Use of Queueing Theory for Problem Solution in Dallas, Tex., Bureau of Vital Statistics

BILLY J. MOORE, PhD

A COMMON PROBLEM that may reflect on the quality of service at a local health department is that of clients' discontent with the long waiting time for service when they visit a service center. Except for some scheduled clinics, health department services are typically offered on a first-come, first-served basis. People generally expect to wait a "reasonable" length of time in a no-appointment facility, but they are dissatisfied if the waiting time becomes excessive.

In an effort to eliminate customer discontent with the services in its Bureau of Vital Statistics, a problem-solving technique was used by the City of Dallas Health Department. The emphasis was on the application of the queueing theory, or the theory of waiting lines. Because the queueing theory is not widely familiar to public health managers, this presentation is at an introductory level. The concepts reviewed here are also applicable to other health department services (for example, venereal disease and tuberculosis clinics) for which patients arrive at random intervals of time.

The Problem and the Objective

By the end of 1972, the Bureau of Vital Statistics was being inundated by demands for birth and death certificates. Long lines of customers were forming at a counter, a backlog of mail was accumulating, and no workload relief was expected. The scenario had much in common with that described by Crandall (*1*). It was apparent that relief could come only from an increase in the bureau's capacity to produce. Furthermore, preliminary consideration of such diverse factors as budgetary difficulties of increasing the work force, the physical limitations of the available workspace, the antiquity of the microfilm system, and the possible calamitous effects of a wrong decision led to a research effort toward solution of the problem.

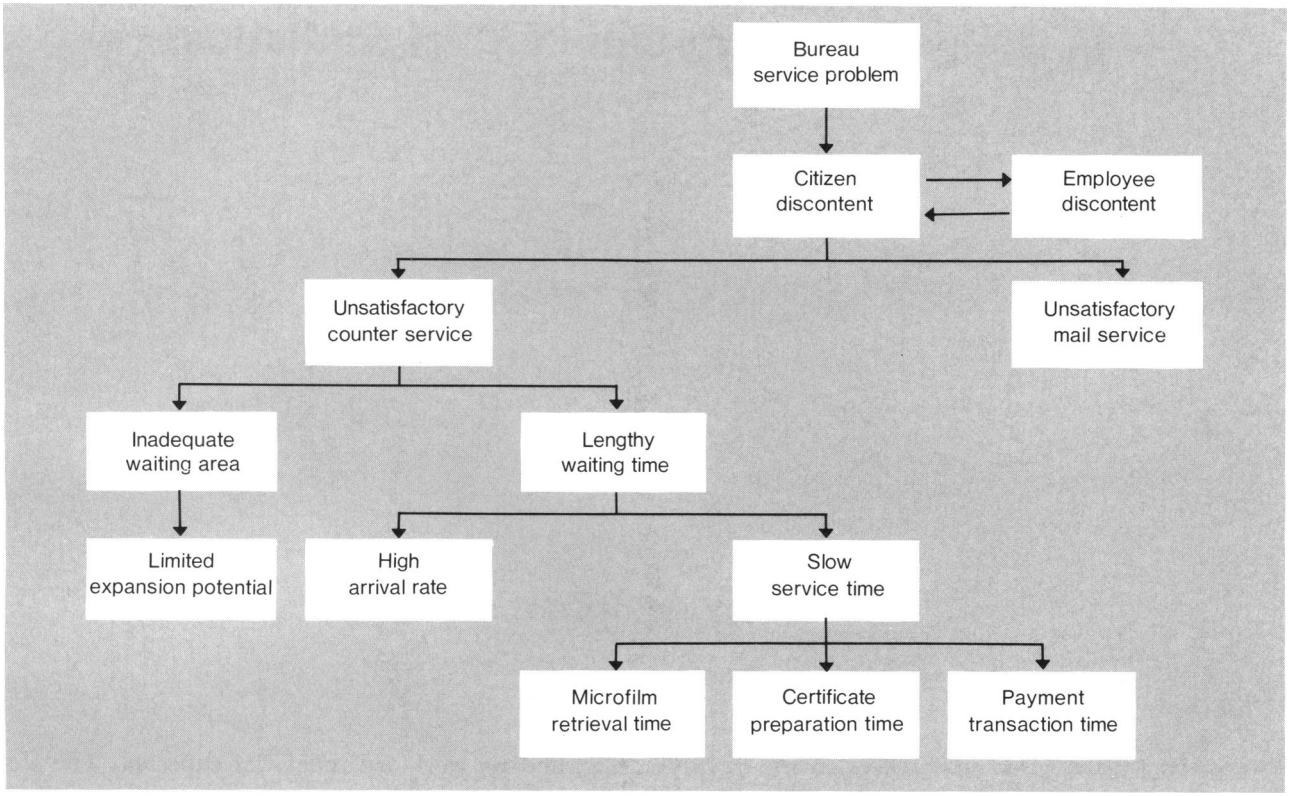
To maintain order in the research and to assure maximum use of effort, statements of the problem and the research objective were prepared. The two statements provided a frame of reference in meetings of bureau staff with technical and administrative consultants.

The problem was stated essentially as follows:

Increased citizen requirements for services of the Bureau of Vital Statistics have resulted in a performance level that is unacceptable to customers who come to the bureau as well as those who mail their requests. This situation has resulted in citizen and employee discontent, and projections offer no

☐ *Tearsheet requests to Dr. Billy J. Moore, Biostatistician, City of Dallas Health Department, 1936 Amelia Court, Dallas, Tex. 75236.*

Figure 1. Problem analysis chart for the Dallas Bureau of Vital Statistics



NOTE: Arrows indicate "is significantly affected by".

decrease in workload in the forseeable future.

For the research objective, a more detailed analysis of the problem was required. Figure 1 depicts the total problem structure as a system of problem components. The connecting arrows indicate "is significantly affected by;" for example, employee discontent is significantly affected by citizen discontent, which, in turn, is significantly affected by both counter service and mail service. The main problem, "unsatisfactory counter service," is jointly affected by "lengthy waiting time" and "inadequate waiting area." Because the bureau office would remain in the same location and little could be done about improving the waiting area, the waiting-area problem was given secondary consideration. Therefore, the specific research objective was:

A service system should be designed and implemented, on a cost-effective basis, such that the average waiting time for customers at the counter is no more than 15 minutes during the nonrush hours. The system should have the potential of meeting these conditions until the year 2000.

For our purposes, waiting time was defined as the length of time a customer spent in the office measured from the time he submitted a completed order form to the time he received his document-payment receipt.

Application of Queueing Theory

The completion of the important preliminaries of stating the problem and the objective led to the consideration of queueing theory as an analysis tool. Queueing theory is concerned with the classic situation of customers arriving at a facility for service, waiting for service to begin, receiving service, and then leaving the facility. Specifically, this theory can help provide insight as to what and how much must be done to decrease customer waiting time to an acceptable level.

Only those mathematical equations required for understanding basic concepts are presented here. More detailed discussions appear in standard textbooks, such as those of Richmond (*2*) and Bhat (*3*).

The four basic inputs to a queueing model are (*a*) the probability distribution of the number of arrivals (*b*) the probability distribution of the service times (*c*) the number of service stations (servers, channels), and (*d*) the queue discipline.

It is assumed that customer arrivals occur randomly throughout the workday and follow a Poisson prob-

ability distribution, and that time taken for service completion is a random variable which follows an exponential distribution. The queue discipline governs the entry into service from the waiting line. The queue discipline assumed here is first come, first served. The preceding are common assumptions in queueing theory.

Throughout the ensuing discussion the following symbols and definitions are used:

W = average waiting time at the counter for a customer, including the service time

λ = arrival rate of customers (number of customers arriving for service per hour)

μ = customer service rate for a single server (number of customers served per hour per server)

$\bar{n}$ = average line length (number of customers being served or waiting for service, on the average)

k = number of servers

$W = \bar{n} \div \lambda$, the formula presented by Little (*4*), shows that, for fixed λ, the waiting time W is directly proportional to the average line length $\bar{n}$. Therefore, a customer's waiting time can be decreased if the length of the average waiting line is decreased.

The question then follows: How does one decrease $\bar{n}$? The logical answer is that the average line length should be decreased by either decreasing the customer arrival rate λ or by increasing the service rate μ, or both. The mathematics of queueing theory affirms that this is indeed the situation. Furthermore, the following general formula for $\bar{n}$ shows that this function depends on λ and μ only through the ratio of λ/μ, and on k:

For k servers, the equation for computing the average line length is

$$\bar{n} = \frac{(\lambda/\mu)^{k+1}}{(k-1)!(k-\lambda/\mu)^2} p + \lambda/\mu, \quad \text{where}$$

$$p = \left[\sum_{n=0}^{k-1} \frac{1}{n!} (\lambda/\mu)^n + \frac{1}{k!} (\lambda/\mu)^k \frac{k\mu}{k\mu - \lambda}\right]^{-1}, \quad k > \lambda/\mu.$$

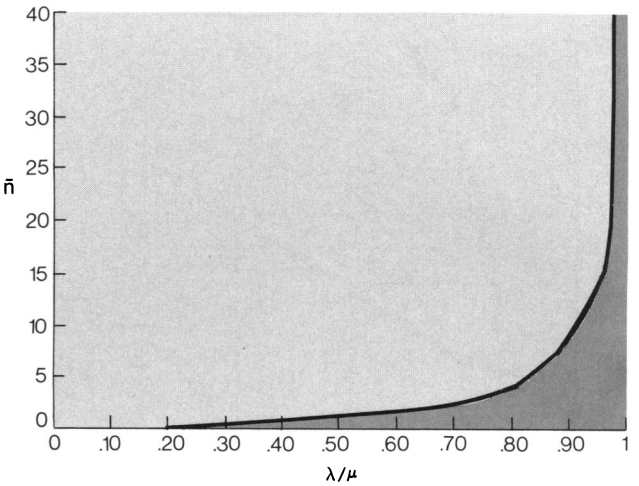
Setting $k = 1$, obtains the relationship between $\bar{n}$ and λ/μ as given by

$$\bar{n} = \frac{\lambda/\mu}{1 - \lambda/\mu},$$

which is plotted in figure 2. (Curves for more than one server are similar.)

Note that as the ratio λ/μ approaches 1, or as the arrival rate approaches the service rate, the average line length progressively increases to an infinite number. The important point here is that, for a given service rate μ existing in a facility, there is some upper limit to the arrival rate of customers the facility can manage; a continued arrival rate that surpasses this upper limit by even a small amount may result in customer saturation of the facility and surprisingly longer expected waiting times.

Figure 2. Average length of line, $\bar{n}$, as a function of the ratio λ/μ for one server

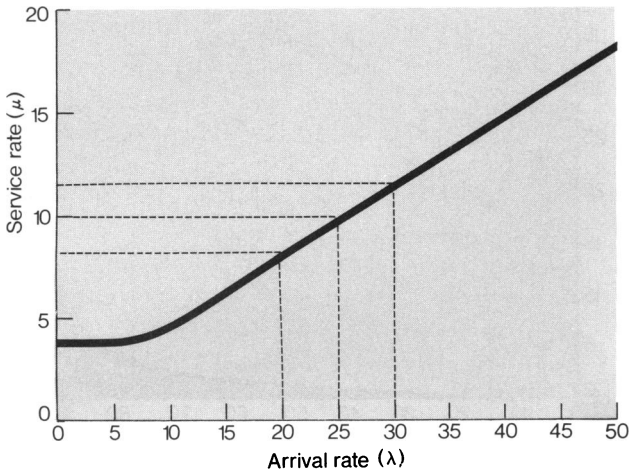


For example, suppose that the service rate for a single-server facility is $\mu = 20$ persons per hour. Then, for an arrival rate of $\lambda = 18$ persons per hour, one obtains by substitution into the equations for $\bar{n}$ and W the average waiting time of $W = 1/2$ hour. Increasing the arrival rate to $\lambda = 19$ produces $W = 1$ hour. Therefore, an increase in the arrival rate by only one person per hour doubles the expected waiting time for customers entering the queue.

The preceding situation was, in essence, that which confronted the Bureau of Vital Statistics. The service rate of the original system had been held relatively constant for years, subjectively estimated at $\mu = 6$, while the customer arrival rate had been steadily increasing.

The task, then, was to estimate the service rate needed to produce an average waiting time of 15 minutes or less. To do this, one may again use the equation for W by setting $W = 1/4$ hour to obtain $\lambda = 4\bar{n}$ where $\bar{n}$ is a function of λ/μ and k. With a fixed value for k, a solution for μ may be obtained for each value of λ. These solutions for $k = 3$, the maximum number of servers desired, are plotted in figure 3. The use and interpretation of this graph is as follows: For any λ value of interest on the horizontal axis, determine the corresponding μ value on the vertical axis. This solution for μ gives the service rate required for each of three servers, when the

Figure 3. Relationship between arrival rate and required service rate for each of three servers in order to obtain average customer waiting time of 15 minutes



customer arrival rate is λ, to produce an average customer waiting time of 15 minutes.

The λ values of primary interest to the bureau were in the 20 to 30 range. Dashed lines in figure 3 show the approximate solutions corresponding to $\lambda = 20$, 25, and 30 to be $\mu = 8.3$, 9.9, and 11.5, respectively. As a consequence of these results, it was decided to pursue a service rate of $\mu = 10$.

Therefore, the problem in the bureau was simplified to increasing the service rate from $\mu = 6$ to $\mu = 10$. The problem could be analyzed and attacked better, however, by transforming from units of service rate (number of customers served per hour) to units of service time (number of minutes to serve one customer). This is easily done by taking the reciprocal of μ and multiplying by 60. For example, when $\mu = 6$, the service time per customer is $60 \div 6 = 10$ minutes. Equivalently, then, the working objective was to decrease the service time from 10 to 6 minutes per customer.

Action Taken and Results

Service time. The components of service time presented in figure 1 and the estimated times required to complete each component in the original system are shown in the table. The remaining research effort would be applied to decreasing the three component times so that their sum would total 6 minutes or less.

The research into possible microfilm retrieval systems was, by far, the most extensive and intensive of the three components. All systems considered were evaluated according to cost of equipment and supplies, speed of retrieval, quality of produced paper

Estimated minutes of service time, by service components for three systems

Service components	System		
	Original	New, immediate	New, potential
Retrieval	6	4	2
Preparation	3	1.5	1.5
Payment	1	0.5	0.5
Total minutes	10	6	4
Service rate (μ)	6	10	15

copies, suitability within the office's physical constraints, and compatibility with the original microfilm system. The compatibility factor was an important issue since it was hoped that the new system would actually be comprised of two subsystems, one delegated to managing past records and one for new records. Therefore, enhanced compatibility between the two subsystems would promote a smoother office operation.

The approval of a proposal containing a detailed cost-benefit analysis resulted in the purchase of a Kodak MICROSTAR (*A*) system for managing past records and the Kodak MIRACODE II (*B*) system for future records. An implementation of MIRACODE I has been reported by Crandall (*1*). Both systems use roll film housed in 4-ounce magazines. The MIRACODE II system does not require an external index. These combined systems would provide an overall savings of at least 2 minutes per retrieval from the outset of system implementation, with promise of time savings to 4 minutes as the request for older documents decreased.

Investigation of the certificate preparation procedure led to such actions as the relocation of office equipment and furniture to minimize the amount of office movement and the modification of copy preparation to expedite the production of a certified copy. Time savings in certificate preparation was estimated to be at least 1.5 minutes per request.

In the payment transaction component, the original cash box and handwritten receipt system was replaced with a cash register. This not only saved at least 0.5 minute per customer request, but it also contributed to saving more than 30 man-hours per month in completing daily cash reports. To increase efficiency, the cash register, microfilm retrieval machines, and a table for processing certificates were

placed in a work area of approximately 70 square feet near the customers' counter.

The required times for each component of service under the new system expected immediately after implementation and the potential of the new system in future years that, if reached, would produce an expected waiting time of only 5 minutes are also shown in the table.

With the use of a printing timeclock, data were obtained on the service times for 1,000 customers 1 year after the system was completely installed. The computed average of the service times, which were measured in minute increments, was 5.8 minutes. Therefore, the working objective of a 6-minute service time was considered to have been achieved.

The frequency distribution of the times was characterized by positive skewness, with a mode at 5 minutes. The service times ranged from 2 to 26 minutes; 69 percent were 6 minutes or less.

Customer waiting times were measured 2 years after the system was installed. During most of the measurement period, the daily arrival rate was such that two clerks were sufficient to serve the counter. Therefore, a direct test of the waiting time criterion had to be made by use of two servers instead of three. Computations show that when $k = 2$ and $\mu = 10$, a waiting time of 15 minutes is expected when $\lambda = 15$. Consequently, data were captured only during work periods when approximately $\lambda = 15$. The results of 265 observations taken over 10 days showed an average waiting time of 12.3 minutes per customer, an encouraging result for queueing theory where one considers the relatively small sample size. Eighty percent of the waiting times were 15 minutes or less. The conclusion was that the objective was being achieved.

Arrival rate. Further efforts were made to decrease waiting time by decreasing the arrival rate of customers to the bureau counter. Because approximately 90 percent of the counter customers requested birth certificates, it was hypothesized that encouraging mail requests for birth certificates would have the following results:

- Decrease the arrival rate of counter customers, with a result of further decreasing waiting time;
- Increase the arrival rate of mail requests, which would result in less time-response pressure;
- Provide more stability to daily office workload; and
- Improve employee morale.

As the result of a suggestion by Harry B. Garrett, registrar of vital statistics, City of Houston Health Department (personal communication, July 1972), an envelope system was adopted. With this system, envelopes are supplied by the Bureau of Vital Statistics to all Dallas hospitals offering obstetrical services. These envelopes, given to parents of new babies before they leave the hospital, were designed for convenient ordering of birth certificates by mail. The bureau currently estimates that more than one-half of the envelopes issued are being used. The full effect on the counter traffic will not be realized until approximately 6 years after implementation of the envelope system.

Another aid in reducing customer arrival rate was a recently passed Texas State law that allows the natural mother of an illegitimate child to obtain a certified copy of the child's birth certificate without a court order. Before this bill was adopted, these mothers had to visit the bureau twice to obtain a certificate, once to obtain the registrar's file number for the court order and once to submit the acquired court order for the certificate. Since only one visit to the bureau is required now in most cases, it is estimated that the overall arrival rate should be reduced by at least 10 percent.

Discussion

A key point repeatedly emphasized during the study was that no single change in service procedures could be expected to solve the bureau's problem completely. A satisfactory solution would require the consideration of all components of the problem. Saving 10 seconds of service time by moving the paper cutter to another location in the office may not seem significant, but six such changes will save 60 seconds per request, which is significant. A similar argument can be made for influencing the arrival rate. Small service time and arrival rate changes do add up, and queueing theory warns that their inclusion or exclusion may mean the difference between a facility's experiencing calm or storm.

References

1. Crandall, J. H.: D. C.'s vital records automated to speed copies. Health Serv Rep 87: 200–204, March 1972.
2. Richmond, S. B.: Operations research for management decicisions. Ronald Press Company, New York, 1968, ch. 14.
3. Bhat, U. N.: Elements of stochastic processes. John Wiley and Sons, Inc., New York, 1972, ch. 11.
4. Little, J. D. C.: A proof for the queueing formula: $L=\lambda W$. Operations Res 9: 383–387, May–June 1961.

Equipment References

A. Recordak MICROSTAR, Eastman Kodak Co., Rochester, N.Y.
B. MIRACODE II, Eastman Kodak Co., Rochester, N.Y.