

APPLICATION OF QUEUING THEORY
TO A NONPROFIT REFERRAL SYSTEM

by

PRAFULLA JOGLEKAR
and
MADJID TAVANA

Applied Research Center
La Salle College, Philadelphia

A paper presented at the Northeast
American Institute for Decision Sciences
April, 1983

Note: For reasons of client confidential-
ity the name of the client
organization is fictitious.

by

Prafulla Joglekar, La Salle College
and
Madjid Tavana, Beaver College

ABSTRACT

Management science techniques are rarely used in social service agencies. It is often assumed that the lack of profit criterion makes it difficult to use the quantitative approach. This paper presents an application of queuing theory to the operations of a non-profit telephone referral system. A model is built to explain the existing system and to estimate the current capacity shortfall, i.e., the number of people who were denied the service. The model is then used to assess the impact of alternative methods of system improvement on the capacity shortfall. Benefits are estimated in quantifiable, but not necessarily monetary, terms.

THE PROBLEM AND ITS SETTING:

In the Spring of 1982, The Attorney Referral Service (ARS) was concerned about a serious capacity shortfall in its telephone referral system. A client could seek legal advice from ARS on phone by calling the Referral (RFL) group of numbers or by calling the DIAL-LAW hot line numbers. ARS's principal concern was the service to callers on the RFL group.

The RFL group consisted of five interconnected numbers. When a client dialed one of these numbers, if that specific line was in use, the call automatically transferred to one of the open lines within the group. If all five lines were in use, the caller got a busy signal. The number of calls coming in on the RFL group differed from day to day and hour to hour. However, detailed statistics were not available. All that ARS staff could say with certainty was that every day some three hours could be seen as "peak hours" with the RFL lines constantly ringing and the other hours were off-peak hours with a more relaxed pace of work. When an RFL line rang a few times, one of the paralegals answered it and put the caller on hold if the paralegal was busy with another client. In general, the caller did not have to be on hold for more than a few seconds, if at all. Next, the paralegal was responsible to interview the caller, ascertain his/her problem and provide the appropriate advice. In a large majority of situations (some 87%), the advice provided was satisfactory to the caller and the call terminated. However, in approximately 13% of the cases, the caller sought the name of a lawyer to contact. In such cases, the paralegal identified the appropriate attorney to contact (a referral), advised the caller of the fees involved, and completed the related paperwork. In short, serving an RFL call involved three basic types of activities: (i) interview and advice, (ii) locating attorney and (iii) completing related paperwork. Estimates of the average length of call differed considerably among the staff interviewed (range: 1½ minutes to 4 minutes if no referral was involved, and 2 to 5 minutes if a referral was involved).

ARS staff felt that activities (i) and (iii) were irreplaceable whereas activity (ii) could be simplified either by obtaining a computer or appropriate clerical help for the paralegals who manned the RFL phones. They felt that locating the attorney took substantial time, since ARS had a list of over 600 attorneys who were admitted to its referral panels of various specialties (e.g., divorce, crime victim, juvenile delinquents, etc.). In giving the referrals, ARS wanted to be as fair and equitable to the panel

members as possible. For example, if a caller wanted a referral to a divorce lawyer, ARS would search its panel to identify a lawyer specializing in divorce who had received the least number of referrals from ARS. Thus, it was necessary to maintain an index card for each attorney with an up-to-date tally of the total number of referrals provided to that attorney. The attorney cards were filed at a central location. Every time a referral was made, the staff had to put a caller on hold, walk up to the card file, locate the appropriate attorney card, take it back to the paralegal's desk, make the referral, update the card to indicate the new referral and return the card to the file.

Until December 1981, the RFL lines were staffed by only two paralegals who felt overwhelmed by the constantly ringing phones, particularly during peak hours. As such, staff attorneys had to help the paralegals by answering a substantial number of calls on the RFL group. The number of calls handled by the attorneys was far in excess of the number necessary for them to "keep in touch." Often during peak hours, all five of the RFL lines were busy and many individuals were unable to contact ARS. The loss of clients meant that ARS was underachieving its community service goals. ARS was also losing potential revenue, since about 35% of all referrals made by ARS eventually yielded a revenue of \$15 per referral from the participating attorney. In January, 1982, ARS hired a third paralegal to man the RFL lines. Even after this addition, the attorneys had to continue to help answer a large number of calls, and often all five lines were busy.

In order to obtain some idea of the number of clients lost because of the capacity shortfall, ARS obtained from the telephone company a study of the number of busy signals on all lines. The study reported that there were an average of 524 busies per day on the RFL group. This was very high, considering that the number of calls answered per day was only 195. Of course, the busies represent the number of times prospective clients tried to call ARS when all five lines were occupied by other callers. The busies do not equate to the number of people who could not reach the ARS office. Nevertheless, the busies were indicative of the large unsatisfied demand for ARS services. ARS was not sure how to estimate the number of people who were being denied their services, and the revenue consequently foregone. It was not clear if the problem was one of capacity or its effective and efficient utilization. ARS considered several courses of action as listed below, but did not know how to evaluate them in terms of their potential implications for service, revenues and costs.

- a) Adding one or more telephone lines to the RFL group of numbers.
- b) Hiring additional paralegals.
- c) Hiring clerical assistance to free the paralegals to do only the interview and advice function.
- d) Providing computer terminals for the paralegals to speed-up clerical tasks involved in a referral.

RESEARCH METHODOLOGY

ARS had no budget to pay for any professional advice on the problem. As part of a grant from a foundation, however, we were authorized to spend up to two days of paid work to help ARS. The methodology described below has to be seen in that light. It is a reasonable and helpful analysis but certainly not thorough, in spite of the fact that we volunteered

several additional days. Secondly, ARS is a nonprofit organization. Consequently, in assessing the effectiveness of alternative system improvement strategies, we had to depend upon such nonmonetary criteria such as reduction in the number of people who are denied the service, or reduction in attorney time spent on answering the phones.

Basically our method was first to build a queuing theory model that explained the number of RFL callers served by the existing system and estimate the number of potential callers denied the service. This model also helped us estimate the number of attorney-hours currently spent in answering the phones. Once the current system was reasonably explained, we used the model to assess the effectiveness of alternative system improvement means. Finally, our recommendations noted that the so called alternative means were not mutually exclusive and it would be best to use a combination of them.

Adaptation of the Text-Book Model:

Queuing theory provides a variety of models depending upon the characteristics of the queuing system. Clearly, since there were three paralegals and five lines we were dealing with a multiple channel, limited queue length situation.

We assumed poisson arrival, infinite source population and exponential service rate (assumptions that have been commonly used and empirically validated by several telephone queue studies). Yet ARS system did not quite fit a standard text-book model for the following reasons:

- 1) Text-book models assume a single queue in front of a group of servers. At ARS, paralegals answered not only the calls on RFL group but also any "overflow" on the DIAL-LAW group. In general, a DIAL-LAW caller sought legal information pertaining to a specific topic (e.g., divorce, small claims, etc.). ARS receptionist picked up the call, asked the topic of the caller's interest and played a pre-recorded message on the subject. In almost 75% of the cases, the message was satisfactory to the caller who hung up at the end. However, the message encouraged the caller to continue holding the line if he/she needed additional information or referral to an attorney specializing in the topic of interest. When a caller held the line a few seconds after the termination of the message, the line began to blink and the paralegals knew that an "overflow" had occurred. The next available paralegal then answered the overflow line in a manner exactly identical to handling an RFL call. Thus, at ARS two different queues were being served. However, with limited time on hand, we decided to assume a fixed ratio of service to the two queues, adjusted the service time per call for the RFL group accordingly, and proceeded to use the standard model.
- 2) Text-book models assume one arrival rate for the system. The peak period at ARS seemed so different from the off-peak period that we decided to analyze the system as a combination of two distinct periods with two distinct arrival rates. We assumed that of the 1882.5 hours of operation in a year, 752 hours (40%) were peak hours with an arrival rate four or five times as large as in the remaining off-peak hours.
- 3) Standard queuing models often consider that the number of servers is the same throughout a period of analysis. Available ARS data indicated that because of other assigned duties, vacations, sick days, lunch hours, etc., the three paralegals were not always available. Using available historical data and a few simple assumptions we estimated

that of the 1882.5 hours of operation in a year, all three paralegals were available for only 546 hours. For another 1129.5 hours, only two paralegals were available, and the remaining 207 hours only one paralegal was on telephone duty.

Of course, when enough paralegals were not available, staff attorneys helped out. Records showed that attorneys were answering 16.8% of RFL calls. Again, using a few simple assumptions, it was estimated that during 1982 attorneys contributed 828 hours to man the telephones, and that contribution resulted in an operation involving 1167 hours (62%) with 3 servers (paralegals + attorneys) and 715.5 hours (38%) with 2 servers.

Thus, to explain the existing system, we used the standard multiple-channel, limited-queue-length model in 4 situations described in Table I. The model and a numerical example for situation B are presented in the Appendix.

TABLE I

No. of Servers \ Arrival Type	Peak Period	Off-Peak Period
Two	Situation A (286.0 hrs/yr)	Situation C (429.5 hrs/yr)
Three	Situation B (466.0 hrs/yr)	Situation D (701.0 hrs/yr)

DATA COLLECTION AND ANALYSIS

To estimate the average service time, we conducted a time and motion study summarized in Table II. The study showed that an average call (RFL or DIAL-LAW overflow) required 2.93 minutes of service time.

TABLE II

Type of Activity \ Type of Call	Average Observed Time in Minutes		
	Involving Referral	Without Referral	Average Call
Interview and Advice	2.73	2.63	2.65
Locating Attorney	0.98	—	0.16
Filling Referral Form	0.82	—	0.13
All Activities	4.53	2.63	2.93

ARS statistics showed that on an average each day 195 RFL calls and 30 DIAL-LAW overflows were serviced. Assuming this proportion would remain unchanged regardless of any changes in the system, we calculated that the effective service time for an RFL call was $2.93 \times 225/195 = 3.38$ minutes. Consequently, the service rate was 17.72 customers per hour per paralegal on duty.

To estimate the arrival rate during peak hours, we picked a known peak period (Monday morning 10:00 AM to 11:30 AM) and observed the number of calls coming in. In 89.8 minutes of observation, RFL lines rang 71 times. Thus average effective arrival rate for the peak period was estimated to be 47.45 calls/hour. The true arrival rate would have been higher but for the limited queue length (5 lines) permitted by the system. To estimate the true arrival rate we also took 30 random observations during a peak period and noted how

many lines were occupied with what frequency. All five lines were occupied in approximately 34% of the observations. Hence, using the theoretical model, we estimated true arrival rate during peak hours to be 72 callers/hour (see Appendix).

Using annual statistics and the theoretical model, we estimated that the off-peak arrival rate was 15.07 callers/hour.

The arrival and service rate estimates enabled us to explain how many callers would be served and how many would be denied service in each of the four types of system situations described in Table I. We estimated that in situation A, 34.63 of the 72 callers would be served per hour. In situation B, 47.45 of the 72 callers will be served. Whereas in situations C and D every one of the 15.07 callers would be served. Thus, if no system changes were made in 1982, the total number of people desiring the service but denied would be 22,187.

CONCLUSION AND RECOMMENDATIONS

With the existing system explained, we were ready to assess the impact of alternative system improvement means. Available space does not permit us to detail the computations. Readers familiar with queuing theory may be able to reconstruct some of our computations by noting that:

- when new telephone lines are added to the RFL group, the principal change is in N , the maximum permissible queue length.
- When a full-time or a part-time paralegal is added, the number of servers increases. But remember that the new paralegal is also entitled to vacations, sick days, lunch hours, etc. Consequently, one must first recompute the number of hours/yr in the various situations of the types described in Table I, and then use the model.
- When clerical assistance is used to locate the attorney and complete the referral, only the interview time from the time motion study would be the pertinent time in determining the service rate. Thus, service rate goes up from 17.72 to 19.285 per hour per paralegal. On the other hand, the line on which referral is given out is occupied for the total time. This situation does not fit a standard model. A careful evaluation of the situation have required substantial modeling efforts. We simply placed upper and lower bounds on the potential effects by using reasonable assumptions.
- If a computer terminal is made available to each paralegal, a portion of the time required for locating the attorney and filling out the form is saved. Average service rate would increase but not substantially.

TABLE III

System Improvement Means	Additional Clients Served/Yr	Attorney Hours Saved/Yr
a.1 Add 1 telephone line	958	0
a.2 Add 2 telephone lines	1,563	0
b.1 Add 1 full-time paralegal	4,371	414
b.2 Add 1 part-time paralegal during peak hours only	4,371	414
c. Provide a full-time clerical assistant	310 to 2,135	30 to 100
d. Provide computer terminals	1,757	50

Table III summarizes our estimates of the annual benefits of each alternative in terms of number of additional clients served and attorney hours saved. Considering the financial costs of each of the alternatives as well, and recognizing that the alternatives are not mutually exclusive, we recommended that ARS add two telephone lines to the RFL group and add a part-time paralegal for the peak hours, to the extent they were predictable.

Appendix*

Multiple-Channel Limited-Queue-Length Model

Let,

- n = Number of Customers in system.
- c = Number of servers.
- N = Capacity of the system, i.e., the maximum number of customers allowed in the system at any time.
- λ = Mean arrival rate per hour.
- λ_e = Effective mean arrival rate per hour (also effective service rate of the system).
- μ = Mean service rate per server per hour.
- ρ = System utilization = (λ/μ)
- P_n = Probability of n customers in system.

Then

$$P_0 = \begin{cases} \left[\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c (1 - (\rho/c)^{N-c+1})}{c!(1 - \rho/c)} \right]^{-1} & \text{if } \rho/c < 1 \\ \left[\sum_{n=0}^{c-1} \frac{\rho^n}{n!} + \frac{\rho^c}{c!} (N - c + 1) \right]^{-1} & \text{if } \rho/c = 1 \end{cases}$$

$$P_n = \begin{cases} \frac{\rho^n}{n!} P_0 & \text{if } 0 \leq n \leq c \\ \frac{\rho^n}{c! c^{n-c}} P_0 & \text{if } c < n \leq N \end{cases}$$

and, $\lambda_e = \lambda(1 - P_N)$

Example: Consider situation B of Table I. Here we know that $\lambda_e = 47.45$ per hour, $\mu = 17.72$ per hour per paralegal, $C = 3$, and $N = 5$. Through trial and error one can determine that $\lambda = 72$ will be consistent with these data. Note that if $\lambda = 72$, $\rho = 4.0632$.

Then, by our formulae $P_0 = .016625$, $P_1 = .0675$, $P_2 = .1372$, $P_3 = .1859$, $P_4 = .2517$, and $P_5 = .3410$

Hence, $\lambda_e = \lambda(1 - P_5) = 72(1 - .3410) = 47.45$

* For a complete description of this model, see Hamdy A. Taha, "Operations Research: An Introduction", MacMillan Publishing Co., Inc., 1982, pp. 611-13.