USE OF BOTH OPTIMIZATION AND SIMULATION MODELS TO ANALYZE COMPLEX SYSTEMS

Dean H. Kropp
Robert C. Carlson
James V. Jucker

ABSTRACT

Recent work dealing with planning of complex facilities has proposed use of a recursive optimization-simulation approach. This technique takes advantage of the best features of both methods while minimizing the disadvantages of each method used alone. Here a mixed integer program generates staffing and facility plans and a simulation model evaluates their day-to-day acceptability. Then a linear regression uses the simulation results to generate non-cost constraints to be added to the optimization model. Results from a hypothetical health care setting demonstrate the value of the recursive method.

1. INTRODUCTION

Researchers modeling complex systems have often rejected optimizing methods because of their lack of realism (such as linearity restrictions) or because the desired "richness of detail" is not available. In such cases, they have used computer simulation because it can handle complex relationships and can provide sufficient detail. However, the costs of simulation are very high. In addition, because simulation is not an optimizing method, the investigators have discovered that they will have to examine many alternatives without any assurance that they are approaching an optimal solution.

An alternative approach to modeling with only an optimizing method or a computer simulation alone uses a recursive optimization-simulation algorithm, such as that described by Nolan and Sovereign [5] or used in modified form by Kolesar et al. [2] With such a method, an optimization technique is used to analyze the system at an aggregate level. The results are then used in a simulation model of the same system, which identifies other information, which can be passed back to the optimization model. The reason for employing both an optimization model and a simulation model is that optimization methods can answer most questions asked by the analyst, whereas simulation cannot. Simulation is largely a descriptive tool. Unless a complete experimental design is used, simulation can answer little more than questions concerning the feasibility of a given alternative.

In the recursive method, the optimization model permits the analyst to reduce the number of variables (and the resultant number of alternatives) to be examined by simulation. The simulation model is then used to handle the system's complex relationships (such as nonlinearities, queueing, and discreteness), which are too cumbersome for the optimization model. Thus, the recursive method attempts to capitalize on the advantages of both approaches while attempting to reduce the disadvantages of either method used by itself.

This work describes the detailed methodology that can be used in the recursive method to link the optimization and simulation models with a system performance constraint. It is shown that the method can be thought of as a two-stage aggregate - disaggregate process. Finally, the paper presents the detailed results of four case studies concerning a hypothetical health care outpatient clinic to demonstrate the value of the recursive method. The method helps identify the least-cost configuration of an outpatient clinic which could not be found by a linear programming method used alone; it also required fewer simulation runs than would have been required by a simulation method used by itself.

2. OPTIMIZATION MODEL

The optimization model used to examine the health care setting is a mixed integer program based on the work of Schneider [6] and of Schneider and Kilpatrick [7]. Given a fixed capitation rate (revenue per person per period of time) and a specified subscriber base, the model's objective is to find the personnel to be hired, the services to be offered, the delegation of services from the physicians to physician's assistants, and the facilities required to minimize total annual cost. The constraints on the problem include meeting annual demand, meeting budget and capital limits, meeting limits on numbers of personnel and on their use, and ensuring that a reasonable number of examination rooms is provided. The mathematical formulation of the problem is presented in reference [4]. The problem is solved using the IBM MPS linear programming computer package and a FORTRAN algorithm for solving mixed integer programs.

Most analytical models of health care have been used for long-range studies. The mixed integer program discussed here is no exception, with the

basic unit of time of one year. Critical variables such as the number and types of personnel assigned to services and the number and types of personnel employed have units of person-years. Clearly, the model is not directed at predicting day-to-day performance. It assumes that the utilization of medical services occurs at a constant level and ignores the variation in time for personnel to perform medical services.

In spite of these limitations, the mixed integer program has been found to be of sufficient accuracy to be used in aggregate planning for health care organizations. In particular, Schneider [6] found that the model closely approximated the operation of a prepaid group practice in central Florida, and that many of the model's results were parallel to those of the clinics of the Kaiser system of hospitals in southern California. Since 1973 the models have been used to plan health care organizations in at least five other areas. For more information about the model see [3].

3. COMPUTER SIMULATION MODEL

The computer simulation model used to evaluate the setting on a day-to-day basis has three major sections - one determining the patient characteristics, one describing the diagnosis and treatment process, and one determining the office and provider availabilities.

PATIENT CHARACTERISTICS

Patient input consists of specifying the patient mix, interarrival times, "no-show" rates, emergency visit rates, and "preference for provider" rates. All of these characteristics may be stochastic in nature. The patient characteristics section can be run independently of the other model sections so that the same patient input stream can be passed through alternative system configurations. This technique ensures that any differences between the results of different configurations are caused only by the configurations themselves.

DIAGNOSIS AND TREATMENT PROCESS

The logical network of the diagnosis and treatment process controls the flow of patients through the facility and specifies the type and amount of resources required at each step in the process. Input parameters determine personnel staffing patterns and working relationships, capabilities of the physician's assistants, number of patient examination rooms, physical configuration of the facility, and patient management decision rules.

A patient is randomly assigned a reason for visit based on the frequency distribution of visits used as input data for the mixed integer program. Depending on the branch probabilities assigned for the visit type, laboratory tests, x-ray examinations, and other procedures may be ordered. Each procedure has associated with it a probability distribution for the procedure time. The choice of the person to perform a procedure depends on the optimal task assignments identified by the mixed integer program and on the dynamic state of the system (e.g., whether the lowest level person capable of performing the task is currently available).

OFFICE AND PROVIDER AVAILABILITY

This section regulates the functioning of the office with respect to its normal opening time, lunch and coffee breaks, and closure at the end of the day. It also provides for provider lateness or absence, telephone calls for the providers, and patient record-keeping.

The simulation model uses the IBM GPSS/360 simulation programming language. For more information about the model see [3].

4. INTEGRATION OF OPTIMIZATION AND SIMULATION MODELS INTO THE RECURSIVE APPROACH

In the recursive method the optimization model first identifies optimal facility characteristics (such as the number of examination rooms) and staffing patterns (including both the numbers of each type of personnel and the assignments of personnel to tasks) for the aggregate problem. Then the simulation model is used to evaluate the day-to-day acceptability of this optimal aggregate solution. In this application, for instance, the simulation model views patient waiting time as the day-to-day measure of acceptability.

Based on previous work [1], one can expect the optimization model to produce a solution which is not acceptable on a day-to-day basis. The patient waiting times will likely be excessive because the simulation considers the variability of actual task times around their mean values used in the aggregate model. In such an event, other possible solutions will have to be examined and the recursive method used in an iterative fashion until an acceptable solution is obtained. Such use of the recursive method requires a transfer of information from the simulation model to the optimization model.

Previous work [4] implied that search techniques could be used in the recursive method to find the optimal solutions. However this proved to be inefficient. To overcome this problem, a method was developed for identifying relationships between the time-based performance measures determined by the simulation (such as patient waiting time) and variables of the simulation which also appear in the optimization model (such as number of physicians). These relationships will be determined by a linear regression performed on the results of a number of simulation runs involving replications of several settings of the simulation input parameters. These linear relationships will then be incorporated as constraints in the optimization model to reflect a

non-cost objective of the facility. An example of such a constraint reflecting the facility's objective to keep the average patient waiting time below 35 minutes is:

$$\text{Constant} - 5 \times \text{Number of Physicians} - 4 \times \text{Number of Physician's Assistants} - 10 \times \text{Number of Examination Rooms} \quad (1)$$

$$\leq \text{35 minute average patient waiting time}$$

In this constraint the minus signs indicate that additional amounts of resources could be expected to reduce waiting time. The coefficients in (1) were determined by the simulation-regression procedure.

This constraint is required to take a form similar to that depicted in (1). Other functional forms, such as those involving powers of the clinic's resources or cross-products of the resources, may also provide accurate representations of the relationship between the time-based measures and the clinic resources. However, such functional forms could not be used within the linear framework of the optimization model.

The linear function produced by the regression will provide a valid estimate of the dependent variable only for the range of independent variables examined in the simulation (such as from 5 to 15 physicians, and 30 to 60 examination rooms). It is reasonable to require that the optimal solution to the revised optimization model (including the new constraint) fall within this range. This requirement results in additional upper and lower bound constraints for the independent variables in the optimization model.

Further, the settings of the independent variables examined in the simulation must be such that the valid range is expected to contain the optimal solution; the waiting times determined in the simulation must bracket the desired waiting time. Otherwise it would not be possible for the recursive method to achieve its objective; as a result the feasible region for the optimization problem would be the null set. Since it is not possible a priori to know which simulation configurations will produce the desired waiting time, this valid range of variables must initially be very large. Thus, the recursive process is based on interpolating between points within the valid range (not extrapolating outside the range) to achieve the desired waiting time.

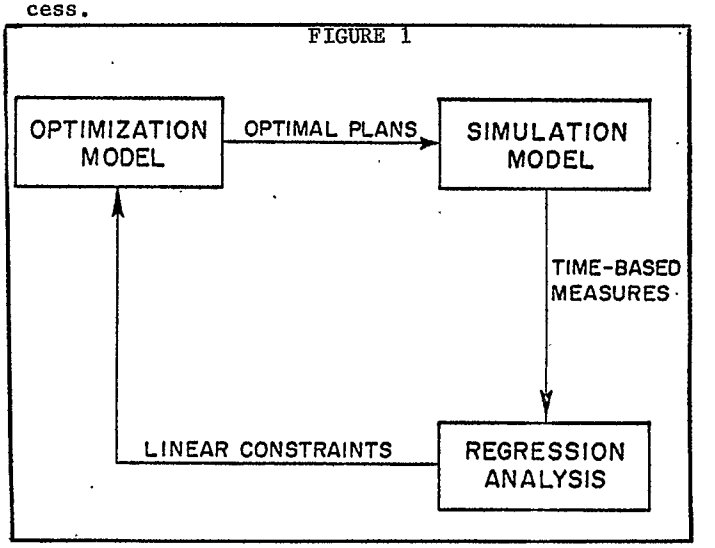
These requirements for the simulation suggest an experimental design in which each of the N independent variables has three settings: low, intermediate, and high. This experimental design could be a 2^N factorial design with a center point.

Based on the above discussion, one can use the recursive method as follows:

1. Identify resources to be used.
2. Given estimated resource productivities and other assumed parameters, solve the optimization model.
3. Run the simulation model using the optimal plan generated in Step 2.
4. Identify a range of values of the variables which is expected to produce values of the time-based performance measure of interest which bracket the desired value.
5. Run the simulation model using the range of values of variables identified in Step 4. A suggested experimental design is a 2^N factorial design with center point. Fractional factorial designs may also be used.
6. Perform a linear regression on the simulation results of Step 5 to express the time-based performance measure of interest as a linear function of variables already present in the optimization model.
7. Incorporate the linear function determined in Step 6 into the optimization model as a constraint reflecting the objective of achieving the desired value of the time-based performance measure. Also incorporate in the optimization model upper and lower bound constraints reflecting the range of variables examined in Step 5.
8. Return to Step 2.
9. Repeat the process until the time-based performance measure of interest determined in Step 3 achieves its desired value.

Figure 1 depicts the recursive nature of this process.

FIGURE 1



5. TWO-STAGE NATURE OF THE RECURSIVE APPROACH

Once the mixed integer program has been solved to determine the least-cost personnel assignments, the assignments can be fixed; they are no longer variables in the problem. This change will cause many

variables in both the objective function and the constraints to become constants. Further, the lower bounds for the independent variables examined in the simulation runs will be at least as great as the mixed integer program's optimal values for the variables. As a result, many of the original constraints in the mixed integer program, such as that of meeting expected patient demand, will automatically be satisfied in the new program which includes the waiting time constraint.

Thus the optimization problem including the waiting time constraint may have many fewer variables and constraints than the original optimization problem. The most simplified version of such a problem would occur for a clinic in the planning stage having no budget constraint, no capital constraint, and no upper bound constraint on physician supervision time.

It is clear, then, that the recursive method can be thought of as involving two levels of problems. The first stage problem involves use of the mixed integer program by itself to achieve an aggregate solution to the clinic's planning problem. On the second level, the optimization problem involving use of the waiting time constraint becomes that of "disaggregating" the aggregate solution. Given that the original aggregate constraints are met, the goal of this second stage problem becomes that of determining the additional resources to be added to the facility to meet the waiting time constraint at minimum cost.

6. PHILOSOPHY OF THE RECURSIVE APPROACH

One of the major premises of the recursive approach is that the relationship between the dependent, time-based variable and the independent variables of the system is nonlinear. For example, one might reasonably expect the system to exhibit diminishing marginal returns: each additional unit of an independent variable might produce successively smaller changes in the dependent variable. The use of different linear relations for each iteration recognizes the potential existence of these nonlinearities. Thus, the method makes successive linear approximations to the assumed nonlinear function.

The recursive method requires use of a linear relation between the dependent variable and independent variables so that the linear form of the optimization model can be retained. Constraints having cross-product or polynomial forms would require more complex solution methodology. One additional justification for the linear restriction is that it is extremely efficient. Even ignoring the difficulties of using a nonlinear constraint in an otherwise linear model, estimation of the coefficients for the cross-product and polynomial terms would require more simulation runs than those required for estimating the coefficients of linear terms. It must be remembered that one of the goals of the recursive method is reducing the total number of simulation runs required. In other words, the method is concerned with achieving a solution meeting both the facility's long-and short-range objectives quickly. The assumed linear functional form meets this need and also allows retention of a simple solution methodology. Since the method relies on use of the successive linear approximations to the nonlinear function of interest, it is difficult, if not impossible, to prove theoretically that the method converges to a solution. This apparent weakness is not a problem. The method depends heavily on human input in determining the simulation experimental designs, in evaluating the regression results, and in interpreting the successive solutions to the optimization problem. This dependence assures a form of practical convergence for the method. Given the wealth of information provided by the method, the analyst can take corrective action quickly (such as by modifying the simulation experimental designs) if the method's results do not exhibit convergence at any point in the process.

Thus, the recursive method takes a pragmatic approach to the postulated problem of disaggregating a facility's aggregate plan. While it does not ignore theory, the method focuses more on the practical problem of meeting the facility's objectives in an efficient manner.

7. EXAMPLE OF THE USE OF THE RECURSIVE METHOD

The recursive optimization-simulation approach was applied to a hypothetical setting - an outpatient clinic providing adult medical care consisting of about 50,000 patient visits per year. The clinic has a staff consisting of physicians, physician's assistants, registered nurses, and licensed practical nurses. In this clinic, each provider spends no more that 35 hours per week in direct contact with patients. Patient services can be performed by alternative sets of personnel technologies that involve one or more providers. For example, a physical examination may be performed by the physician alone or by the physician with the help of one of the support personnel.

The mixed integer program is formulated to minimize the facility's total cost for a specified number of subscribers. Necessary additional inputs to the model include the types of services available, the distribution of service requirements, the alternative personnel technologies, the average time required for each service, the salary of each type of provider, the annual number of visits per subscriber, and the amount of equipment and floor space required per provider. In addition, the model places realistic lower and upper bounds on the number of providers and the number of examination rooms. The number of providers and examination rooms and the personnel technologies that will minimize total annual cost are obtained as the solution of the model.

The simulation model includes more detailed information, including daily operating hours, schedules for patient appointments, varied arrivals of walk-in and nonemergency patients, variability in patient arrivals, required services and service times, and sequencing of tasks performed to meet patient

service requirements. The practice is open five days a week, with regular hours from 8:00 AM to 5:00 PM, a lunch break from 12:00 noon to 1:00 PM, and ten minute coffee breaks at 10:30 AM and 3:00 PM. Appointment patients are scheduled at 15 minute intervals from 9:00 AM to 12:00 noon and from 1:00 PM to 4:00 PM. Although equal numbers of patients are scheduled for each of the appointment intervals, approximately 10% of them are "no-shows". The actual arrival time of these appointment patients is considered to be normally distributed, with the average arrival time being 2.5 minutes early. In addition, walk-in patients and acutely ill near-emergency patients come to the clinic, with their mean interarrival times depending on the time of day.

Patient service requirements are randomly determined based on the frequency distribution of the different services. Once the patient's medical needs are known, the optimal personnel technology identified by the mixed integer program is used to identify which provider or providers the patient will see. Service time is considered to be gamma distributed, with the mean time equal to the time requirements used in the mixed integer program. For services associated with the assessment of the patient's condition, the tasks are performed serially. For services associated with a treatment, the providers must see the patient concurrently.

Although the example was analyzed to demonstrate the use of the recursive method, another important goal of the work was to investigate the use of physician's assistants in the hypothetical setting. Accordingly, the example included four case studies, each having the same patient requirements and clinic financial structure. The only difference between the case studies resulted from use of different skill levels of the physician's assistants: no physician's assistants, low-skill physician's assistants, intermediate-skill physician's assistants, and high-skill physician's assistants. These skill levels of the physician's assistants reflected different levels of training and responsibility, and hence different salaries.

For the case studies, the operational goal was to achieve an average waiting time per patient of 35 minutes, approximately half of the total time the patient is busy. The average waiting time per patient was determined from simulation of five days of operation of the clinic. Further, in the linear regression program used in the recursive method, the minimum acceptable level of significance for both the regression equation and each of its coefficients was set at 0.05. The results of the case studies are identified in Table 1.

TABLE 1

PHYSICIAN'S ASSISTANT SKILLS	ANNUAL COST	AVERAGE PATIENT WAITING TIME	MDs	PAs	RNs	LPNs	ERs
INITIAL SOLUTION							
No PAs	$1,507,326	263.1 Min.	12	0	8	2	12
Low	1,453,366	253.3 Min.	9	6	6	1	15
Medium	1,382,275 (1,316,275)	246.8 Min.	4	11	6	2	15
High	1,405,149 (1,301,649)	246.9 Min.	4	9	8	2	13
FINAL SOLUTION							
No PAs	$3,237,602	33.9 Min.	40	0	30	7	55
Low	2,486,365	34.0 Min.	23	6	25	9	50
Medium	2,707,336 (2,539,336)	35.2 Min.	15	28	28	6	54
High	2,813,659 (2,468,659)	33.5 Min.	13	30	23	9	53

Note: Costs in parentheses are costs of solutions when all physician's assistants' salaries are kept constant at $13,500, regardless of their skill level.

MDs: Physicians
PAs: Physician's Assistants
RNs: Registered Nurses
LPNs: Licensed Practical Nurses
ERs: Examination Rooms

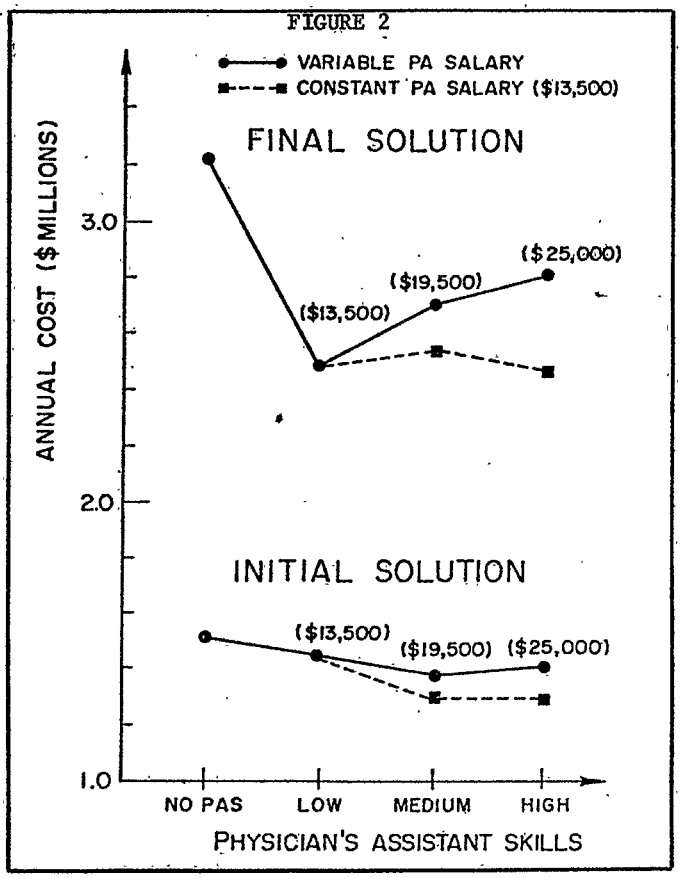
8. Discussion

Although the studies assumed that physician's salaries were proportional to their skills, it was also possible to determine the costs of the solutions under the assumption that the salaries remained constant at $13,500. Table 1 identifies both sets of costs.

The average four-hour waiting times of the initial solutions clearly reveal the day-to-day problems that would result from blind implementation of the optimal aggregate solutions. Because of such waiting times, it is evident that the clinic would not be attractive either to patients or to providers. In addition to not meeting the facility's own criterion, the excessive patient waiting times would affect the physicians' availability for patient care activities outside the clinic, as well as the morale of the clinic staff. Further, the patient waiting time and resulting congestion would likely cause patients to go to other practices involving less waiting.

One of the goals of the example was to investigate the use of physician's assistants in the hypothetical facility. Figure 2 provides a graphical representation of the results of such an investigation.

FIGURE 2



If the initial aggregate solution were used as a guide to decide whether physician's assistants should be used and, if so, what their skills should be, the minimum cost recommendation would be to use physician's assistants having either medium or high skills. The final solution indicates the weakness of such a recommendation. It is reasonable to expect that physician's assistants' salaries would be commensurate with their skills. The figure indicates that this salary would have to be only slightly higher than $13,500 (actually $14,000) for the high-skill physician's assistants to be no longer the least costly. Further, if their salary were $25,000, the use of low-skill physician's assistants would be cheaper by over $350,000 per year! Thus, the recursive method as used here has demonstrated the potential failings of an aggregate method even in deciding which skill levels of physician's assistants should be used.

The recursive method is a heuristic method. Nevertheless, its results across the four case studies exhibit a great deal of consistency. The numbers of registered nurses (23 to 30), licensed practical nurses (6 to 9), and examination rooms (50 to 55) do not vary considerably between case studies. Further, the total numbers of support personnel (32 to 37) and primary personnel (40 to 43, considering physician's assistants only when they can lead teams) remain relatively constant.

In addition to providing consistent results, the recursive method also has given results that are intuitively appealing. When physician's assistants' skills are increased, the results indicate that physicians will be increasingly replaced by physician's assistants. When physician's assistants are not used (and their beneficial effect of reducing physician's task times is eliminated) the number of physicians must be increased significantly from 23 to 40. In addition, as was postulated previously, the nonlinear relationship between the dependent variable (waiting time) and the independent variables (the clinic resources) has exhibited diminishing marginal returns. For example, when no physician's assistants are used, the waiting time reduction for each additional physician decreased from 3.49 minutes to 1.66 minutes as the average patient waiting times approached the goal of 35 minutes. Similar reductions for examination rooms (2.52 minutes to 0.33 minutes) and registered nurses (7.82 minutes to 0.95 minutes) can also be noted.

Based on the experience of these four case studies, other comments are also in order. In each of the studies the only independent variables considered were the four types of providers and the examination rooms. In other settings other independent variables might well be examined, such as the level of annual patient demand or the amounts of other clinic resources such as x-ray machines. Indeed, in the hypothetical setting the single x-ray machine eventually proved to be the largest single bottleneck. Additional x-ray machines could well have provided an inexpensive way of further reducing waiting times, and the experimental designs could have included settings for the number of x-ray machines. Of course, the addition of other variables would significantly increase the number of simulations and the cost of the analysis.

The use of only one short-term performance measure, patient waiting time, forced low waiting times and low resource utilizations and high costs. In real world clinics it may be necessary to recognize the tradeoff between waiting times and resource utilizations by using both considerations as short-term performance measures. It is possible to perform linear regressions to develop one constraint for each measure; however, the mixed integer program might have to be replaced by a multicriteria optimization algorithm because the two constraints could conflict to produce a null feasible region.

9. Conclusion

The primary objective of this paper has been to discuss how planning of health care facilities can be improved through use of a recursive modeling approach. The example was used only to demonstrate this point. It is important, therefore, to recognize that the medical practice analyzed in the example is hypothetical and based on data from a variety of sources. Parameter values were chosen to be representative of a typical practice, not to dramatize the limitations of aggregate planning techniques. Similarly, the reader is cautioned not to draw general conclusions based on the example. The results for any given facility will depend on the specific characteristics of the facil-

ity.

Nevertheless, it is clear that this new recursive approach can be of great value to health care planners because it focuses on the entire time horizon faced by management and because it can overcome many of the disadvantages of either an optimization model or a simulation model used alone.

REFERENCES

1. Hershey, J. C., D. H. Kropp and I. M. Kuhn. "Physician's Assistants in Ambulatory Health Care Settings: Need for Improved Analysis," Research Paper Series, Health Services Administration, Stanford University School of Medicine (February 1976).
2. Kolesar, P. J., K. L. Rider, T. B. Crabill, and W. E. Walker. "A Queueing - Linear Programming Approach to Scheduling Police Patrol Cars," Operations Research, Volume 23, No. 6, (November/December 1975), pp. 1045-1062.
3. Kropp, D. H. "Recursive Modeling of Ambulatory Health Care Settings," Unpublished Ph.D. Dissertation, Stanford University (1977).
4. Kropp, D. H. and R. C. Carlson. "Recursive Modeling of Outpatient Health Care Settings", Journal of Medical Systems, Vol. 1, No. 2, 1977, pp. 123-135.
5. Nolan, R. L. and M. G. Sovereign, "A Recursive Optimization and Simulation Approach to Analysis With an Application to Transportation Systems," Management Science, Vol. 18, No. 12, pp. B676-B690, August 1972.
6. Schneider, D. P. "A Systems Analysis of Optimal Manpower Utilization in Health Maintenance Organizations," Unpublished Ph.D. Dissertation, University of Florida (1973).
7. Schneider, D. P. and K. E. Kilpatrick. "An Optimum Manpower Utilization Model for Health Maintenance Organizations", Operations Research, Vol. 23, No. 5 (September/October 1975) pp. 869-889.