

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

Received: 17 April 2008 / Accepted: 17 October 2008 / Published online: 22 November 2008
© Springer Science + Business Media, LLC 2008

Abstract This research develops a novel data-integrated *simulation* to evaluate *nurse–patient assignments* (SIMNA) based on a real data set provided by a northeast Texas hospital. Tree-based models and kernel density estimation (KDE) 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 the northeast Texas hospital are discussed.

Keywords Nurse–patient assignment · Nurse workload · Simulation

1 Introduction

The health care system in the United States has a shortage of nurses. According to the U.S. Department of Health and Human Services (DHHS), the national shortage for registered nurses was 110,000 or 6% in 2000. 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 [23]. These statistics show 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 [2]. 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 [11]. Such restrictions may reduce nurses' workloads, but they will unlikely resolve the issue because differences in workload among nurses depend upon the 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 resources across different hospitals. For these reasons,

D. Sundaramoorthi (✉)
Steven L. Craig School of Business,
Missouri Western State University,
Saint Joseph, USA
e-mail: dsundaramoorthi@missouriwestern.edu

V. C. P. Chen · J. M. Rosenberger · S. B. Kim
Department of Industrial & Manufacturing Systems
Engineering, University of Texas at Arlington,
Arlington, USA

V. C. P. Chen
e-mail: vchen@uta.edu

J. M. Rosenberger
e-mail: jrosenbe@uta.edu

S. B. Kim
e-mail: sbkim@uta.edu

D. F. Buckley-Behan
School of Nursing, University of Texas at Arlington,
Arlington, USA
e-mail: dbehan@uta.edu

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 [3, 22, 43]. All of 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 [1, 5–7, 10, 20, 25, 29, 35, 51] and does not address the nurse-to-patient assignment issue. Apart from the proposed model in this paper, Vericourt and Jennings [50] and Punnakitikashem et al. [37] are two other contemporary research papers that address the nurse-to-patient assignment issue. However, these research papers did not use real data as extensively as our approach in modeling 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 the northeast Texas hospital 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, factorial designs and screening methods were used to reduce the number of simulation variables [8, 14, 42]. 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, we present a new methodology to reduce the number of combinations and find transition probabilities for stochastic simulation models using data from

the northeast Texas Hospital. Tree-based models and kernel density estimates (KDE) were utilized to extract important knowledge about the workload of nurses from an encrypted data set provided by the northeast Texas hospital 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 [9], a data mining tool for prediction and classification, was applied to the northeast Texas hospital 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.

1.1 Contribution

There are two major contributions made in this research:

- This research introduces a tool to evaluate nurse-to-patient assignment policies, such as the ones described in Section 5, to identify good assignment policies well ahead of a shift. Prior to a shift at the northeast Texas Hospital, nurse supervisors assess the expected workload of nurses for the given set of patients and nurses in the care unit. The SIMNA model can aid them in their decisions by providing a tool to evaluate nurse-to-patient assignments from different preferred assignment policies.
- 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.

The rest of this paper is organized as follows: In Section 2, a literature review on nursing research, data mining, and simulation is provided. 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 model is

validated by comparing simulation results with the actual data. Section 7 presents a discussion on adaptability of the simulation model to a different hospital. In Section 8, we provide concluding remarks including discussions on a possible simulation-optimization approach to optimize nurse-to-patient assignments and other opportunities for future work.

2 Literature review

There are three major components in this research—nurse planning, data mining, and simulation modeling. This section 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 [1, 5–7, 10, 20, 25, 29, 35, 51]. The focus of this research is the nurse assignment stage of nurse planning. Mullinax and Lawley [36] 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. [37] formulated and solved a two-stage stochastic integer programming nurse assignment problem to minimize excess workload of nurses. Vericourt and Jennings [50], using a queuing approach, showed that the 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 did not differentiate assigned and unassigned patients of nurses, which is discouraged in practice for maintaining continuity of care. 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 actual 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. In unsupervised learning or clustering, one wants to observe only the features and have no measurements of the outcome. Supervised learning is the subject of interest in this research as

we deal with predicting the time spent and location for nurses. Classification and Regression Trees [9]—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. Application of data mining tools to health care problems is quite common and has produced a significant amount of literature. For instance, recently, Ceglowski and Churilov [12], and Ceglowski et al. [13] used self organizing maps, a clustering technique, to determine treatment paths of emergency room patients and Ramon et al. [38] used decision trees, first order random forests, naive Bayes, and tree augmented naive Bayes to predict patients' length of stay, patient survival, and endangering states.

2.3 Simulation modeling in health care

Studying industrial systems using simulation was prevalent as early as the late 1950s and early 1960s. Simulation modeling has been used to study a wide range of problems in health care [15, 17, 28, 34, 46]. In recent years, Zenios et al. [52], Kreke et al. [31], and Shechter et al. [41] utilized simulation models to study organ allocation systems. A comprehensive review of health care simulation models can be found in Klein et al. [30] and Jun et al. [27]. 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. By contrast, 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.

3 Data description

At the northeast Texas hospital, each nurse wears a locating device that transmits data to a repository, where the data automatically expire after 1 month. The hospital 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. The hospital 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 [33] was developed and applied to the data before our analysis. The 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.

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 related research, presented in Sundaramoorthi et al. [48, 49], seven variables were created to hold information on the previous seven locations. The simulation models developed with seven previous locations were found to overfit the pattern based on movements and, hence, were insensitive to other practically important variables. For this reason, the simulation presented here is unlike the one in Sundaramoorthi et al. [48, 49] because it 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 variables, 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 had 570,660, 418,683, 315,997, and 210,457 observations, respectively. Following the conclusions in Sundaramoorthi et al. [47] and further similar analysis presented in Sundaramoorthi et al. [48], 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 two dummy categories for empty patient rooms and non-patient locations. See INGENIX [24] for more details on diagnosis codes.
4. Shift: three weekday shifts (8 h each) and two weekend shifts (12 h each).
5. Hour: 24 h 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 later and X_{P2L} being the older between 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}, \dots, 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}, \dots, X_{DLR} are the variables representing primary diagnosis of patients in rooms 1 to R .

4 Data mining for simulation

From a methodological perspective, this research introduces a novel approach to the simulation community for constructing efficient simulation models based on data mining. This section introduces a technique to obtain transition probabilities for simulation from tree models that are discussed in Section 4.1. 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 call center with one hundred agents using a queuing network.

Table 1 Encryption example

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

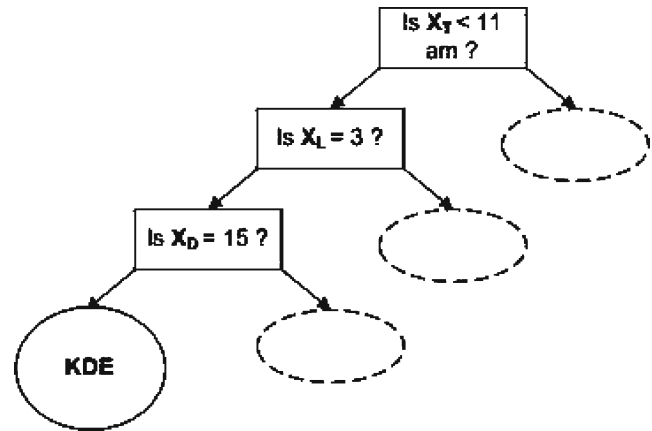
There is a potential dependence between service times of call center agents because of resource sharing and other human interactions among them. Consequently, assuming these service times are independent may lead to an inaccurate simulation model. To model the call center system with dependent service times, the modeler would need to determine the dependencies of each pair of service times using statistical tests of independence. This would require 10,000 such tests with an appropriate significance level. If multiple agents' service times were found to be statistically dependent, then the modeler would have to group the agents into sets in which the agents are dependent. Then, the modeler would have to develop multivariate distributions for each group that may have tens of variables. In practice though, the modeler would likely make potentially inaccurate 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 the northeast Texas hospital, 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. 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, it is not easy to justify ignoring or combining them without using a data mining tool like trees.

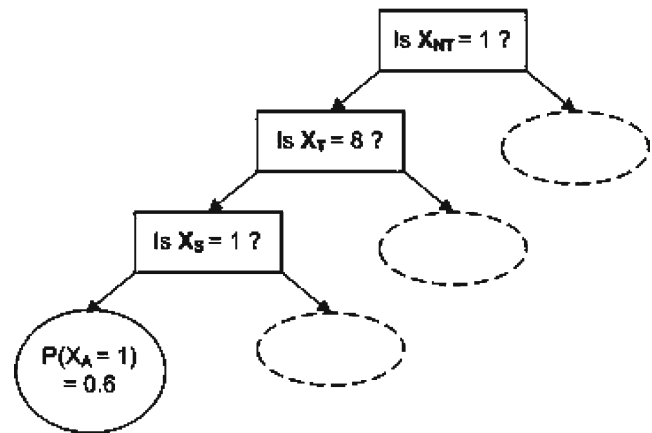
4.1 Classification and regression trees

Classification and Regression Trees (CART) are data mining tools for prediction and classification [9, 21]. CART utilizes recursive binary splitting to uncover structure in a high-dimensional space. On an application to a data set, CART will partition the input space into many disjoint sets in which 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} , ..., X_{DLR} , X_A , X_{P1L} , and X_{P2L} , and (b) a regression tree to predict

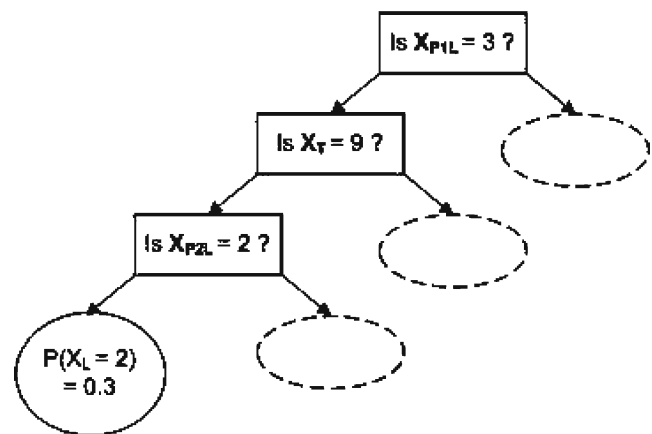
the amount of time a nurse will spend in a location based on the levels of X_S , X_T , X_{NT} , X_L , X_D , and X_A . A hypothetical regression tree is shown in Fig. 1a to illustrate a prediction of the amount of time a nurse would spend in a location. A question is asked at each



(a) A Hypothetical Regression Tree.



(b) A Hypothetical "Location Type Tree".



(c) A Hypothetical "Location Tree".

Fig. 1 Regression and classification tree structures

node of the tree. A data point that satisfies the question will belong to the left branch, but it will belong to the right branch 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 Fig. 1b and another “location tree” in Fig. 1c, 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, such as an unassigned patient room (0), an assigned patient room (1), or a non-patient room (2), is obtained from the location type classification tree based on the levels of X_S , X_T , and X_{NT} . Every data point in a “location tree” ends up in one of the terminal nodes of the tree depending on the levels of X_S , X_T , X_{NT} , X_{DL1} , ..., X_{DLR} , X_A , X_{PIL} , and X_{P2L} . In the terminal node, 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, $j = 1, \dots, J$ are the terminal nodes of a “location tree”; $n(1), \dots, n(J)$ are the numbers of observations in terminal nodes 1, ..., J , respectively; $l = 1, \dots, N_L$ 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} represents the number of terminal nodes of a “location type tree.” For a “location type tree”, $l = 0, \dots, 2$, representing unassigned patient rooms (0), assigned patient rooms (1), and non-patient rooms (2), are the levels of X_A .

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}, \dots, X_{ALR} and X_{DL1}, \dots, 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 of these intricate and subtle differences in the system without using a tool like CART.

While growing the trees, ten-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.

4.2 Estimation of time spent distribution

For each terminal node of the regression trees, kernel density estimation 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 $y_1, \dots, y_{n(j)}$ 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 Eq. 2 [45] as follows:

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

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 because this method is one of the best for optimizing bandwidth [26, 39, 40]. However, it should be noted that bandwidth selection is not precise and is often an “art.” Tuning of the bandwidths based on our desired criteria is discussed in Section 4.2.2. Random variables $Y(1), \dots, Y(J_R)$ denote the time spent (Y) in terminal nodes 1, ..., J_R , respectively. Kernel density estimates

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

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 Fig. 2.

4.2.1 Kernel choice

Kernel functions include uniform, Gaussian, triangular, Epanechnikov, quadratic, and cosine. 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 [40] were calculated for each terminal node of the regression tree using SAS[®]. Figure 2 and the normal probability plots in Sundaramoorthi et al. [47] 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., $(0, M/2]$, $(M/2, M]$, $(M, (M + M/2)]$, $((M + M/2), \infty)$, 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 of the $J_R \times 4$ 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 kernels were considered to be

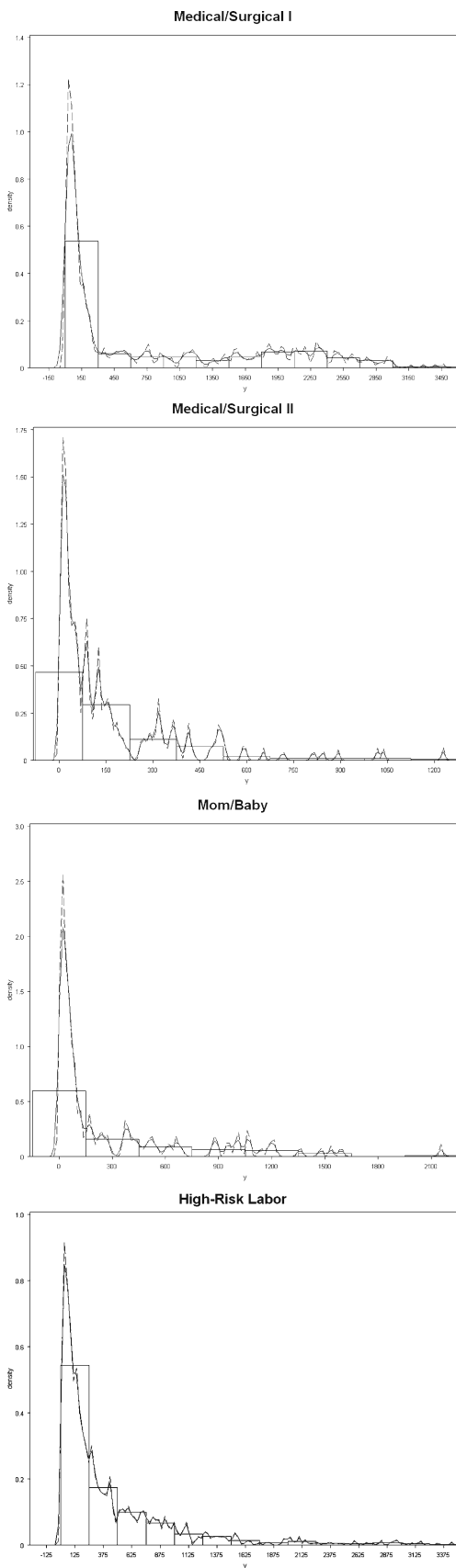


Fig. 2 Kernel density estimates (Solid-Gaussian, and Broken-Triangular)

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%

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 northeast Texas hospital data.

4.2.2 Bandwidth tuning

The accuracy of estimates depends more on choosing an appropriate bandwidth than on the choice of kernels [16, 44]. Bandwidth selection methods, including SJPI bandwidth estimates [40], 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 6-month period and have hundreds of thousands of observations for each care unit. With data collected over months, the different possible characteristics of the northeast Texas hospital 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 ± 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 I shown in Table 4 has realizations that violated the ± 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 of 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 the nurse spend there? After an initial simulation run in which nurses visit their assigned patients for an initial assessment, transition probabilities obtained by Eq. 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 the nurse 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 procedure of sampling location type, location, and time spent, shown in Algorithm 1, is repeated until the shift ends.

It has to be noted that dependencies of a nurse's visit to a new location with shift (X_S), hour of the day (X_T), nurse type (X_{NT}), diagnosis of patients on floor (X_{DL1}, \dots, X_{DLR}), and her/his assigned patients (X_{AL1}, \dots, X_{ALR}) are explicitly captured by "location type trees" and "location trees." These trees also implicitly capture low and high demand circumstances using assignment variables (X_{AL1}, \dots, X_{ALR}) and their combination with other variables, for instance, diagnoses (X_{DL1}, \dots, X_{DLR}) and time of the day (X_T) variables, while determining a new location for the nurse. Extraction of such structures by trees from the actual data would send nurses to locations appropriately based on demand. Once the nurse is in a new location, it is implicitly assumed that the task performed at the location, such as patient care in patient rooms, medical refills at medical-supply room, and charting at the nurses' station, and in turn the amount of time consumed for performing the task is independent of the demand in other locations. The amount of time to be spent in the new location is determined by kernel densities that use bandwidths that assure less than 1.5% deviation from the actual data. In addition, kernel density estimates in the regression tree terminal nodes implicitly model the variation of time spent with the uncaptured demand within the location. Repeated sampling from kernel density estimates for time spent data during multiple simulation scenarios would produce enough representation of different regions of the kernel density estimates, which would reflect the actual variation found in the real system. However, there are other potential dependencies, such as a dependency between time spent by a nurse at a location and the nurse's cumulative assigned and unassigned direct care, walk time, and a dependency between nurses, not explicitly captured with variables in this research. Inclusion of such factors explicitly in the tree models has the potential to improve modeling of demand factors and interactions.

The more efficient the simulation, the more useful it will be for making real-time decisions. For example, a charge nurse will assess the balance of nurse workload for a given nurse-to-patient assignment prior to a

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

Algorithm 1 Simulation procedure

Step 0:

1. Initialize all the variables specified in Section 3 to reflect the starting state at a care unit.
2. Start the clock time.
3. Nurse visits her / his patients and spends a constant amount of time for an initial assessment at the beginning of a shift.
4. Update X_{P1L} , X_{P2L} , clock time and, if necessary, X_T (every time the nurse is about to leave a location).
5. Nurse returns to nurses’ station.
6. Nurse spends a constant amount of time at nurses’ station.
7. Update X_{P1L} , X_{P2L} , clock time and, if necessary, X_T .

Step 1: Sample a location type (1—assigned patient rooms, 0—unassigned patient rooms, and 2—non-patient locations) to be visited by the nurse from the “location type tree.”

Step 2: Sample a specific new location to be visited by the nurse from an appropriate “location tree” for the given location type.

Step 3: Determine a deterministic walk time based on the distance between the current and the new location to be visited.

Step 4: Update clock time by adding the walk time and, if necessary, X_T .

Step 5: Move the nurse to the new location.

Step 6: Determine a random sample of time spent y from the kernel density estimate at the terminal node of the current state in the regression tree.

Step 7: Nurse spends y amount of time at the location.

Step 8: Update X_{P1L} , X_{P2L} , clock time and, if necessary, X_T

Step 9: If clock time is less than shift duration, Go To Step 1 else Stop.

shift. The simulation model could assist in this process provided its run time is sufficiently fast. The simulation model developed using trees, discussed in Section 4.1, requires only J_{LT} terminal nodes for sampling a loca-

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

Variable level	Care unit			
	Med/surg I	Med/surg II	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_{it}	480	960	960	840
J_{LT}	145	259	322	196
NPC_1	$> 10^{46}$	$> 10^{47}$	$> 10^{47}$	$> 10^{17}$
J_1	397	440	271	69
J_2	1,816	1,554	1,194	96
J_0	262	268	118	38

tion type and $J_0 + J_1 + J_2$ terminal nodes for sampling a location in simulation based on the patterns extracted from the data. Differences between NPC_{it} and J_{LT} , NPC_1 and $J_0 + J_1 + J_2$ given in Table 5, demonstrate that our approach is significantly more efficient. Also, the simulation procedure developed in this research, listed in Algorithm 1, shows that once tree models are built, there is no subjective input needed for the simulation. 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 actual system over a long period of time.

5 SIMNA experiments

A 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 I with a thousand different random seeds. A test problem with four nurses and twenty one patients was considered. SIMNA tested four assignment policies: a clustered assignment and three assignments from Punnakitikashem et al. [37]—the random assignment, the heuristic assignment, and the optimal assignment using Benders’ decomposition on a stochastic programming model. In the heuristic assignment, all of the nurses get the same number of patients when the number of nurses divides into the number of patients evenly. 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 of 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. [37], which seeks to balance the expected workload of RNs, by modeling the estimated direct and indirect care of individual patients, provided assignments for the fourth policy. In real life, nurses often perform indirect care, such as charting and medication preparation for a patient, at non-patient locations. It was practically impossible to break down the consolidated indirect care data at locations like the nurses' station and the medical-supply room for individual patients from the real data set. Hence, in the simulation, unlike

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	1,107	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	1,132	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	1,107	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	1,132	446	363

Simulations were performed in seconds but results are reported here in minutes. Due to rounding, minor discrepancies between sums and individual components are present

in Punnakitikashem et al. [37], indirect care of all the assigned patients of a nurse is modeled together. It should be noted that this does not affect the ability to evaluate the optimized assignment in SIMNA.

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 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 from the test problem are given in Table 7. For balancing TADC, the heuristic assignment performed best and the random assignment performed worst among the policies considered. For balancing TDC, the heuristic assignment performed the worst, and the other three were similar to each other. For balancing TDC for RNs, the heuristic and optimal assignments performed best, and the random assignment performed worst. Finally, for balancing walking time, the clustered assignment performed 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. To quantify the differences between the four policies, the squared differences between the individual ratios and the best ratio are provided in Table 8. The sum of the squared differences across all four performance measures for each policy is shown in the last column. Based on this measure, the clustered assignment is the best, and the random assignment, not surprisingly, is the least desirable. It should be noted that the above conclusions about the performance of policies are confined to the test problem and could differ for other problems.

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

Table 8 Squared differences between the individual ratios and the best ratio

Assignment policy	TADC	TDC	TDC (RNs)	Walk time	Sum
Random	0.4624	0.0000	0.0064	0.0400	0.5088
Heuristic	0.0000	0.0196	0.0001	0.0576	0.0773
Cluster	0.0100	0.0000	0.0036	0.0000	0.0136
Stochastic prog.	0.0484	0.0004	0.0000	0.0676	0.1164

Prior to a shift, SIMNA results can aid the charge nurse in determining appropriate nurse-to-patient assignments. In theory, perfect workload balance could be achieved even with nurses assigned to a significantly fewer or higher number of patients than others. However, at the northeast Texas hospital an effort is made to assign the same number of patients to each nurse. Occasionally nurses have one or even two patients more than other nurses due to patient admissions and discharges, and divisibility issues. If the balance in workload is not satisfactory, the nurse supervisor can reassign patients or hire an agency nurse for that shift to redistribute and balance the workload while maintaining the same number of patients for nurses. Hiring agency nurses would likely cost more in the short term, but would yield better patient care and retention of nurses in the longer run. Thus, SIMNA upon installation in hospitals will aid charge nurses and management to make decisions about nurse-patient assignments based on the dynamics learned from the system itself.

6 Simulation validation

Among different steps in a simulation modeling process, 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 verification and validation steps were among those, performed as part of the validation process in this data-integrated simulation modeling approach.

1. **Tree Structure:** The tree structures were checked 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 that the entire shift duration is within reason.
3. **Kernel Density:** The kernel and bandwidth validations, presented in Section 4.2, ensured that the

regression trees provided a reliable approximation of the data.

- 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 create a tool to identify policies that provide 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 6 and 7. As part of the main validation, the actual TADC, TDC, TNPL, and WALK TIME of 15 nurses arbitrarily chosen to represent the entire data set were compared with that of simulated data. An effort was made to select the 15 nurses from different parts of the real data. For instance, three nurses were chosen from each of the five shifts to avoid any bias towards a shift. The 15 arbitrarily chosen nurses with their assigned patients' and shift information were simulated over 1,000 different scenarios. The comparison between the mean values of the performance measures from a thousand simulated scenarios and the actual data of the 15 nurses, chosen from different shifts, are plotted in Fig. 3. It should be noted that the comparison of the actual and the simulated max–min ratios of the performance measures of the 15 shifts would be more appropriate as it measures the balance in workload directly. While composing max–min ratio data for individual shifts, it was found that for certain day and evening shifts, even when a shift started with a specific set of patients and nurses, the set often changed because of nurses working half shifts, float units, and relief time, and patients getting admitted or discharged. Also, nurse–patient assignments were slightly altered when nurses or patients changed during the shift. It was practically impossible to consolidate the data for the entire shift to calculate max–min ratios, especially the 'min' part, without any bias towards nurses, shifts, and time of the day. For this reason, it was decided to compare performance measures of nurses working full shifts directly instead of ratios of performance measures. The current validation approach, even with a bias towards the nurses working the full shift, should be better than ratios that either include approximations or totally ignore nurses who worked a fraction of an entire shift. An alternate way to tackle this issue would be to incorporate patient admissions and discharges, floating nurses, break times, and half shifts in the simulation model to reflect the reality, which would allow a comparison of max–min ratios from the simulation to the actual data set. Unfortunately, the present data set

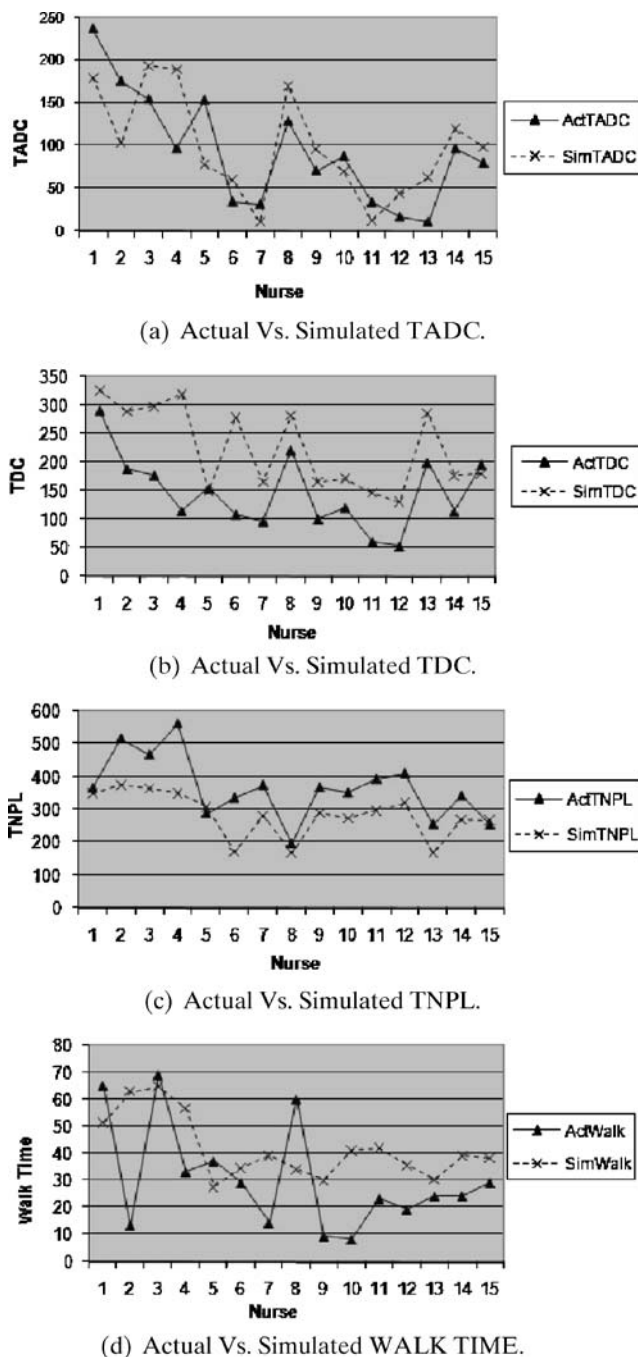


Fig. 3 Comparison of actual data with simulated data

from the northeast Texas hospital does not explicitly reveal admits, discharges, half shifts, float nurses, break times, and relief nurses information. Incorporation of these features and events in subsequent versions of this simulation would be useful if such data or standards become available.

Figure 3a specifically shows the comparison of actual and simulated TADC. In the TADC comparisons, as

well as the TDC, TNPL, and WALK TIME comparisons shown in Fig. 3b, c, and d, dotted curves represent the mean from the 1,000 simulation scenarios, while solid curves represent actual data. Ideally, it is desirable to have the dark solid curve overlapping with the dotted curve. In the figure, different nurses were joined by curves, as if nurses were continuous, just to visualize the overall difference between the actual performance and the simulated performance. In the TADC comparisons, the mean of the simulation scenarios estimates 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 Fig. 3b, 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 the simulation estimates the pattern of actual data closely. However, the plots show that the TDC from the simulation over-estimates the TDC of the 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 should 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 the simulation matches with the actual data as shown for TDC in Fig. 3b. 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 should not alter the optimal decision.

Figure 3c 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 ratios for TNPL. However, the plots show that the TNPL from the simulation under-estimates the TNPL of the 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 Fig. 3d, 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 estimates the magnitude of real walking time reasonably.

The above discussion shows that performance measures of the simulation model estimate 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 Sensitivity and adaptability

Sensitivity analysis would be needed if the simulation input involves either uncertain parameters or uncertain functional forms. In the traditional simulation input modeling, the uncertain parameter(s) or uncertain distribution(s) would provide realizations to one or more interrelated simulation “events.” Such a simulation model is also complex in construction and execution because of its interdependence of individual “events” and their outcomes.

There is a subtle difference between this research and the traditional simulation input modeling. This research seamlessly integrates the knowledge gained from the data mining algorithm, and in turn the real information from the system to the simulation. Once tree structures are built, simulation is merely an exercise of sampling repeatedly from alternate tree structures until the end of the shift condition is met. Hence, this research has identified a technique to represent a dynamically evolving system using repeated sampling from static tree structures. Unlike input parameters and distributions in traditional input modeling, the tree structures are not uncertain for a given splitting rule. CART software provides six splitting rules for classification trees: Gini, Symmetric Gini, Entropy, Class Probability, Twoing, and Ordered Twoing. As per the CART manual, depending upon the choice of splitting rule, the accuracy may differ as much as 10%. The Class Probability splitting rule was chosen in this research as it forces CART to build class probability trees instead of classification trees. In a traditional classification tree model with two classes, for example, a terminal node with 51% of Class One data and 49% of Class Two data will be classified as Class One. For future predictions, the model will predict Class One for every instance of that state. In this research, the focus is to obtain transition probabilities for the simulation from the terminal node rather than the classification itself. In theory, the

probability distributions from the tree built with the Class Probability rule are more accurate than the other choices. CART software provides two splitting rules for regression trees: Least Squares and Least Absolute Deviation (LAD). The default method is Least Squares. Least Squares penalizes deviations away from the mean more, proportional to their squares, than the deviations closer to the mean. This way of penalizing is preferred in this research over LAD to find states that would tend to have identical time spent data. It should be noted that in theory, the Class Probability rule and the Least Squares rule are the best for this application. Once the choice of splitting rules is justified, the tree structures are not uncertain. Hence, sensitivity analysis on simulation output based on different tree structures is unnecessary.

To simplify the complexity of the simulation, a deterministic simulation might be preferred in some applications rather than the stochastic simulation discussed in this research. If such a deterministic simulation is preferred, one would build traditional classification trees and use the classified location types/locations at terminal nodes to determine nurse transitions. Similarly, the average value of the observations in each terminal node would be used instead of sampling from kernel density functions. In that case, quantifying the differences in modeling assumptions in terms of the accuracy of simulation models developed from different tree structures would be an interesting direction for research. However, the comparison of different modeling assumptions is not necessary for this application because of the stochastic nature of the actual system.

As mentioned earlier, the simulation model developed in this research is hospital specific and has to be adapted accordingly to use in different hospitals. To adapt this research, the most accurate approach is to install a data collection system similar to the one used in this research and build tree models from it. Hill-Rom, a major hospital equipment supplier is one of several companies that supply this type of system. The primary purpose of having such a system in a hospital is for the unit secretary to locate the closest nurse to a patient in urgent need, which often leads to unassigned direct care. Nurses' current locations can be viewed by the secretary on her desk top computer. If a patient needs immediate assistance, the secretary would page the nearest nurse seeking for assistance. From this system, the data is transferred and stored in a repository continuously. With recent innovations and the proliferation of RFID technology, installing such a system in hospitals has become easier and cheaper and has found to be useful for different purposes. 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 or disinterest in certain variables. Even for the same variables, it is likely that the number of categories will vary at different hospitals. 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, the simulation should be performed as explained in Section 4.3. The impact of factors/variables on the simulation can be judged based on variable importance scores given in Table 2. The variable that receives a 100 score indicates the most influential variable for prediction (higher impact on the simulation), followed by other variables based on their relative importance to the most important one. However, care should be taken to avoid overfitting based on certain artificial variables that could mask other important variables. Using the concept of structures and pointers [18, 32], the C++ simulation code written in this research can read any tree structure when the same or a subset of variables from this research are used, and then simulate by sampling repeatedly from the trees until the entire shift period ends. This way of coding makes it easy to adapt the simulation code to different hospitals.

The input data set, which was collected continuously for a long period in this research, is treated as the population data. The tree models are used to extract patterns from the input data for the simulation. When new data are available and hence new trees are built in CART, the simulation model can update itself by reading and simulating from the new tree structures. As a result, this research introduces a readily adaptable simulation model to different data sets even though the simulation developed in this paper is hospital specific. If it is impossible to install such a data collection method, the same tree structures developed in this research can be used if there is a reasonable justification for similar work dynamics in the new hospital. It is also possible to append "If ... Then ..." rules to the tree structure if there are additional restrictions. However, walk time between different locations should be adjusted to reflect the new hospital layout.

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 the nurse will spend at the new location. Simulation models developed with this approach will be significantly more efficient than the simulation models that consider all possible combinations of state variables. Optimal nurse–patient assignments can be identified by applying simulation-optimization methods, such as Atlason et al. [4] and Fu and Hu [19], 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, retention of nurses, and cost savings for hospitals can be achieved.

Acknowledgements This research was supported by the Robert Wood Johnson Foundation grant number 053963. We thank Terry Clark from the northeast Texas hospital 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, Dowsland KA (2003) An indirect genetic algorithm for a nurse scheduling problem. *Comput Oper Res* 31(5):761–778
- Aiken LH, Clarke S, Sloane D, Sochalski J, Silber J (2002) Hospital nurse staffing and patient mortality, nurse burnout, and job dissatisfaction. *JAMA* 288:1987–1993
- AONE (2003) Policy statement on mandated staffing ratios. http://www.aone.org/aone/docs/ps_ratios.pdf. Accessed September 2007
- Atlason J, Epelman MA, Henderson SG (2004) Call center staffing with simulation and cutting plane methods. *Ann Oper Res* 127:333–358
- Azaiez MN, Sharif SSA (2005) A 0-1 goal programming model for nurse scheduling. *Comput Oper Res* 32:491–507
- Bard J, Purnomo HW (2005) Preference scheduling for nurses using column generation. *Eur J Oper Res* 164:510–534
- Beddoe GR, Petrovic S (2006) Selecting and weighting features using a genetic algorithm in a case-based reasoning approach to personnel rostering. *Eur J Oper Res*, 175: 649–671
- Bettonvil B, Kleijnen JPC (1997) Searching for important factors in simulation models with many factors: sequential bifurcation. *Eur J Oper Res* 96:180–194
- Breiman L, Friedman, JH, Oishen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont, California
- Burke EK, Cowling P, Caumaecker PD (2001) A memetic approach to the nurse rostering problem. *Appl Intell* 15:199–214, special issue on Simulated Evolution and Learning
- CDHS (2005) Nurse-to-patient staffing ratio regulations. <http://www.dhs.ca.gov/lnc/NTP/default.htm>. Accessed January 2006
- Ceglowski A, Churilov L (2008) Using self organising feature maps to unravel process complexity in a hospital emergency department: a decision support perspective. In: Intelligent decision making: an AI-based approach, Springer Berlin, Heidelberg, pp 365–385
- Ceglowski A, Churilov L, Wassertheil J (2005) Knowledge discovery through mining emergency department data. In: Proceedings of the 38th annual Hawaii international conference on system sciences, Hawaii, USA
- Cheng RCH (1997) Searching for important factors: sequential bifurcation under uncertainty. In: Proceeding of the 1997 winter simulation conference, Piscataway, New Jersey, USA
- Dumas MB (1985) Hospital bed utilization: an implemented simulation approach for adjusting and maintaining appropriate levels. *Health Serv Res* 20:43–61
- Epanechnikov VA (1969) Nonparametric estimation of a multivariate probability density. *Theory Probab Appl* 14: 153–158
- Evans, GW, Gor TB, Unger E (1996) A simulation model for evaluating personnel schedules in a hospital emergency department. In: Proceedings of the 1996 winter simulation conference, Coronado, California, USA
- Foster LS, Foster WD (2003) C by discovery. Galgotia, Daryaganj, New Delhi
- Fu MC, Hu JQ (1997) Conditional Monte Carlo: gradient estimation and optimization applications. Kluwer, Norwell, Massachusetts
- Gutjahr WJ, Rauner MS (2007) An aco algorithm for a dynamic regional nurse-scheduling problem in Austria. *Comput Oper Res* 34:642–666
- Hastie T, Tibshirani R, Friedman JH (2001) The elements of statistical learning: data mining, inference, and prediction. Springer, New York
- 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, Vovor T (1998) A generalized linear programming model for nurse scheduling. *Eur J Oper Res* 107(1):1–18
- Jones MC, Marron JS, Sheather SJ (1996) A brief survey of bandwidth selection for density estimation. *J Am Stat Assoc* 91(433):401–407
- Jun JB, Jacobson SH, Swisher JR (1999) Application of discrete event simulation in health care clinics: a survey. *J Am Stat Assoc* 50(2):109–123
- Kim SC, Horowitz I, Young KK, Buckley TA (2000) Flexible bed allocation and performance in the intensive care unit. *J Oper Manag* 18(4):365–385
- Kirkby MP (1997) Moving to computerized schedules: a smooth transition. *Nurs Manage* 28:42–44
- Klein RW, Dittus RS, Roberts SD, Wilson JR (1993) Simulation modeling and health-care decision making. *Med Decis Mak* 13(4):347–354
- Kreke J, Schaefer AJ, Angus D, Bryce C, Roberts M (2002) Incorporating biology into discrete event simulation models of organ allocation. In: Proceedings of the 2002 winter simulation conference, San Diego, California, USA

32. Lafore R (2000) Object-oriented programming in Turbo C++. Galgotia, Daryaganj, New Delhi
33. Law AM, Kelton WD (2001) Simulation modeling and analysis. McGrawHill, New York
34. Lim T, Uyeno D, Vertinsky I (1975) Hospital admission systems: a simulation approach. *Simul Games* 6:188–201
35. Miller HE, Pierskalla WP, Rath GJ (1996) Nurse scheduling using mathematical programming. *Oper Res* 24(5):857–870
36. Mullinax C, Lawley M (2002) Assigning patients to nurses in neonatal intensive care. *J Oper Res Soc* 53:25–35
37. Punnakitikashem P, Rosenberger JM, Behan DF (2008) Stochastic programming for nurse assignment. *Comput Optim Appl* 40:321–349
38. Ramon J, Fierens D, Guiza F, Meyfroidt G, Blockeel H, Bruynooghe M, Berghe VDG (2007) Mining data from intensive care patients. *Adv Eng Inf* 21:243–256
39. Sheather SJ (2004) Density estimation. *Stat Sci* 19(4): 588–597
40. Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. *J R Stat Soc, Ser B* 53(3):683–690
41. Shechter SM, Bryce C, Alagoz O, Kreke JE, Stahl JE, Schaefer AJ, Angus D, Roberts M (2005) A clinically based discrete event simulation of end-stage liver disease and the organ allocation process. *Med Decis Mak* 25(2): 199–209
42. Shen H, Wan H (2005) Controlled sequential factorial design for simulation factor screening. In: Proceedings of the 2005 winter simulation conference. Orlando, Florida, USA
43. SHS (2005) Nurse-to-patient staffing ratio regulations. http://iienet2.org/uploadedFiles/SHS/Resource_Library/Details/positionPaper.pdf. Accessed September 2007
44. Silverman BW (1978) Choosing window width when estimating a density. *Biometrika* 65(1):1–11
45. Silverman BW (1986) Density estimation for statistics and data analysis. Chapman and Hall, London
46. Smith EA, Warner HR (1971) Simulation of a multiphasic screening procedure for hospital admissions. *Simulation* 17:57–64
47. Sundaramoorthi D, Chen VCP, Rosenberger JM, Green DFB (2005) Knowledge discovery and mining for nurse activity and patient data. In: Proceedings of the 2005 IIE annual conference, Atlanta, Georgia, USA
48. Sundaramoorthi D, Chen VCP, Rosenberger JM, Kim SB, Behan DFB (2006) A data-integrated nurse activity simulation model. In: Proceedings of the 2006 winter simulation conference
49. Sundaramoorthi D, Chen VCP, Rosenberger JM, Kim SB, Behan DFB (2006) Using classification and regression trees for a nurse activity simulation. In: Proceedings of the 2006 IIE annual conference
50. Vericourt FD, Jennings OB (2006) Nurse-to-patient ratios in hospital staffing: a queuing perspective. <http://faculty.fuqua.duke.edu/%7Efdv1/bio/ratios3.pdf>. Accessed July 2006
51. Warner DM (1976) Scheduling nursing personnel according to nursing preferences: a mathematical approach. *Oper Res* 24:842–856
52. Zenios SA, Wein LM, Chertow GM (1999) Evidence-based organ allocation. *Am J Med* 107(1):52–61