SIMULATING THE CUSTOMER PURCHASING / PERSONAL SELLING PROCESS IN RETAILING WITH A QUEUING APPROACH

R. Dale VonRiesen
Lester W. Jacobs

ABSTRACT

This paper presents a simulation model developed for a nonstandard retail queuing system. The type of system for which the model and simulation results are appropriate is one utilizing both open display and personal selling to sell shopping goods. Two characteristics of this kind of system differentiate it from operations generally considered amenable to waiting line analysis. First, an arrival to the system does not necessarily constitute immediate demand for service. Second, there are multiple points in the system at which a customer may leave without having service. The model may be used to determine optimum numbers of servers and to experiment with different sets of marketing strategies designed to alter the behavior of customers in the system. The model also may be extended for use in a Bayesian decision framework by treating simulation-generated expected values as conditional values. The findings of experimentation on three factors--arrival intensity, time of unassisted search, and transitions following unassisted search--are presented. It is also shown that management may use measures of either capacity utilization or the maximum expected number of services in a congestion-free system to test the effectiveness of present policy for an ongoing system without solving the model.

INTRODUCTION

A major problem in retail systems which utilize both open display and sales personnel to sell shopping goods is determining optimum numbers of clerks. There is evidence that this type of system is likely to be understaffed [1]. This situation, and the resulting loss of potential revenue, is largely due to a failure on the part of management to view the system as a queuing process. The purpose of this paper is to present a model developed for the simulation of this type of nonstandard queuing process.

THE MODEL

The model of the selected retail queuing system has a variable phase, variable channel structure. The variable phase aspect of the model is required to account for (1) different customer desires regarding the form of search, (2) departures following the unassisted examination of merchandise due to information obtained, (3) balking, and (4) reneging. The number of channels is treated as variable to adjust for staffing changes made in response to different arrival rates and for temporary absences of personnel for meals and breaks.

PHASE ASPECTS

With open display, customers may examine merchandise without the aid of sales personnel. Fitting and/or information-providing operations, as well as the completion of a transaction, however, require the assistance of a clerk. Thus, by expanding the general definition of a queuing phase to include an operation performed on, for, or by an element, the model incorporates two possible phases--namely, unassisted and assisted search. Given two phases, there are three possible states in which a customer may spend time--unassisted search, assisted search, and waiting. The possible transitions between states are shown in Figure 1.

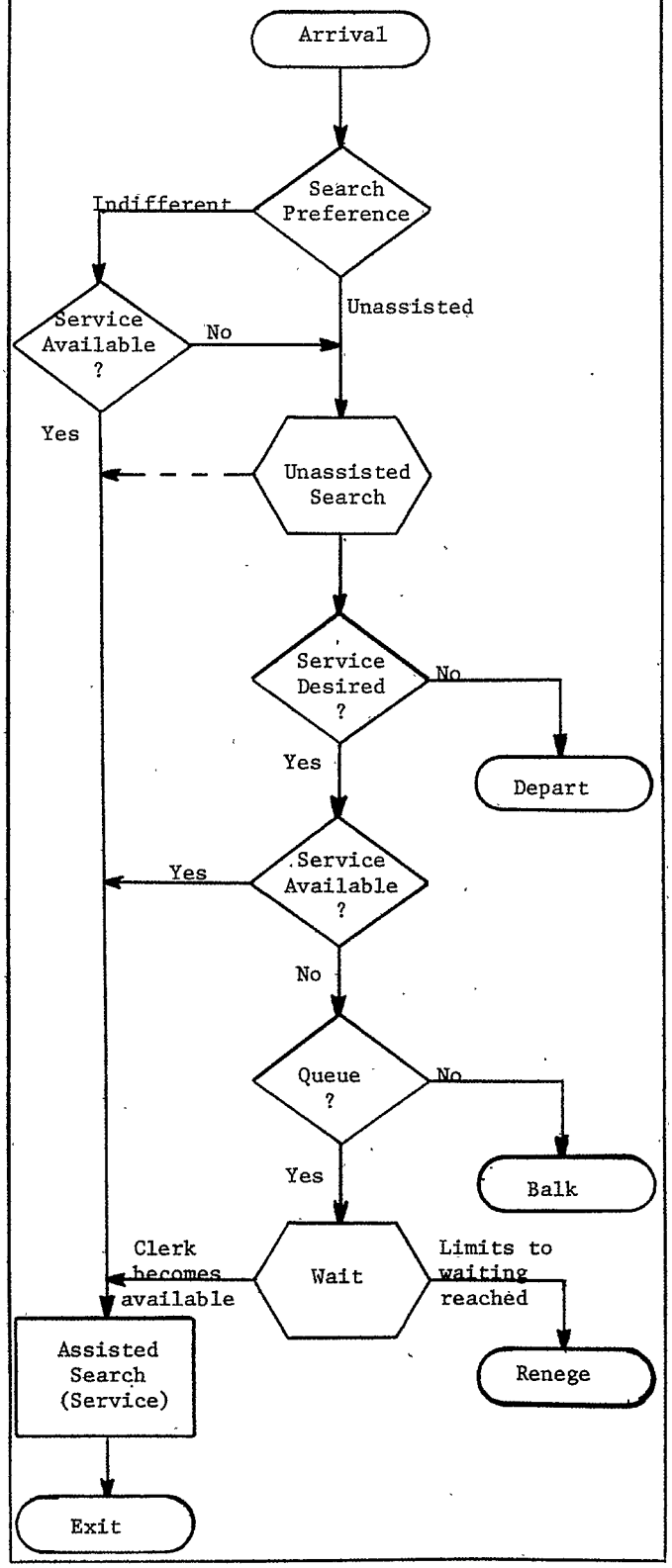
Unassisted search includes merchandise examination activities which customers perform in the absence of a clerk. It differs from waiting (time in queue) in that it is assumed to cause no dissatisfaction with the system's inability to provide a server, and it is an active part of the customer's information gathering process. Assisted search includes face-to-face activities of customer and clerk in a joint search process--that is, the usual service or holding action. The state of waiting is assumed to produce dissatisfaction, and it does not involve an active examination of merchandise.

The initial state occupancy of an arrival to the system depends on the customer's preference regarding the form of search and, in some cases, on the availability of sales personnel. Arrivals may (1) prefer unassisted search initially, (2) prefer to start the process with assisted search, or (3) be indifferent to the initial form of search.

Customers preferring unassisted search start with this state regardless of the availability of clerks and will not accept service until they have completed the process. Those who prefer to start with assisted search start with this state, if sales

FIGURE 1

Customer Activity Patterns



personnel are available upon arrival. If personnel are not available, these customers actually have the choice of not entering the system or joining the queue. However, since not entering the system has much the same influence on the system's output as a balk following unassisted search, an operational simplification treats these customers in the same manner as those who are indifferent to the initial form of search. Customers who are indifferent to the form of search start with unassisted search, if clerks are not available. They accept service, if it is available upon arrival.

Behavior at the completion of unassisted search depends on the results of unassisted search and on the availability of sales personnel. Some customers conclude from unassisted search that further search is not desired. These customers leave the system regardless of clerk availability. Others do desire service. When personnel are available, these customers transfer directly from unassisted to assisted search. When clerks are not available, a choice is required. Some customers refuse to wait and balk (leave the system). Others enter the queue. Those in queue either transfer to assisted search or renege (leave the system).

CHANNEL ASPECTS

As indicated above, the number of channels is not fixed since staffing changes are made in response to varying arrival rates during a single period of operation and due to personnel absences for meals and breaks. The system is also nonstandard in the sense that service time and idle time do not constitute 100 percent of the clerk's floor, or work, time. Four additional time-expenditure categories (listed below) are used to account for the in-department time of personnel.

The utilization of personnel time is accounted for by eight states. They are:

1. Time moving from one location in the department to another in order to initiate service for a customer in queue.
2. Time moving from one location in the department to another in order to contact a customer in unassisted search or a new arrival to the system.
3. Offering service to a customer who declines in favor of unassisted search.
4. Service time.
5. Time required to return products shown during service to inventory.
6. Breaks.
7. Meals.
8. Idle time.

The behavior of sales personnel is responsive to customer state conditions and is governed by seven general rules. The rules are:

1. Customers in queue have priority over those in the unassisted search state.

2. Customers completing unassisted search have priority over other customers in the unassisted search state regardless of arrival times.

3. Customers in queue are contacted on a first waiting, first contact basis.

4. Arrivals are contacted on a first arrival, first contact basis.

5. A customer in the unassisted search state is contacted a maximum of one time prior to voluntary completion of unassisted search.

6. All customer contacts have priority over personnel departures for breaks and meals.

7. Personnel departures for the purpose of terminating a period of employment have priority over any customer contact not already initiated.

PROBABILISTIC NATURE OF THE MODEL

The model basically is probabilistic in nature. Interarrival times, unassisted search times, and service times are treated as negative exponential functions. With each factor, the mean of the distribution is conditional on external and/or policy circumstances. For example, the mean interarrival time is conditional on the season, day of the week, time of day, and advertising policy. The mean unassisted search time is conditional on the effectiveness and extent of the open display. The mean service time is conditional on inventory policy and promotional strategy.

Other uncontrollable variables--such as customer search form preference, state transition, and the outcome of service--also are treated stochastically. Like the above, these variables are conditional on other circumstances such as managerial policy regarding promotional activities and the nature of the open display.

All aspects of personnel behavior other than service and idle time are handled in a deterministic manner. Constants are assigned to time expenditures in the first three personnel states above. The reason for this treatment is that the amount of time required by each state is small and, consequently, the effects of the states on the output of the system are primarily due to their cumulative utilization. The time assigned to the fifth personnel state--returning products to inventory--is a constant proportion of the time of the just-completed service on the assumption that the number of items to be returned to stock, and the time required to do so, is related to the length of service. The model includes exception routines for each constant in order to prohibit the generation of needless balking and reneging. For example, if a salesperson who has just completed service is needed to contact a customer in queue, the time of the fifth state is absorbed in the following service.

COMPUTER MODEL STRUCTURE

The basic methodology of the simulation model utilizes a next event time chain in variable-time increments. The modular structure of the model makes it extremely flexible as the choice of varying routines may be established at load time, with additional changes in model parameter values being indicated at run time. During a simulation run the following variants may be introduced:

1. Alter the interarrival time mean.

2. Increase/decrease the sales force size.

3. Initiate a meal/break routine (allow up to n total salesmen to leave for h hours with no more than m salesmen out at any given time).

4. Take a "snap shot" of the system accumulating run statistics with the option to reinitialize all statistics (i.e. a "warm up" period).

5. Terminate the run with the option to save the current status of the system.

Item five is a special case of item four and is incorporated as a separate function in the model to facilitate changes in parameters that are restricted to be constants for any particular simulation run. By saving the current status at the end of a run, the following parameters may be altered and the simulation picked up in the middle of a simulated day on the subsequent run:

1. The mean unassisted search time for customers.

2. The mean service (assisted search) time.

3. The customer search form preference on arrival.

4. The customers' propensity to depart on completion of unassisted search.

5. The propensity of customers to balk.

6. Salesman movement times to initiate a contact and time spent offering service.

7. The proportion of service time required to return products to inventory following service.

TESTS OF THE MODEL

The model was tested for workability and sensitivity in a two-stage process. First-stage testing involved comparisons between values from a hand-computed simulation and those obtained from measurements taken during four days of observation

of an actual system. Second-stage testing involved computer simulations.

The data used for testing were obtained from the womens' shoe department of a large department store. The policy of the department was to have at least one of each pattern in stock on display. During the data collection period, approximately 300 different patterns were available for customer examination.

First-stage testing involved four tests of workability--the ability of the model to approximate observed behavior--and one test of sensitivity. These tests were made at the .05 level. The major test of workability involved the difference between simulated and observed ratios of the number of services to the number of arrivals. Supplemental tests involved comparisons of simulated and observed ratios of departures to arrivals, balks to arrivals, and customers who transitioned from queue to service to the number of arrivals. Nonsignificant differences in all four cases indicated workability in the sense that the model is capable of approximating the observed behavior of the actual retail system. A test of sensitivity was made using two subsets of the simulation data. The expected number of arrivals per clerk in one set was about 60 percent greater than in the other. A ratio of services to arrivals in the first set which was significantly less than that for the second set provided evidence of the model's responsiveness to changes in arrival intensity and/or sales force size.

While the second-stage testing is a continuous process and has not been completed, the tests that have been performed with the aid of computer simulation have been much more extensive than those of the first stage. Thirty days of operation have been simulated for each of several sets of conditions. The results have been consistently supportive of the first-stage findings regarding both workability and sensitivity. To date, the model has been found sensitive to changes in four variables--number of clerks, arrival intensity, mean unassisted search time, and state transition following unassisted search. Data on these variables and the managerial implications of the simulation results are summarized in a followup section. Preliminary tests of different policies regarding the temporary reduction of the number of available channels due to clerk absences for meals and breaks have shown the model to be insensitive to different methods of staggering personnel departures.

MEASURES OF EFFECTIVENESS

Due to the nonstandard nature of the queuing system under study, somewhat atypical measures of effectiveness are required for performance analysis. The two measures having the greatest managerial relevance are the cost of congestion and capacity utilization.

THE COST OF CONGESTION

With many systems that are treated with waiting line analysis, the purpose of the analysis is to minimize the sum of the cost of a service and the cost of waiting. However, with many marketing systems, there are customer-related costs to the system other than the cost associated with waiting. In fact, the cost associated with waiting may be a minor part of the total cost of congestion [2]. If competitive conditions are such that the customer's alternatives are perceived to offer similar values, any behavior resulting in departure prior to service can be costly. Models of systems of this type require the utilization of a cost of congestion concept which is more extensive than the concept of waiting time cost.

The short-run cost of congestion for the type of system considered is defined as lost immediate sales opportunity. The loss of sales opportunity occurs at three points in the system. First, sales opportunity is reduced by an inability to contact some customers who are indifferent to the form of search. All customers who are indifferent to the initial form of search will accept service, if it is offered before they complete unassisted search. However, when left to conduct the search process on their own, some conclude at the completion of the process that service is not desired. The departures of these customers represent unassisted search cost. Second, sales opportunity is reduced by the system's failure to have clerks available when some customers complete unassisted search with a desire to have service, but leave the system due to an aversion to waiting. This is the cost of balking. Finally, opportunity is reduced by not having sales personnel available to service some customers in queue before the customers' limits to waiting are reached. This is the cost of reneging.

The cost of congestion may be measured in terms of either number of services or dollars. If, as was the case during the first-stage testing, the probability of purchase and average purchase amount are independent of state patterns leading to service, the cost of congestion is the difference between the maximum number of services expected in a congestion-free system and the number provided by a given number of clerks. A congestion-free system would be one having a theoretically infinite number of clerks in order to meet every demand at every point in the system. The maximum number of services (MES) is estimated as follows:

$$MES = N \times P(A) + N \times P(R) \times P(S|R)$$

where

N = The number of arrivals during a specified time period

P(A) = The probability of a customer accepting service upon arrival

P(R) = The probability of a customer refusing service upon arrival

P(S|R) = The probability of a customer accepting service at the completion of unassisted search given that the customer had refused service upon arrival

CAPACITY UTILIZATION

Capacity utilization is defined as the ratio of total service time to floor time. Floor time excludes the time clerks are away from the system for breaks and meals. As was indicated above, floor time is somewhat greater than the sum of service time and idle time. Four other personnel activities--categories 1, 2, 3, and 5 from the list of eight personnel states--also account for part of floor time. These activities are excluded in order to center attention on the major purpose of capacity--namely, the provision of service. Also, the service function is more visible to a manager than are the other four activities. Given an optimum number of servers, the excluded activities comprise about 6 percent of floor time. Thus, as defined, the maximum capacity utilization is about 94 percent.

SIMULATION RESULTS

Experimentation with the model has produced results which have interesting managerial implications. The basis of comparison for the experimentation was thirty simulations of a nine-hour day using parameters that are representative of those for a typical day in the shoe department used for testing.

The average interarrival time is conditional on the time of day. The normal-day mean interarrival times for three periods--9:00 to 12:00, 12:00 to 4:30, and 4:30 to 6:00--are 1.88, 1.40, and 2.17 minutes. The search form preference probabilities are .664 for an unassisted search preference and .336 for indifference to the initial form of search. The unassisted and assisted search time means are 1.72 and 8.61 minutes. State transition probabilities for customers completing unassisted search are shown in Table 1. The function governing the act of reneging is specified in intervals of one minute. As examples, the probabilities of reneging after 5, 10, 15, 20, and 25 minutes in queue are .143, .621, .925, .994, and 1.0. The probability of purchase for customers having service is .309, and the average gross margin per transaction is $13.60.

Payroll costs are based on a $5.00 per hour cost for the manager, who also performs as one of the clerks, and $3.00 per hour for all other clerks. All clerks are paid for eight hours. With a nine-hour day and one hour off for lunch plus two 15-minute breaks per clerk, total floor time is equal to 7.5 hours multiplied by the number of clerks employed for the day. The constants used for the two personnel states involving location change are 10 seconds. The constant for the state of offering service is 20 seconds. Time required to return products to inventory is 10 percent of the last service time.

TABLE 1

State Transition Probabilities

Event	Probability
If clerk is available at completion of unassisted search, accept service	.5175
If clerk is available at completion of unassisted search, depart as result of unassisted search	.4825
If clerk is not available at completion of unassisted search, depart as result of unassisted search	.4825
If clerk is not available at completion of unassisted search, enter queue	.2640
If clerk is not available at completion of unassisted search, balk	.2535

Experimentation conducted to date has been concentrated on the effects of variation in the arrival intensity, mean unassisted search time, and state transition probabilities on the optimum number of servers, the cost of congestion, and capacity utilization. The system has been simulated using two additional sets of interarrival time means. With one set, the interarrival means are 50 percent greater than those of the typical situation above. In the other case, the means are two-thirds the value of the normal-day means. Two additional mean unassisted search times have been used. One of these is two-thirds the value of the normal-day mean and the other is 50 percent greater than the normal. The modifications of the transition probabilities are comparable to those for interarrival and unassisted search times, but the changes involve ratios. With one set of probabilities, the ratio of the probability of departure to the probabilities of other actions is two-thirds the ratio for the normal day. In the other case, the ratio is 50 percent greater than the normal ratio.

Of the three variables examined, arrival intensity has the greatest impact on the number of servers required for optimal operation. Increasing the rate of customer flow from light to normal changes the optimum number of clerks from six to eight. The increase in arrival intensity from normal to heavy increases the optimum number to 12.

An interesting finding is that the percentage decrease in the cost of congestion exceeds the percentage decrease in capacity utilization as the sales force size is increased. In the area of the optimum sales force size, the slope of the congestion cost curve is approximately 2.5 to 3.5 times the slope of the capacity utilization curve, depending on the rate of customer flow.

It is not surprising that the above factor of 2.5 is associated with the light arrival flow and the 3.5 with the heavy since queuing systems are expected to operate more effectively as they become more heavily loaded. Two other indications of the loading phenomenon are the increase in capacity utilization and a decrease in the cost of

congestion as a percent of MES as the arrival intensity increases. Given an optimum number of servers, capacity utilization for the light rate of customer flow is 41.8 percent. The heavy flow utilization figure for the optimum number of servers is 49.0 percent. Given optimum numbers of servers again, the cost of congestion as a percent of MES decreases from 9.8 to 6.7 percent with the change from light to heavy arrival intensity.

Increasing the average time of the unassisted search process was found to have no effect on the optimum number of servers. These increases also resulted in only negligible increases in capacity utilization. However, they did produce reductions in the cost of congestion ranging from 10.9 to 13.0 percent.

Decreasing the probability of departure relative to other actions was found to increase the optimum sales force size by a factor of one clerk. These increases in the optimum sales force size produced only negligible decreases in capacity utilization, but the decreases in the cost of congestion range from 8.7 to 11.4 percent.

The explanation for the findings on unassisted search times and transition probabilities is that reducing the probability of departure following unassisted search has a more direct and more pronounced effect on the system's load than does increasing unassisted search time. Reducing the probability of departure increases MES. In spite of this increase, the decrease in the cost of congestion as a percent of MES is greater in the case of the reduced probability of departure than it is with increased unassisted search time and a constant MES.

There are several managerial implications of the above findings. A major implication is that more attention should be given to methods of increasing customer traffic than to decreasing the probability of departure or increasing the average unassisted search time. However, management must be prepared to staff in response to an increased arrival intensity. Not to do so is to accept a substantial opportunity cost. For example, given the revenue and cost data used in the study, the expected contribution margin (gross margin less payroll cost) for eight clerks in the normal flow case is $639 per day. With a 50 percent increase in the arrival intensity, the expected daily contribution margin for eight clerks would be $914. However, with 12 clerks--the optimum number--the contribution margin would be $990. A daily opportunity loss of $76 translates to an annual figure in excess of $22,500 for a department which is open six days a week.

Another implication is that marketing strategies designed to reduce the probability of departure following unassisted search should receive more attention than those intended to increase the average unassisted search time. Again, management must be prepared to staff accordingly, since successful strategy would be expected to increase the number of clerks required for optimal operation.

A final implication is that there is little excuse for management staffing in a manner which is not nearly optimal. Either MES or measures of capacity utilization may be used to estimate how close a particular staffing decision is to optimal. Given the model parameters and revenue and cost data used in this study, the number of services performed as a percent of MES ranges from 90.2 to 93.3 percent. The data required for the calculation of MES are easy to obtain. The study also shows that optimal staffing may be expected to result in capacity utilization of less than 50 percent. Capacity utilization also is easy to measure. Thus, management has its choice of two guidelines, the application of which should produce staffing decisions close to those that would result from a simulation of the particular system.

SUMMARY

This paper treats a simulation model developed for a nonstandard retail queuing system. While the major purpose of the model is the determination of optimum numbers of clerks, a benefit of a simulation rather than an analytic model is the ease with which experimentation may be conducted. The model allows for the testing of strategies other than sales force size and the analysis of interactions among proposed strategies. For example, it is possible to experiment with advertising strategies designed to increase customer flow and/or modify search form preferences and display strategies intended to increase the average unassisted search time and/or modify transition probabilities following unassisted search. The effects of changes such as these on the optimum number of servers and profitability may be estimated easily.

Experimentation with the model indicates that arrival intensity has a pronounced impact on both the number of servers required for optimal operation and profitability. In fact, change in any of three factors examined--arrival intensity, time of unassisted search, and transition probabilities--can have a substantial effect on profitability.

While the solution of the model is required to test alternative strategies in advance of actual implementation, the simulation experiments conducted suggest that management should be able to estimate closely the extent to which staffing in an ongoing system approaches optimality for the set of circumstances in existence. This may be done readily with the aid of either of two measures--capacity utilization or the maximum expected number of services in a congestion-free system.

An interesting extension of the model is to use the simulation outputs in a Bayesian decision analysis. Using this approach, expected outputs generated by simulation are treated as conditional

values associated with different sets of circumstances resulting from alternative combinations of marketing strategies. Then the probabilities of the circumstances occuring with various strategies are used to determine the best combination of strategies.

REFERENCES

[1] VonRiesen, R. Dale, "Toward Staffing Optimality in Retail Selling," Journal of Retailing 49 (Winter, 1973-74), pp. 37-47, 95.

[2] VonRiesen, R. Dale, "Some Findings on the Cost of Congestion in a Retail Application of Waiting Line Analysis," Barnes, Jim D., and Heflin, Thomas L., Proceedings and Abstracts, American Institute for Decision Sciences Sixth Annual Meeting, Western Regional Conference (Phoenix: AIDS, 1977) pp. 270-2.