A data-integrated simulation model to evaluate nurse-patient assignments

Durai Sundaramoorthi*, Victoria C. P. Chen†, Jay M. Rosenberger†, Seoung Bum Kim†,

Deborah F. Buckley-Behan‡

September 21, 2007

Abstract

This research develops a novel data-integrated simulation to evaluate nurse-patient assignments (SIMNA) based on a real data set provided by Baylor Regional Medical Center (Baylor) in Grapevine, Texas. Tree-based models and kernel density estimation were utilized to extract important knowledge from the data for the simulation. Classification and Regression Tree models, data mining tools for prediction and classification, were used to develop five tree structures: (*a*) four classification trees, from which transition probabilities for nurse movements are determined; and (*b*) a regression tree, from which the amount of time a nurse spends in a location is predicted based on factors such as the primary diagnosis of a patient and the type of nurse. Kernel density estimation is used to estimate the continuous distribution for the amount of time a nurse spends in a location. Results obtained from SIMNA to evaluate nurse-patient assignments in medical/surgical unit I of Baylor are discussed.

1 Introduction

The health care system in the United States has a shortage of nurses. In 2000, according to the U.S. Department of Health and Human Services (DHHS), the national shortage for registered nurses was 110,000 or 6%. DHHS anticipates that the shortage will grow relatively slowly until it reaches 12% around 2010. From then, it is expected to worsen at a faster rate and reach a 20% shortage by 2015. A shortage of 3% or more was observed in 30 states during 2000, and similar shortages are predicted to occur in 44 states by 2020 (HRSA, 2002). These statistics show

*Department of Engineering Management and Systems Engineering - University of Missouri-Rolla - 223 Engineering Management Rolla, MO 65409.

†Dept. of Industrial & Manufacturing Systems Engineering - The University of Texas at Arlington - Campus Box 19017 Arlington, TX 76019-0017.

‡School of Nursing - The University of Texas at Arlington - Arlington, TX 76019.

that the severity of this shortage is widespread. As a consequence of the nurse shortage, it is natural to expect issues such as job burnout and poor patient care (Aiken et al., 2002). In an attempt to ease the health care system from such issues, California has set a limit on the number of patients that can be assigned to nurses at the same time (CDHS, 2005). Such restrictions may reduce nurses' workload, but will unlikely resolve the issue because differences in workload among nurses depend on amount of care required and the physical location of the patients to which a nurse is assigned. Static nurse-to-patient ratios ignore the differences in patient mix, care unit, hospital layout, and nurse resource across different hospitals. For these reasons, professional organizations such as, the American Organization of Nurse Executives (AONE), the Society for Health Systems (SHS), and the Healthcare Information and Management Systems Society (HIMSS) oppose the mandatory static ratios (AONE, 2003; SHS, 2005; HIMSS, 2006). All these organizations, in their position statements, either implicitly or explicitly call for models that consider hospital specific factors to address nurse-to-patient assignments. Thus, instead of statically limiting the number of patients per nurse, it is important to optimize the nurse-patient assignments for a balanced workload with a hospital specific model. In the literature, most of the relevant research focuses on nurse budgeting, nurse scheduling (rostering), and nurse re-scheduling methodologies (Aickelin and Dowsland, 2003; Burke, Cowling and Caumaecker, 2001; Jaumard et al., 1998; Kirkby, 1997; Miller et al., 1996; Warner, 1976; Bard and Purnomo, 2005b; Azaiez and Sharif, 2005; Beddoe and Petrovic, 2006; Gutjahr and Rauner, 2007) and did not address nurse-to-patient assignment issue. Apart from the proposed model in this paper, Vericourt and Jennings (2006) and Punnakitikashem et al. (2006) are two other contemporary research that addresses nurse-to-patient assignment issue. However, these researches did not use real data as extensively as it would require to model nurse-to-patient assignments at a care unit level for a given hospital. By contrast, our research considers hospital and care unit specific factors and develops a data-integrated simulation to evaluate nurse-patient assignments (SIMNA) that utilizes patterns in a real data set to balance workload among nurses. The data set for this research was provided by Baylor Regional Medical Center (Baylor) and hence the results are confined to it. However, the simulation model could be easily adapted to other hospitals once similar data analysis is performed. The mechanism for adapting our simulation model to other hospitals is briefly explained in section 7.

In traditional stochastic simulation models, transition probabilities are obtained either subjectively or by looking at all possible combinations of the levels of the simulation state variables. If the system under consideration is complex, such as nurse movement, then a subjective approach is unlikely to be accurate, and an approach using all possible combinations of the states will be impractical. In the past, in order to reduce the number of simulation variables, factorial designs and screening methods were used (Bettonvil and Kleijnen, 1997; Cheng, 1997; Shen and Wan, 2005). Even after eliminating some of the variables, a few remaining variables could lead to a huge number of combinations for the simulation. For instance, six categorical variables with ten categories each will lead to a million possible states in the simulation. Obtaining accurate transition probabilities for such a huge simulation model is still difficult. In this paper, using data from Baylor Regional Medical Center (Baylor) in Grapevine, Texas, we present a new methodology

to reduce the number of combinations and find transition probabilities for stochastic simulation models. Tree-based models and kernel density estimates were utilized to extract important knowledge about the workload of nurses from an encrypted data set provided by Baylor for four care units. The four units include two medical/surgical units, one mom/baby unit, and one high-risk labor-and-delivery unit. Classification and Regression Trees (Breiman et al., 1984), a data mining tool for prediction and classification, was applied to the Baylor data to develop five tree structures: (a) four classification trees, from which transition probabilities for nurse movements are determined; and (b) a regression tree, from which the amount of time a nurse spends in a location is predicted based on factors such as the primary diagnosis of a patient and the type of nurse. Simulation models developed with this approach will be much more representative of actual systems and more efficient than those that consider all possible combinations.

Following are two major contributions made in this research:

- This research introduces a novel approach, discussed in section 4, to the simulation community for constructing efficient simulation models based on data mining. This way of simulation modeling avoids misrepresentation of system dynamics and characteristics because it is entirely based on the pattern learned from a real data set collected from the system over a long period of time. Moreover, this approach reduces simulation states and is consequently more efficient to run.

- This research introduces a tool, discussed in section 5, to evaluate nurse-to-patient assignments and enable decisions in real time. At Baylor, prior to a shift, the decision to hire agency nurses is determined by nurse supervisors, who assess whether the set of scheduled nurses is sufficient for that shift. The SIMNA model can aid them in their decisions by providing a tool to test nurse-to-patient assignments.

The rest of this paper is organized as follows: In Section 2, a literature review on nursing research and the contributions of this research are given. In Section 3, a brief introduction is given on data and notation. Section 4 describes the data mining tree structures used to build the simulation model, kernel density estimation, and the simulation structure. Section 5 presents results from SIMNA for a set of sample assignments from medical/surgical unit I. In Section 6, the simulation modeled is validated by comparing simulation results with the actual data. Section 7 presents a discussion on adaptability of the simulation model to a new hospital. In Section 8, we provide concluding remarks, discussion on a possible simulation-optimization approach to optimize nurse-to-patient assignments, and other opportunities for future work.

2 Literature and contribution

There are three major components in this research, i.e, nurse planning, data mining, and simulation modeling. This chapter gives a brief literature review on each of these topics.

2.1 *Nurse planning*

Nurse planning typically has four stages: nurse budgeting, nurse scheduling, nurse rescheduling, and nurse assignment. In the literature, most of the relevant research focuses on the first three stages of planning.

In nurse budgeting: Kao and Queyranne (1985) showed that a single-period demand estimate gives a good approximation for nurse budgeting cost. Trivedi (1981) used mixed-integer goal programming to optimize the expenses for nurse personnel. Kao and Tung (1980) used a linear programming-based approach to assess needs for regular, overtime, and agency workforce levels for a given time period.

In nurse scheduling: Warner and Prawda (1972) optimized nurse schedules by formulating a mixed-integer quadratic programming problem. Later, Warner (1976) formulated and solved another multiple-choice math programming scheduling problem incorporating nursing preferences. Miller et al. (1996) minimized an objective function that balanced the trade-off between staffing coverage and preferences of nurses. Burke, Cowling and Caumaecker (2001) and Burke, Caumaecker and Petrovic (2001) used a combination of tabu search, genetic algorithm, and steepest descent improvement heuristics to solve a nurse rostering problem. Aickelin and Paul (2004) formulated the nurse scheduling as an integer programming problem and compared solutions from different algorithms using statistical techniques. Azaiez and Sharif (2005) computerized the nurse-scheduling problem for Riyadh Al-Kharj hospital (in Saudi Arabia) using a 0-1 goal programming that incorporated nurses' preferences and hospital objectives. Wong and Chan (2004) introduced a probability-based ordering method for a nurse rostering problem that considered twelve nurses. It reported its solution time as half a second. Beddoe and Petrovic (2006) used genetic algorithm to solve another nurse rostering problem by considering violations made in prior rosters. Gutjahr and Rauner (2007) used ant colony optimization to schedule nurses for four weeks among different hospitals in a region.

In nurse rescheduling: Benton (1994) showed how the scheduled nursing scenario changes when the patient acuity and number of patients change. Walts and Kapadia (1996) developed a patient classification system to redistribute nursing personnel across different care units based on patient acuity. Bard and Purnomo (2005a,b) formulated a nurse rescheduling integer programming problem and solved it using branch and price considering the resource shortage, demand drop, and nurse preferences. CDHS (2005) required health care providers to maintain certain nurse-to-patient ratios for improving quality of care. Vericourt and Jennings (2006), using a queuing approach, showed that same set of ratios for different sizes of care units lead to inconsistent amounts of care. Alternatively, they proposed a heuristic-based policy to provide better care. However, their model allowed nurses to serve unassigned patients, which is discouraged in practice for maintaining continuity of care.

In nurse assignment: Mullinax and Lawley (2002) formulated and solved an integer programming problem using heuristics to assign nurses to patients by balancing workload for nurses based on patient acuity in a neonatal intensive care. Punnakitikashem et al. (2006) formulated and solved a two-stage stochastic integer programming nurse

assignment problem to minimize excess workload of nurses. None of the methods discussed above provides a tool to evaluate nurse-patient assignments to make decisions in real time. Also, other methods did not use real data to reflect the real system as extensively as the approach presented in this research.

2.2 Data mining

Data Mining can be broadly classified into two groups: supervised learning and unsupervised learning. In supervised learning, an outcome variable is present to guide the learning process. Whereas, in unsupervised learning or clustering, one wants to observe only the features and have no measurements of the outcome. Data Mining can be viewed as statistical learning from data or more generally as an approach that seeks to uncover patterns in data. Typically, learning could be an outcome measurement, quantitative (like the amount of time spent by nurses in a given location) or categorical (like different locations a nurse visits), that one wants to predict based on set of features (like type of the nurse, diagnosis of the patient, and time of the day) if available (Hastie et al., 2001). Supervised learning is the subject of interest in this research as we deal with predicting the time spent and location for nurses. Regression, kernel methods, tree based models, neural networks, and support vector machines are some popular supervised learning methods. Regression methods are one of the traditional tools used for prediction (Neter et al., 1996; Hastie et al., 2001; Walpole et al., 2002). Multivariate Adaptive Regression Splines (MARS), a spline based prediction model (Friedman, 1991) was recently applied to different prediction problems (Chen et al., 1999; Tsai et al., 2003; Chen et al., 2003; Siddappa et al., 2006; Pilla et al., 2005). Neural networks, a nonlinear statistical model (Ripley, 1996; Haykin, 1999), often represented by a network diagram, can be used for prediction or classification. Le Cun et al. (1990) applied neural networks to identify handwritten zip code digits. Cervellera, Chen and Wen (2006) and Cervellera, Wen and Chen (2006) approximated stochastic dynamic programming value functions of an inventory forecasting problem and a water reservoir problem with neural networks. Classification and Regression Trees (Breiman et al., 1984), a data mining tool for prediction and classification, is used in this research for its applicability to regression and classification problems, and its readily usable tree structures in simulation.

2.3 *Simulation modeling in health Care*

Studying industrial systems using simulation was prevelant as early as the late 1950's and early 1960's. Youle et al. (1959) and Clementson (1966) discuss simulations of different industrial processes available at that time. In health care, simulation modeling has been used to study a wide range of problems. Bailey (1952) and Kachhal et al. (1981) studied patient queues and waiting times. Smith and Warner (1971), Lim et al. (1975), and Hancock and Walter (1984) studied patient admission and its impact. Zilm et al. (1983) and Dumas (1984, 1985) modeled and analyzed patient bed planning and utilization under different scenarios. Kumar and Kapur (1989), Draeger (1992) and Evans et al. (1996)

evaluated nurse schedules for the emergency care department. In recent years, Zenios et al. (1999), Kreke et al. (2002), and Shechter et al. (2005) utilized simulation models to study organ allocation systems. A comprehensive review of health care simulation models can be found in Klein et al. (1993) and Jun et al. (1999). In the literature, most of the health care staffing simulations analyzed the emergency departments in hospitals. Moreover, the simulation modeling approaches in the literature, both deterministic and stochastic, required the knowledge of experts to estimate parameters and order of events in the simulation. If the system under consideration is complex, such as nurse movement in hospitals, then it is impossible even for the experts to comprehend the intricacies of the system by observation. Whereas, the simulation modeling technique introduced in this research captures the system dynamics from a real data set collected from the system and requires only minimal input from the experts.

2.4 *Contribution*

There are two major contributions made in this research:

- This research introduces a novel approach to the simulation community for constructing efficient simulation models based on data mining. This way of simulation modeling avoids misrepresentation of system dynamics and characteristics because it is entirely based on the pattern learned from a real data set collected from the system over a long period of time. Moreover, this approach reduces simulation states and is consequently more efficient to run.

- This research introduces a tool to evaluate nurse-to-patient assignments and enable decisions in real time. At Baylor, prior to a shift, the decision to hire agency nurses is determined by nurse supervisors, who assess whether the set of scheduled nurses is sufficient for that shift. The SIMNA model can aid them in their decisions by providing a tool to test nurse-to-patient assignments.

3 Data description

At Baylor, each nurse wears a locating device that transmits data to a repository, where the data automatically expire after one month. Baylor provided data for this research from four care units: Medical/Surgical unit I, Medical/Surgical unit II, Mom/Baby unit, and High-Risk Labor unit. These *nurse data* contain information on month, day, shift, time, location, nurse, nurse type and time spent for the location visited by the nurse. Baylor also provided *patient data*, which contain information on admit date, discharge date, room number and diagnosis code for each patient. These two data sets were merged by matching the date and location information and are referred to as the *merged data*. The resulting *merged data* have all the variables from the nurse and patient data sets. To preserve the confidentiality of nurses, patients and the medical center, an encryption code using the U16807 method (Law and Kelton, 2001) was

developed and employed to the data before our analysis. U16807 method was chosen for encryption because of its efficiency to handle cycling. An example for date and location variables in our data before and after encryption is shown in Table 1.

Table 1: Encryption Example

Variable	**Before**	**After**
Date	4/5/04	2/15/73622
Room	442	704

Two new variables were created to hold information on the previous two locations visited for each location entered by nurses to predict patterns in their movements. In a related research, presented in Sundaramoorthi, Chen, Rosenberger, Kim and Behan (2006) and Sundaramoorthi, Chen, Kim, Rosenberger and Behan (2006), seven variables were created to hold information on previous seven locations. The simulation models developed with seven previous locations were found to overfit the pattern based on movements and hence insensitive to other practically important variables. For this reason, unlike Sundaramoorthi, Chen, Rosenberger, Kim and Behan (2006) and Sundaramoorthi, Chen, Kim, Rosenberger and Behan (2006), the simulation presented here includes location variables that specify only two previous locations and the current location to avoid overfitting patterns based purely on nurse movements. Furthermore, a variable was created to indicate the nurse-patient assignments. To create nurse-patient assignment variable, it is assumed that the nurse who spent the most time in a patient's room during a shift is the nurse assigned to that patient for that shift. After processing the data, medical/surgical unit I, medical/surgical unit II, mom/baby unit, and high-risk labor-and-delivery unit have about 570,660, 418,683, 315,997, and 210,457 observations, respectively. Following the conclusions in Sundaramoorthi et al. (2005) and further similar analysis presented in Sundaramoorthi, Chen, Rosenberger, Kim and Behan (2006), the following types of variables with their specific levels are considered significant for the methodology presented here.

1. Location : patient rooms, nurse station, break room, reception desk, and medical room.
2. Nurse Type: registered nurse (RN), licensed vocational nurse (LVN), and nurse aide (NA).
3. Diagnosis Code : 19 categories covering the range of diagnosis codes, and 2 dummy categories for empty patient rooms and non-patient locations. See INGENIX (2003) for more details on diagnosis codes.
4. Shift: 3 weekday shifts (8 hours each) and 2 weekend shifts (12 hours each).
5. Hour: 24 hour ranges covering a complete day.
6. Assignment: An assigned nurse entering a patient room (1), an unassigned nurse entering a patient room (0), and a nurse entering any location other than patient rooms (2).

7. Time Spent: Time Spent is the dependent variable that denotes the amount of time a nurse spends in a given location.

Data from different care units were handled separately as the number of categorical levels of the considered variables, listed above, differed slightly among different care units. In this research, we maintain the following notations: X_S, X_T, X_{NT}, X_L, X_A, and X_D are the variables representing shift, hour, nurse type, current location, assignment, and primary diagnosis of the patient in a current location, respectively. N_S, N_T, N_{NT}, N_L, N_A, and N_D are the number of levels of X_S, X_T, X_{NT}, X_L, X_A, and X_D, respectively. X_{P1L}, and X_{P2L} are the variables representing the two previous locations with X_{P1L} being the latest and X_{P2L} being the oldest among the two locations visited before any current location. X_{P1L} and X_{P2L} have the same number of levels (N_L) as of X_L. For each nurse, X_{AL1}, $\ldots$, X_{ALR} are the binary variables indicating patients assigned to her/him in a shift. R is the number of patient rooms in a care unit. $X_{DL1}, \ldots, X_{DLR}$ are the variables representing primary diagnosis of patients in rooms 1 to R.

4 Data mining for simulation

4.1 *Classification and regression trees*

Classification and Regression Trees (CART) are data mining tools for prediction and classification (Breiman et al., 1984; Hastie et al., 2001). CART utilizes recursive binary splitting to uncover structure in a high-dimensional space. CART, on application to a data set, will partition the input space into many disjoint sets, where values within a set have a more similar response measure than values in different sets. Salford Systems' CART® software (www.salfordsystems.com) was used to obtain our tree structures. In particular, five tree structures were developed: (a) four classification trees from which transition probabilities for nurse movement are determined based on the levels of X_S, X_T, X_{NT}, $X_{DL1}, \ldots, X_{DLR}$, X_A, X_{P1L}, and X_{P2L}; and (b) a regression tree to predict the amount of time a nurse will spend in a location based on the levels of X_S, X_T, X_{NT}, X_L, X_D, and X_A. A hypothetical regression tree is shown in Figure 1(a) to illustrate a prediction of the amount of time a nurse would spend in a location. At each node of the tree, a question is asked; a data point that satisfies the question will go left in the branching; and right if it fails to meet the criterion. Based on the levels of X_S, X_T, X_{NT}, X_L, X_D, and X_A, every data point ends up in one of the terminal nodes of the tree. Two hypothetical classification trees, one "location type tree" in Figure 1(b) and another "location tree" in Figure 1(c), are shown to illustrate the estimation of the probability that a location would be visited by a nurse. At each node of these trees, similar to the regression tree, a question is asked; data that satisfy the question will go left in the branching; and right if they fail to meet the criterion. The probability of going to a location type, i.e, unassigned patient room (0), assigned patient room (1), and non-patient room (2) is obtained from the location type classification tree based on the levels of X_S, X_T, and X_{NT}. In the "location tree," depending on the levels of X_S,

X_T, X_{NT}, X_{DL1}, …, X_{DLR}, X_A, X_{P1L}, and X_{P2L}, every data point ends up in one of the terminal nodes of the tree, where transition probabilities are estimated as follows:

$$\hat{p}(l/j) = \frac{1}{n(j)} \sum_{i=1}^{n(j)} I(i \in l), \tag{1}$$

where, j = 1, …, J are the terminal nodes of a "location tree"; $n(1)$, …, $n(J)$ are the numbers of observations in terminal nodes 1, …, J, respectively; l = 1, …, N_L are the levels of X_L, i.e., the different locations in a given care unit, and I is an indicator function. The number of terminal nodes (J) differ for each tree. To be precise, J_0, J_1, and J_2 represent the number of terminal nodes of "location trees" for location types 0, 1, and 2, respectively. J_{LT} represent the number of terminal nodes of a "location type tree." For a "location type tree", l = 0, …, 2 are the levels of X_A, i.e., unassigned patient room (0), assigned patient room (1), and non-patient room (2).

One useful outcome from using tree-based models is the variable importance scores that provide information on the influence of each variable to predict a response. Variable importance scores for all the trees are shown in Table 2. Variable importance scores for the regression trees estimating the amount of time a nurse will spend in a location are given in the first row. It can be seen that location is the most important variable. Primary diagnosis and assignment play a relatively more important role in medical/surgical II and high-risk Labor units than mom/baby and medical/surgical I units, and time (hour) of the day is more important than shift. Nurse type has about the same magnitude of importance across all the care units. Variable importance scores for the "location type trees" predicting a nurse$'$s next location type are shown in the second row of Table 2. It can be observed that nurse type for mom/baby and high-risk labor units, and time (hour) of the day for medical/surgical I & II units are the most important factors to predict the location type. Similar to the regression trees, time (hour) of the day is more important than shift. Variable importance scores of selected variables in the "location trees" predicting a nurse's next location for different location types are shown in the last three rows of Table 2. It can be seen that the previous locations are the most important variables to predict the next location. Once again, time (hour) of the day is more important than shift. Variable importance scores of the variables X_{AL1}, …, X_{ALR} and X_{DL1}, …, X_{DLR} in the "location trees" are not presented here to make the table concise. As mentioned earlier, it is impossible even for a health care expert to observe all these intricate and subtle differences in the system without using a tool like CART.

While growing the trees, 10-fold cross validation was used for testing; class probability and least squares splitting rules were used for creating branching decisions of classification trees and regression trees, respectively. Developing theories and models for justifying the choice of testing and splitting rules for data-integrated simulations would be an interesting direction for future research.

4.2 *Estimation of time spent distribution*

For each terminal node of the regression trees, kernel density estimation (KDE) is used to estimate the probability density function for time spent (Y) by a nurse (under the conditions specified by that terminal node). Assume we have $n(j)$ independent observations $y_1, \ldots, y_{n(j)}$ for the random variable $Y(j)$ in the terminal node j. Let $K(\cdot)$ be a kernel function. Then the kernel density estimator $\hat{f}_{j,h}(y)$ at a point y is defined by equation (2) (Silverman, 1986), as follows:

$$\hat{f}_{j,h}(y) = \frac{1}{h \times n(j)} \sum_{i=1}^{n(j)} K(\frac{y_i - y}{h}), \tag{2}$$

where, h is the bandwidth, which controls the "window" of neighboring observations that will highly influence the estimate at a given y. Sheather and Jones plug-in (SJPI) bandwidth estimates for h are used, as this method is one of the best for optimizing bandwidth (Jones et al. (1996); Sheather and Jones (1991); Sheather (2004)); however, it should be noted that bandwidth selection is not precise and often an "art." Tuning of the bandwidths based on our desired criteria is discussed in Section 4.2.2. Random variables $Y(1), \ldots, Y(J_R)$ denote the time spent (Y) in terminal nodes $1, \ldots, J_R$, respectively. Kernel density estimates with SJPI bandwidths were obtained for each terminal node of the regression trees. A typical plot with Gaussian and triangular kernels for each of the four care units is shown in Figure 2.

4.2.1 *Kernel choice*

Kernel functions include uniform, Gaussian, triangular, Epanechnikov, quadratic, and cosinus. Gaussian and triangular kernels were chosen for this research as they are common among modelers. Moreover, it is relatively easy to draw samples from Gaussian and triangular distributions, which is required for sampling the time spent random variable. SJPI bandwidth estimates (Sheather and Jones, 1991) were calculated for each terminal node of the regression tree using SAS®. Figure 2 and the normal probability plots in Sundaramoorthi et al. (2005) show that the time spent data have a long right tail, and a major portion of the data is concentrated near the left end of the distribution. Gamma distributions provided inadequate density estimates, motivating the use of KDE. To assess how well KDE represents the time spent distribution, 100,000 realizations of time spent data were generated from Gaussian and triangular kernel density estimates. The simulated data were compared with the actual data in four different ranges, i.e., (0, $M/2$], ($M/2$, M], (M, ($M + M/2$)], (($M + M/2$), ∞), where, M is the median of the actual data. Results from 100,000 simulated realizations of Gaussian and triangular kernels are shown in Table 3. There were 181, 109, 123 and 49 terminal nodes in the regression trees of medical/surgical I, medical/surgical II, mom/baby and high-risk labor units, respectively. The table shows that the triangular kernel wins more often than the Gaussian kernel irrespective of the care units and ranges. Among all the competitions i.e., $J_R \times 4$ competitions, the triangular won 75%, 80%, 82% and 78% of the competitions in medical/surgical I, medical/surgical II, mom/baby and high-risk labor units, respectively.

A *terminal-node-win* was considered to be achieved if a kernel managed to win at least three ranges out of the four considered. Both the kernels were considered to be *tied* if they won two ranges each. The results on terminal node wins shown on the last two rows of Table 3 for each care unit further indicate that the triangular kernel is a better choice to model the Baylor data.

4.2.2 *Bandwidth tuning*

The accuracy of estimates depends more on choosing an appropriate bandwidth than on the choice of kernels (Epanechnikov, 1969; Silverman, 1978). Bandwidth selection methods, including SJPI bandwidth estimates (Sheather and Jones, 1991), try to find the optimal bandwidth that compromises a tradeoff between oversmoothness and undersmoothness of the estimated density. After obtaining bandwidths, we can decide to either decrease or increase the bandwidth size depending on the knowledge of the system. Data used in this project were collected over more than a six-month period and have hundreds of thousands of observations for each care unit. With data collected over months, the different possible characteristics of the Baylor system will be well reflected in the simulation if the bandwidths are tuned to prefer a less smooth density estimate that reflects the data more accurately. In this research, if the fraction of simulated realizations in the ranges given in the previous section goes beyond $\pm$ 0.015 of the actual fraction of data, the bandwidth was iteratively decreased by one until this criterion was met. For example, the ninth terminal node of medical-surgical unit I shown in Table 4 has realizations that violated the $\pm$ 0.015 limit. After forty four iterations of bandwidth tuning, all four ranges have fractions within the limit. This leads to a change of bandwidth at this particular terminal node to 8.46 from 52.46 and thus yields a less smooth kernel density estimate that is more representative realizations of the time spent data.

4.3 *Data-driven simulation model*

To drive a nurse activity simulation, three essential questions are asked: (1) Which location type will a nurse go to next given her nurse type, shift, and time (hour) of the day? (2) Where will a nurse go next given her two past locations, next location type, shift, hour, nurse type, assignments, and diagnoses of all the patients? (3) How much time will she spend there? After an initial simulation run in which nurses visit their assigned patients for an initial assessment, transition probabilities obtained by equation (1) from the location type and location trees determine the next location a nurse will visit. Once a location type and in turn a location has been sampled for a given nurse, the amount of time she spends there is determined by a random sample of time spent y from the kernel density estimate at the appropriate terminal node in the regression tree. Clock time and the location variables are then updated. The level of X_T is changed if the updated time enters a new category. The levels of variables X_S and X_{NT} associated with a nurse remain unchanged throughout the shift. This process of sampling location type, location, and time spent is repeated until the shift ends.

Traditionally, in stochastic simulations, transition probabilities are obtained either subjectively or by looking at all the possible combinations of variable levels. In practice, simulation modelers combine states by making a variety of assumptions on their models. For instance, suppose a simulation expert were to model a system using a queuing network with one hundred servers. To model the system accurately, the modeler would need to determine whether the service times of each pair of servers were independent. This would require ten thousand tests of independence. If multiple servers were found to be dependent, then the modeler would have to group the servers into sets in which the servers are dependent. Then, the modeler would have to develop enormous multivariate distributions for each group that may consider tens of variables. In practice though, the modeler would likely make assumptions about the independence of these variables to limit the dimensionality of the multivariate distributions. If the system under consideration is complex, such as the care units in Baylor, then a subjective approach is unlikely to be accurate, and it will be impractical to implement an approach using all possible combinations of the levels of the simulation variables. In the latter approach, the number of possible combinations (NPC) grows exponentially with the number of variables. In our problem, there are $N_S \times N_T \times N_{NT}$ combinations, denoted as NPC_{lt}, for sampling a location type and $N_S \times N_T \times N_{NT} \times N_A \times {N_L}^2 \times {N_D}^R \times 2^R$ combinations, denoted as NPC_l, for sampling a location. On the other hand, simulation models developed using trees, discussed in Section 4.1, require only J_{LT} terminal nodes for sampling a location type and $J_0 + J_1 + J_2$ terminal nodes for sampling a location based on the patterns extracted from the data. The more efficient the simulation, the more useful it will be for making real-time decisions. For example, prior to a shift, a charge nurse will determine whether the set of scheduled nurses is sufficient for the shift. If there is a shortage, nurse supervisor will call a nurse agency to hire nurses for that shift. The simulation model can assist in this decision provided its run time is sufficiently fast. Differences between NPC_{lt} and J_{LT}, NPC_l and $J_0 + J_1 + J_2$ given in Table 5, demonstrate that our approach is significantly more efficient. All locations in the care units under consideration can be visited from any other location of that care unit. Even though some of these combinations of locations are unlikely to be visited in succession, without using a data mining tool like trees, it is not easy to justify ignoring or combining them.

5 SIMNA experiments

A generic C++ program was written to rebuild the tree structures given by CART and to run the simulation procedure explained in Section 4 for medical/surgical unit I with a thousand different random seeds. A test problem with four nurses and twenty one patients was considered. SIMNA tested four assignment policies, i.e., a clustered assignment and three assignments from Punnakitikashem et al. (2006)–the random assignment, the heuristic assignment, and the optimal assignment using Benders' decomposition on a stochastic programming model. In the heuristic assignment, when the number of nurses divides into the number of patients evenly, all of the nurses get the same number of patients.

The patient with the highest expected direct care time is arbitrarily assigned to a nurse. The patient with the second highest expected direct care time is then arbitrarily assigned to a second nurse, and so on. After assigning one patient for each nurse, in the second cycle of assignments, the patient with the lowest expected direct care time is assigned to the first nurse. The patient with the second lowest expected direct care time is assigned to the second nurse, and so on. This process of assignment is repeated until all the patients are assigned. In the test problem, each nurse was assigned to five patients by the heuristic method and the left over patient was arbitrarily assigned to the first nurse. In the clustered assignment, patients are assigned by location; that is, patients in consecutive rooms are assigned to the same nurse. In the test problem, the nurse assigned to the cluster closest to the nurses' station was assigned six patients, while the other nurses were assigned to five patients. Finally, the optimized assignment from Punnakitikashem et al. (2006) seeks to balance the expected direct and indirect care provided by RNs. It should be noted that indirect care cannot be quantified from our data and is not represented in our simulation.

The tested assignments and their results are shown in Table 6. Total assigned direct care (TADC), total unassigned direct care (TUADC), total direct care (TDC), total time spent in non-patient locations (TNPL), and the walking time (Walk Time) are shown in the last five columns. TADC is the total duration of time a nurse spent with her assigned patients in the entire shift. TUADC is the total duration of time a nurse spent with unassigned patients. TDC is the sum of TADC and TUADC. TNPL is the the total time spent at locations other than patient rooms (e.g., the medical supply rooms, the charting rooms, the nurses' station, etc). In order to assess the balance of workload, we consider the ratios of maximum to minimum values for TADC, TDC, TDC for RNs, and walking time. Ratios closer to one indicate better balance. These ratios are given in Table 7. For balancing TADC, the heuristic assignment performs best and the random assignment performs worst. For balancing TDC, the heuristic assignment is worst, and the other three are similar to each other. For balancing TDC for RNs, the heuristic and optimal assignments perform best, and the random assignment performs worst. Finally, for balancing walking time, the clustered assignment performs better than the others. In particular for the optimal assignment, the sum of all nurses' TADC and TDC is higher than the other assignments, while the total walking time of the optimal assignment is less than that of the other assignments. Overall, the random assignment, not surprisingly, is the least desirable.

Prior to a shift, SIMNA results can aid the charge nurse in determining appropriate nurse-to-patient assignments. If the direct care time and balance in workload are not satisfactory, nurse supervisor can call a nurse agency to hire nurses for that shift. Thus, SIMNA upon installation in hospitals will aid charge nurses and management to make decisions about assignments and the nurse work force based on the dynamics learned from the system itself.

6 Simulation validation

Interestingly, it was observed that the "40-20-40" rule McKay et al. (1986); Sheppard (1983) still holds well in our data-integrated simulation modeling. According to this rule, 40% of the effort in a simulation project is devoted to understanding, conceptualizing the system, and formulating the model; 20% of the effort is devoted to make the actual simulation model, and the last 40% of the effort includes analysis, calibration and validation of the simulation model. Most of the first and last 40% of the project, i.e., understanding, formulation, conceptualization, calibration and validation, are conducted through data mining.

Among different steps in the traditional simulation modeling, validation is an important step in which accuracy of the model is verified by comparing it to the actual system. Depending on the magnitude of the discrepancy, if needed, the simulation model would be calibrated based on the insights gained by the modeler from the simulation output analysis. The following were among several common validation steps performed as part of the validation process in this data-integrated simulation modeling approach.

1. Tree Structure: The tree structures were printed before the first scenario of simulation run to ensure accurate building of trees for simulation runs.
2. Shift Duration: TDC, TNPL, and WALK TIME were added for each nurse to check with the entire shift duration.
3. Kernel Density: The kernel and bandwidth validations, presented in section 4.2, ensured a reliable approximation of data in regression trees.
4. Cumulative Density: The cumulative densities of kernel distributions in each terminal node were printed to check if they were close to one.

The primary objective of this research is to provide a tool to aid charge nurses in making balanced nurse-patient assignments. In this research, the balance of workload and performance of nurses were judged based on performance measures TADC, TDC, TNPL, and WALK TIME that were introduced in section 5 and shown in tables 7 and 6. As part of the main validation, actual TADC, TDC, TNPL, and WALK TIME of fifteen arbitrarily chosen nurses were compared with that of simulated data. The fifteen arbitrarily chosen nurses with their assigned patients' and shift information were simulated over one thousand different scenarios. The comparison between mean values of performance measures from a thousand scenarios and the actual data are plotted in figure 3.

Figure 3(a) specifically shows the comparison of actual and simulated TADC. In the TADC comparisons, as well as TDC, TNPL, and WALK TIME comparisons shown in figures 3(b), 3(c), and 3(d) , purple curves represent the mean from the one thousand simulation scenarios while dark blue curves represent actual data. Ideally, it is desirable to have the dark blue curve overlapping with the pink curve. In TADC comparisons, the mean of the simulation scenarios approximates the actual data closely by picking up the pattern as well as the magnitude. Among the different

performance measures used in this research, TADC is the most important as it measures the amount of assigned direct care provided by nurses and directly impacts patient care and continuity of care.

Simulated and actual TDCs, shown in figure 3(b), compare another important performance measure in terms of nurse work load as well as patient care. It can be seen that, the mean TDC from simulation approximates the pattern of actual data closely. However, the plots show that TDC from simulation over-estimates the TDC of actual data. If the objective were to predict the TDC of nurses in isolation without any comparison, it would be desired to calibrate the simulation to reduce the magnitude of TDC. However, this research seeks only the balance, as shown in table 7, by comparing the maximum of a performance to the corresponding minimum. The resultant max-min ratio will not be altered by the discrepancy in the magnitude, neither by an over-estimation nor an under-estimation, as long as the pattern of the performance measure in simulation matches with the actual data as shown for TDC in figure 3(b). Also, if optimization of the system with respect to TDC, either minimization of nurse-workload or maximization of patient care, were the final goal, the discrepancy in the magnitude of the objective will not alter the optimal decision.

Figure 3(c) shows the comparison of actual and simulated TNPL. It can be seen from the figure that the simulation model provides TNPL that matches the pattern of actual data and hence provides reliable max-min ratio for TNPL. However, the plots show that TNPL from simulation under-estimates the TNPL of actual data and should not be used to interpret the magnitude of TNPL of individual nurses in isolation. Simulated and actual WALK TIME, shown in figure 3(d), compare the performance measure that accounts for the amount of time a nurse walks during the entire shift. In this research, a deterministic time is added depending on the distance between two locations a nurse walks in the simulation. In reality, these walk-times are stochastic as different nurses at different times would spend different amounts of time walking between the same locations. As expected, it can be observed that simulated WALK TIMEs have less variability across the nurses. It also shows that the simulation approximates the magnitude of real walking time reasonably.

The above discussion shows that performance measures of the simulation model approximate the pattern of real data, and to a certain extent the magnitude. Hence, it represents the actual system well enough to arrive at conclusions about the nurse work load balance in terms of the ratios introduced in table 7 without further calibration of the simulation.

7 Simulation adaptability

As mentioned earlier, the simulation model developed in this research is hospital specific and has to be adapted accordingly to use in different hospitals. Section 3 introduced the variables used in this research. The number of variables for the data mining and in turn for the simulation would depend on the availability of data in a given hospital. Apart from the variables discussed in section 3, other variables such as, experience level and education level of nurses,

secondary diagnosis, length of stay, and age of patients, would be interesting to consider. For some hospitals, there could be fewer variables than in this model due to unavailability of data. Even for the same variables, it is very likely that the number of categories will be different at a different hospital. In any case, CART should be applied on the hospital specific data set to fit the five tree structures discussed in section 4. The choice of independent variables for each tree can differ from the ones used in this research. The selection of independent variables can be made based on the variable importance scores from CART and practical significance of the variables to the hospital. Once data mining is completed simulation is performed as explained in section 4.3. The generic C++ simulation code written in this research can read any tree structure using the concept of structures and pointers (Foster and Foster, 2003; Lafore, 2000) and simulate by sampling repeatedly from trees until the entire shift period is exhausted. This way of coding makes it easy to adapt the simulation code to different hospitals. Even for the same hospital, when new data is available and hence new trees are built in CART, the simulation model could update itself by reading and simulating from the new tree structures. As a result, this research introduces a hospital specific and yet an easily adaptable simulation model to hospitals.

8 Conclusions and future work

A novel approach to construct a nurse activity simulation model from real data was developed using classification and regression trees. Classification trees provide transition probabilities to determine where a nurse will go next. Regression trees combined with kernel density estimates determine the amount of time she will spend once she goes to a new location. Simulation models developed with this approach will be significantly more efficient than the simulation models that consider all possible combinations. Optimal nurse-patient assignments can be identified by applying simulation-optimization methods, such as Atlason et al. (2004) and Fu and Hu (1997), to our resulting simulation model. Implementing this methodology as an information technology tool in hospitals will help charge nurses make better decisions on nurse-patient assignments for a shift. As a result, better care for patients, balanced work loads for nurses, and cost savings for hospitals can be achieved.

In the SIMNA model presented in this paper, it was assumed that there are no patient admissions and discharges. However, it is common to have a discharge or/and admission during a given shift. Incorporating patient admissions and discharges to evaluate nurse-patient assignments would be an interesting topic for future work.

9 Acknowledgements

This research is supported by the Robert Wood Johnson Foundation grant number 053963. We thank Terry Clark from Baylor Medical Center at Grapevine TX and Patricia G. Turpin from the School of Nursing at The University of Texas

at Arlington, for providing us data for this research.

References

Aickelin, U. and Dowsland, K. A. (2003). An indirect genetic algorithm for a nurse scheduling problem, *Computing and Operational Research* **31**(5): 761 – 778.

Aickelin, U. and Paul, W. (2004). Building better nurse scheduling algorithms, *Annals of Operations Research* **124**(1 - 4): 159 – 177.

Aiken, L. H., Clarke, S., Sloane, D., Sochalski, J. and Silber, J. (2002). Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction, *The journal of the American Medical Association* **288**: 1987 – 1993.

AONE (2003). Policy statement on mandated staffing ratios, http://www.aone.org/aone/docs/psratios.pdf (accessed September 2007).

Atlason, J., Epelman, M. A. and Henderson, S. G. (2004). Call center staffing with simulation and cutting plane methods, *Annals of Operations Research* **127**: 333 – 358.

Azaiez, M. N. and Sharif, S. S. A. (2005). A 0-1 goal programming model for nurse scheduling, *Computers & Operations Research* **32**: 491 – 507.

Bailey, N. T. (1952). A study of queues and appointment systems in hospital outpatient departments, with a special reference to waiting times, *Journal of Royal Statistics Society* **A14**: 185 – 199.

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

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

Beddoe, G. R. and Petrovic, S. (2006). Selecting and weighting features using a genetic algorithm in a case-based reasoning approach to personnel rostering, *European Journal of Operational Research* **175**: 649 – 671.

Benton, W. C. (1994). A decision modes for shift scheduling of nurses, *European Journal of Operational Research* **74**(3): 519 – 527.

Bettonvil, B. and Kleijnen, J. P. C. (1997). Searching for important factors in simulation models with many factors: sequential bifurcation, *European Journal of Operational Research* **96**: 180 – 194.

Breiman, L., Friedman, J. H., Oishen, R. A. and Stone, C. J. (1984). *Classification And Regression Trees*, Wadsworth, Belmont, California.

Burke, E. K., Caumaecker, P. D. and Petrovic, S. (2001). Variable neighbourhood search for nurse rostering problems, *Proceedings of 4th Metaheuristics International Conference*, Porto, Portugal.

Burke, E. K., Cowling, P. and Caumaecker, P. D. (2001). A memetic approach to the nurse rostering problem, *Applied Intelligence special issue on Simulated Evolution and Learning* **15**: 199 – 214.

CDHS (2005). Nurse-to-patient staffing ratio regulations, http://www.dhs.ca.gov/lnc/NTP/default.htm (accessed January 2006).

Cervellera, C., Chen, V. C. P. and Wen, A. (2006). Optimization of a large-scale water reservoir network by stochastic dynamic programming with efficient state space discretization, *European Journal of Operational Research* **171**: 1139 – 1151.

Cervellera, C., Wen, A. and Chen, V. C. P. (2006). Neural network and regression spline value function approximations for stochastic dynamic programming, *Computers and Operations Research* **34**: 70 – 90.

Chen, V. C. P., Gnther, D. and Johnson, E. L. (2003). Solving for an optimal airline yield management policy via statistical learning, *Journal of the Royal Statistical Society, Series* **C**(20): 1 – 12.

Chen, V. C. P., Ruppert, D. and Shoemaker, C. A. (1999). Applying experimental design and regression splines to high dimensional continues state stochastic dynamic programming, *Operations Research* **47**: 38 – 53.

Cheng, R. C. H. (1997). Searching for important factors: Sequential bifurcation under uncertainty, *Proceeding of the 1997 Winter Simulation Conference*, Piscataway, New Jersey, USA.

Clementson, A. T. (1966). Simulation applied to industry, *The Statistician* **16**(4): 339 – 350.

Draeger, M. A. (1992). An emergency department simulation model used to evaluate alternative nurse staffing and patient population scenarios, *Proceedings of the 1992 Winter Simulation Conference*, Arlington, Virginia, USA.

Dumas, M. B. (1984). Simulation modeling for hospital bed planning, *Simulation* **43**: 69 – 78.

Dumas, M. B. (1985). Hospital bed utilization: An implemented simulation approach for adjusting and maintaining appropriate levels, *Health Services Research* **20**: 43 – 61.

Epanechnikov, V. A. (1969). Nonparametric estimation of a multivariate probabilty density, *Theory Prob. Applic.* **14**: 153 – 158.

Evans, G. W., Gor, T. B. and Unger, E. (1996). A simulation model for evaluating personnel schedules in a hospital emergency department, *Proceedings of the 1996 Winter Simulation Conference*, Coronado, California, USA.

Foster, L. S. and Foster, W. D. (2003). *C by Discovery*, Galgotia, Daryaganj, New Delhi.

Friedman, J. H. (1991). Multivariate adaptive regression splines,, *The Annals of Statistics* **19**(1): 1 – 141.

Fu, M. C. and Hu, J. Q. (1997). *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*, Kluwer, Norwell, Massachusetts.

Gutjahr, W. J. and Rauner, M. S. (2007). An aco algorithm for a dynamic regional nurse-scheduling problem in austria, *Computers & Operations Research* **34**: 642 – 666.

Hancock, W. and Walter, P. (1984). The use of admissions simulation to stabilize ancillary workloads, *Simulation* **43**: 88 – 94.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag, New York.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*, Prentice Hall, New Jersey.

HIMSS (2006). Himss position statement, http://www.himss.org/content/files/PositionStatements/ AdvancedPositionOnMandatedNurseRatio.pdf (accessed September 2007).

HRSA (2002). Projected supply, demand, and shortages of registered nurses: 2000-2020, ftp://ftp.hrsa.gov/bhpr/nationalcenter/rnproject.pdf (accessed January 2006).

INGENIX (2003). *ICD-9-CM Professional For Hospitals: Volumes 1, 2 & 3*, St. Anthony Publishing/Medicode, Salt Lake City, UT.

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

Jones, M. C., Marron, J. S. and Sheather, S. J. (1996). A brief survey of bandwidth selection for density estmation, *Journal of the American Statistical Association* **91**(433): 401 – 407.

Jun, J. B., Jacbson, S. H. and Swisher, J. R. (1999). Application of discrete event simulation in health care clinics: A survey, *The Journal of the Operational Research Society* **50**(2): 109 – 123.

Kachhal, S. K., Klutke, G. A. and Daniels, E. B. (1981). Two simulation applications to outpatient clinics, *Proceedings of the 1981 conference on Winter simulation*, Atlanta, Georgia, USA.

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

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

Kirkby, M. P. (1997). Moving to computerized schedules: A smooth transition, *Nurse Management* **28**: 42 – 44.

Klein, R. W., Dittus, R. S., Roberts, S. D. and Wilson, J. R. (1993). Simulation modeling and health-care decision making, *Medical decision making* **13**(4): 347 – 354.

Kreke, J., Schaefer, A. J., Angus, D., Bryce, C. and Roberts, M. (2002). Incorporating biology into discrete event simulation models of organ allocation, *Proceedings of the 2002 Winter Simulation Conference*, San Diego, California, USA.

Kumar, A. P. and Kapur, R. (1989). Discrete simulation application-scheduling staff for the emergency room, *Proceedings of the 1989 Winter Simulation Conference*, Washington DC, USA.

Lafore, R. (2000). *Object-Oriented Programming in Turbo C++*, Galgotia, Daryaganj, New Delhi.

Law, A. M. and Kelton, W. D. (2001). *Simulation Modeling and Analysis*, McGrawHill, New York.

Le Cun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W. and Jackel, L. (1990). Handwritten digit rcognition with a back propogation network,, *Advances in Neural Information Processing Systems* **2**.

Lim, T., Uyeno, D. and Vertinsky, I. (1975). Hospital admission systems: A simulation approach, *Simulation and Games* **6**: 188 – 201.

McKay, K., Buzacott, J. and Strang, C. (1986). Software engineering applied to discrete event simulation, *Proceedings of the 1986 Winter Simulation Conference*, USA.

Miller, H. E., Pierskalla, W. P. and Rath, G. J. (1996). Nurse scheduling using mathematical programming, *Operations Research* **24**(5): 857 – 870.

Mullinax, C. and Lawley, M. (2002). Assigning patients to nurses in neonatal intensive care, *Journal of the Operational Research Society* **53**: 25 – 35.

Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996). *Applied Linear Statistical Models*, WCB McGraw-Hill, Boston, Massachusetts.

Pilla, V. L., Rosenberger, J. M., Chen, V. C. P. and Smith, B. C. (2005). A statistical computer experiments approach to airline fleet assignment, *Technical Report COSMOS-05-03*, The University of Texas at Arlington, Department of Industrial and Manufacturing Systems Engineering (Available from http://ieweb.uta.edu/TechReports/COSMOS-05-03.pdf).

Punnakitikashem, P., Rosenberger, J. M. and Behan, D. F. (2006). Stochastic programming for nurse assignment, *Computational Optimization and Applications* p. to appear.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, University Press, Cambridge.

Sheather, S. J. (2004). Density estmation, *Statistical Science* **19**(4): 588 – 597.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estmation, *Journal of Royal Statistical Society.Series B* **53**(3): 683 – 690.

Shechter, S. M., Bryce, C., Alagoz, O., Kreke, J. E., Stahl, J. E., Schaefer, A. J., Angus, D. and Roberts, M. (2005). A clinically based discrete event simulation of end-stage liver disease and the organ allocation process, *Medical Decision Making* **25**(2): 199 – 209.

Shen, H. and Wan, H. (2005). Controlled sequential factorial design for simulation factor screening, *Proceedings of the 2005 Winter Simulation Conference*, Orlando, Florida, USA.

Sheppard, S. (1983). Applying software engineering to simulation, *Simulation* **10**(1): 13 – 19.

SHS (2005). Nurse-to-patient staffing ratio regulations, http://iienet2.org/uploadedFiles/SHS/Resource_Library/Details/positionPaper.pdf (accessed September 2007).

Siddappa, S., Gnther, D., Rosenberger, J. M. and Chen, V. C. P. (2006). A statistical modeling approach to airline revenue management, *Technical Report IMSE-06-04*, The University of Texas at Arlington, Department of Industrial and Manufacturing Systems Engineering (Available from http://ieweb.uta.edu/TechReports/IMSE-06-04.pdf).

Silverman, B. W. (1978). Choosing window width when estimating a density, *Biometrika* **65**(1): 1 – 11.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Chapman and Hall, London.

Smith, E. A. and Warner, H. R. (1971). Simulation of a multiphasic screening procedure for hospital admissions, *Simulation* **17**: 57 – 64.

Sundaramoorthi, D., Chen, V. C. P., Kim, S. B., Rosenberger, J. M. and Behan, D. F. B. (2006). A data-integrated nurse activity simulation model, *Proceedings of the 2006 Winter Simulation Conference*, Monterey, California, USA.

Sundaramoorthi, D., Chen, V. C. P., Rosenberger, J. M. and Green, D. F. B. (2005). Knowledge discovery and mining for nurse activity and patient data, *Proceedings of the 2005 IIE Annual Conference*, Atlanta, Georgia, USA.

Sundaramoorthi, D., Chen, V. C. P., Rosenberger, J. M., Kim, S. B. and Behan, D. F. B. (2006). Using classification and regression trees for a nurse activity simulation, *Proceedings of the 2006 IIE Annual Conference*, Orlanda, Florida, USA.

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

Tsai, J. C. C., Chen, V. C. P., Lee, E. K. and Johnson, E. L. (2003). Parallelization of the mars value function approximation in a decision-making framework for wastewater treatment, *Technical Report COSMOS-03-02*, The University of Texas at Arlington, Department of Industrial and Manufacturing Systems Engineering (Available from http://ieweb.uta.edu/TechReports/COSMOS-03-02.pdf).

Vericourt, F. d. and Jennings, O. B. (2006). Nurse-to-patient ratios in hospital staffing: a queuing perspective, http://faculty.fuqua.duke.edu/%7Efdv1/bio/ratios3.pdf (accessed July 2006).

Walpole, R. E., Myers, R. H., Myers, S. L. and Ye, K. (2002). *Probability & Statistics for Engineers & Scientists*, Prentice Hall, Upper Saddle River, New Jersey.

Walts, L. M. and Kapadia, A. S. (1996). Patient classification system: an optimization approach, *Health Care Management Review* **21**(4): 75 – 82.

Warner, D. M. (1976). Scheduling nursing personnel according to nursing preferences: A mathematical approach, *Operations Research* **24**: 842 – 856.

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

Wong, G. Y. C. and Chan, A. H. W. (2004). Constraint-based rostering using meta-level reasoning and probability-based ordering, *Computers & Operations Research* **17**: 599 – 610.

Youle, P. V., Tocher, K. D., Jessop, W. N. and Musk, F. I. (1959). Simulation studies of industrial operations, *Journal of Royal Statistical Society, Series A (General)* **122**(4): 484 – 510.

Zenios, S. A., Wein, L. M. and Chertow, G. M. (1999). Evidence-based organ allocation, *American Journal of Medicine* **107**(1): 52 – 61.

Zilm, F., Arch, D. and Hollis, R. B. (1983). An application of simulation modeling to surgical intensive care bed need analysis in a university hospital, *Hospital and Health Services Administration* **28**: 82 – 101.

Biographies

DURAI SUNDARAMOORTHI is a Lecturer of Engineering Management and Systems Engineering at the University of Missouri-Rolla. He holds a B.E. in Mechanical Engineering from Bharathiar University, and M.S. and Ph.D in Industrial Engineering from The University of Texas at Arlington. His research interests are data mining, simulation modeling and simulation optimization. He is interested in health care, supply chain, environmental engineering, and financial engineering applications. When he was a graduate student, he interned with Global Logistics group of FedEx, Control Systems group of General Electric, and Quality Control group of Thomas & Betts. He worked as a graduate research associate at the Center on stochastic modeling, optimization and statistics during his dissertation. He is a member of INFORMS, IIE, Tau Beta Pi, and Alpha Pi Mu.

Dr. VICTORIA C.P. CHEN is an Associate Professor of Industrial and Manufacturing Systems Engineering at The University of Texas at Arlington. From 1993-2001, she was on the Industrial and Systems Engineering faculty at the Georgia Institute of Technology. She holds a B.S. in Mathematical Sciences from The Johns Hopkins University, and M.S. and Ph.D. in Operations Research and Industrial Engineering from Cornell University. Dr. Chen's primary research interests utilize statistical methodologies to create new methods for operations research problems appearing in engineering and science. She has expertise in the design of experiments and statistical modeling, particularly for computer experiments and stochastic optimization. She has studied applications in inventory forecasting, airline optimization, water reservoir networks, wastewater treatment, and air quality. Through her statistics-based approach, she has developed computationally-tractable methods for continuous state stochastic dynamic programming, yield management, and environmental decision-making.

Dr. JAY M. ROSENBERGER is an Assistant Professor of Industrial and Manufacturing Systems Engineering at The University of Texas at Arlington. He has a B.S. in Mathematics from Harvey Mudd College, an M.S. in Industrial Engineering from University of California at Berkeley, and a Ph.D. in Industrial Engineering from the Georgia Institute of Technology. His research interests include mathematical programming and simulation in transportation, defense, and health care. He is the original developer of SimAir, a simulation of airline operations, which is currently used by many airlines and airline-consulting firms around the world. Dr. Rosenbergers graduate research on airlines won the First Place 2003 Pritsker Doctoral Dissertation award. Prior to joining the faculty at UTA, Dr. Rosenberger worked in the Operations Research and Decision Support (ORDS) Department at American Airlines.

Dr. SEOUNG BUM KIM is an Assistant Professor of Industrial and Manufacturing Systems Engineering at the University of Texas at Arlington. He received an M.S. in Industrial and Systems Engineering in 2001, an M.S. in

Statistics in 2004, and a Ph.D. in Industrial and Systems Engineering in 2005 from the Georgia Institute of Technology. He was awarded the Jack Youden Prize as the best expository paper in Technometrics for the Year 2003. His research interests include data mining, bioinformatics, environmetrics, and multiple hypotheses testing.

DEBORAH F. BUCKLEY-BEHAN is a Clinical Instructor in the School of Nursing at the University of Texas at Arlington, a Nurse Researcher and is a medical-surgical certified registered nurse. Her clinical experiences include medical-surgical, and critical care. Her research interests are 'Health of Nurses'. She received an associate degree in nursing from University of Arkansas, Fayetteville, AR; a bachelor's degree from Southwest State Missouri University, Springfield, MO; Master's from the University of Oklahoma, Tulsa, OK; and is currently writing dissertation at Texas Woman's University, Denton, TX..

Table 2: Variable importance scores for regression and classification trees

Tree Type	Med/Surg I	Med/Surg II	Mom/Baby	High-Risk Labor
Regression Tree				
X_L	100.00	100.00	100.00	100.00
X_D	11.20	60.02	7.54	70.42
X_{NT}	17.17	17.70	16.76	14.78
X_T	29.76	13.83	24.48	8.64
X_S	10.35	6.82	9.82	4.75
X_A	13.43	73.03	10.25	65.36
"Location Type" Tree				
X_{NT}	41.92	70.66	100.00	100.00
X_T	100.00	100.00	40.60	16.47
X_S	33.46	95.07	15.59	4.88
"Location" Tree ($X_A = 1$)				
X_{P1L}	100.00	68.36	100.00	100.00
X_{P2L}	67.21	100.00	72.95	76.26
X_{NT}	0.86	3.11	7.63	2.75
X_T	4.52	8.16	17.84	14.97
X_S	3.03	3.22	11.96	12.08
"Location" Tree ($X_A = 2$)				
X_{P1L}	100.00	100.00	100.00	100.00
X_{P2L}	52.56	48.53	66.37	82.15
X_{NT}	3.08	10.68	3.42	34.14
X_T	5.79	6.17	4.10	4.57
X_S	2.26	3.39	1.39	2.12
"Location" Tree ($X_A = 0$)				
X_{P1L}	100.00	96.47	100.00	100.00
X_{P2L}	65.35	100.00	68.35	94.09
X_{NT}	5.50	11.69	6.33	9.54
X_T	6.59	16.22	9.57	28.22
X_S	2.38	6.67	2.81	10.87

Table 3: Performance of Gaussian and triangular kernels

Care Unit	Gaussian	Triangular	Tie
MED/SURG I			
J_R=181			
Range I wins	26	155	
Range II wins	45	136	
Range III wins	77	105	
Range IV wins	36	145	
% wins	25%	75%	
Ter. node wins	13	135	33
% Ter. node wins	7%	75%	18%
MED/SURG II			
J_R=109			
Range I wins	15	94	
Range II wins	24	85	
Range III wins	31	78	
Range IV wins	18	91	
% wins	20%	80%	
Ter. node wins	7	92	10
% Ter. node wins	6%	85%	9%
MOM/BABY			
J_R=123			
Range I wins	13	110	
Range II wins	25	98	
Range III wins	31	92	
Range IV wins	18	105	
% wins	18%	82%	
Ter. node wins	9	104	10
% ter. node wins	7%	85%	8%
HIGH-RISK			
J_R=49			
Range I wins	9	40	
Range II wins	13	36	
Range III wins	19	30	
Range IV wins	3	46	
% wins	22%	78%	
Ter. node wins	3	38	8
% ter. node wins	6%	78%	16%

Table 4: Bandwidth tuning for terminal node 9 of medical/surgical unit I

Bandwidth Tuning	Sim. Fraction	Actual Fraction	Diff.
BEFORE h=52.46			
range I	0.070110	0.278986	0.208876
range II	0.083750	0.244842	0.161092
range III	0.075310	0.086039	0.010729
range IV	0.770830	0.390133	-0.380697
AFTER h=8.46			
range I	0.266580	0.278986	0.012406
range II	0.234510	0.244842	0.010332
range III	0.094890	0.086039	-0.008851
range IV	0.404020	0.390133	-0.013887

Table 5: Numerical values of levels in different care units and number of combinations

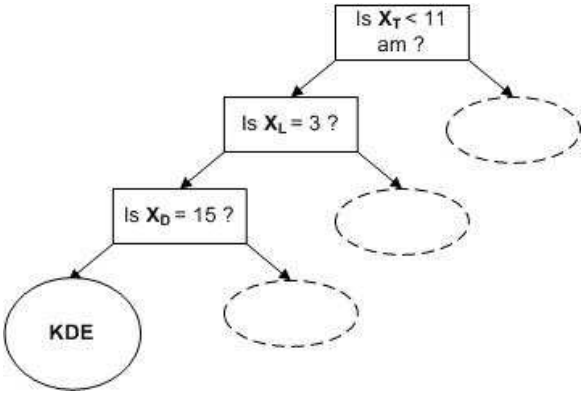
Variable Level	Care Unit Med/SurgI	Med/SurgII	Mom/Baby	High-Risk
N_S	5	5	5	5
N_T	24	24	24	24
N_{NT}	4	8	8	7
N_D	19	21	10	8
N_L	34	32	52	52
R	26	26	32	10
N_A	3	3	3	3
NPC_{lt}	480	960	960	840
J_{LT}	145	259	322	196
NPC_l	$> 10^{46}$	$> 10^{47}$	$> 10^{47}$	$> 10^{17}$
J_1	397	440	271	69
J_2	1816	1554	1194	96
J_0	262	268	118	38

Table 6: SIMNA assignment policy results for medical/surgical unit I

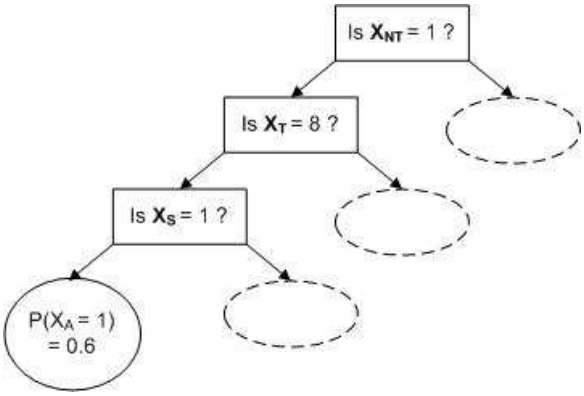
Assignment Policy	Assigned Patient Locations	Assigned Patient Diagnoses	TADC (min)	TUADC (min)	TDC (min)	TNPL (min)	Walk Time (min)
RANDOM							
Nurse1 (LVN)	4, 6, 10, 17, and 18	1, 6, 16, 8 and 14	92	119	211	158	116
Nurse2 (RN)	3, 13, 15, 19, and 26	9, 16, 13, 12 and 15	152	127	279	118	87
Nurse3 (RN)	1, 7, 14, 16, and 20	14, 10, 3, 4 and 8	220	84	304	94	87
Nurse4 (RN)	2, 5, 8, 9, 23, and 24	13, 8, 3, 6, 8, and 15	185	127	312	83	88
Total			651	459	1107	455	379
HEURISTIC							
Nurse1 (LVN)	9, 10, 13, 14, 23, and 26	6, 16, 16, 3, 8, and 15	122	74	196	173	115
Nurse2 (RN)	5, 7, 15, 16, and 20	8, 10, 13, 4 and 8	209	95	304	93	87
Nurse3 (RN)	2, 4, 6, 8, and 19	13, 1, 6, 3 and 12	163	149	312	83	89
Nurse4 (RN)	1, 3, 17, 18, and 24	14, 9, 8, 14 and 15	192	126	318	83	84
Total			688	446	1132	434	376
CLUSTER							
Nurse1 (LVN)	1, 4, 14, 17, 20, and 24	14, 1, 3, 8, 8, and 15	194	16	210	171	102
Nurse2 (RN)	3, 6, 8, 10, and 13	9, 6, 3, 16 and 16	172	139	311	83	90
Nurse3 (RN)	2, 16, 19, 23, and 26	13, 4, 12, 8 and 15	125	158	283	106	94
Nurse4 (RN)	5, 7, 9, 15 and 18	8, 10, 6, 13 and 14	107	195	302	89	94
Total			600	520	1107	451	381
STOCHASTIC PROGRAMMING							
Nurse1 (LVN)	10, 13, 14, 16 and 17	16, 16, 3, 4 and 8	164	45	209	172	104
Nurse2 (RN)	3, 7, 20, 24 and 26	9, 10, 8, 15 and 15	222	85	307	101	75
Nurse3 (RN)	1, 2, 4, 6, 8, and 23	14, 13, 1, 6, 3, and 8	193	120	313	82	89
Nurse4 (RN)	5, 9, 15, 18 and 19	8, 6, 13, 14 and 12	115	187	302	89	94
Total			696	441	1132	446	363

Table 7: Maximum-to-minimum ratios for TADC, TDC, TDC of RNs, and Walking time

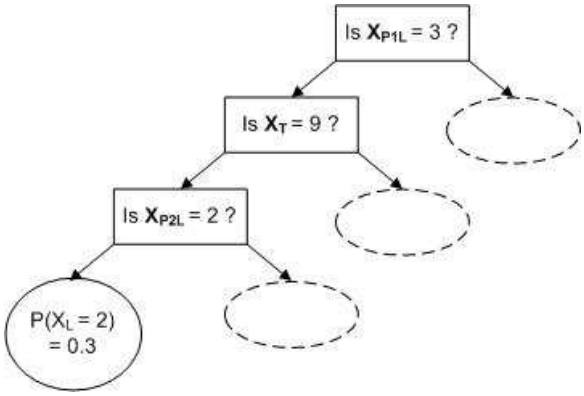
Assignment Policy	TADC	TDC	TDC (RNs)	Walk Time
Random	2.39	1.48	1.12	1.33
Heuristic	1.71	1.62	1.05	1.37
Cluster	1.81	1.48	1.10	1.13
Stochastic Prog.	1.93	1.50	1.04	1.39



(a) A Hypothetical Regression Tree.



(b) A Hypothetical "Location Type Tree".



(c) A Hypothetical "Location Tree".

Figure 1: Regression and classification tree structures.

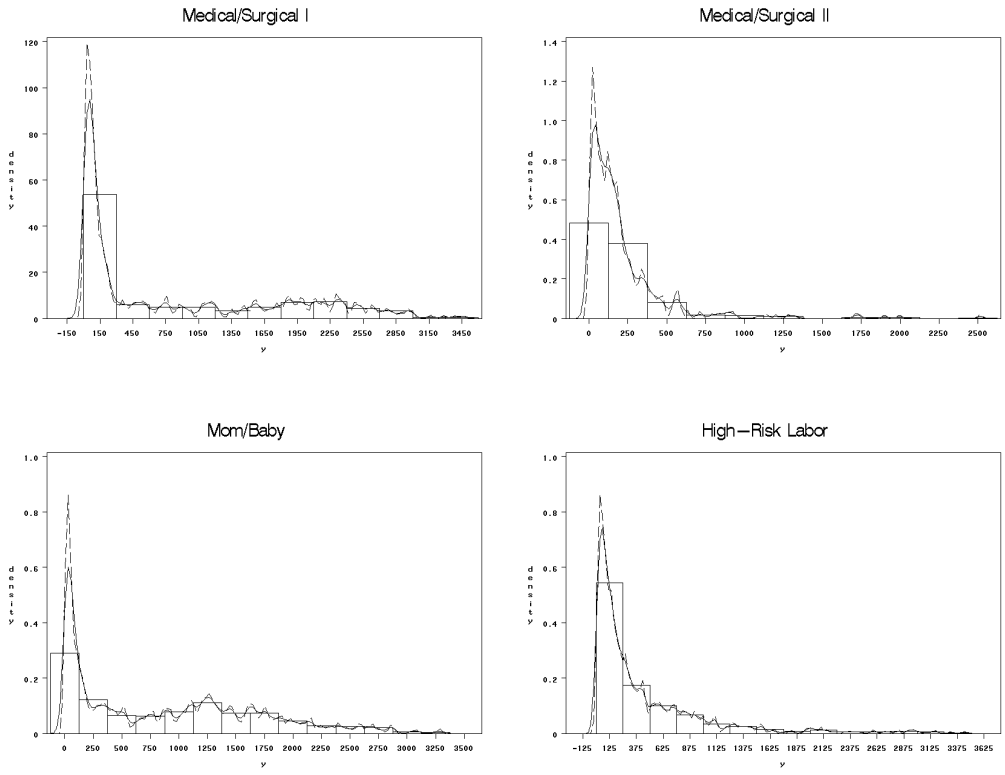
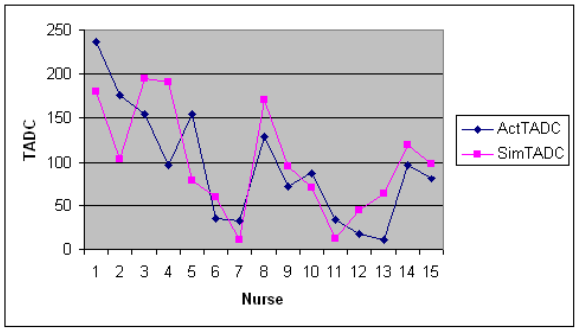
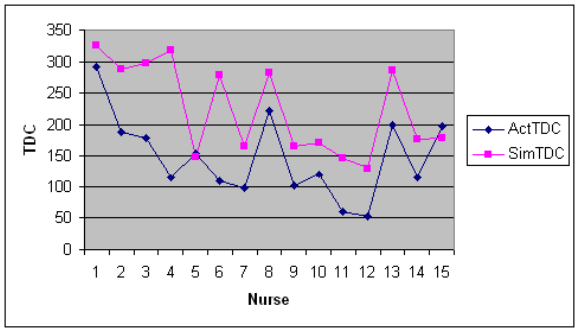


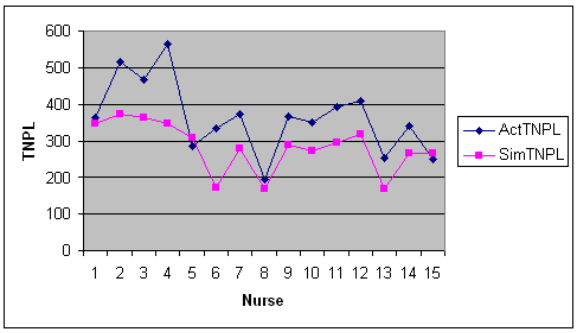
Figure 2: Kernel density estimates (Solid-Gaussian, and Broken-Triangular).



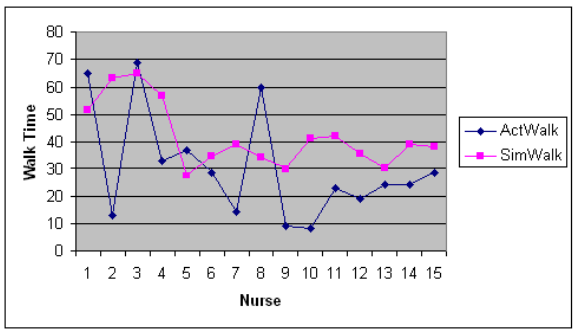
(a) Actual Vs. Simulated TADC



(b) Actual Vs. Simulated TDC



(c) Actual Vs. Simulated TNPL



(d) Actual Vs. Simulated WALK TIME

Figure 3: Comparison of actual data with simulated data.