The Application of Queueing Theory to Study Optimum Balance Point Between Waiting Cost and Idle Cost

Nivarti Narayan Bharkad
Research Scholar, Department of Mathematics,
M G ARTS, SCIENCE & Late N P Commerce college, Armori, Dist- Gadchiroli(MS)
Email: nnbharkad@gmail.com
Dr. Lalsingh H. Khalsa
Principal, M G ARTS, SCIENCE & Late N P Commerce college, Armori, Dist- Gadchiroli(MS)
Email : lalsinghkhalsa@yahoo.com

Abstract: Early days to get any service from service provider we have to wait long time in a queue, means waste of time that is waste of money. Because peoples or any vehicle or machine that are waiting for service are losing their earning time in waiting line. Its happen with every one and every arriving unit. Hence waiting time in a queueing system of all arriving units is so much. Their waiting cost is also so large. sometimes it may also happen with the servers that providing services due to no arriving unit in the system they become idle.

In the present study we have studied different parameters that affect both waiting cost and idle cost and how to maintain equilibrium between these two cost to reduce the total cost of the queueing system.

Key words: Queueing model, Equilibrium, Waiting cost, Ideal cost, Servers, Customers.

1.Introduction: Queueing theory deals with study of arriving rate, arriving pattern, behavior, waiting time in queue, waiting time in system, waiting cost of customer and service rate, idle time, servicing cost of sever. A basic queueing organization is a service organization at which "customers" arrive to a "servers" and need some service from of them. It is important to know that a "customer" is whatever arriving unit is waiting to get service and not necessary to be a person. Likewise, a "server" is the person or device that delivers the service. If all attendants are busy when a customer's arrival, then they should join a queue. Hence queues are physical lines of persons or objects, they may be invisible for example cell phone calls waiting on hold. The rule that decides the rule in which queued clients are served is known as the queue *discipline*. The most usual discipline is the first-in, first-out(FIFO) rule, but remaining disciplines are usually used to increase productivity or minimize the delay for extra time-crucial customers. For instance, in an ICU of hospital. In most queueing system, the assumption is there is infinite number of clients that can be waiting to get service. This is a good supposition when customers do not really join a queue, as in a cell phone call center, or if the physical space where clients wait is large related to the number of clients who are waiting to get service. in this cases new arriving customers who look a long waiting line may "balk" and do not join queue. This might occur in a shop of hair salon. Other behavior that is occurred in queueing systems is "reneging" it happens when clients become inpatient and left the queue before being attended. For example, this conduct is found in some hospital where patients who renege are often called as "left without medical treatment".

Finally, queues may be planned in various ways. In many cases, we will assume a *single line* that fodders into all servers. But sometimes every server has his/her own waiting line as may be the situation for a primary care workplace at which patients have their own doctor. This design is frequently referred to as waiting line in *parallel*. In other circumstances, we may want to assume a *network* policy in which customers get service from different kinds of servers in a successive manner. For instance, a surgical patient requires an

operation theater, then a bed in recovery unit, after wards a bed in a surgical ICU, and/or other place of the hospital. A queueing model is mathematical explanation of a queuing system that makes some specific assumptions of the probabilistic nature of arrival and service manners, the number and kind of servers, and the discipline of queue and organization. There are countless possible variations, but some queueing models are more widely used and we will focus on these in this chapter. For these models, as well as many others, there are formulae available that enable the fast calculation of various performance
measures that can be used to help design a new service system or improve an existing one.

2.Charecteristics of Queueing Models

Many queueing theories deals with *steady-state* system. That is, most queueing theories assume that the system has been working with the same arrival rate, expected service time and other features for a sufficiently long period that the probabilistic nature of performance measures for instance, queue length and customer delay is not dependent of when the system is observed. In this study, we will consider that we are seeing at systems in steady-state only. For specifying a queueing model, we should make assumptions about the probabilistic behavior of the arrival and service progressions. The most common assumption about arrivals is that they follow the *Poisson* process. The name arises from the fact that number of arrivals in given time period has the Poisson distribution. Thus, if *N(t)* is number of arrivals during the time period of duration t then N(t) has a
Poisson distribution:

Probability of $\{N(t) = n\} = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$

where λ is the arrival *rate* and it is defined as the average number of arrivals per unit of time.
Another way to describe the Poisson process is the time between successive arrivals, known as the inter arrival time, have an *exponential* distribution. Thus if *IA* is an inter arrival time of the Poisson process with arrival rate λ then:

Probability of $\{IA \leq t\} = 1 - e^{-\lambda t}$

where1/λ is the average time among arrivals. A significant property of exponential distribution is "memoryless". It means that the time of the succeeding arrival is not dependent of when the previous arrival occurred. This property too leads to the fact that when the arrival process is Poisson, then the number of arrivals in any time interval is independent of the number of arrivals in any other non-overlapping interval of time. Conversely, it can be proved analytically that when the arriving customers are independent from each another, then arrival process is the Poisson process. Because of this, the Poisson process is assumed the most "random" arrival process. In identifying whether Poisson process is a realistic model for coming unit in a particular service
system, it is helpful to consider its three properties:
1. Consumers arrive one at a time.
2. The probability that a consumer arrives at any time is not dependent on other consumers arrived.
3. The probability that a consumer arrives at a specified time is independent of time.

In more situations, customers usually do arrive one at a time. However, there may be occasions, such as major accident, that generate multiple simultaneous arrivals, it is likely to be an exceptional context which will not considerably affect the effectiveness of the modeling supposition. Certainly, the second property is also frequently a reasonable assumption. For instance, in an emergency room, wherever the population of potential patients is so large, it is unlikely that somebody arriving with a cracked arm has something to do with someone else's or illness or injury, or that the fact that the number of patients who arrived between 10 a.m. and11 a.m. was five provides some information about number of patients that are expected to arrive between 11am and 12am. Again, there may be infrequent exceptions, such as a flu outbreak, that violate this assumption, yet in the aggregate, it's expected to be reasonable. Though, the third property may be more certain.

3.The M/M/s model

The most commonly applied queueing model is the *M/M/s* model. This model assumes a Single waiting line with unlimited waiting apartment that feeds into *s* equal servers. Customers arrive as per to the Poisson process with a constant arrival rate, and the service time has an exponential distribution. The advantage of Appling the *M/M/s* model is it only needs three parameters and therefore it can be used to find performance estimates with very small data. The average arrival rate is λ, an average service time is1/μ and number of servers is *s*, to determine performance measures easy-to-compute formulae are existing like the probability of an arrival will experience positive delay P_D, and the average delay, *Wq* etc;

$$p_D = 1 - \sum_{n=0}^{s-1} P_n \quad (1)$$

$$W_q = \frac{P_D}{[(1-\rho)s\mu]} \quad (2)$$

where $\rho = \frac{\lambda}{s\mu}$ (3)

Also

$$P_n = \begin{cases} \frac{\lambda^n}{n!\mu^n} p_0; & 1 \le n \le s \\ \frac{\lambda^n}{s^{n-s} s!\mu^n} p_0; & n \ge s \end{cases} \quad (4)$$

Where

$$P_0 = \left[\sum_{n=0}^{s-1} \frac{(\rho s)^n}{n!} + \frac{\rho^s s^{s+1}}{s!(s-\rho s)} \right]^{-1} \quad (5)$$

Note that ρ is the expected utilization for the queueing system and this equation is only valid if the utilization is quite less than one. Furthermore, note that the average delay rises as utilization tends to one.

4. Identifying Costs into the queueing system

To determine the best number of severs essential in the model, the two different Costs would be assumed in decision making: Service costs and Waiting costs. The service cost means the cost suffered in providing of expected service, it is represented by C_s. It includes payments paid to staffs, price of equipment and utensils applied, worth of service space, rent, deliveries, cost of *ICU* etc. On the other hand, waiting cost contains cost suffered by the patients due to the waiting in a queue. It also consists of cost of losing life due to the waiting. This cost is known as waiting cost and indicated by W_s the understanding of these costs support in obtaining balance point between service cost and waiting costs.

In a heath services, the most dangerous time is the waiting time in waiting line before service starts and it is indicated by W_q. The expected waiting cost per patients per unit of the time is;

$$C_q = \lambda W_q C_w = L_q C_w \quad (6)$$

Where $W_q = \frac{L_q}{\lambda}$ (7)

The average service cost suffered by the health services is symbolized by C_s. Hence total service cost is sC_s

Where s means the number of servers. The total cost is calculated by adding the waiting cost and service cost that gives

$$TC = sC_s + C_q \quad (8)$$

5. Loss function of queue

The measure of quality, as associated to both the product and the facility, is often difficult to exactly quantify because of different viewpoints of individuals, but quality includes short waiting time, cleanliness, acceptable service, affordable and friendly. The low level of service can be inexpensive, in the little run but the service supplier may incur high cost of customer displeasure such as lose of future trade, loss of a potential sale, growth of poor reputation, loss of kindness and increased competition by businesses in the same industry. Nonconformity from the expected quality of facility leads to the situation where the customer suffers opportunity price. The level of cost suffered can be determined by a loss function that links the cost and the level of nonconformity from the expected level. The succeeding two loss functions have been applied previously in finding the opportunity cost the customer suffers when the product or service flops to meet the target requirement value. That are traditional loss function, Taguchi loss functions explained below.

5.1. Traditional quality loss function

As expected, customers suffer costs if the services delivered are not meeting the expected level, means the services are either very low to meet the essential expectations, or very high that the customer is not able to pay the cost. The traditional quality loss function is a square function as shown in Fig. 1. In this function, the consumers are equally satisfied, and hence they do not suffer any loss, as long as quality of facilities meet the specifications within LSL and USL. It is not realistic, and hence, a better Taguchi loss function displayed in Fig. 2 was framed applying a quadratic function [25]. Principles of framing Taguchi loss function supposes that, there is no cost suffered by the service providing association or by the customer except the product or service goes outside its Upper Specification Limits (USL) or Lower Specification Limits (LSL).

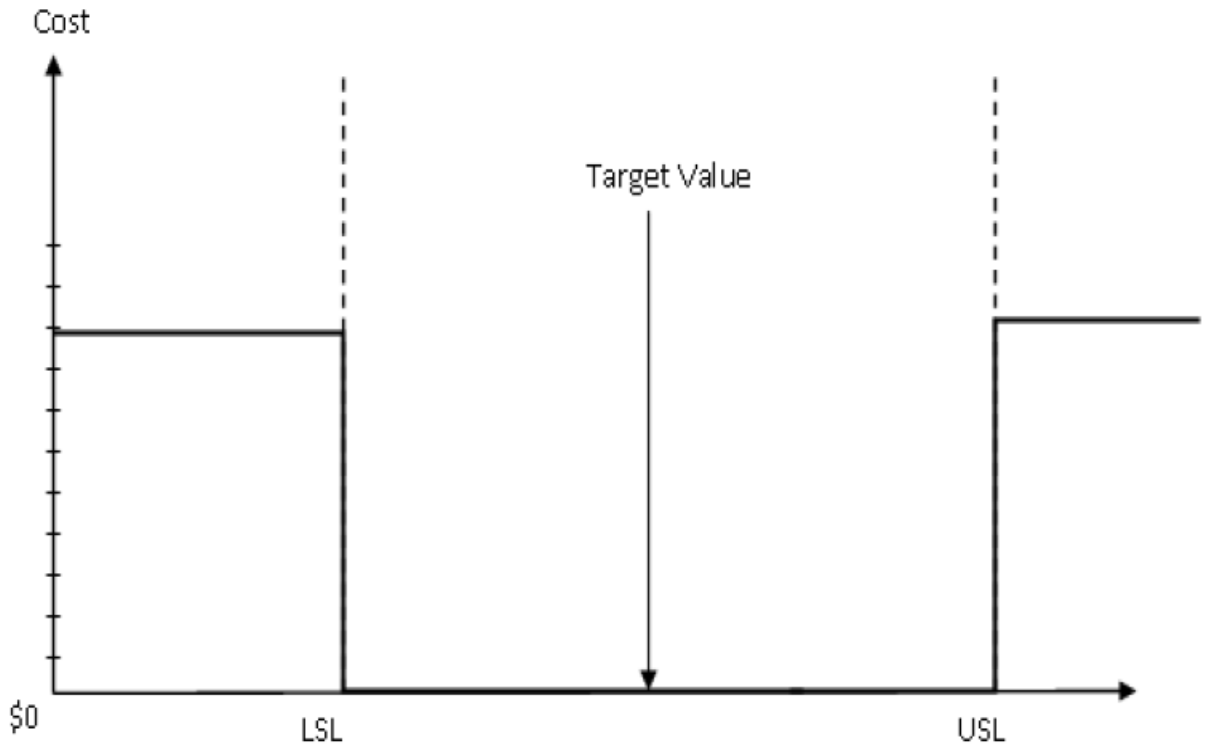

Fig. 1. Traditional loss function displaying USL and LSL

5.2. Taguchi loss function

The Taguchi Loss Function proceeds with different perspective on if the cost of low quality are suffered. Taguchi imagined that rather than suffer costs beginning from two finite points that are plus or minus a specific level of tolerance from target value, prices are actually suffered as soon as the price moves from its target value [25]. In addition, rather than continue at a constant rate, these costs are suffered at the square of deviation from target value, and hence continue to rise the farther the specification diverges from the targeted value. The point in the system at which no loss is suffered is at actual targeted value of Z.

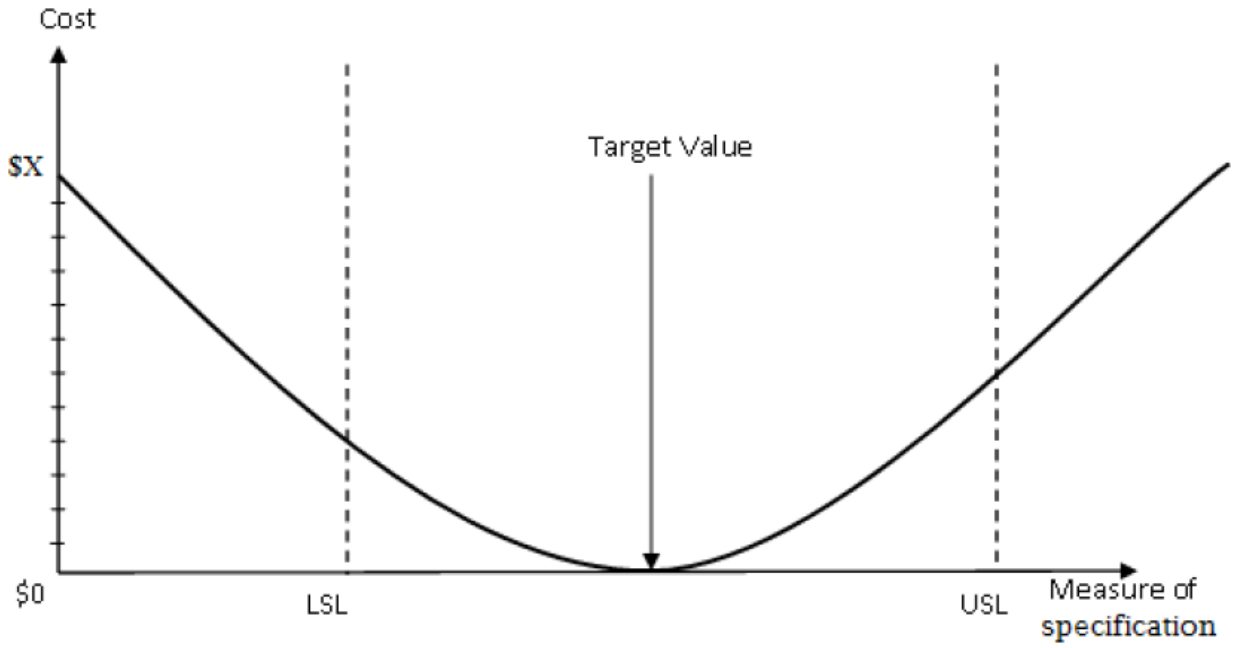

Fig. 2. Taguchi function displaying calculation of loss function

In contrast with traditional loss function, the Taguchi Loss Function explain the failure to meet desired requirement cost C suffered as;

(9)

$$C = \begin{cases} 0, if LSL \geq m \geq USL \\ k(m-T)^2 ; otherwise \end{cases}$$

Where T is the target requirement, m is the unit amount of quality requirement and the constant *K* is obtained from the cost of refusing the item at a requirement limit from the relation;

$$K = \frac{R}{(USL - T)^2} \quad (10)$$

Where *R* is the cost of refusing the item at requirement limit. Obviously, the Taguchi loss function is quadratic function that hits the zero cost line when the requirement limit is same to the target value *L=T*. This also has a identical gradient for all customers and unlimited cost on extreme deviance from the target value. It does not put into account the individual tolerance differences on levels of fulfilment of the item or service requirement.

5.3. Improved normal loss function

In present study, it is expected that individual preference is normally spread with a mean is $\bar{x}$ and a standard deviation is σ. A normal distribution table is applied to decide the proportion of the cost suffered if the measure of product or facility quality deviates from the nominal requirement value. In this study, the loss function $f(x)$ has a graph similar to inversed normal curve, but a gap of 2τ in between as displayed in fig.3

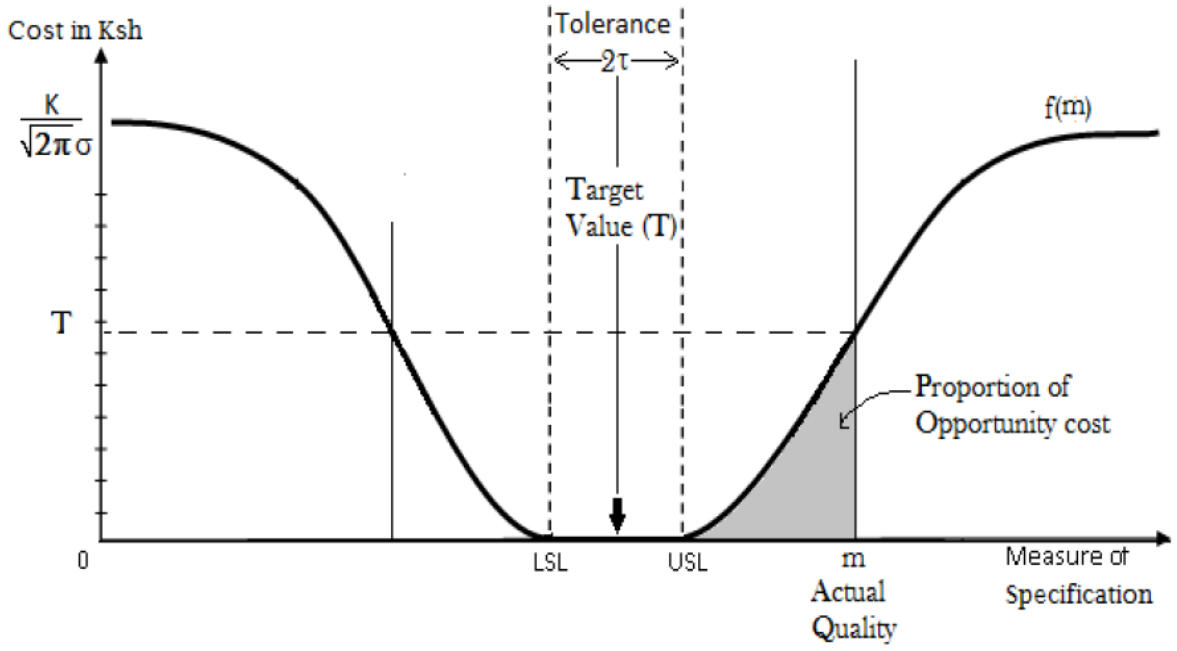

Fig. 3. Normal function displaying calculation of loss function

The cost function $f(x)$ is equal to one sided normal distribution function that assigns a numerical

value proportionate to the amount at which quality of service or product diverges from the individual requirement target. The individual target quality shall be range of values in interval $I = -\tau \le T \le \tau$. If the actual quality of product or facility $x = m$ is outside the interval I, the customers will start suffering costs due to dissatisfaction. Area under the curve, x-axis and the lines $x = USL$ and $x = m$ measures proportion of opportunity cost individual shall suffer due to unsatisfactory standards. The cost function is defined by,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\bar{x}}{\sigma}\right)^2} \qquad (11)$$

where $\bar{x}$ is the mean of quality requirement value, that is equivalent to the standards delivered by the service provider and σ is the standard deviation quality specification , while τ is the individual tolerance levels and x is the measure of actual quality requirement.

5.4. Tolerance and opportunity cost

The opportunity cost suffered due to delay of service or low quality of products is inversely proportionate to individual level of tolerance. If lesser the tolerance, then higher the opportunity cost. The quality tolerance is defined as the measure of deviancy from quality nominal requirement without suffering any opportunity cost. It should be a normal stretch without any impact. This characteristic is measured by parameter τ , that accounts for individual alterations on preferences or the expected standards. Difference in individual favorites or tastes or tolerance is due to the individual lifestyle, social status, profession, financial status, cost of other similar service elsewhere, emergency of the required service, risk related with delay of the essential service, reason of the product required, etc.

Let C_w is the cost of rejection at the requirement limit $x = m$ and let T is the mean of target requirement value with the individual level of the tolerance of $\tau \ge T$. Then actual cost of rejection suffered by the customer fulfils the condition

$$C_w = \begin{cases} kf(m); if m > \left|\bar{x} + \tau\right| \\ 0; if m < \left|\bar{x} + \tau\right| \end{cases} \qquad (12)$$

Where k be the highest opportunity cost suffered as $m \to \pm\infty$ and $\bar{x} + \tau = SL$ =Specification (requirement) limit.

6.Conclusion

As per above derivations waiting cost, idle cost can be determined and by evaluating the Optimum Balance Point Between Waiting Cost and Idle Cost. The total cost of queuing system can be reduced according optimal number of servers used. The Loss function of queue can be calculated

On the basis traditional loss function and Taguchi loss function. The tolerance and opportunity cost also play an important role in study of queuing theory and to decide appropriate number of servers.

REFERENCES

1. Allen, A.O., 1978, *Probability, statistics and queueing theory, with computer science applications*. New York, Academic Press.
2. Brecher, C. and Speizio, S., 1995, *Privatization and Public Hospitals*, Twentieth Century Fund Press, N.Y.
3. Brewton, J.P., 1989, Teller staffing models, *Financial Manager's Statemen*t, July-August: 22-24.
4. Brigandi, A.J., Dargon, D.R., Sheehan, M.J. and Spencer III, T., 1994, AT&T's call processing simulator (CAPS) operational design for inbound call centers, Interfaces 24: 6-28.
5. Brockmeyer, E., Halstrom, H.L., and Jensen, A., 1948, The life and works of A.K. Erlang, *Transactions of the Danish Academy of Technical Science* 2.
6. Brusco, M.J., Jacobs, L.W., Bongiorno, R.J., Lyons, D.V. and Tang, B., 1995, Improving personnel scheduling at airline stations,Operations Research, 43: 741-751.
7. Chelst, K. and Barlach, Z., 1981, Multiple unit dispatches in emergency services, *Management Science*, 27: 1390-1409.

8. Cobham, A., 1954, Priority assignment in waiting line problems, *Operations Research*, 2: 70-76.
9. Freeman, R.K., and Poland, R.L., 1997, Guidelines for Perinatal Care, 4th ed., American College of Obstetricians and Gynecologists,Washington, D.C.
10. Green, L.V., Giulio, J., Green, R., and Soares, J., 2005, Using queueing theory to increase the effectiveness of physician stiffening in the emergency department, *Academic Emergency Medicine*, to appear.
11. Green, L.V., 2003, How many hospital beds? *Inquiry*, 39: 400-412.
12. Green, L.V., Kolesar, P.J., Svoronos, A., 2001, Improving the SIPP approach for staffing service systems that have cyclic demands, *Operations Research*, 49: 549-564.
13. Green, L.V. and Nguyen, V., 2001, Strategies for cutting hospital beds: the impact on patient service. *Health Services Research*,36: 421-442.
14. Green, L.V., Kolesar, P.J., and Svoronos, A., 1991, Some effects of no stationarity on multi-server Markovian queueing systems.*Operations Research*, 39: 502-511.
15. Green, L.V., and Kolesar, P.J., 1984, The feasibility of one-officer patrol in New York City, *Management Science* 20: 964-981.
16. Hall, R.W., 1990, *Queueing Methods for Service and Manufacturing*. New Jersey: Prentice Hall.
17. Holloran, T. J. and Byrne, J.E., 1986, United Airlines station manpower planning system, *Interfaces*, 16: 39-50.
18. Green, L.V., 2003, How many hospital beds? *Inquiry*, 39: 400-412.
19. Kaplan, E.H., Sprung, C.L., Shmueli, A., and Schneider, D., 1981. A methodology for the analysis of comparability of services and financial impact of closure of obstetrics services. *Medical Care*, 19: 395-409.
20. Kim, S., Horowitz, I., Young, K.K., and Buckley, T.A., 1999, Analysis of capacity management of the intensive care unit in a hospital, *European Journal of Operational Research* 115: 36-46.
21. Kolesar, P.J., Rider, K., Crabill, T., and Walker, W., 1975, A queueing linear programming approach to scheduling police cars,*Operations Research*, 23: 1045-1062.
22. Larson, R.C., 1972, *Urban Police Patrol Analysis*, MIT Press, Cambridge.
23. McCaig, L.F., and Burt, C.W., 2004, National hospital ambulatory medical care survey: 2002 emergency department summary.*Advance Data from Vital and Health Statistics*, 340: 1-35.
24. Stern, H.I. and Hersh, M., 1980, Scheduling aircraft cleaning crews, *Transportation Science*, 14: 277-291.
25. Taylor, P.E. and Huxley, S.J., 1989, A break from tradition for the San Francisco police: patrol officer scheduling using an optimization-based decision support system, Interfaces, 19: 4-24.
26. Worthington, D.J., 1987, Queueing models for hospital waiting lists. Journal of the *Operations Research Society*, 38: 413-422.
27. Young, J.P., 1965, Stabilization of inpatient bed occupancy through control of admissions, *Journal of the American Hospital Association*, 39: 41-48.