') to obtain an assignment $\overline{X}$ and an anticipated objective value $\overline{\eta}$. (On the first iteration, let $\overline{\eta} \leftarrow -\infty$, and let $\overline{X}$ be a feasible assignment.)

for all $n \in N$, $\xi \in \Xi$ **do**

Solve the dual subproblem (D_n^ξ) to obtain extreme point $(\tilde{\pi}_n^\xi, \tilde{Y}_n^\xi, \tilde{\rho}_n^\xi)$.

end for

if $\overline{\eta} < \sum_{p \in P} \sum_{n \in N(P)} \sum_{\xi \in \Xi} \sum_{t \in T} \phi^\xi \left[\left(\tilde{\pi}_{tn}^\xi e_{tpn} + \tilde{Y}_{tn}^\xi d_{tpn} \right) \overline{X}_{pn} + \sum_{i=1}^{k} (m_{tni} - m_{tn(i+1)}) \tilde{\rho}_{\tau ni}^\xi \right]$ **then**

$\overline{\Delta} \leftarrow \overline{\Delta} \cup \left\{ (\tilde{\pi}, \tilde{Y}, \tilde{\rho}) \right\}$, where $(\tilde{\pi}, \tilde{Y}, \tilde{\rho})$ is the combination of the vectors $(\tilde{\pi}_n^\xi, \tilde{Y}_n^\xi, \tilde{\rho}_n^\xi)$.

else

$STOP \leftarrow TRUE$.

end if

end while

Assumption A5 The nondecreasing piecewise linear penalty is the same for each time period; that is, $\alpha_{1i} = \alpha_{2i} = \cdots = \alpha_{|T|i}$ and $m_{1i} = m_{2i} = \cdots = m_{|T|i}$ for all $i = 1, \ldots, k$,

The intuitive explanation for Assumption A5 is that workload is equally penalized throughout a shift.

Consider the greedy algorithm (GAPS) for solving the primal subproblem PS, displayed as Algorithm 2. GAPS uses a solution $(\tilde{A}, \tilde{E})$ that satisfies constraints in (6), (7), and (11), and it increases $\tilde{A}$, $\tilde{E}$, and the objective value as little as possible until constraints in set (10) are satisfied. First, GAPS introduces a counter $l(\tau)$ such that a marginal increase in workload for time period τ will increase the objective value by $\alpha_{l(\tau)}$. All direct care is assigned to its given time period, and $\tilde{A}$ is increased appropriately. On every iteration, GAPS considers the time periods in which some indirect care on or prior to these time periods is unassigned. Among these time periods, GAPS examines those with the smallest counter l (equivalently the least marginal penalty α), and it selects the latest such time period τ. Then GAPS finds the latest time period $t \leq \tau$ that has remaining unassigned indirect care. Next $\tilde{A}_{\tau l(\tau)}$ and $\tilde{E}_{t\tau}$ are increased until either $\tilde{A}_{\tau l(\tau)}$ reaches its upper bound (7) or all indirect care from time period t is assigned. The counter $l(\tau)$ is incremented if $\tilde{A}_{\tau l(\tau)}$ is increased to its upper bound.

Theorem 2. *GAPS finds an optimal solution* $(\tilde{A}, \tilde{E})$.

The proof Theorem 2 is given in the appendix. We now describe a dual solution $(\tilde{\pi}, \tilde{Y}, \tilde{\rho})$ to DS that is complimentary

Algorithm 2 Greedy Algorithm for the Primal Subproblem (GAPS).

for all $\tau \in T$ **do**

Let the counter $l(\tau)$ be such that $m_{\tau l(\tau)} \leq \overline{d}_\tau < m_{\tau l(\tau)+1}$.

$$\tilde{A}_{\tau i} \leftarrow \begin{cases} m_{\tau(i+1)} - m_{\tau i} & 1 \leq i < l(\tau), \\ \overline{d}_\tau - m_{\tau i} & i = l(\tau), \\ 0 & l(\tau) < i \leq k. \end{cases}$$

$\tilde{E}_{t\tau} \leftarrow 0, \forall t \leq \tau$

end for

while $\sum_{\widetilde{\tau}=t}^{|T|} \tilde{E}_{t\widetilde{\tau}} < \overline{e}_t, \forall t \in T$ **do**

$\tau \leftarrow \max\left\{\arg\min_{\widehat{\tau} \in T}\left\{l(\widehat{\tau}) \,\middle|\, \exists \widehat{t} \leq \widehat{\tau}, \sum_{\overline{\tau}=\widehat{t}}^{|T|} \tilde{E}_{\widehat{t}\overline{\tau}} < \overline{e}_{\widehat{t}}\right\}\right\}$.

$t \leftarrow \max\left\{\widehat{t} \in T \,\middle|\, \widehat{t} \leq \tau, \sum_{\widehat{\tau}=\widehat{t}}^{|T|} E_{\widehat{t}\widehat{\tau}} < \overline{e}_{\widehat{t}}\right\}$.

$\delta \leftarrow \min\left\{\overline{e}_t - \sum_{\widetilde{\tau}=t}^{|T|} \tilde{E}_{t\widetilde{\tau}}, m_{\tau(l(\tau)+1)} - \tilde{A}_{\tau l(\tau)}\right\}$.

$\tilde{A}_{\tau l(\tau)} \leftarrow \delta + \tilde{A}_{\tau l(\tau)}$

$\tilde{E}_{t\tau} \leftarrow \delta + \tilde{E}_{t\tau}$

if $\tilde{A}_{\tau l(\tau)} = m_{\tau(l(\tau)+1)}$ **then**

$l(\tau) \leftarrow l(\tau) + 1$.

end if

end while

to a solution from GAPS $(\tilde{A}, \tilde{E})$. The complementary slackness conditions of PS and DS are

$$(-A_{\tau i} - m_{\tau i} + m_{\tau(i+1)})\rho_{\tau i} = 0 \qquad \forall i = 1, \ldots, k, \forall \tau \in T, \tag{19}$$

$$(\alpha_{\tau i} - Y_\tau + \rho_{\tau i})A_{\tau i} = 0 \qquad \forall i = 1, \ldots, k, \forall \tau \in T, \tag{20}$$

$$(\pi_t - Y_\tau)E_{t\tau} = 0 \qquad \forall t, \tau \in T, t \leq \tau. \tag{21}$$

For each time period τ, consider the following two sets of time periods:

$$\mathcal{T}(\tau) = \{\tilde{\tau} \in T | \exists t_1, \ldots, t_{q-1}, \tau_1, ..., \tau_q, \tau_1 = \tau, \tau_q = \tilde{\tau}, t_1 \leq \tau_2, t_2 \leq \tau_3, t_{q-1} \leq \tau_q, \\ \tilde{E}_{t_1\tau_1}, \tilde{E}_{t_2\tau_2}, ..., \tilde{E}_{t_{q-1}\tau_{q-1}} > 0\} \cup \{\tau\}, \tag{22}$$

$$\mathcal{T}^{-1}(\tau) = \{t | E_{t\tilde{\tau}} > 0, \forall \tilde{\tau} \in \mathcal{T}(\tau)\}. \tag{23}$$

Let time period $\tilde{\tau} \in \mathcal{T}(\tau)$. Consider the dual solution $(\tilde{\pi}, \tilde{Y}, \tilde{\rho})$ given by

$$\tilde{Y}_\tau = \begin{cases} \min_{\tilde{\tau} \geq \min \mathcal{T}^{-1}(\tau)} \{\alpha_{l(\tilde{\tau})}\} & \text{if } \mathcal{T}^{-1}(\tau) \neq \emptyset \\ \alpha_{l(\tau)} & \text{otherwise} \end{cases} \qquad \forall \tau \in T, \tag{24}$$

$$\tilde{\pi}_t = \min_{\tau \geq t} \tilde{Y}_\tau \qquad \forall t \in T, \tag{25}$$

$$\tilde{\rho}_{\tau i} = \max\{\tilde{Y}_\tau - \alpha_i, 0\} \qquad \forall i = 1, \ldots, k, \forall \tau \in T. \tag{26}$$

Theorem 3. *Let $(\tilde{A}, \tilde{E})$ be an optimal solution from GAPS. The dual solution given by (24)—(26) is a complimentary optimal dual solution.*

The proof of Theorem 3 is similarly in the appendix.

3.3 *Patient-to-Nurse Ratio Constraint*

Many states limit the number of patients that can be assigned to a nurse for certain units in a hospital. For instance, California mandates nurse-to-patient ratio regulations that allow no more than six patients assigned to any one nurse for medical/surgical unit [10]. Typically, the total number of nurses for a shift are obtained from the nurse rescheduling stage. Based upon the number of nurses and the number of patients, we introduce the following *patient-to-nurse ratio constraints*

$$\sum_{p \in P(n)} X_{pn} \leq \left\lceil \frac{|P|}{|N|} \right\rceil \qquad \forall n \in N, \tag{27}$$

where $\lceil x \rceil$ represents the ceiling of a value x. Constraint set (27) serves the two important purposes:

- It reduces the feasible region of SPA and improves the performance of the algorithm.
- It prevents an assignment with an uneven number of patients, which would not be popular with the nurses even if it were balanced in terms of required care.

Although constraint set (27) can lead to suboptimal solutions, we were unable to construct such a solution in practice.

Instance	Shift	Pat	RN	LVN
1	Day	19	2	1
2	Day	15	4	0
3	Evening	15	2	1
4	Night	11	1	1

Table 1: *Instances generated from Baylor data over ten months*

4 Computational Study

In this section, we provide a computational study on nurse assignment. Problem instances were generated based upon data from Baylor Regional Medical Center in Grapevine, Texas as described in Section 4.1. These instances, however, cannot be solved exactly within 30 minutes. Consequently, the focus of the computational study is to find good solutions within the time limit. Moreover, we considered several alternative assignment methods. Finally, we compared the solutions from these methods with those from executing the Benders' approach for 30 minutes.

4.1 *Problem Instances*

Each nurse at Baylor wears a badge that locates the nurse in the hospital unit. The purpose of the locator is so a charge nurse can inform a nurse immediately when one of her patients calls the nurses' station. The locator system stores data on the location of the nurses for one month. In addition to these data, Baylor provided encrypted patient data for a medical surgery unit to study for this research from March 2004 - December 2004.

We generated four random instances based upon these data. The first two instances were day shifts from 8:00 AM to 4:00 PM, while instances 3 and 4 were evening and night shifts from 4:00 PM to 12:00 AM and 12:00 AM to 8:00 AM, respectively. Sundaramoorthi et al. [27] noted that patients' diagnoses and locations are the most significant factors affecting the amount of time nurses spend with patients. For each instance, we sampled a random set of patients from an empirical distribution of patients with similar diagnoses and patient rooms. We used a census matrix from a medical/surgical unit to determine the number and type of nurses for the shift. Although we assumed nurses in our computational experiments were identical, SPA allows for nurses with different skills. Table 1 displays characteristics of the four instances. The column labelled "instance" is the random instance, "shift" is the time of the shift, "pat" is the number of patients, and "RN" and "LVN" are the number of registered and licensed vocational nurses on duty, respectively.

We partitioned the shift into eight one-hour time periods for T, and we generated 100, 200, 500, and 3000 random scenarios for Ξ. For each time period $\tau \in T$, each patient $p \in P$, and each scenario $\xi \in \Xi$, the direct care $d^{\xi}_{\tau p}$, was sampled from a gamma distribution. Each gamma distribution was fitted by the moment estimator method [35] from

the amount of time during time period τ that nurses were in the rooms of patients with diagnoses and rooms similar to those of patient p. Because indirect care can be performed in several locations, it cannot be estimated from the data from Baylor. However, in some patient classification systems for similar medicial/surgical units, total indirect care is 32% of direct care. Consequently, we estimated indirect care $e^{\xi}_{\tau p} = 0.32 \times d^{\xi}_{\tau p}$, $\forall \tau \in T$, $\forall p \in P$, and $\forall \xi \in \Xi$. In addition, we implicitly assumed that that direct care engenders indirect care.

4.2 *Alternative Assignments*

In this section, we describe several alternative approaches to finding an assignment. With many scenarios, stochastic integer programming problems are often computationally intractable, but the Mean Value Problem (MVP) often provides a good solution [6]. For each of the four instances from Section 4.1, we replaced the direct and indirect care random variables with their mean, and we solved the deterministic integer programming problem. In all four instances, solving MVP required less than one minute of CPU time, so finding a good solution is computationally tractable.

In addition to MVP, we also used a heuristic that balanced workload based upon the expected total required care of the patients. When the number of nurses divides the number of patients evenly, the heuristic assigns the patients with the greatest and least required care time to the same nurse. Otherwise, the heuristic assigns the patients with greatest required care to the nurses who are assigned to fewer patients. Finally, we randomly divided the patients evenly among the nurses without considering workload.

In practice, charge nurses often intuitively assign patients to nurses. More sophisticated hospitals use patient classification systems that only consider the expected total care and ignore the fluctuations and uncertainty of care. Consequently, assignments in practice are often similar to those of the heuristic or random assignment.

4.3 *Computational Results*

In this section, we compare the performance of assignments from five different assignment methods. We present the appropriate numbers of scenarios for solving SPA. We study the implementation of the patient-to-nurse ratio constraints. Finally, the efficiency of using the GAPS versus the simplex method to solve the recourse subproblems is discussed.

We evaluated the performance of five different nurse assignment methods–the random assignment method, the heuristic, the mean value problem solution (MVP), and solving SPA with and without using the Benders' approach, denoted as SPA-IP and SPA-BA, respectively. If a method required more than 30 minutes to solve, we considered the best solution found within the time limit.

Table 2 compares the expected excess workload for assignments given five different assignment approaches. MVP, SPA-IP, and SPA-BA were implemented in ANSI C and processed by a Dual 3.06-GHz Intel Xeon Workstation using CPLEX 8.0 software. All assignments from SPA-IP and SPA-BA were obtained by optimizing the four patient

instances with 500 scenarios. To find an initial solution for SPA-BA, we used the MVP for less than one minute and then used SPA-BA for the remaining time. In other studies we found that averages estimated under 3000 scenarios were within one minute of the true mean. Consequently, after having obtained solutions from each approach, GAPS calculated the excess workload of each assignment in each problem instance with 3000 scenarios. Table 2 displays the expected workload in minutes, the average excess workload, and the average excess workload as a percentage of the expected workload minutes.

In all four instances, SPA-BA found the best solution within the time limit. Assignments from SPA-BA reduced the average excess workload for nurses between 2 minutes and 18 minutes over the random assignment, up to 15 minutes over the heuristic assignment, and up to 13 minutes over MVP. Considering there are 1095 8-hour shifts per year, SPA-BA could save up to 273 hours of excess workload each year in each unit of a hospital. Thus, a nurse-assignment decision-support system that used SPA-BA would reduce the burden of the nursing shortage.

We examined the number of scenarios that gives the best SPA results. We obtained assignments by optimizing based upon the four patient instances with 100, 200, and 500 scenarios and evaluated those assignments with 3000 scenarios with GAPS. Table 3 compares the average excess workload and the percentage of average excess workload to the expected workload minutes of optimizing SPA-IP and SPA-BA with different numbers of scenarios. Optimizing using SPA-BA with 500 scenarios found solutions within one minute of solution quality of the best known solution in each of the problem instances. Therefore, we used SPA-BA and SPA-IP with 500 scenarios in the remainder of this study.

The computational effects of applying patient-to-nurse ratio constraints are in Table 4. Assignments were obtained by solving MVP, SPA-IP, and SPA-BA with and without the patient-to-nurse ratio constraints. Solving SPA-IP and SPA-BA were also based upon optimizing instances with 500 scenarios and evaluating those assignments with 3000 scenarios. Table 4 shows the average excess workload of assignments from the three methods with and without the ratio constraints. Adding the patient-to-nurse ratio constraints to MVP and SPA-IP reduces average excess workload. For both MVP and SPA-IP, there is only one problem instance in which the ratio constraints weakened the solution quality. Thus, the patient-to-nurse ratio constraints improve overall performance of MVP and SPA-IP. SPA-BA provided good solutions without patient-to-nurse ratio implying that solving SPA with only Benders decomposition algorithm provides well-balanced patient loads for nurses.

We compared the efficiencies of GAPS and the simplex method. Assignments were obtained by optimizing SPA-BA with the four patient instances using both GAPS and CPLEX 8.0 to solve the linear subproblems. Table 5 displays the average excess workload and the number of cuts added to the restricted master problem by solving subproblems with GAPS and simplex. We can add more cuts using GAPS than using simplex in all instances, suggesting that GAPS is faster than the simplex method. In three of the four patient instances, adding more Benders optimality cuts to the restricted master problem improved the quality of solutions. GAPS is computationally efficient because it is faster than a current commercial linear programming solver. Applying GAPS to problems also offers potentially better solutions.

Instance	Algorithm	Expected Total Workload	Average Excess Workload	Percent
1	Random	1136	51.0	4.49
1	Heuristic	1136	50.0	4.40
1	MVP	1136	48.4	4.26
1	SPA-IP	1136	39.9	3.51
1	SPA-BA	1136	35.4	3.11
2	Random	1083	36.9	3.41
2	Heuristic	1083	31.2	2.88
2	MVP	1083	37.5	3.46
2	SPA-IP	1083	24.8	2.29
2	SPA-BA	1083	24.1	2.23
3	Random	927	59.0	6.36
3	Heuristic	927	47.9	5.17
3	MVP	927	51.4	5.54
3	SPA-IP	927	43.0	4.64
3	SPA-BA	927	40.9	4.41
4	Random	368	8.1	2.21
4	Heuristic	368	6.1	1.65
4	MVP	368	5.9	1.61
4	SPA-IP	368	6.4	1.73
4	SPA-BA	368	5.8	1.58

Table 2: *The computational results comparing solutions from 5 methods on instances 1, 2, 3, and 4*

Instance	Algorithm	Expected Patient Workload	100 scenarios optimized		200 scenarios optimized		500 scenarios optimized	
			Average Excess Workload	%	Average Excess Workload	%	Average Excess Workload	%
1	SPA-IP	1136	35.6	3.13	35.5	3.12	39.9	3.51
1	SPA-BA	1136	34.2	3.01	34.4	3.03	35.4	3.11
2	SPA-IP	1083	24.5	2.26	25.4	2.35	24.8	2.29
2	SPA-BA	1083	25.7	2.37	25.8	2.39	24.1	2.23
3	SPA-IP	927	42.8	4.61	41.4	4.46	43.0	4.64
3	SPA-BA	927	41.8	4.51	40.8	4.40	40.9	4.41
4	SPA-IP	368	5.8	1.58	5.9	1.61	6.4	1.73
4	SPA-BA	368	5.8	1.57	5.9	1.61	5.8	1.58

Table 3: *The computational results comparing average excess workload from solving SPA-IP and SPA-BA with different numbers of scenarios*

Instance	Algorithm	Average Excess Workload with Ratio Constraints	Average Excess Workload without Ratio Constraints
1	MVP	48.4	49.4
1	SPA-IP	39.9	59.5
1	SPA-BA	35.4	34.0
2	MVP	37.5	42.8
2	SPA-IP	24.8	28.7
2	SPA-BA	24.1	24.2
3	MVP	51.4	46.9
3	SPA-IP	43.0	89.0
3	SPA-BA	40.9	42.3
4	MVP	5.9	13.9
4	SPA-IP	6.4	5.8
4	SPA-BA	5.8	5.8

Table 4: *The computational results comparing average excess workload from 3 methods with and without patient-to-nurse ratio constraints*

Instance	Algorithm	Average Excess Workload with GAPS	no. of cuts	Average Excess Workload with simplex	no. of cuts
1	SPA-BA	35.4	428	34.4	385
2	SPA-BA	24.1	739	24.8	668
3	SPA-BA	40.9	658	41.0	585
4	SPA-BA	5.8	14	5.8	12

Table 5: *The computational results comparing average excess workload from solving subproblems with GAPS and the simplex method*

5 Conclusions and Future Research

In this paper, we developed a two-stage stochastic integer programming model for nurse assignment (SPA) with a recourse penalty function to minimize excess workload for nurses. We employed the L-shaped method to solve our problem and demonstrated how it could save up to 273 hours of excess workload on nurses per year in each medical/surgical unit. However, decisions made in earlier stages of nurse planning can have a dramatic effect on nurse assignment. Solutions for early stages that anticipate their consequences on nurse assignment would likely further reduce the burden of the nursing shortage. One interesting topic of future research is to integrate some of the earlier stages of nurse planning, such as nurse rescheduling, with the stochastic programming model for nurse assignment.

6 Acknowledgement

We would like to thank Terry Clark from Baylor Medical Center at Grapevine TX, for providing us data for this research. We also thank Patricia G. Turpin from the School of Nursing at The University of Texas at Arlington for providing additional patient classification data.

References

[1] AARP, “A profile of older Americans: 2000,” http://research.aarp.org/general/profile_2000.pdf, 2000.

[2] U. Aickelin and K. A. Dowsland, “An indirect genetic algorithm for a nurse scheduling problem,” to appear in Computing and Operational Research, 2003.

[3] U. Aickelin and P. White, “Building better nurse scheduling algorithms,” submitted to Annals of Operations Research special issue on Rostering, 2002.

[4] R. N. Bailey, K. M. Garner, and M. F. Hobbs, “Using simulated annealing and genetic algorithms to solve staff scheduling problems,” Asia-Pacific Journal of Operational Research, vol. 14, pp. 27–433, 1997.

[5] J. F. Bard and H. W. Purnomo, “Hospital-wide reactive scheduling of nurses with preference considerations,” IIE transaction, vol. 35, pp. 589–608, 2005.

[6] J. R. Birge and F. Louveaux, Introduction to Stochastic Programming. New York, New York: Springer, 1997.

[7] P. I. Buerhaus, D. O. Staiger, and D. I. Auerbach, “Implications of an aging registered nurse workforce,” The Journal of the American Medical Association, vol. 283, pp. 2948–2954, 2000.

[8] E. K. Burke, P. D. Causmaecker, S. Petrovic, and G. V. Berghe, “Variable neighbourhood search for nurse rostering problems,” in Proceedings of 4th Metaheuristics International Conference, Porto, Portugal, pp. 755–760, 2001.

[9] E. K. Burke, P. Cowling, P. D. Caumaecker, and G. V. Berghe, “A memetic approach to the nurse rostering problem,” Applied Intelligence special issue on Simulated Evolution and Learning, vol. 15, pp. 199–214, 2001.

[10] California Department of Health Services, “State health director announces proposed changes to nurse-to-patient ratio regulations,” http://www.applications.dhs.ca.gov/pressreleases/store/PressReleases/04-70.html, 2004.

[11] C. C. Caroe and J. Tind, “L-shaped decomposition of two-stage stochastic programs with integer recourse,” Mathematical Programming, vol. 83, pp. 451–464, 1998.

[12] A. H. W. Chun, S. H. C. Chan, G. P. S. Lam, F. M. F. Tsang, J. Wong, and D. W. M. Yeung, “Nurse rosternig at the hospital authority of Hong Kong,” in Proceedings of the 11th Conference on Innovative Applications of Artificial Intelligence, Austin, TX, 2000.

[13] A. J. Dieck, “Forcasting in-patient admissions to public health facilities: a comparison of Box-Jenkins and Winters’ approaches,” in Proceedings of the Annual Pittsburgh Conference, pp. 1321–1325, ISA, 1984.

[14] GAO, “Nursing workforce: Emerging nurse shortage due to multiple factors,” Tech. Rep. GAO-01-944, General Accounting Office, 2001.

[15] B. Jaumard, F. Semet, and T. Vovor, “A generalized linear programming model for nurse scheduling,” European Journal of Operations Research, vol. 107, pp. 1–18, 1998.

[16] E. P. C. Kao and M. Queyranne, “Budgeting costs of nursing in a hospital,” Management Science, vol. 31, pp. 608–621, 1985.

[17] E. P. C. Kao and G. G. Tung, “Forecasting demands for inpatient services in a large public health care delivery system,” Socio-Economic Planning Science, vol. 14, pp. 97–106, 1980.

[18] M. P. Kirkby, “Moving to computerized schedules: A smooth transition,” Nurse Management, vol. 28, pp. 42, 44, 1997.

[19] G. Lynn, “The nursing shortage: Causes, impact and innovative remedies,” http://edworkforce.house.gov/hearings/107th/fc/nurses92501/lynn.htm, 2001. Testimony of the American Hospital Association before the United State House of Representatives Committee on Education and the Workforce.

[20] H. E. Miller, W. P. Pierskalla, and G. J. Rath, “Nurse scheduling using mathematical programming,” Operations Research, vol. 24, pp. 857–870, 1976.

[21] C. Mullinax and M. Lawley, “Assigning patients to nurses in neonatal intensive care,” Journal of the Operational Research Society, vol. 53, pp. 25–35, 2002.

[22] T. Osogami and H. Imai, “Classification of various neighborhood operations for the nurse scheduling problem,” to appear in The Institute of Statistical Mathematics Coorperative Research Report, 2000.

[23] J. Powers, “Accepting and refusing assignments,” Nursing Management, vol. 24, pp. 64–73, 1993.

[24] S. Shaha and C. Bush, “Fixing acuity: a professional approach to patient classification and staffing.,” Nursing Econ, vol. 14, pp. 346–356, 1996.

[25] S. P. Siferd and W. C. Benton, “Decision modes for shift scheduling of nurses,” European Journal of Operational Research, vol. 74, pp. 519–527, 1994.

[26] M. Stanton and M. K. Rutherford, “Hospital nurse staffing and quality of care,” Research in Action Issue 14 04-0029, Agency for Healthcare Research and Quality, 2004.

[27] D. Sundaramoorthi, V. C. P. Chen, and J. M. Rosenberger, “Knowledge discovery and mining for nurse activity and patient data,” IE Research Conference, Atlanta, GA, 2005.

[28] The Kaiser Family Foundation and Agency for Healthcare Research and Quality and Harvard School of Public Health, “National survey on consumers experiences with patient safety and quality information,” 2004.

[29] V. M. Trivedi, “Mixed-integer goal programming model for nursing service budgeting,” Operations Research, vol. 29, pp. 1019–1034, 1981.

[30] C. R. Victor and I. Higginson, “Effectiveness of care for older people: a review,” Quality Health Care, vol. 3, pp. 210–216, 1994.

[31] L. Walts and A. Kapadia, “Patient classification system: An optimization approach,” Health Care Management Review, vol. 21, pp. 75–82, 1996.

[32] D. M. Warner, "Forecasting the demand for nursing personnel and its optimal scheduling by ward," in Proceeding of a conference held in Cambridge, U.K., under the aegis of the NATO Science Committee and the NATO Advisory Panel on Operational Research, pp. 465–474, Engl Univ Press, 1974.

[33] D. M. Warner, "Scheduling nursing personnel according to nursing preferences: A mathematical programming approach," Operations Research, vol. 24, pp. 842–856, 1976.

[34] D. M. Warner and J. Prawda, "A mathematical programming model for scheduling nursing personnel in a hospital," Management Science, vol. 19, pp. 411–422, 1972.

[35] D. P. Wiens, J. Cheng, and N. C. Beaulieu, "A class of method of moments estimators for the two-parameter gamma family," Pakistan Journal of Statistics, vol. 19, pp. 129–141, 2003.

Appendix

In the appendix, we give some properties of GAPS solutions and prove Theorems 2 and 3. Let problem $PS(\overline{e}_1, \overline{d}_1, \ldots, \overline{d}_{|T|})$ be a special instance of PS in which $\overline{e}_t = 0$ for all $t = 2, \ldots, |T|$.

Lemma 4. *Let $(\tilde{A}, \tilde{E})$ be a solution found by GAPS on $PS(\overline{e}_1, \overline{d}_1, \ldots, \overline{d}_{|T|})$. Then $(\tilde{A}, \tilde{E})$ is an optimal solution to $PS(\overline{e}_1, \overline{d}_1, \ldots, \overline{d}_{|T|})$.*

Proof. Suppose to the contrary that $(\tilde{A}, \tilde{E})$ is not an optimal solution. Let $\tilde{l}(\tau)$, $\forall \tau \in T$ be the counters defined in GAPS. Let (A^*, E^*) be an optimal solution to $PS^{\xi}_n(\overline{e}_1, \overline{d}_1, \ldots, \overline{d}_{|T|})$ that minimizes the distance $||E^* - \tilde{E}||$. Let $l^*(\tau)$, $\forall \tau \in T$, be the counters defined in Proposition 1. If $A^*_{\tau i} = m_{i+1} - m_i$ and $A^*_{\tau(i+1)} = 0$, then $l^*(\tau) = i + 1$. Because $(A^*, E^*) \neq (\tilde{A}, \tilde{E})$ and $\sum_{\overline{\tau}=1}^{|T|} E^*_{1\tau} = \sum_{\overline{\tau}=1}^{|T|} \tilde{E}_{1\tau} = \overline{e}_1$, there exist time periods $\tilde{\tau}, \tau^* \in T$ such that $\overline{d}_{\tilde{\tau}} \leq \sum_{i=1}^{\tilde{l}(\tilde{\tau})} \tilde{A}_{\tilde{\tau}i} = \overline{d}_{\tilde{\tau}} + \tilde{E}_{1\tilde{\tau}} < \sum_{i=1}^{l^*(\tilde{\tau})} A^*_{\tilde{\tau}i} = \overline{d}_{\tilde{\tau}} + E^*_{1\tilde{\tau}}$ and $\overline{d}_{\tau^*} \leq \sum_{i=1}^{l^*(\tau^*)} A^*_{\tau^* i} = \overline{d}_{\tau^*} + E^*_{1\tau^*} < \sum_{i=1}^{\tilde{l}(\tau^*)} \tilde{A}_{\tau^* i} = \overline{d}_{\tau^*} + \tilde{E}_{1\tau^*}$. Now consider the following cases:

Case 1: Suppose $\overline{d}_{\tau^*} + \tilde{E}_{1\tau^*} \leq m_{\tilde{l}(\tilde{\tau})+1}$. Then $\overline{d}_{\tau^*} + E^*_{1\tau^*} < \overline{d}_{\tau^*} + \tilde{E}_{1\tau^*} \leq m_{\tilde{l}(\tilde{\tau})+1}$ and $\overline{d}_{\tilde{\tau}} + E^*_{1\tilde{\tau}} > \overline{d}_{\tilde{\tau}} + \tilde{E}_{1\tilde{\tau}} \geq m_{\tilde{l}(\tilde{\tau})}$. By Assumption A5, increasing $E^*_{1\tau^*}$ and decreasing $E^*_{1\tilde{\tau}}$ does not increase the objective value of (A^*, E^*). Consequently, it is not an optimal solution to $PS(\overline{e}_1, \overline{d}_1, \ldots, \overline{d}_{|T|})$ that minimizes the distance $||E^* - \tilde{E}||$.

Case 2: Suppose $\overline{d}_{\tau^*} + \tilde{E}_{1\tau^*} > m_{\tilde{l}(\tilde{\tau})+1}$. Consider the last iteration of GAPS in which $\tilde{E}_{1\tau^*}$ was increased. By the definition of GAPS, $\tilde{l}(\tau^*) \leq \tilde{l}(\tilde{\tau})$, so $\overline{d}_{\tau^*} + \tilde{E}_{1\tau^*}$ would have increased to at most $m_{\tilde{l}(\tilde{\tau})+1}$, in contradiction to the assumption that $\overline{d}_{\tau^*} + \tilde{E}_{1\tau^*} > m_{\tilde{l}(\tilde{\tau})+1}$.

Thus, $(\tilde{A}, \tilde{E})$ is an optimal solution to $PS(\overline{e}_1, \overline{d}_1, \ldots, \overline{d}_{|T|})$. □

Now consider two primal problems $PS(\overline{e}_1, \overline{d}_1, \overline{d}^1_2, \ldots, \overline{d}^1_{|T|})$ and $PS(\overline{e}_1, \overline{d}_1, \overline{d}^2_2, \ldots, \overline{d}^2_{|T|})$. Let primal problems $PS(\overline{d}^1_2, \ldots, \overline{d}^1_{|T|})$ and $PS(\overline{d}^2_2, \ldots, \overline{d}^2_{|T|})$ be special instances in which $\overline{e}_1 = \overline{d}_1 = 0$ and $\sum_{t=2}^{|T|} \overline{d}^1_t = \sum_{t=2}^{|T|} \overline{d}^2_t$ and let

$z_{PS(\overline{d}_2^1,\ldots,\overline{d}_{|T|}^1)}$ and $z_{PS(\overline{d}_2^2,\ldots,\overline{d}_{|T|}^2)}$ be their optimal objective values, respectively. Without loss of generality, suppose $z_{PS(\overline{d}_2^1,\ldots,\overline{d}_{|T|}^1)} < z_{PS(\overline{d}_2^2,\ldots,\overline{d}_{|T|}^2)}$.

Lemma 5. $z_{PS(\overline{e}_1,\overline{d}_1,\overline{d}_2^1,\ldots,\overline{d}_{|T|}^1)} \leq z_{PS(\overline{e}_1,\overline{d}_1,\overline{d}_2^2,\ldots,\overline{d}_{|T|}^2)}$.

Proof. Let (A^2, E^2) be an optimal solution to $PS(\overline{e}_1, \overline{d}_1, \overline{d}_2^2, \ldots, \overline{d}_{|T|}^2)$. Construct the following solution (A^1, E^1). Let the set of time periods $T^1 \subset T$ be such that $\forall t^1 \in T^1$, $\overline{d}_{t^1}^1 > \overline{d}_{t^1}^2 + E_{1t^1}^2$. For each time period $t \in T \setminus T^1$, increase the value of E_{1t}^1 such that $\overline{d}_t^1 + E_{1t}^1 = \overline{d}_t^2 + E_{1t}^2$. Since $\overline{e}_1 + \sum_{t\in T} \overline{d}_t^1 = \overline{e}_1 + \sum_{t\in T} \overline{d}_t^2 = \sum_{t\in T\setminus T^1} \overline{d}_t^2 + E_{1t}^2 + \sum_{t\in T^1} \overline{d}_t^2 + E_{1t}^2 < \sum_{t\in T\setminus T^1} \overline{d}_t^1 + E_{1t}^1 + \sum_{t\in T^1} \overline{d}_t^1$, then $\sum_{t\in T} E_{1t}^1 > \overline{e}_1$. Let $t^1 \in \arg\max\{\overline{d}_t^1 + E_{1t}^1 | E_{1t}^1 > 0\}$, let $l^1 = \max\{i = 1, \ldots, k | \overline{d}_t^1 + E_{1t}^1 > m_i\}$, and reduce E_{1t}^1 until either $E_{1t}^1 = 0$, $\overline{d}_t^1 + E_{1t}^1 = m_i$, or $\sum_{t\in T} E_{1t}^1 = \overline{e}_1$. Repeat the selection of t^1 and reduction of E_{1t}^1 until $\sum_{t\in T} E_{1t}^1 = \overline{e}_1$. Consider the subset of time periods $T^2 \subset T$ for which a time period $t^2 \in T^2$, $\overline{d}_{t^2}^2 > \overline{d}_{t^2}^1$. Reducing the most penalized $\overline{d}_{t^2}^2$ in time periods $t^2 \in T^2$ and increasing $\overline{d}_{t^1}^2$ in time periods $t^1 \in T^1$ does not increase the objective penalty because $z_{PS(\overline{d}_2^1,\ldots,\overline{d}_{|T|}^1)} < z_{PS(\overline{d}_2^2,\ldots,\overline{d}_{|T|}^2)}$. By definition $T^2 \subseteq T \setminus T^1$, so reducing $\overline{d}_{t^2}^1 + E_{1t^2}^1$ in the most penalized time periods $t^2 \in T \setminus T^1$ to account for $\sum_{t^1\in T^2} \overline{d}_{t^1}^1 - \overline{d}_{t^1}^2 + E_{1t^1}^2$ will not increase the objective penalty. Thus the objective function value of (A^1, E^1) is less than that of (A^2, E^2), so $z_{PS(\overline{e}_1,\overline{d}_1,\overline{d}_2^1,\ldots,\overline{d}_{|T|}^1)} \leq z_{PS(\overline{e}_1,\overline{d}_1,\overline{d}_2^2,\ldots,\overline{d}_{|T|}^2)}$. □

Consider the revised greedy algorithm (RGAPS) for PS as given by Algorithm 3.

Algorithm 3 Revised Greedy Algorithm (RGAPS)

$t \leftarrow |T|$.

while $t \geq 1$ **do**

Solve $PS(\overline{e}_t, \overline{d}_t, \ldots, \overline{d}_{|T|})$ using GAPS.

$\overline{d}_{\hat{t}} \leftarrow \overline{d}_{\hat{t}} + E_{t\hat{t}}, \forall \hat{t} = t, \ldots, |T|$.

$t \leftarrow t - 1$.

end while

Lemma 6. *Let $(\tilde{A}, \tilde{E})$ be a solution found by RGAPS. Then $(\tilde{A}, \tilde{E})$ is an optimal solution for PS.*

Proof. By induction and Lemmas 4 and 5, $(\tilde{A}, \tilde{E})$ is an optimal solution for PS. □

Lemma 7. *Let $(\tilde{A}, \tilde{E})$ be a solution found by GAPS. Let time period $\tau \in T$ be such that there exists time periods $t_1, t_2 \in T$, where $t_1 < t_2 \leq \tau$ and $\tilde{E}_{t_1\tau} > 0$ and $\tilde{E}_{t_2\tau} > 0$. Then GAPS increases $\tilde{E}_{t_2\tau}$ to its final value before it increases $\tilde{E}_{t_1\tau}$.*

Proof. Consider the first iteration in which $\tilde{E}_{t_1\tau}$ was increased. By the definition of GAPS, $\tilde{E}_{t_2\tau}$ would have been selected unless $\sum_{\tilde{\tau}=t_2}^{|T|} \tilde{E}_{t_2\tilde{\tau}} = \overline{e}_{t_2}$. Consequently, $\tilde{E}_{t_2\tau}$ must have been increased its final value before the iteration. □

Theorem 2. *GAPS finds an optimal solution* $(\tilde{A}, \tilde{E})$.

Proof. By Lemma 6, it remains to be proven that RGAPS and GAPS return equivalent solutions. Consider the following induction proof on the number of time periods $|T|$. (*Base Case*) For $|T| = 1$, RGAPS has one iteration, which uses GAPS, so they are equivalent algorithms. (*Induction Hypothesis*) Suppose RGAPS and GAPS are equivalent algorithms for a problem instance PS in which $|T| = \mathbf{T}$. Let $(A^{\mathbf{T}}, E^{\mathbf{T}})$ the optimal solution given by both algorithms with counters $l^{\mathbf{T}}(t)$, $\forall t \in T$. Consider an instance of PS in which $|T| = \mathbf{T}+1$ and $\overline{d}_{t+1}^{\mathbf{T}+1} = \overline{d}_t^{\mathbf{T}}$ and $\overline{e}_{t+1}^{\mathbf{T}+1} = \overline{e}_t^{\mathbf{T}}$, $\forall t = 1, \ldots, \mathbf{T}$. Let $(A^{\mathbf{T}+1}, E^{\mathbf{T}+1})$ be the solution given by GAPS. Let $\hat{\tau} \in T$ be such that $E_{1\hat{\tau}}^{\mathbf{T}+1} > 0$, and consider the iteration in which $E_{1\hat{\tau}}^{\mathbf{T}+1}$ was first increased. Prior to the iteration, $E_{\hat{t}\hat{\tau}}^{\mathbf{T}+1}$ had been increased to its final value and $\sum_{\overline{\tau}=\hat{t}}^{|T|} E_{\hat{t}\overline{\tau}}^{\mathbf{T}+1} = \overline{e}_{\hat{t}}^{\mathbf{T}+1}$ for all time periods $\hat{t} = 2, \ldots, \hat{\tau}$ by Lemma 7 and the definition of GAPS. Since $\forall \hat{t} = 2, \ldots, \hat{\tau}$, $\sum_{\overline{\tau}=\hat{t}}^{|T|} E_{\hat{t}\overline{\tau}}^{\mathbf{T}+1} = \overline{e}_{\hat{t}}^{\mathbf{T}+1}$, $E_{\hat{t}\overline{\tau}}^{\mathbf{T}+1}$ must have been its final value, so the value of $E_{1\hat{\tau}}^{\mathbf{T}+1}$ has no effect on the value $E_{\hat{t}\overline{\tau}}^{\mathbf{T}+1}$. The iteration then increases $E_{1\hat{\tau}}^{\mathbf{T}+1}$ and updates $l(\hat{\tau})$ if necessary but makes no changes to $l(\overline{\tau})$ for $\overline{\tau} \neq \hat{\tau}$. Hence the order of the selection of a time period τ in GAPS is not changed for $\overline{\tau} \neq \hat{\tau}$. Thus the value of $E_{1\hat{\tau}}^{\mathbf{T}+1}$ has no effect on the value $E_{\hat{t}\overline{\tau}}^{\mathbf{T}+1}$, $\forall \hat{t} = 2, \ldots, |T|$, and by the induction hypothesis, $E_{\hat{t}\overline{\tau}}^{\mathbf{T}+1}$ must be the same in the solution found using RGAPS. Moreover, prior to the iteration that first increased $E_{1\hat{\tau}}^{\mathbf{T}+1}$, the counter $l(\hat{\tau})$ must be equal to the equivalent counter in RGAPS after the iteration in which $t = \mathbf{T}$. Since the magnitude of an increase in $E_{1\hat{\tau}}^{\mathbf{T}+1}$ uses the same rule in both GAPS and RGAPS, the selection and changes in the counters are the same. Thus GAPS and RGAPS are equivalent algorithms. □

Lemma 8. *Let* $(\tilde{A}, \tilde{E})$ *be an optimal solution found by GAPS with objective value* z. *Let* $(\tilde{Y}, \tilde{\pi}, \tilde{\rho})$ *be the dual solution given by (24)-(26). With a sufficiently small* $\varepsilon > 0$ *increase in* $\overline{d}_\tau$ *for some* $\tau \in T$, *there exists a primal feasible solution with an objective function value* $z + \varepsilon\tilde{Y}_\tau$.

Proof. Consider the following two cases:

Case 1: Suppose $\mathcal{T}^{-1}(\tau) = \emptyset$. If $\overline{d}_\tau$ is increased by $\varepsilon \leq m_{l(\tau)+1} - m_{l(\tau)} - \tilde{A}_{\tau l(\tau)}$, then a feasible solution in which $\tilde{A}_{\tau l(\tau)}$ is increased by ε can be constructed. Since the penalty on $\tilde{A}_{\tau l(\tau)}$ is $\alpha_{l(\tau)}$, the increase in the objective value is $\varepsilon\alpha_{l(\tau)} = \varepsilon\tilde{Y}_\tau$.

Case 2: Suppose $\mathcal{T}^{-1}(\tau) \neq \emptyset$. Let $\hat{\tau} \in \arg\min_{\tilde{\tau} \geq \min \mathcal{T}^{-1}(\tau)}\{\alpha_{l(\tilde{\tau})}\}$, and let $\hat{t} = \min \mathcal{T}^{-1}(\tau)$. By the definition of $\mathcal{T}^{-1}(\tau)$, $\exists \tilde{\tau} \in \mathcal{T}(\tau)$ such that $\tilde{E}_{\hat{t}\tilde{\tau}} > 0$. By definition of $\mathcal{T}(\tau)$, $\exists t_1, \ldots, t_{q-1}, \tau_1 = \tau, \ldots, \tau_q = \tilde{\tau}$ such that $t_1 \leq \tau_2, t_2 \leq \tau_3, \ldots, t_{q-1} \leq \tau_q$ and $\tilde{E}_{t_1\tau_1}, \tilde{E}_{t_2\tau_2}, \ldots, \tilde{E}_{t_{q-1}\tau_{q-1}} > 0$. Now suppose

$$\varepsilon \leq \min(\tilde{E}_{t_1\tau_1}, \tilde{E}_{t_2\tau_2}, \ldots, \tilde{E}_{t_{q-1}\tau_{q-1}}, \tilde{E}_{\hat{t}\tilde{\tau}}, m_{l(\hat{\tau})+1} - m_{l(\hat{\tau})} - \tilde{A}_{l(\hat{\tau})}).$$

If $\overline{d}_\tau$ were increased by ε, a feasible solution can be constructed in which both $\tilde{E}_{t_1\tau_1}, \tilde{E}_{t_2\tau_2}, ..., \tilde{E}_{t_{q-1}\tau_{q-1}}$ and $\tilde{E}_{\hat{t}\tilde{\tau}}$ were decreased by ε, and $\tilde{E}_{t_1\tau_2}, \tilde{E}_{t_2\tau_3}, ..., \tilde{E}_{t_{q-1}\tau_q}$, $\tilde{E}_{\hat{t}\hat{\tau}}$, and $\tilde{A}_{l(\hat{\tau})}$ were increased by ε. Since the penalty on $\tilde{A}_{l(\hat{\tau})}$ is $\alpha_{l(\hat{\tau})} = \tilde{Y}_\tau$, the increase in the objective value is $\varepsilon\tilde{Y}_\tau$.

□

Let $(\tilde{A}, \tilde{E})$ be a primal solution found by GAPS, and let $(\tilde{Y}, \tilde{\pi}, \tilde{\rho})$ be the dual solution given by (24)-(26).

Lemma 9. *$(\tilde{Y}, \tilde{\pi}, \tilde{\rho})$ satisfies the complementary slackness conditions (19).*

Proof. Suppose to the contrary, there exists a time period $\tau \in T$ such that $\tilde{A}_{\tau i} < m_{i+1} - m_i$ and $\tilde{\rho}_{\tau i} > 0$. If $\overline{d}_\tau$ is increased by a sufficiently small $\varepsilon > 0$, then a primal feasible solution in which the objective value is increased by $\varepsilon\alpha_i$ can be constructed by Case 1 of Lemma 8. Consequently, $\tilde{Y_\tau} \leq \alpha_i$, in contradiction to the assumption that $\tilde{\rho}_{\tau i} = \tilde{Y}_\tau - \alpha_i > 0$. Hence no such $\tau \in T$ exists. □

Lemma 10. *$(\tilde{Y}, \tilde{\pi}, \tilde{\rho})$ satisfies the complementary slackness conditions (20).*

Proof. Suppose to the contrary, there exist a time period $\tau \in T$ such that $\tilde{A}_{\tau i} > 0$ and $\tilde{Y}_\tau - \tilde{\rho}_{\tau i} < \alpha_i$. The index $i \leq l(\tau)$ since $\tilde{A}_{\tau i} = 0, \forall i \geq l(\tau) + 1$. This implies $\tilde{\rho}_{\tau i} = 0$ and $\tilde{Y}_\tau < \alpha_i$ by definition (26) and $\tilde{Y}_\tau < \alpha_i \leq \alpha_{l(\tau)}$ by the definition of α. Since $\tilde{Y}_\tau < \alpha_{l(\tau)}$, the set $\mathcal{T}^{-1}(\tau) \neq \emptyset$ by the definition (24). For a sufficiently small $\varepsilon > 0$ increase in $\overline{d}_\tau$, a primal feasible solution can be constructed in which the objective value is increased by $\varepsilon\tilde{Y}_\tau$ by case 2 in Lemma 8. Similarly, for a small $\varepsilon' = \min(\varepsilon, \tilde{A}_{\tau i}) > 0$ decrease in $\tilde{A}_{\tau i}$, a primal feasible solution can be constructed in which the objective value is decreased by $\varepsilon'(\alpha_i - \tilde{Y}_\tau) > 0$. The assumption that $(\tilde{A}, \tilde{E})$ is optimal is contradicted, so no such $\tau \in T$ exists. □

Lemma 11. *$(\tilde{Y}, \tilde{\pi}, \tilde{\rho})$ satisfies the complementary slackness conditions (21).*

Proof. Suppose there exist time periods $t \leq \tau$ in which $\tilde{E}_{t\tau} > 0$. By definition (25), let $\tilde{\tau} \in \arg\min_{\bar{\tau} \geq t} \left\{\tilde{Y}_{\bar{\tau}}\right\}$, so $\tilde{\pi}_t = \tilde{Y}_{\tilde{\tau}}$ and $\tilde{\tau} \geq t$. By Lemma 8, for a sufficiently small $\varepsilon > 0$ increase in $\overline{d}_{\tilde{\tau}}$, a primal feasible solution in which the objective value is increased by $\varepsilon\tilde{Y}_{\tilde{\tau}}$ can be constructed. Similarly, for a small $\varepsilon' = \min(\varepsilon, \tilde{E}_{t\tau}) > 0$ increase in $\overline{d}_{\tilde{\tau}}$, a primal feasible solution in which the objective value is increased by $\varepsilon\tilde{Y}_{\tilde{\tau}}$ can be constructed by decreasing $\tilde{E}_{t\tau}$, increasing $\tilde{E}_{t\tilde{\tau}}$, and changing the same variables as done for an increase in $\overline{d}_{\tilde{\tau}}$ by ε'. Since case 2 of Lemma 8 includes all such general constructions of primal feasible solutions, the increase in the objective function value $\varepsilon'\tilde{Y}_{\tilde{\tau}}$ is no less than $\varepsilon'\tilde{Y}_\tau$. Hence, $\tilde{Y}_\tau = \tilde{Y}_{\tilde{\tau}} = \tilde{\pi}_t$. □

Theorem 3. *Let $(\tilde{A}, \tilde{E})$ be an optimal solution from GAPS. The dual solution given by (24) - (26) is an complimentary optimal dual solution.*

Proof. By definitions of $(\tilde{Y}, \tilde{\pi}, \tilde{\rho})$ in equations (24) - (26), the dual feasibility constraints (13) - (16) are satisfied. By Lemmas 9 - 11, $(\tilde{Y}, \tilde{\pi}, \tilde{\rho})$ satisfies the complemnentary slackness conditions (19) - (21). □