TIME-ORIENTED MODELS AND THE LEGAL PROCESS: REDUCING DELAY AND FORECASTING THE FUTURE*

STUART S. NAGEL**
MARIAN NEEF***

CONTENTS

---

* This article completes the second part of a series of law review articles and books dealing with applications of operations research, management science, and related methods to improving the legal process. Other items in the series include S. NAGEL & M. NEEF, THE LEGAL PROCESS: MODELING THE SYSTEM (1977); S. NAGEL & M. NEEF, OPERATIONS RESEARCH METHODS: AS APPLIED TO POLITICAL SCIENCE AND THE LEGAL PROCESS (1976); Nagel & Neef, *Legal Policy Optimizing Models*, 29 J. LEGAL EDUC. 31 (1977).

The authors would like to thank Eli Noam of the Columbia University Law and Economics Program and Nancy Munshaw of the University of Illinois Department of Urban and Regional Planning for their helpful suggestions. Thanks are also owed to the Ford Foundation Public Policy Committee, the Illinois Law Enforcement Commission, and the University of Illinois Research Board for financing various aspects of the legal policy research. The ideas advanced in this article do not necessarily represent the views of these organizations.

** Professor of Political Science, University of Illinois. B.S., 1957, J.D., 1958, Ph.D., 1961, Northwestern University. Member, Illinois Bar.

*** Ph.D. Candidate, Department of Political Science, University of Illinois. B.A., 1973, M.A., 1974, University of Illinois.

Processing delays are a serious problem in both civil and criminal cases. Civil cases often take many years to process from the time the litigation is filed until the case goes to trial. Delay injures plaintiffs who must wait to be compensated for their injuries. It also adversely affects the judicial system because of the increased likelihood that witnesses will forget, disappear, or die before a case reaches trial. Although criminal cases generally do not consume as much time as civil cases, the injury caused by delay may be even greater. If the defendant is jailed while awaiting trial, he may be confined for a longer period than the maximum sentence he could have received if convicted. If the defendant is not jailed while awaiting trial, long delay may increase the likelihood that he will commit further crimes or fail to appear at his trial.[1]

A related legal process problem is the lack of planning for the effects of changes in the legal system. These changes include, but are not limited to, increases or decreases in certain kinds of cases, resources available to the courts, discretion in legal decisionmaking, severity in sentences or damages, and concepts of due process. Lack of planning often means increased delay, undesirable outcomes, and disruption of individual rights. The legal system can avoid these results if legal process planners had models like those that are helpful in planning business activities, government budgets, municipal facilities, and adaptations to technological change.[2]

This article describes a series of models for analyzing the legal proc-

1. For thorough analyses of the problem of delay in the legal process, see H. JAMES, CRISIS IN THE COURTS (rev. ed. 1971); H. ZEISEL, H. KALVEN, JR., & B. BUCHHOLZ, DELAY IN THE COURT (1959) [hereinafter cited as H. ZEISEL]; WALTER E. MEYER RESEARCH INSTITUTE OF LAW, DOLLARS, DELAY AND THE AUTOMOBILE VICTIM (1968); THE COURTS, THE PUBLIC, AND THE LAW EXPLOSION (H. Jones ed. 1965); SELECTED READINGS: COURT CONGESTION AND DELAY (G. Winters ed. 1971).

2. For thorough analyses of the problem of lack of planning in the legal process, see THE COUNCIL OF STATE GOVERNMENTS, JUDICIAL PLANNING IN THE STATES (1976); D. GIBBONS, J. THIMM, F. YOSPE, & G. BLAKE, CRIMINAL JUSTICE PLANNING (1977); D. GLASER, STRATEGIC CRIMINAL JUSTICE PLANNING (1975); NATIONAL CENTER FOR STATE COURTS, PLANNING IN STATE COURTS: A SURVEY OF THE STATE OF THE ART (1976); ISSUES IN CRIMINAL JUSTICE: PLANNING AND EVALUATION (M. Riedel & D. Chappel eds. 1976).

ess that emphasize either time-saving or the prediction of future events from past events. A model or analytic perspective describes a set of concepts, methods, and basic principles to deduce prescriptive statements of what should be done (prescriptive or optimizing models) or to describe or predict future consequences (descriptive models). The prescriptive models primarily seek to reduce time consumed in the legal process; the descriptive or predictive models are primarily concerned with predicting future legal process events from prior events.[3] Further, optimizing models may be valuable predictive tools, and predictive models may be valuable in making time-saving recommendations. The article is divided into two parts with the first emphasizing prescriptive models and the second emphasizing descriptive models. The prescriptive models include: (1) queueing theory, which emphasizes time reduction by reducing arrivals, increasing processors, and shortening processing; (2) dynamic programming or optimum sequencing, which emphasizes time reduction through more efficient ordering of cases; (3) PERT or critical path analysis, which emphasizes time reduction through concentration on those events that are responsible for delaying the occurrence of subsequent events; and (4) optimum level analysis, which attempts to find an optimum level of effort or expenditures to minimize the sum of the delay costs and the speed-up costs. The descriptive models include: (1) Markov chain analysis, which emphasizes

---

3. On prescriptive and descriptive modeling in management science and operations research, see D. ANDERSON, D. SWEENEY, & T. WILLIAMS, AN INTRODUCTION TO MANAGEMENT SCIENCE: QUANTITATIVE APPROACHES TO DECISION MAKING (1976); S. RICHMOND, OPERATIONS RESEARCH FOR MANAGEMENT DECISIONS (1968); H. TAHA, OPERATIONS RESEARCH: AN INTRODUCTION (2d ed. 1976); R. THIERAUF, DECISION MAKING THROUGH OPERATIONS RESEARCH (1970); H. WAGNER, PRINCIPLES OF OPERATIONS RESEARCH WITH APPLICATIONS TO MANAGERIAL DECISIONS (2d ed. 1975). Management science and operations research develop methods to determine policies for maximizing goals under varying conditions.

On modeling, particularly as applied to governmental and legal problems, see E. BELTRAMI, MODELS FOR PUBLIC SYSTEMS ANALYSIS (1977); J. BYRD, OPERATIONS RESEARCH MODELS FOR PUBLIC ADMINISTRATION (1975); M. GREENBERGER, M. CRENSON, & B. CRISSEY, MODELS IN THE POLICY PROCESS (1976); W. HELLY, URBAN SYSTEMS MODELS (1975); S. NAGEL & M. NEEF, THE LEGAL PROCESS: MODELING THE SYSTEM (1977); M. WHITE, M. RADNOR, & D. TANSIK, MANAGEMENT AND POLICY SCIENCE IN AMERICAN GOVERNMENT (1975); A GUIDE TO MODELS IN GOVERNMENTAL PLANNING AND OPERATIONS (S. Gass & R. Sisson eds. 1975); SYSTEMS ANALYSIS FOR SOCIAL PROBLEMS (A. Blumstein, M. Kamrass, & A. Weiss eds. 1970). Management science models as applied to the legal process have dealt with delay reduction, but those applications have emphasized flow chart models and queueing theory rather than the fuller range of potentially applicable models.

For a discussion of the elementary mathematical aspects of modeling, see M. BRENNAN, PREFACE TO ECONOMETRICS (1973); C. DINWIDDY, ELEMENTARY MATHEMATICS FOR ECONOMISTS (1968).

the prediction of subsequent events through knowledge of the probability of one event leading to another event; (2) time series analysis, which describes the inductive prediction of variables or legal process characteristics at a future time from variables at a prior time; and (3) difference equations, which deductively predict an event at a future time from an event at a prior time. These models are clarified throughout by legal process illustrations. No methodological knowledge beyond simple arithmetic or high school algebra is presupposed. The only presupposition is that the reader is interested in predicting key events and reducing delay in the legal system.

I. PRESCRIPTIVE OR OPTIMIZING MODELS

The following prescriptive or optimizing models are primarily concerned with saving time. In that sense they have a normative, prescriptive, or optimizing goal. They also have descriptive or predictive elements in the sense that they often attempt to describe or predict the time, if any, that can be saved by alternative procedures.

A. *Queueing Theory*

1. *The Basic Model*

Queueing theory uses a set of mathematical models or formulae which take as their main inputs the number of cases arriving in a system per day and the number of days or other time units needed to process each case. From these inputs and the resulting models and formulae, one can deduce such predictive outputs as the average time spent in the system and such prescriptive outputs as the methods to reduce that average time.[4] For example, if we are concerned with the legal process from arrest through arraignment in misdemeanor cases in middle-sized cities, we might collect data for a sample of ten separate working days. As a result of arrests on four of those days, ten cases per day arrived in the system, and on the other six days, twenty cases per day arrived in the system. Thus on the average day, sixteen cases arrived in the system ($16 = [(4)(10) + (6)(20)] \div 10$). This figure represents the first key input: the arrival rate (A).

Within the same period, eighteen arrestees were given arraignment

4. For a discussion of queueing theory, see J. BYRD, *supra* note 3, at 198-208; D. GROSS & C. HARRIS, FUNDAMENTALS OF QUEUEING THEORY (1974); A. LEE, APPLIED QUEUEING THEORY (1966); S. RICHMOND, *supra* note 3, at 405-38; T. SAATY, ELEMENTS OF QUEUEING THEORY (1961).

hearings on each of seven days while twelve cases were similarly serviced on each of three days. Thus on the average day, the system serviced 16.2 cases (16.2 = [(7)(18) + (3)(12)] ÷ 10). This figure represents the service rate (S), which is the second key input. The service rate (S) should be distinguished from the service time (T_S), which in this context represents the average amount of time needed for an arraignment hearing—about 30 minutes. The two figures are related, however, because a shorter service time will usually result in a higher service rate. Although a sample of only ten days is probably too small to calculate an accurate average arrival or service rate, we will assume for the purposes of illustration that those figures ($A = 16$, $S = 16.2$) represent the true averages. From these figures we can deduce or predict an expected or likely amount of time an average case will spend in the system. Time in this context includes both waiting time (*i.e.*, time spent awaiting the arraignment hearing) and servicing time (*i.e.*, time actually consumed by the arraignment hearing). The formula for calculating the average time spent in the system is $T = 1/(S\text{-}A)$. In the preceding hypothetical, the average time would be five working days. ($T = 1 \div (16.2\text{-}16) = 1 \div .2 = 5$). This formula operates under assumptions that have been repeatedly validated in case processing based on the distribution around the arrival and service rates. Knowledge of the mathematics supporting these assumptions is not necessary to make use of this and other queueing formulae. Intuitively, queueing formulae seem logical. For example, if 16.2 cases are processed on an average day, the total time formula indicates that an average case takes 1/16.2 days to process (or six percent of an eight-hour day, which is about thirty minutes). The formula also confirms the intuitive assumption that if the arrival rate (A) is equal to or greater than the service rate (S), an infinitely long backlog would develop and new cases would not be serviced. This phenomenon is reflected in the denominator of the preceding formula.

2. *Variations and Implications*

Queueing theory also offers a formula, $T_W = T(A/S)$, which determines the average waiting time before service begins. This formula indicates that waiting time equals total time multiplied by the arrival/service ratio.[5] Given our hypothetical data for A and S and the

5. Queueing theory considers time consumed to be a function or effect of the arrival rate and the service rate. In reality, however, there may be dynamic reciprocal causation; if the arrival rate declines, time consumed will decline. But if time consumed declines, it may cause more

previously calculated T, waiting time is 4.9 working days ($4.9 = (5)(16)/16.2$). Because total time equals waiting time plus servicing time ($T = T_W + T_S$), servicing time can be predicted by subtracting 4.9 from 5. This computation reveals that servicing time is approximately one percent of a working day.

In addition to time consumption formulae, queueing theorists have developed formulae dealing with backlog. For example, the formula used to predict the number of cases backlogged in the system is $N = (A/S)/(1\text{-}A/S)$. The size of the backlog varies directly with the arrival/service ratio and inversely with the complement of that ratio. In our hypothetical system, we would expect eighty cases to be backed up ($80 = 16/16.2)/(1\text{-}16/16.2)$. This means that on an average day there are eighty arrested defendants who have not yet been arraigned. Of these, one defendant is in the process of being arraigned or serviced while the remaining seventy-nine are in the waiting line. In other words, the total backlog (N) equals the backlog being serviced (N_S) plus the backlog awaiting servicing (N_W).[6]

Other queueing models have been developed in which the outputs are in the form of probabilities. For example, one formula uses the average arrival rate to calculate the probability that a specific number of cases will arrive in the system on a particular day. Another formula uses the average service rate to calculate the probability of servicing a specific number of cases on a given day. Still another formula utilizes both A and S as inputs to yield outputs showing the probable number

---

people to enter the line, thereby causing the arrival rate to increase at least marginally. If, however, the arrival rate then increases, this change will cause time consumed to increase, which in turn is likely to decrease the willingness of people to enter the line (*i.e.*, decrease the arrival rate) and thereby decrease the time consumed. If the arrival rate has a positive effect on time consumed equal to the negative effect that time consumed has on the arrival rate, they would offset each other and the arrival rate and time consumed would tend to remain in equilibrium. The extent that the two relations approximate each other is, however, an empirical question.

A dynamic reciprocal causation may exist between the service rate and time consumed: when the service rate increases, time consumed declines. But this occurrence may cause the processors to relax their efforts, which may cause the service rate to decline again, offsetting the previous reduction in time consumed. If time consumption thereby increases as a result of the increase in the service rate, this change may cause the processors to accelerate their efforts, which means the service rate increases, the time consumed thus increases, and the cycle continues. Another empirical question is the extent that the inverse effect of the service rate on time consumed approximates the strength of the positive effect of time consumed on the service rate.

6. In the same way that T_w can be calculated with the formula $T_w = T(A/S)$, N_w can be calculated with the formula $N_w = N(A/S)$. The latter formula also yields the same result ($79 = 80\ (16/16.2)$).

of cases backlogged in the system.[7]

The preceding models have assumed that the city has only one arraignment court. The basic formulae, however, can be modified for any number of arraignment courts operating simultaneously or for unusual distributions around the average arrival or service rates.[8] They also can be modified to consider: (1) queue discipline, or rules providing priority servicing of certain kinds of cases; (2) jockeying, or procedures that allow lawyers to choose the court or judge that will hear their cases; (3) truncating arrivals, or rules that specify the maximum number of cases or kinds of cases that can be processed; and (4) multiple stages of processing, *i.e.*, waiting and processing in the preliminary hearing, arraignment, and trial stages.[9]

Queueing formulae are also useful for making reasonably accurate estimates of reductions in time and backlog that would result from changes in the average arrival rate, the average service rate, the number of courts, the system of priorities, and other queueing variables. For example, we can reduce the initial arrival rate by having the police resolve more complaints without making arrests. Further, we can reduce the arrival rate at the processing or servicing stage by encouraging more settlements between arrest and arraignment. We can reduce the service time and thus increase the service rate by having arraignment hearings follow a more standardized script that avoids unnecessary matters.[10] In that regard, if an average two-day trial can be reduced to

---

7. *See* S. RICHMOND, *supra* note 3.

8. The usual distribution for arrival rates, the Poisson distribution, comprises a few days that have extremely low arrivals, many days that have low to moderate arrivals, and a few days that have high arrivals. This distribution is peaked to the left, quickly rising and then gradually falling. The usual distribution for service rates, the exponential distribution, comprises many days that have relatively low service rates and a few days that have moderate or high service rates. This distribution, when plotted, goes continuously in a downward direction.

9. Standard queueing theory formulae assume that the arrival rate and service rate are independent. In reality, this assumption may not be valid. For example, an increase in the service rate may encourage those who might otherwise settle out of court to bring their disputes to court. Thus by speeding the servicing, the arrival rate may increase, thereby offsetting some of the time-saving benefits of the improved service rate. Similarly, reduction of the arrival rate may encourage processors to reduce their individual processing rates. Thus by decreasing the arrivals, the service rate may decline, thereby offsetting some of the time-saving benefits of the improved arrival rate. It also follows that a poor service rate may favorably decrease the arrivals, and increasing the arrival rate (*i.e.*, encouraging more arrivals) may stimulate faster service. The extent that A and S correlate cannot be determined deductively, but only through empirical study of the collected data.

10. A recent study of delay in the courts indicates that shortening case processing time might be fruitless because cases with short processing times require almost the same total times as cases with long processing times. K. Portnoy, Per Curiam Opinions and Appellate Court Delay: A

one day and there are 500 cases waiting in line, then the 500th case will be heard 250 working days sooner because of the one day saved per case. In other words, total time (T) saved equals not just one day (*i.e.*, the reduction in T_S), but rather the reduction in T_S multiplied by the number of cases in the waiting line (N_W). The queueing model also points up the need for more courts, judges, and judge time per year since T and N are influenced by the availability of processing channels. Time and backlog can also be reduced through systems of priorities for particular kinds of cases, by reducing the number of stages in the total process, and by establishing a central administration that directs cases through the most efficient paths.[11]

B. *Dynamic Programming: Optimum Sequencing*

1. *Sequencing of Cases*

Dynamic programming[12] seeks to minimize the total time consumed by ordering events in the most efficient sequence. For example, assume that three cases arrive in a single-court system: case A requires twenty days through the trial stage, case B requires ten days, and case C requires five days. At first glance, one might conclude that processing time will be thirty-five days regardless of the order of the cases. The flaw in that conclusion, however, is that it does not consider that total time (T) equals waiting time (T_W) plus processing or servicing time.[13] If

---

Research Note (1978) (unpublished paper available from the author at the Florida State University Political Science Department). One, however, would expect both short and long cases to require about the same total time because total time equals waiting time plus processing time, and in some court systems waiting time may be almost 100 percent of the total time. Thus, a relatively long processing time has little incremental detriment. The amount of waiting time, however, is a function of the amount of processing time. By reducing an average processing time of two days by half, the waiting time for cases can also be reduced by half, thereby reducing total time by half.

11. For additional examples of the application of queueing theory to the legal process, see J. CHAIKEN, T. CRABILL, L. HOLLIDAY, D. JAQUETTE, M. LAWLESS, & E. QUADE, CRIMINAL JUSTICE MODELS: AN OVERVIEW (1975) [hereinafter cited as J. CHAIKEN & T. CRABILL]; J. CHAIKEN & P. DORMONT, PATROL CAR ALLOCATION MODEL: EXECUTIVE SUMMARY (1975); J. REED, THE APPLICATION OF OPERATIONS RESEARCH TO COURT DELAY (1973); H. Bohigian, The Foundations and Mathematical Models of Operations Research with Extensions to the Criminal Justice System 191-209 (1971) (unpublished Ph.D. dissertation). *See also* H. ZEISEL, *supra* note 1, which includes discussion of "Reducing the Trial Time" (*i.e.*, increasing servicing rates), "Increasing Settlements" (*i.e.*, decreasing arrival rates), and "More Judge Time" (*i.e.*, increasing the channels or processors) and is organized in accordance with queueing theory.

12. *See* J. BYRD, *supra* note 3, at 139.

13. On dynamic and sequential programming, see K. BAKER, INTRODUCTION TO SEQUENCING AND SCHEDULING (1974); J. BYRD, *supra* note 3, at 139-56; R. CONWAY, W. MAXWELL, & L. MILLER, THEORY OF SCHEDULING (1967); B. GLUSS, AN ELEMENTARY INTRODUCTION TO DYNAMIC PROGRAMMING (1972); A. KAUFMAN, GRAPHS, DYNAMIC PROGRAMMING, AND FINITE

the cases are processed in inverse order of their length, A will consume twenty days ($T = 0 + 20$); B will consume thirty days because it must wait twenty days for A to be processed ($T = 20 + 10$); and C will consume thirty-five days because it must wait thirty days for A and B to be processed ($T = 30 + 5$). The total processing time for the three cases when so sequenced is thus eighty-five days; the average processing time is twenty-eight days per case. The goal of dynamic programming is to develop an optimum order for these three cases and then to generalize to rules applicable to larger samples and more complex variations on this basic example.[14]

In our three case sample,[15] there are six possible sequences, each yielding a different average processing time. Table 1 indicates that the most efficient sequence is order 6, which orders the cases from the shortest to the longest. This sequence reduces average processing time from twenty-eight to eighteen days. Assuming that the only goal is to minimize the average time per case,[16] and there are no maximum time constraints, we could formulate a general rule that cases be processed in inverse order of their expected length. If, however, we more realistically provide that no case be allowed to consume more than a certain amount of time, this sequence would be unworkable because the longer cases might never be processed. For example, if we specify that no case should be allowed to take more than thirty days, order 6 is no longer feasible because case A requires thirty-five days and thus violates the thirty-day maximum constraint. Regardless of the ordering, the total waiting and servicing times for the case will equal the sum of the servicing times for all of the cases. In our hypothetical, that sum is thirty-five days; therefore all six orders are infeasible because the last case of

GAMES (1967); G. NEMHAUSER, INTRODUCTION TO DYNAMIC PROGRAMMING (1966); S. RICHMOND, *supra* note 3, at 461-80.

14. Dynamic programming may also be used to minimize lateness and total cost. *See* J. BYRD, *supra* note 3, at 152-55.

15. The model assumes that the three cases arrive on the same day.

16. In other contexts, an appropriate goal might be something other than minimizing the average time spent per case with or without a maximum constraint. For example, in handling jobs in a business firm, the goal might be to minimize the number of late jobs, the average lateness of the jobs, the average cost of lateness, the maximum lateness, or the maximum lateness cost. Those alternative goals are discussed in J. BYRD, *supra* note 3, at 139-56. Alternative procedures for achieving one's goals might include: (1) a rule of first come first served; (2) a rule of shortest cases first where a set of cases arrive during the same week; (3) a rule that gives top priority to the job that has the earliest due date; or (4) a more complicated business rule that schedules last those jobs that have the largest ratio between the time required and the lateness consequences.

TABLE 1. WAYS OF ORDERING THREE CASES

Order 1

Case	T_W +	T_S =	T
A	0	20	20
B	20	10	30
C	30	5	35
		SUM =	85
		AVG. =	28

Order 2

Case	T_W +	T_S =	T
B	0	10	10
A	10	20	30
C	30	5	35
		SUM =	75
		AVG. =	25

Order 3

Case	T_W +	T_S =	T
C	0	5	5
A	5	20	25
B	25	10	35
		SUM =	65
		AVG. =	22

Order 4

Case	T_W +	T_S =	T
A	0	20	20
C	20	5	25
B	25	10	35
		SUM =	80
		AVG. =	27

Order 5

Case	T_W +	T_S =	T
B	0	10	10
C	10	5	15
A	15	20	35
		SUM =	60
		AVG. =	20

Order 6

Case	T_W +	T_S =	T
C	0	5	5
B	5	10	15
A	15	20	35
		SUM =	55
		AVG. =	18
		(Optimum)	

T_w = Waiting time, T_S = Servicing time, T = Total time.

each violates the maximum constraint.[17]

Compliance with the thirty-day maximum constraint could be achieved by reducing the number of arrivals, reducing the service time for some or all of the cases, or adding additional channels or courts.[18] Table 2 shows the possible orderings when two courts are used to process cases. Because there are still three cases, there are six possible orderings. With the additional court, however, the average time per case is reduced because two cases can be processed simultaneously with only one case at the waiting stage. The optimum order now shifts from order 6, which was optimum without constraints, to order 4 or 5, both of which satisfy the new thirty-day constraint with the relatively low average time per case of thirteen days. Because order 4 has a lesser maximum time (twenty days) than order 5 (twenty-five days), it is the optimum of the Table 2 orders. From that analysis, we can generalize that within each court the shorter cases should be heard first unless exigent circumstances require that longer cases be moved up.[19]

The preceding optimum sequencing principles can be applied by computer programs when the volume of cases is too large to calculate all the possible permutations. Each new case entering the system can be analyzed to estimate the amount of trial time required. In criminal cases, these estimates can be based on a statistical analysis that consid-

17. The shortest-case-first rule should not be considered fallacious merely because it does not work without maximum constraints. Use of maximum constraints is simply a recognition that one cannot apply management science or operations research principles to the legal process without considering constitutional, statutory, and precedential constraints. These constraints sometimes include speedy trial rules.

The hypothetical constraint could have been changed from 30 to 40 days. This change, however, would eliminate the possibility of discovering the alternative options queueing theory provides for satisfying a maximum constraint, which, in the absence of alternative options, would be violated given the data in Table 1.

18. With one court, judge, or channel, any maximum constraint will either be violated by all the orders or by none of them. This results because the last case consumes an amount of time equal to the sum of the servicing times for all the cases being considered. This phenomenon is apparent in Table 1 where the last case of each order consumes 35 days. The notion of a maximum constraint, however, is quite meaningful in the legal process because almost all court systems have more than one judge. The notion is also meaningful with only one judge; the constraint can conceivably be satisfied by reducing the arrival rate or increasing the servicing rate.

19. Gillespie, *Economic Modeling of Court Services, Work Loads, and Productivity*, in MODELING THE CRIMINAL JUSTICE SYSTEM 175 (S. Nagel ed. 1977); S. Flanders, District Court Studies Project Interim Report (June 1976) (report to the Federal Judicial Center); USDA Statistical Reporting Service, The 1969-70 Federal District Court Time Study (June 1971) (report to the Federal Judicial Center).

TABLE 2. WAYS OF ORDERING THREE CASES WITH TWO COURTS

Order 6		3		5		2		4		1	
Case	Time (days)	Case	Time (days)	Case	Time (days)	Case	Time (days)	Case	Time (days)	Case	Time (days)
Court 1											
C(5)	5	C(5)	5	B(10)	10	B(10)	10	A(20)	20	A(20)	20
Court 2											
B(10)	10	A(20)	20	C (5)	5	A(20)	20	C (5)	5	B(10)	10
A(20)	30	B(10)	30	A(20)	25	C (5)	25	B(10)	15	C (5)	15
SUM = 45		SUM = 55		SUM = 40		SUM = 55		SUM = 40		SUM = 45	
AVG. = 15		AVG. = 18		AVG. = 13		AVG. = 18		AVG. = 13		AVG. = 15	

ers such variables as whether the crime is severe, whether the defendant has asked for a jury or bench trial, and whether the defendant has a private counsel or a public defender. In personal injury cases, time predictions can be made from such variables as the plaintiff's latest settlement demand, the defendant's latest settlement offer, and the kind of personal injury claimed. Regression equations can be developed by computerized regression analysis of data based on these and other variables.

2. *Sequencing of Stages*

Optimum sequencing of stages of cases seeks to determine the most efficient order of processing the various separable parts of a case. For example, optimum sequencing of stages might in a particular case indicate that the liability and damages determinations should be made at separate trials. Determining liability and damages in one trial would seem to ensure optimal efficiency. The split trial system, however, can result in greater time savings than might at first glance be expected. For example, in a hypothetical sample of 100 personal injury cases, only about sixty-four will typically result in a judgment for the plaintiff, and thus thirty-six cases require only half a trial because they do not reach the damages question. Furthermore, about thirty-two cases or half of those in which the defendant is found liable are likely to be settled before the damages trial. If we assume that the average combined trial takes ten days, the separate liability trial seven days, and the separate damage trial six days, then under the combined system, 100 cases would comsume 1,000 trial days. Under the split trial system, however, 100 cases might consume only 892 trial days. Of those 892 days, 252 are consumed by the 36 cases in which liability is not established (36 × 7); 448 are consumed by the 64 cases in which liability is established (64 × 7); and 192 are consumed by the 32 cases in which liability is established and a second hearing rather than a settlement is needed to determine the damages (32 × 6). The split trial system thus potentially offers a substantial savings over the combined system.

Other factors must be considered when determining the optimum sequence of stages. Under a split trial system, the recovery rate may decline because the jury cannot temper its verdict with a reduced damages recovery. This may be most apparent in cases involving contributory negligence. But the split trial system provides an unforeseen time benefit because fewer cases are eligible for the second trial and many of those cases will be settled between the liability trial and the damages trial. The split trial system thus may have a substantial effect on the outcome of a case.[20] In analyzing whether this reform should be adopted, the desirability of this impact should be considered.

A second and more common kind of optimum sequencing is the ordering of the stages of a case in relation to the stages of other cases with the goal of reducing the overall processing time of the cases in the system. It typically considers whether and to what extent the early stages of one case should be processed before the later stages of other cases. In a single-court system, assume two cases, each of which has a pleading stage and a trial stage. Assume further that pleading for case A takes one hour (P_1), pleading for case B takes two hours (P_2), trial for case A takes three hours (T_1), and trial for case B takes four hours (T_2). Table 3 shows the possible ways the stages of those cases could be sequenced without violating the inherent limitation that pleadings must precede trial. Each order shows the servicing, waiting, and total times for each stage. The servicing time remains constant regardless of the order, but the waiting time, which represents the time that a stage of a case must wait for a stage of another case to be processed, varies. The optimum sequence, therefore, is the one with the least waiting time.

Table 3 shows that order 1, the optimum sequence, proceeds with all the stages of case A before proceeding with any of the stages of case B. Case A is preferred because it is a shorter case. It is also preferable in

20. For a statistical study of a sample of cases comparing the split trial method with integrated trials (producing findings like those described above), see Zeisel & Callahan, *Split Trials and Time Saving: A Statistical Analysis*, 76 HARV. L. REV. 1606 (1963).

TABLE 3. WAYS OF ORDERING TWO STAGES OF TWO CASES

Order 1			Order 2				Order 3				Order 4				Order 5				Order 6			
T_S	T_W	T	Stage	T_S	T_W	T	Stage	T_S	T_W	T	Stage	T_S	T_W	T	Stage	T_S	T_W	T	Stage	T_S	T_W	T
1	0	1	P_B	2	0	2	P_A	1	0	1	P_B	2	0	2	P_A	1	0	1	P_B	2	0	2
3	0	3	T_B	4	0	4	P_B	2	1	3	P_A	1	2	3	P_B	2	1	3	P_A	1	2	3
2	4	6	P_A	1	6	7	T_A	3	2	5	T_A	3	0	3	T_B	4	0	4	T_B	4	0	4
4	0	4	T_A	3	0	3	T_B	4	3	7	T_B	4	4	8	T_A	3	6	9	T_A	3	4	7
	SUM = 14				SUM = 16				SUM = 16				SUM = 16				SUM = 17				SUM = 16	
	AVG. = 7				AVG. = 8				AVG. = 8				AVG. = 8				AVG. = 8.5				AVG. = 8	
	(Optimum)																					

this simple illustration to complete all the stages of each case without interruption because this will avoid unnecessary waiting time. These conclusions assume that the only stages in each case are pleading and trial. If the parties are not ready to go to trial immediately after pleading, but instead require a preparation for trial stage, it would seem logical to schedule the case *B* pleading between the case *A* pleading and trial. The case *B* pleading could then be processed during the case *A* preparation stage. This allows the productive use of what would otherwise be waiting time, thereby reducing total time.

Although it might seem more efficient, even with only two stages, to interrupt P_1 and T_1 with P_2 if case *A* had shorter pleadings and a longer trial than case *B*, Table 3 reveals that the average time per case would be minimized by processing them in uninterrupted succession. If, however, the trial time of case *B* is shortened so that it is less than the total time of case *A*, all stages of case *B* should be completed before beginning the processing of case *A*.

The rule of processing the stages of a given case in uninterrupted succession applies when using one or more courts, provided each court is capable of processing cases at both the pleading and trial stages. If one court specializes in pleadings and a second court specializes in trials, the optimum sequencing requires finding the shortest time unit among all stages and cases. If, for example, the shortest time unit among all the stages is the pleading stage of a case, it should be scheduled first. If the shortest time unit is the trial stage of a case, it should be scheduled last. After finding the shortest time unit, one looks for the next shortest and continues to follow the rule that a pleading unit is heard next to the top while a trial unit is heard next to the bottom. By following these rules, courts might effectively minimize the average time consumed by each case.[21] To avoid the cumbersome task

21. This method, however, would not save time if trial judges have to spend substantial time familiarizing themselves with cases in which they did not participate in the pleading stage.

For additional examples of dynamic and sequential programming applied to the legal process, see H. ZEISEL, *supra* note 1, at 201-05; Hausner, Lane, & Oleson, *Automated Scheduling in the Courts*, in OPERATIONS RESEARCH IN LAW ENFORCEMENT, JUSTICE, AND SOCIETAL SECURITY 217 (S. Brounstein & M. Kanrass eds. 1976); H. Bohigian, *supra* note 11, at 171-90; J. Jennings, Evaluation of the Manhattan Criminal Court's Master Calendar Project: Phase 1—February 1-June 30, 1971 (Jan. 1972) (Rand Institute study); R. Nimmer, The System Impact of Criminal Justice Reforms: Judicial Delay as a Case Study 62-86 (1974) (unpublished thesis); Programming Methods Inc., Justice: A Judicial System to Increase Court Effectiveness (Apr. 1971) (a system design study of the Criminal Court of the City of New York).

of enumerating all possible orders, stage sequencing, like case sequencing, can be computerized using as its inputs time consumption figures for the stages based on characteristics of the cases that have been found to correlate with the amount of time required for each stage.

C. *Critical Path Method and Flow Chart Models*

1. *Critical Path*

Critical Path Method (CPM) or Program Evaluation and Review Technique (PERT) seeks to reduce total time by reducing the time of those stages that most directly influence it rather than by reducing the time of all stages. CPM and PERT, therefore, focus on two kinds of stages: (1) a stage essential to a subsequent stage, and (2) a stage that requires more time than other stages but is essential to a subsequent stage. Figure 1 is illustrative. Preparation by the prosecutor and public defender are usually essential before proceeding to the trial stage. If, however, the public defender requires an average of three weeks to prepare for trial, while the prosecutor requires only two weeks, the critical path from pleading to trial is through the lower arrow which represents preparation by the public defender. Reducing the prosecutor's preparation time, therefore, would not allow the earlier commencement of trials, but reducing the public defender's preparation time by providing him with additional resources would allow this. Note that if we reduce the public defender's preparation time to less than two weeks, the prosecutor's preparation replaces it as a critical path.[22]

Critical path analysis can be expanded to include the entire criminal justice process from arrest to parole and the civil justice process from complaint to recovery on the judgment. Some stages of the criminal and civil justice systems require the completion of two or more procedures as prerequisites to a subsequent stage. Examples of these jointly converging stages include information gathering by the defense and prosecutor for pretrial release, and information gathering by the defense, prosecutor, and probation department for post-conviction sentencing.[23] The probation department's presentence report normally

22. On critical path method and flow chart modeling, see R. ARCHIBALD & R. VILLORIA, NETWORK-BASED MANAGEMENT SYSTEMS (PERT/CPM) (1967); J. BYRD, *supra* note 3, at 115-38; H. EVARTS, INTRODUCTION TO PERT (1964); B. HANSEN, PRACTICAL PERT (1964); S. RICHMOND, *supra* note 3, at 481-500; G. WHITEHOUSE, SYSTEMS ANALYSIS AND SYSTEMS DESIGN (1973).

23. If one prepares a flow chart from arrest to sentencing showing all the connecting and

follows the conviction of the defendant. Substantial time, therefore, might be saved by having the probation department prepare presentence reports on all defendants before the verdict. Although it may seem inefficient to prepare reports for the hypothetical thirty percent of defendants who are not convicted, the cost of preparing these reports might prove to be less than the cost of following the usual practice of preparing reports after conviction while the defendant sits in jail. Specifically, the additional cost of reporting on all defendants would be approximately .30 of the number of defendants multiplied by the average cost per report. The additional cost of not reporting on all the defendants includes: (1) the cost of incarcerating those defendants who subsequently will be released on probation, and (2) the cost to the local jurisdiction attributable to the delay in sending to prison those defendants who subsequently will be denied probation.

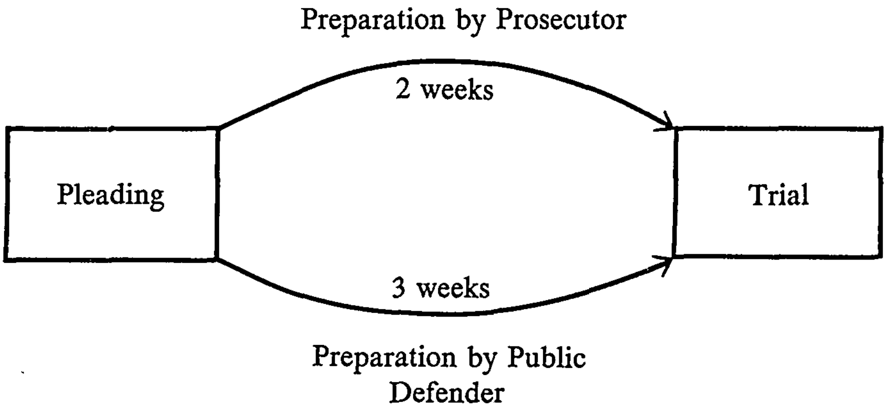
Figure 1 illustrates a hypothetical average preparation time for the prosecutor and public defender. In practice this data may be calculated on the basis of three subjective estimates reported by the attorney: (1) likely preparation time (comparable to the mode in statistical analysis); (2) an optimistic estimate of preparation time (which occurs about once in 100 cases); and (3) a pessimistic estimate of preparation time (which also occurs about once in 100 cases). From these estimates, a mean time can be computed using the formula: T_E (expected time) equals T_O (optimistic time) plus four times T_L (likely time) plus T_P (pessimistic time), with the sum divided by six. This formula is based on the assumption that although people have difficulty estimating an average outcome, they can accurately estimate optimistic, modal, and pessimistic figures. The formula is also based on assumptions about the tendency of averages to relate to those figures and the usefulness of those input figures in PERT-CPM outputs.

---

converging paths, then the overall critical path would be the combination of all the longest paths starting from sentencing and working back to arrest. A more sophisticated approach would also recognize that if one comes to a fork when determining a critical path, one should not automatically choose path *A* over path *B*. Instead, look to the sums of all the connecting paths that converge on a stage in determining the alternative path that is critical for reducing the total time consumed from arrest to sentencing.

We could extend the hypothetical data of Figure 1 to show this kind of flow chart. Unfortunately, the real data available on time consumption at various stages in the criminal justice process does not deal with such converging matters as the time needed by various participants to prepare for various stages. Figure 2, which contains real data for connecting paths, but not for converging paths, illustrates this situation. However, information about their preparation times can be obtained from the participants by examining files and other records or by questionnaires and interviews.

FIGURE 1. A SIMPLE CRITICAL PATH MODEL



Entering information about the ordering of the stages and the optimistic, modal, and pessimistic estimates for each stage into a PERT or CPM computer program yields a variety of useful outputs: (1) estimates of the date by which each stage is likely to be completed; (2) estimates of the accumulated time as of each stage; (3) the stages that constitute the critical path; (4) estimates of the time spent waiting for an adjacent stage to be completed so that a dependent stage can begin; and (5) the probability that a subsequent stage will have to wait for a prior stage that is not on the critical path. These informational outputs can be helpful in planning both complex and routine cases.[24] Such planning, however, should not encompass all stages on the critical path, but rather only those that consume the most time or are most subject to time reduction as indicated by the spread between the optimistic, mo-

24. A variation of PERT analysis measures the time and cost to conduct each activity in the routine manner and the time and cost to conduct them as expeditiously as possible (*i.e.*, crashing). The cost of crashing in dollars per day or other time unit can then be determined for each activity. With this information, one can then determine the activities and the extent that those activities are worth crashing or accelerating. Activities are not worth accelerating when the completion of the total task would still have to await the completion of other activities. Activities that can be accelerated at relatively little cost and, when completed, move the entire job forward, however, are worth accelerating. *See generally* W. GREENWOOD, DECISION THEORY AND INFORMATION SYSTEMS (1969).

dal, and pessimistic estimates. The greater the spread at a given stage, the greater the probability that the stage may be subject to time reduction, provided the correlates of that spread across cases, across courts, or over time can be determined.[25]

2. *Flow Chart Models*

Flow chart models, which illustrate the critical path method, represent the stages of the legal process through a series of rectangles and connecting arrows. The rectangles indicate the time needed to complete each stage, the interrelationships or flow of the stages, and alternative possibilities including the possibility or probability that a stage will drop out of the system. Flow chart models are a useful visual aid for understanding general case processing and often suggest ideas for reducing time. They also can be entered into a computer to show the output effects of changes in the times, case quantities, stages, or other inputs.

A simple hypothetical illustrates the usefulness of the flow chart model. In City *A* the average felony case takes one hundred days to complete, twenty days of which involve waiting for a one-day grand jury proceeding after the preliminary hearing. The flow chart indicates

25. *See* Nagel, *Measuring Unnecessary Delay in Administrative Proceedings: The Actual versus the Predicted*, 3 POLICY SCI. 81 (1972).

Closely related to critical path method is the interesting problem of choosing the court that will minimize time consumption and other costs. For example, suppose a personal injury plaintiff could choose between the paths leading to a federal court (diversity jurisdiction) or to a state court. Suppose further that if the plaintiff goes to the federal court, his case will be heard within one year, and if successful he will collect about $15,000, but there is only a 20 percent chance of winning. If, on the other hand, he goes to the state court, his case will be heard within two years, and if successful he will collect $10,000, but he has a 40 percent chance of winning. Which path should he follow?

The expected value of the federal path is $3,000 (*i.e.*, $15,000 discounted by the .20 probability of success) without considering the time element, and the expected value of the state path is $4,000 (*i.e.*, $10,000 discounted by the .40 probability of success). If, however, we consider that one has to wait two years for the $10,000 from the state court, its value substantially decreases. More specifically, the present value of a future amount is calculated by the formula $P = A/(1{+}r)^t$, where r is the interest rate that could be obtained by depositing the cash value in a savings account for t years. If we assume the interest rate is 6 percent then the present value of the state path's $10,000 two years from now is $8,900. If we now discount that present value by the .40 probability of obtaining it, the expected value of the state case becomes $3,560. Applying the same formula to the federal path, the present value of its $15,000 award would be $14,151 because $14,151 = $15,000/(1+.06). If we now discount that present value by the .20 probability of obtaining it, the expected value of the federal path becomes $2,830, which is still less than the state path, but by a smaller margin, considering differences in time consumption. We, of course, could have offered a hypothetical example where considering the time consumption reverses the rank order of the better path.

that elimination of the grand jury proceeding could save twenty-one days or twenty-one percent of the total time. This falsely assumes, however, that what replaces grand jury indictment—a formal complaint issued by the prosecutor—requires no further time expenditure. Another flow chart, which considers the prosecutor's formal complaint, might show that the defendant waited ten days for the complaint, indicating the savings from replacing the grand jury with the prosecutor would be eleven days. But the savings probably will be less than eleven days if the prosecution is overburdened or the cases are more complex than those previously handled.

Table 4 provides previously unpublished data for constructing a flow chart (shown in Figure 2) for a nationwide sample of 11,000 state criminal cases compiled in 1961 by the American Bar Foundation. Unlike most flow chart models, which are based on single-court jurisdictions, Table 4 shows the average number of days from one event to another for the subset of the 11,000 cases in which the two events occurred and information was available. It also shows the standard deviation for each of these time consumption figures. If the actual figures for the cases at any given time passage have a normal distribution, approximately two-thirds should be within one standard deviation of the mean. For practical purposes the standard deviation as a measure of spread is most useful for indicating which time passages have the greatest variation and are thus most subject to having their excessive cases pushed toward the mean.[26] The coefficient of variation, *i.e.*, the ratio between the standard deviation and the mean, is a better measure of this occurrence since one would expect a larger spread where there is a larger mean.[27] This measure shows that on percentage of days saved, improvement can be made most readily at the arrest to bail release stage where the coefficient of variation is almost two to one. Many days can be saved simply because that time passage is the sum of all the component time passages.

In the flow chart, the nodes or events in the solid rectangles are generally required events (*e.g.*, in felony cases the required arrest, indictment, arraignment, and either a nontrial disposition or a trial). The

26. The standard deviation is calculated by: (1) summing the time consumption for the cases at a given passage and dividing by the number of cases, (2) subtracting that mean from the actual time consumed for each case, (3) squaring those differences, (4) dividing the sum of those squares by the number of cases, and (5) taking the square root of that quotient.

27. *See* H. BLALOCK, SOCIAL STATISTICS (1972); J. MUELLER, K. SCHUESSLER, & H. COSTNER, STATISTICAL REASONING IN SOCIOLOGY (1970).

TABLE 4. TIME CONSUMPTION AT VARIOUS STAGES IN STATE CRIMINAL CASES ACROSS THE U.S.

Time Number	Start Node	End Node	Time Description	# of Cases with Info.	Mean Days	Standard Deviation	Coefficient of Variation
1	1	3	Arrest to counsel	2114	29	31	1.07
2	2	3	Preliminary to counsel	1147	25	33	1.32
3	3	4	Counsel to bail release	1145	24	34	1.42
4	1	8	Arrest to nontrial disposition	2907	95	96	1.01
5	1	9	Arrest to trial begins	409	160	128	.80
6	3	8	Counsel to nontrial disposition	2355	61	61	1.00
7	3	9	Counsel to trial begins	338	93	69	1.03
8	7	8	Arraignment to nontrial disposition	3491	48	57	1.19
9	7	9	Arraignment to trial begins	507	81	70	.86
10	1	4	Arrest to bail release	1313	22	40	1.82
11	1	2	Arrest to preliminary hearing	1516	12	13	1.08
12	1	7	Arrest to arraignment	3003	46	60	1.30
13	6	8	Indictment to nontrial disposition	3815	71	88	1.24
14	6	9	Indictment to trial begins	544	114	101	.89

nodes or events in the dashed rectangles are optional events because many of the cases do not require bail release, preliminary hearing, or counsel. The numbers on the arrows indicate the average time consumed between events. The numbers in the rectangles indicate the event number arranged in the usual chronological order. Time consumption figures without asterisks are calculated from the raw data supplied by the American Bar Foundation.[28] This data provides information for the fourteen time passages or variables in Table 4 and Figure 2. Time consumption figures with asterisks (X times) are determined by subtracting the shorter time (S time) from the longer time (L time), where $X + S$ should equal L when there is perfect consistency in the data. For example, arrest to indictment is calculated at twenty-four days for nontrial disposition because the period from arrest to nontrial disposition is ninety-five days, and from indictment to nontrial disposition, seventy-one days. The numbers are consistent because certain events are optional and information is missing from some cases. It thus takes a composite average of twenty-nine days from arrest to counsel with or without a preliminary hearing, but an average of thirty-seven days (12+25) from arrest to counsel in those cases that had a preliminary hearing.

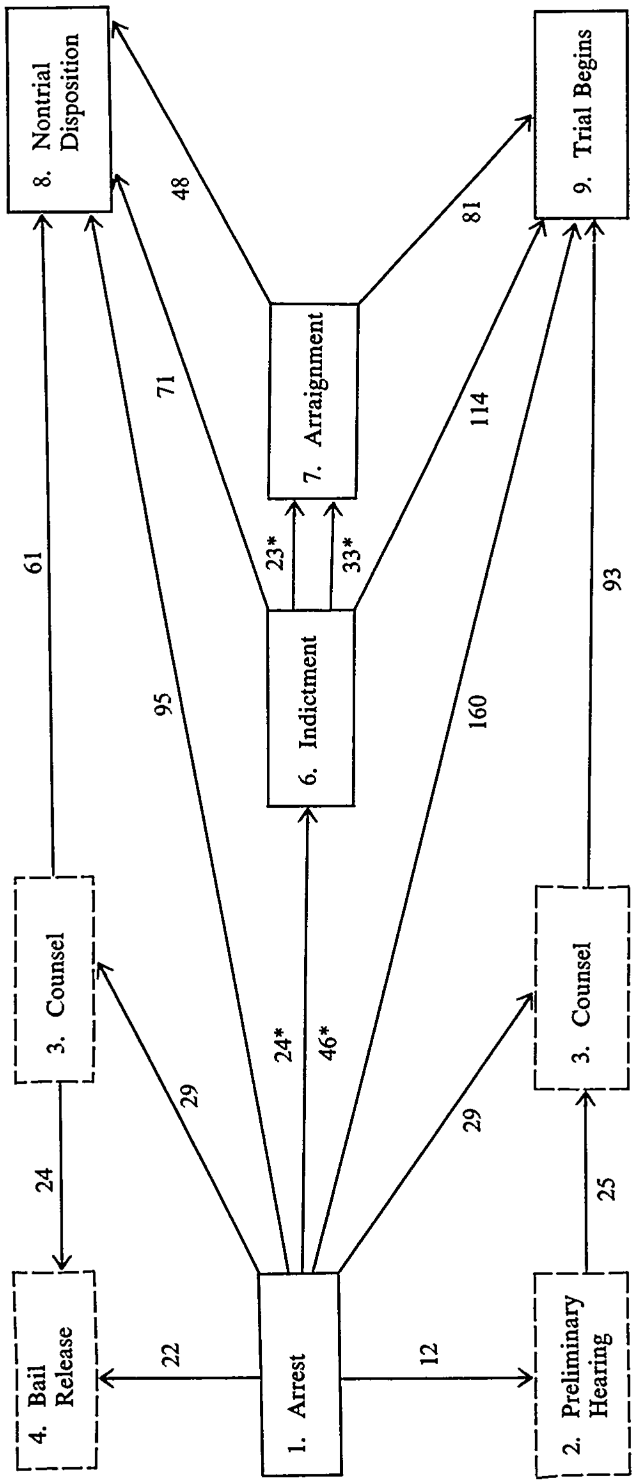
The flow chart shows that trial disposition takes 160 days while nontrial disposition takes ninety-five days. This indicates that each case in which a trial could be eliminated would save sixty-five days. The flow chart also shows that a defendant who is released on bail usually encounters a rather long three-week waiting period between arrest and the securing of defense counsel. Further, the period between indictment and arraignment consumes almost a month, although during this time important plea bargaining may occur. Arrest to indictment tends to take about a month in nontrial cases followed by more than two months for nontrial disposition; in trial cases, arrest to indictment tends to take one and a half months followed by three months awaiting trial.[29]

Processed data like that shown in Table 4 and Figure 2 has been generated from the raw data in the American Bar Foundation files for six different kinds of crime: armed robbery, aggravated assault, grand larceny, rape, burglary, and auto theft. Separate tables and figures also have been generated for cases from metropolitan counties with more than 400,000 population, urban counties with populations between 100,000 and 400,000, and rural counties with populations under 100,000. A separate table and figure is available for each of the fifty states. This data may prove especially useful for determining the pro-

28. For a description of the data-set, see L. SILVERSTEIN & S. NAGEL, AMERICAN BAR FOUNDATION: STATE CRIMINAL COURT CASES (ICPR 1974). The same IBM card was analyzed in Nagel, *Disparities in Criminal Procedure*, 14 U.C.L.A. L. REV. 1272 (1967).

29. These time estimates are derived from Table 4 and Figure 2, which in turn are derived from the data-set cited in note 28 *supra*.

FIGURE 2. A FLOW CHART OF TIME CONSUMPTION AT VARIOUS STAGES IN CRIMINAL CASES ACROSS THE U.S.



cedures, demographic characteristics, and governmental characteristics that correlate with high or low time consumption at various stages in the criminal justice process.[30] The data is still being analyzed for this purpose; when more detailed data is available, more complex flow charts can be developed. Variations include showing average time consumption for each passage, measures of spread or distribution, optimistic or desired time consumption, the proportion of cases that move from one event to another when there is provision for branching or dropping out, and the dollar cost of each event or time passage to the legal system or the parties. In addition to rectangles and arrows, one can use a great variety of geometric forms to show: (1) events or nodes that begin or end the process; (2) time passages that always occur or that occur with given probabilities; and (3) queueing, critical path, and other information. This mass of information can then be used as an input into a computerized program to provide a variety of outputs showing how the numbers change as the volume of cases entering the system or the proportion of cases that take one turn rather than another at a branching point changes.[31]

D. *Optimum Level, Mix, and Choice Analysis*

1. *Optimum Level Analysis*

Queueing theory indicates that cases can be processed faster if we decrease arrivals, increase the servicing rate, and increase the number of processors. This is not meant to imply that we should decrease arrivals to zero or increase the servicing rate or number of processors to the point where time consumption becomes virtually zero. On the contrary, the speed-up costs may be so great that it is better to keep the

30. The more detailed versions of Table 4 and Figure 2 are available from the senior author of this article upon request.

31. This analysis is closely related to Markov chain analysis. *See* note 41 *infra* and accompanying text. Additional applications of flow chart modeling to legal process include J. CHAIKEN & T. CRABILL, *supra* note 11, Blumstein, *A Model to Aid in Planning for the Total Criminal Justice System*, in QUANTITATIVE TOOLS FOR CRIMINAL JUSTICE PLANNING 129 (L. Oberlander ed. 1975); Navarro & Taylor, *Data Analyses and Simulation of a Court System for the Processing of Criminal Cases*, in THE PRESIDENT'S COMMISSION ON LAW ENFORCEMENT AND ADMINISTRATION OF JUSTICE, TASK FORCE REPORT: SCIENCE AND TECHNOLOGY (1967); Cassidy, *A Systems Approach to Planning and Evaluation in Criminal Justice Systems*, 9 SOCIO-ECON. PLAN. SCI. 301 (1975); W. Biles, A Simulation Study of Delay Mechanisms in Criminal Courts (unpublished paper presented at the meeting of the Operations Research Society of America, at New Orleans, La., 1972); G. Hogg, R. DeVor, & M. Handwerker, Analysis of Criminal Justice Systems via Stochastic Network Simulation (unpublished paper presented at the Workshop on Operations Research in the Criminal Justice System, at San Diego, Cal., 1973).

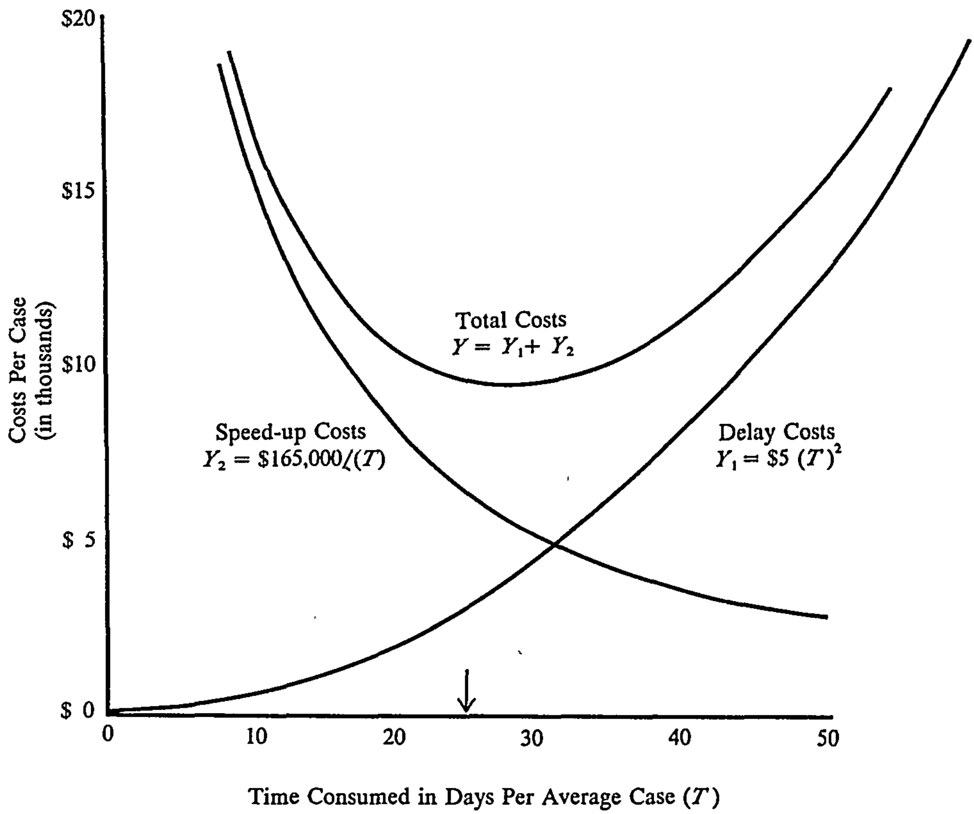
delay. Optimum level analysis of time consumption reveals the level of delay that will minimize the sum of the delay costs (Y_1) and the speed-up costs (Y_2).[32]

Figure 3 in graphic form shows optimum level analysis for a hypothetical metropolitan court system. To fully apply the analysis we need to develop an equation showing the relation between delay costs and time consumed. For the sake of simplicity we will hypothesize that each extra day consumed in completing a criminal case costs the system about $7 for each jailed defendant who receives an acquittal, dismissal, or probation; $2 represents wasted jail maintenance costs, and $5 represents lost national product due to defendant's nonproductivity. The $2 is calculated by approximating a cost of $6 per day to maintain a defendant in jail, noting that one-third of the defendants receive nonjail dispositions upon trial. The $5 is calculated by approximating that a defendant in this metropolitan area can earn about $15 per day when not in jail, and that approximately one-third of the defendants would not be in jail if the system could process their acquitted or dismissed cases sooner. There might also be a cost of $3 per day for each released defendant. This represents waste generated by the release of those who will be jailed when they are eventually tried and convicted but who during the delay commit a crime or must be rearrested for failure to appear in court. The $3 figure is determined by calculating: (1) the crime-committing cost or the rearresting cost for the average released defendant, (2) multiplied by the low probability of crime-commission or rearrest, (3) multiplied by the moderate probability of conviction and incarceration if the case were to come to disposition, and (4) divided by the number of days released. If half the arrested defendants are jailed and half are released, the $7 delay cost per day per *jailed* defendant is $3.50 per day per arrested defendant, and the $3 delay cost per day per *released* defendant is $1.50 per day per arrested defendant. Thus the total delay cost per day per case would be $5 ($3.50+$1.50). If this is a constant figure, the delay cost (Y_1) would equal $5 times T days, or $Y_1 = \$5\,(T)$.

Since the likelihood of crime commission and the need for rearrest increase as delay increases, the relation between Y_1 and T might be better expressed by an equation of the form $Y_1 = \$5(T)^2$. This equation

32. On optimum level analysis, see M. BRENNAN, *supra* note 3, at 111-92; J. BYRD, *supra* note 3, at 183-98; S. RICHMOND, *supra* note 3, at 87-126; J. SHOCKLEY, THE BRIEF CALCULUS: WITH APPLICATIONS IN THE SOCIAL SCIENCES (1971).

FIGURE 3. OPTIMUM LEVEL ANALYSIS APPLIED TO TIME CONSUMPTION



indicates that when T is one day, Y_1 is $5; but when T is X days, Y_1 is not $5 times X, but rather Y_1 expands at an increasing rate. More specifically, as T rises 1%, Y_1 rises 2%. We can determine the values of the multiplier and exponent of T by performing a log-linear regression analysis provided that data is available that shows: (1) the amount of time each case consumed; (2) the approximate cost of each case in terms of jail maintenance and lost national product for those held and crime-committing and rearresting costs for those released; and (3) the proportion or probability of cases in which nonjail sentences were imposed and the proportion of crime commission and rearrest of released defendants. The more time consumed the higher the delay costs become, possibly at an increasing rate. The more we rush cases to a disposition, however, the greater the speed-up costs might be. These costs are predominantly the monetary cost of hiring additional personnel or introducing new facilities and procedures.

Suppose that either through deductive queueing theory or the compilation of empirical data we find that with twenty judges an average case takes seventy-five days, with sixty judges, twenty-five days, and with eighty judges, nineteen days. We can assume that with no judges the number of days would rise to infinity. Conversely, it would require an infinite number of judges to reduce the number of days to zero.

The speed-up costs curve shown in Figure 3 incorporates the preceding data and assumptions. A curve of this kind can be expressed by the equation $J=a/T$, where J represents the number of judges, and T represents time in days per average case. If $J=a/T$, then $T=a/J$. The a in the former equation represents the number of judges needed to reduce time to one day per case (*i.e.,* $T=1$) while the a in the latter equation represents the number of days consumed when there is only one judge (*i.e.,* $J=1$). From the preceding data and a computerized regression analysis we can determine that $a=1500$. According to our data, this means that $J=1500/T$ and $T=1500/J$.

Instead of thinking in terms of the relation between the number of judges and the number of days consumed, it would be more practical to think in terms of the cost of judges and the number of days consumed. If one judge costs \$40,000 per year, the daily cost is \$110. The equation $J=1500/T$ thus should be changed to $Y_2=\$165{,}000/T$. Y_2 represents the speed-up costs or the additional judge costs and the \$165,000 is simply \$110 times the previous a (scale coefficient) of 1500, which shows the scale has been increased by \$110 per judge per day. The equation $Y_2=\$165{,}000/T$ is the algebraic equivalent of the equation shown in Figure 3, *i.e.*, $Y_2=165{,}000(T)^{-1}$.

Given the relation between delay costs and time consumed of $Y_2=\$5\,(T)^2$ and the relation between speed-up costs and time consumed of $Y_1=\$165{,}000(T)^{-1}$, the relation between total costs (Y) and time consumed is $Y=\$5(T)^2+\$165{,}000(T)^{-1}$. We now can calculate the optimum level of time consumed, *i.e.*, the value of T where the total cost curve hits bottom. This calculation recognizes that the total cost curve has a negative slope before it hits bottom, a positive slope after it hits bottom, and a zero slope when it bottoms out. It, therefore, is necessary to know the slope of Y to T. We then can set that slope at zero and solve for T.

In elementary calculus, one learns that in an equation of the form $Y = aX^{b2}$ the slope of Y to X is baX^{b-1}. Accordingly, in our total cost equation the slope of Y to T is (2) (\$5) $(T)^{2-1}$+ (−1) (\$165,000) $(T)^{-1-1}$. If we set that expression at zero and solve for T, we get $\$10(T) - \$165{,}000/(T)^2 = 0$, or $T = (\$16{,}500)^{.33}$, which means $T = 25$ days where the total costs hit bottom. This result indicates that twenty-five days or slightly less than one month is the optimum level of time consumption to minimize the sum of the delay and speed-up costs. Furthermore, the optimum number of judges is sixty since $J = 1500/T$, or $60 = 1500/25$. Our court system thus could minimize its total costs with about sixty full-time judges.

This optimum level analysis can be made more complex by recognizing that speed-up costs (Y_2) can only be indicated accurately as a combination of the cost of judges, prosecutors, public defenders, and other miscellaneous court expenses. The methodology, however, is similar: (1) obtaining an empirical equation that relates speed-up and delay costs to time, (2) finding the slope of the sum of these two equations, (3) setting that slope at zero and solving for T to determine the optimum number of days per average case for minimizing total costs, and (4) thereby indirectly determining the optimum number of judges, prosecutors, public defenders, and other miscellaneous court expenses.[33]

2. *Optimum Mix Analysis*

Queueing theory reveals that we can process cases faster if we decrease arrivals and increase the service rate and number of processors. Focusing first on the number of processors, we might conclude that adding more judges, prosecutors, and public defenders would solve the time consumption problem. The previous section, however, demonstrated that to reduce time consumption to a level approaching zero might require so many judges that the cure becomes more expensive than the problem. We therefore calculated the level of judges that would optimize the sum of the delay and judge costs. A similar analy-

---

33. For additional examples of optimum level analysis applied to delay reduction and other aspects of the legal process, see S. NAGEL & M. NEEF, LEGAL POLICY ANALYSIS: FINDING AN OPTIMUM LEVEL OR MIX (1977); S. NAGEL, P. WICE, & M. NEEF, TOO MUCH OR TOO LITTLE POLICY: THE EXAMPLE OF PRETRIAL RELEASE (1977); H. ZEISEL, *supra* note 1, at 169-220; Phillips & Voety, *An Economic Basis for the Definition and Control of Crime*, in MODELING THE CRIMINAL JUSTICE SYSTEM 89 (S. Nagel ed. 1977); Merrill & Schrage, *Efficient Use of Jurors: A Field Study and Simulation Model of a Court System*, 1969 WASH. U.L.Q. 151; G. Munstermann & W. Pabst, Operating an Efficient Jury System (unpublished paper presented at the International Meeting of the Institute of Management Sciences, 1975).

sis could determine the optimum level of lawyers in the prosecutor's or the public defender's offices. Performing three separate optimum level analyses, however, would be inefficient because the determination of the optimum level for each of these activities requires consideration of the other activities. A more efficient approach is to perform an optimum mix analysis.[34]

This kind of analysis allocates a given budget among judges, prosecutors, and defense counsel with the goal of minimizing time consumption. Suppose an optimum level analysis reveals the optimum level of expenditure is $700,000 for judges, $400,000 for prosecutors, and $200,000 for defense counsel. If the total budget is only $1,000,000, it would be quite inefficient to blindly slash each activity by $100,000.[35] It would also be inefficient to allocate the reduced budget in the same proportions as the preceding budget. Either method of bringing expenditures within the total budget fails to consider the marginal rates of return on each activity. By taking that perspective, we can arrive at a meaningful optimum mix analysis.

In our hypothetical court system, we previously determined that the relation between time consumed and the number of judges is $T = 1500/J$. Given that the cost of a judge's salary is $40,000 per year, the relation between time consumed and the cost of judges is $T = 165{,}000/\$J$, where $\$J$ represents dollars spent on the judiciary per day. Thus, if $400,000/365, or $1,096, is spent on ten judges per day, the equation indicates the average case will consume 150 days ($165{,}000/1096 = 150$). Using a similar analysis, we might determine that the relation between time consumed and the number of prosecutors is $T = 1200/P$. This result indicates that if there is only one prosecutor in our court system, cases would average 1200 days—ignoring for the moment the number of judges and defense counsel. If prosecutors are paid $30,000 per year, the cost equation becomes $T = 98{,}400/\$P$,

34. On optimum mix analysis, see J. BYRD, *supra* note 3, at 85-114; P. KOTLER, MARKETING DECISION MAKING: A MODEL BUILDING APPROACH (1971); R. LLEWELLYN, LINEAR PROGRAMMING (1963); C. MCMILLAN, MATHEMATICAL PROGRAMMING (1970); S. RICHMOND, *supra* note 3, at 314-404.

35. It would be inefficient to reduce each of the three activities by $100,000 to bring the $1,300,000 down to the available $1,000,000 because the marginal rate of return may differ substantially for each activity. For example, if an extra dollar given to judges produces a greater saving of time than an extra dollar given to prosecutors or defense counsel, then we would not want to take equally from each of the three activities, but rather would take more from the allocation to prosecutors and defense counsel. Likewise, it would be inefficient to reduce each activity in proportion to the size of its separate optimum budget without considering the marginal rates of return for an extra dollar spent or taken away for each activity relative to the other activities.

where $\$P$ represents dollars spent on prosecution ((1200) (30,000)/365 = 98,400). A similar analysis for public defenders might reveal an equation of $T = 1000/D$. If they are paid \$20,000 per year, the third equation becomes $T = 55{,}000/\$D$ because (1000) (20,000)/365 = 55,000.

The judicial equation— $T = 165{,}000/\$J$—can also be written as $T = 165{,}000(\$J)^{-1}$. This reveals that the slope or marginal rate of T to $\$J$ is $-165{,}000(\$J)^{-2}$, which follows from the related equation of $Y = aX^{b}$, where the slope of Y to X is baX^{b-1}. Similarly, the slope of T to $\$P$ is $-98{,}400(\$P)^{-2}$ and the slope of T to $\$D$ is $-55{,}000(\$D)^{-2}$. With this information we can solve for $\$J$, $\$P$, and $\$D$ in a set of three simultaneous equations to allocate our \$1,000,000 budget optimally:

$$-165{,}000(\$J)^{-2} = -98{,}400(\$P)^{-2}$$
$$-98{,}400(\$P)^{-2} = -55{,}000(\$D)^{-2}$$
$$\$J+\$P+\$D=\$1{,}000{,}000$$

By solving for the three unknowns we are equalizing the marginal rates of return across the three activities so that nothing can be gained by shifting dollars among activities. Simultaneously, spending does not exceed the total budget.

A defect in the preceding analysis is that it fails to consider the overlapping effect of the time consumed by judges, prosecutors, and defense attorneys. We need an equation that provides a good fit to our data and shows the average time consumed for various combinations of judges, prosecutors, and defense counsel. This equation would show the time consumed when there are different numbers and combinations of judges, prosecutors, and public defenders. Placing this data into a log-linear regression analysis would generate an equation of the form $T = a(\$J)^{b1} (\$P)^{b2} (\$D)^{b3}$. Applying the same rule for finding the slope of Y to X, this equation indicates the slope of T to $\$J$ is $(b_1)(a)(\$P)^{b2} (\$D)^{b3} [\$J]^{b1-1}$. Accordingly, the slope of T to $\$P$ is $(b_2)(a)(\$J)^{b1} (\$D)^{b3} [\$P]^{b2-1}$, and the slope of T to $\$D$ is $(b_3)(a)(\$J)^{b1} (\$P)^{b2} [\$D]^{b3-1}$. With this new information, we can more meaningfully determine an optimum mix among $\$J$, $\$P$, and $\$D$ by setting the three slopes equal to each other in the first two equations, setting the sum of the three unknowns equal to \$1,000,000 in the third equation, and solving for the three unknowns. The multivariate approach is actually simpler because it shows the optimum mix procedure and then allocates to each i activity in accordance with the equation $b_iG/(b_1 + b_2 + b_3)$, where b_i is the elasticity coefficient or exponent of activity i, and G is the total available for allocation.[36]

Optimum mix analysis can also determine the maximum average time consumption that the system can tolerate and the minimum total expenditures for the three activities to achieve that time consumption. In other words, optimum mix analysis can include determining the optimum mix that will minimize expenditures subject to a maximum time constraint or that will minimize time subject to a maximum expenditure constraint. The expenditure minimization approach requires solving for $\$J$, $\$P$, and $\$D$ in the following equations by using the slopes that do not consider the overlapping interaction among the activities:

$$-165{,}000(\$J)^{-2} = -98{,}400(\$P)^{-2}$$
$$-98{,}400(\$P)^{-2} = -55{,}000(\$D)^{-2}$$
$$165{,}000/\$J + 98{,}400/\$P + 55{,}000/\$D = 120$$

The first two equations result in an optimum mix in which each activity is given the dollar amount that at those three points equalizes the marginal rates of return. The third equation indicates that the sum of the three separate time consumptions should be less than 120 days, if that is the maximum time we are willing to tolerate for the average case. A more meaningful third equation would be in the form $a(\$J)^{b1}(\$P)^{b2}(\$D)^{b3} = 120$. If we use the more meaningful multivariate equation to express the maximum time constraint, we should also use the more meaningful slopes that are based on that equation in the first two equations. Wherever a $\$J$, $\$P$, or $\$D$ appears we can also substitute $\$J + M_1$, $\$P + M_2$, and $\$D + M_3$, where M represents the minimum amount to be allocated to the activity before the remaining dollars are allocated in accordance with the marginal rates of return.[37]

---

36. As an alternative to inductive processing of data at various points in time showing the number of days consumed for various combinations of judges, prosecutors, and public defenders to obtain the parameters for a multivariate regression equation, one could in part deductively reason the interrelations between changes in the budgets of judges, prosecutors, and defense attorneys. This partially deductive approach is used in Noam, *The Criminal Justice System: An Economic Model*, in MODELING THE CRIMINAL JUSTICE SYSTEM 41 (S. Nagel ed. 1977). For example, through a chain of deductions combined with empirical data from the District of Columbia, Noam finds that if an extra \$100 is allocated to the judge's budget, \$110 must be allocated to the prosecutor's budget to maintain the existing amount of trials per judge per time period. In other words, 10 percent more prosecutors would be needed to keep the increased pool of judges as busy as the previous pool. The multivariate regression equation approach determines the required increase or decrease in $\$P$ for a \$100 increase in $\$J$ by simply solving for $\$P$ in the equation when: (1) T and $\$D$ are set like the previous T and $\$D$, (2) the a's and b's are those from the previous computerized regression analysis, and (3) $\$J$ is increased by \$100.

37. One could arrive at the minimum figure for each activity by analyzing data from many court systems and finding the minimum number of dollars spent on judging, prosecuting, and defending cases. One could then multiply these figures by the estimated case quantity for the next

The optimum allocation can also be based on a composite goal in which delay reduction is just one of the variables to be optimized.[38]

3. *Optimum Choice Analysis*

Both optimum level and optimum mix analyses require working with an optimizing variable that has a continuum of categories similar to our monetary system. Optimum choice analysis, in contrast, uses a variable that has discrete categories such as yes-no or do-don't. This analysis might be valuable in analyzing what judges, prosecutors, and defense counsel should do to achieve settlements and reduce service time and average time consumed per case.[39]

Optimum choice analysis operates on the assumption that when individuals choose an action they are implying that they expect the benefits minus the costs of the chosen action to be greater than the benefits minus the costs of the rejected actions. The expected benefits equal the benefits to be received from an action discounted or multiplied by the probability that the event upon which the benefits are contingent will occur. The expected costs equal the costs of an action discounted by the probability that contingent events will occur.

Figure 4 represents the general decision theory often involved in optimum choice analysis. It shows prosecutors how to accelerate slow and difficult cases so that they will not exceed a maximum time threshold. These methods include: increasing the benefits and decreasing the

---

year of the court system to arrive at the three minimums. The minimum figures could also be found by taking the lowest monetary allocation each activity received during the year or by asking knowledgeable people to estimate the minimum cost of each activity and averaging their responses.

Note, however, that each division will likely receive more than the minimum if: (1) the court system spends its total budget, (2) the sum of the minimums is less than the total budget, (3) each activity has a favorable marginal rate of return because increased expenditures resulted in time reduction, and (4) the reduction benefits tend to taper off making it efficient to switch expenditures from the most generally efficient activity to a less efficient activity when a substantial decrease in benefits occurs. In other words, if assumptions 1, 2, and 3 are met, then the most efficient allocation, after satisfying the minimums, is to allocate the remainders to the activities in proportion to their exponents in an equation of the form $(\$J)^{b1} (\$P)^{b2} (\$D)^{b3} = 120$.

38. For additional examples of optimum mix analysis applied to the legal process, see W. HIRSCH, THE ECONOMICS OF STATE AND LOCAL GOVERNMENT 217-54 (1970); S. NAGEL, MINIMIZING COSTS AND MAXIMIZING BENEFITS IN PROVIDING LEGAL SERVICES TO THE POOR (1973); D. SHOUP & S. MEHAY, PROGRAM BUDGETING FOR URBAN POLICE SERVICES (1971).

39. On optimum choice analysis, see R. BEHN & J. VAUPEL, ANALYTICAL THINKING FOR BUSY DECISION MAKERS (1978); R. MACK, PLANNING ON UNCERTAINTY: DECISION MAKING IN BUSINESS AND GOVERNMENT ADMINISTRATION (1971); H. RAIFFA, DECISION ANALYSIS: INTRODUCTORY LECTURES ON CHOICES UNDER UNCERTAINTY (1968); S. RICHMOND, *supra* note 3, at 301-60.

costs of making time-saving decisions; decreasing the benefits and increasing the costs of making time-lengthening decisions; and increasing or decreasing the probabilities of relevant contingent events. To

FIGURE 4. OPTIMUM CHOICE ANALYSIS APPLIED TO TIME CONSUMPTION

ALTERNATIVE DECISIONS	ALTERNATIVE OCCURRENCES: Being Penalized for Lengthening Time (P)	ALTERNATIVE OCCURRENCES: Not Being Penalized for Lengthening Time (1-P)	BENEFITS MINUS COSTS
Time-Saving Decision (S)	B_S Benefits from S	C_S Costs from S	$B_S - C_S$
Time-Lengthening Decision (L)	C_L Costs from L	B_L Benefits from L	$(B_L)(1-P) - (C_L)(P)$

Abbreviations: P = probability of being penalized. B = benefits. C = costs. S = time-saving decision. L = time-lengthening decision.

To increase the likelihood that prosecutors will make time-saving decisions:

1. Increase the benefits of making time-saving decisions (*i.e.*, increase B_S).

   For example, reward prosecutors with salary increases and promotions for reducing the average time consumption per case.

2. Decrease the costs of making time-saving decisions (*i.e.*, decrease C_S).

   For example, establish a computerized system that informs prosecutors of actual and predicted times at various stages for all cases to minimize the problems of keeping track of cases and to provide more investigative and preparation resources.

3. Increase the costs incurred of making time-lengthening decisions (*i.e.*, increase C_L).

   For example, provide under the speedy trial rules for absolute discharge of defendants whose cases extend beyond the time limit.

4. Decrease the benefits of making time-lengthening decisions (*i.e.*, decrease B_L).

   For example, increase releases on recognizance so that lengthening the pretrial time will not make the jailed defendant more vulnerable to pleading guilty.

5. Raise the probability that the decisionmaker will be penalized for lengthening time (*i.e.*, increase P).

   For example, allow fewer exceptions to the speedy trial rules.

encourage favorable time consumption decisions, prosecutors can be given monetary rewards (to increase the benefits) and work-saving resources (to decrease the costs). An alternative method of encouraging favorable time consumption decisions is to punish prosecutors by providing for an absolute discharge of defendants whose prosecution has been excessively delayed or by depriving prosecutors of the plea bargaining benefits of lengthy pretrial incarceration. These devices may incur substantial speed-up costs, which may outweigh the delay costs, but it stimulates speculation about how the system can influence decisionmakers to make time-saving decisions.

A similar optimum choice analysis could be applied to the decisions of public defenders or private defense attorneys. The suggestions for encouraging time-saving decisions in these contexts may, however, conflict with the suggestions applicable to the prosecutor. For example, we might recommend more pretrial release to decrease the benefit the prosecutor receives by holding a defendant in jail, which facilitates obtaining a guilty plea. We might, however, recommend less pretrial release to increase the cost to the defendant of delaying his case. We must resolve this conflict on the basis of criteria other than time-savings. This analysis also stimulates benefit-cost suggestions applicable to the defense that do not conflict with the previous suggestions applicable to the prosecution. For example, providing additional monetary rewards and resources to public defenders does not conflict with the suggestions for improving the prosecutor's efficiency unless there is a fixed quantity of resources in the system.

Judicial decisions affecting time consumption also are subject to optimum choice analysis. For example, judges now incur virtually no personal costs by granting repeated continuances or making other delaying decisions. If, however, the system kept public records of the length of time each judge takes to process various cases, this visibility might cause slow judges to decrease their time delaying decisions rather than incur the cost of adverse peer pressure or publicity. Similar records of prosecutors and public defenders in a court system or across court systems can also be kept.[40]

---

40. For additional examples of choice theory applied to the legal process, see S. NAGEL & M. NEEF, DECISION THEORY AND THE LEGAL PROCESS (1978); Stover & Brown, *Reducing Rule Violations by Police, Judges, and Corrections Officials*, in MODELING THE CRIMINAL JUSTICE SYSTEM 297 (S. Nagel ed. 1977); Nagel, Neef, & Schramm, *Decision Theory and Pretrial Release Decision in Criminal Cases*, 31 U. MIAMI L. REV. 1433 (1977).

II. Descriptive or Predictive Models

The following descriptive or predictive time-oriented models primarily predict future consequences from changed procedures or variables that are not necessarily related to time reduction. They are time-oriented models only in the sense that they forecast future events on the assumption that we can not prescribe methods to reduce delay unless we can predict those events. Like optimizing models, which have predictive aspects, predictive models often provide essential inputs for prescription or optimization. As before, these models will be illustrated with legal process examples.

A. *Markov Chain Analysis*

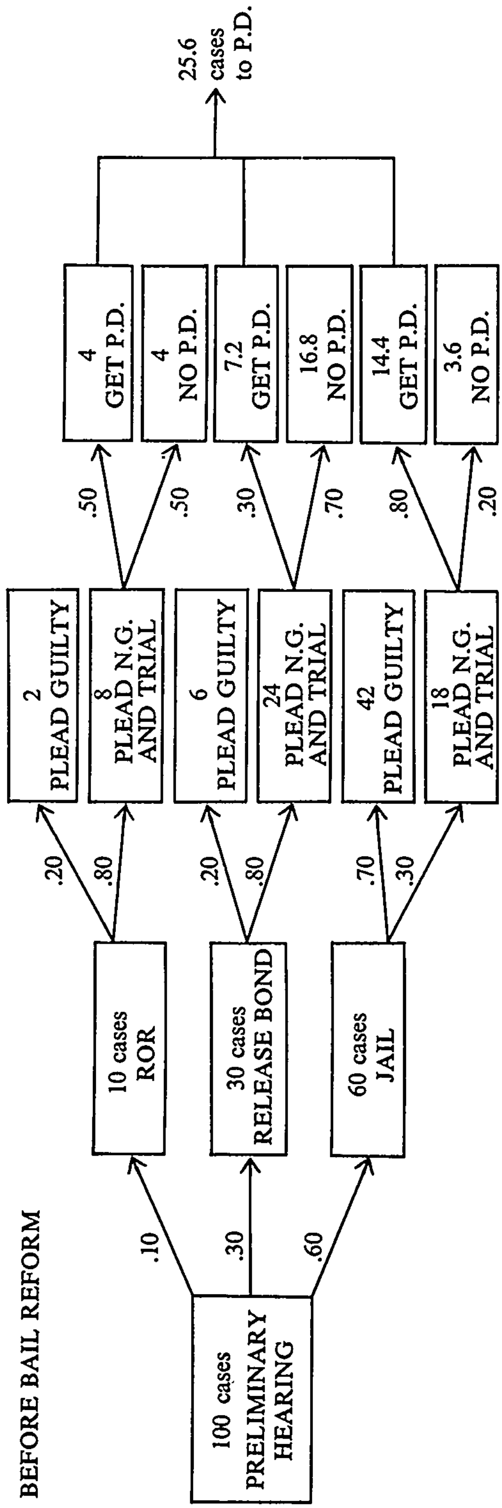
Markov chain analysis predicts subsequent events by determining the probability that one event will follow another. A simple example is illustrative: If we know that sixty percent of the convicted defendants in a court system are imprisoned and forty percent receive probation, we can roughly predict the effect of increasing convictions from 200 to 300 per year. Specifically, before the increase the prison caseload was 120 per year (.60 × 200), and the probation caseload was eighty cases per year (.40 × 200); after the increase the prison caseload probably will be 180 cases per year (.60 × 300), and the probation caseload, 120 cases per year (.40 × 300). This simple example would be much more interesting if it were part of a chain of branching events in which a change in an early event has a domino effect on the branch events.[41] The application of Markov chain analysis to the preceding illustration allows prediction of the effects of the conviction rate increase on those branch events.

Figure 5 illustrates the use of Markov chain analysis to predict the effect on a public defender's caseload of a change in the probability that a defendant will be held in jail pending trial. In this example of 100 defendants entering the system, ten percent were released on their own recognizance, thirty percent were released on bond, and sixty percent were held in jail pending trial. Markov chain analysis reveals the net increase or decrease in the public defender's caseload if the probabilities are changed respectively to .40, .20, and .40 (*e.g.*, as a re-

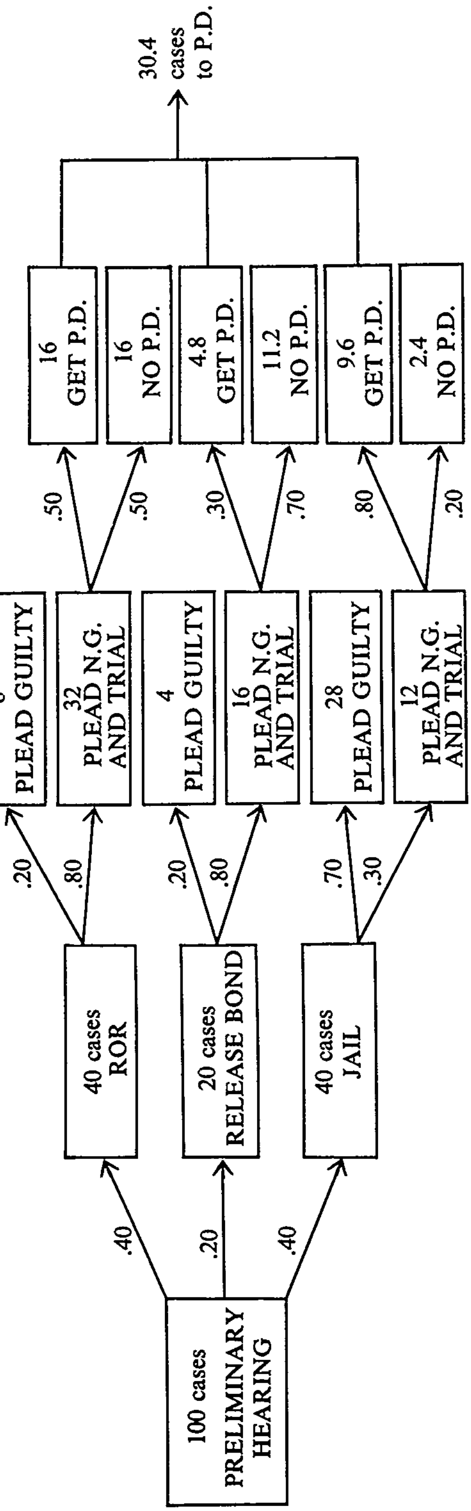
41. On Markov chain analysis, see D. Isaacson & R. Madsen, Markov Chains Theory and Applications (1975); J. Kemeny & J. Snell, Finite Markov Chains (1960); S. Richmond, *supra* note 3, at 439-60; Ulmer, *Stochastic Process Models in Political Analysis*, in Mathematical Applications in Political Science (J. Herndon & J. Bernd eds. 1971).

FIGURE 5. APPLYING MARKOV CHAIN ANALYSIS TO PREDICTING THE EFFECT ON PUBLIC DEFENDER CASELOADS OF CHANGES IN PRETRIAL RELEASE

BEFORE BAIL REFORM



AFTER BAIL REFORM



sult of intervening bail reform). To be accurate, the Markov statistician first must determine the probability that the defendant will plead not guilty and, for those defendants who plead not guilty, the probability that the court will appoint a public defender to represent them.

The hypothetical data indicates that defendants in jail pending trial are more likely to plead guilty and that jailed defendants (who plead not guilty) are more likely to be indigent and thus eligible for a public defender. Among released defendants, those who post bond are less likely to qualify for a court appointed public defender than those released on their own recognizance (ROR). Further, for the purpose of this hypothetical we assume that approximately half of the ROR cases involve middle-class defendants who are ineligible for public defender representation and half involve indigent defendants who are eligible.

The analysis simply requires the allocation of the 100 cases in accordance with the first tier of probabilities to the second column of events; then allocating this outcome with the second tier of probabilities to the third column of events; and then allocating this outcome with the third tier of probabilities to the fourth column of events. The last step in the analysis requires summing the number of cases allocated to the public defender in the various rows of the fourth column of events to determine the total public defender caseload. The same methodology is applied to the "before" probabilities and the "after" probabilities. This indicates that for the 100 cases the predicted "before" caseload is twenty-five and one-half to the public defender and the predicted "after" caseload is thirty and one-half, resulting in an increase of nineteen percent in the public defender's caseload. If the public defender's office is to continue to assign the same caseload per attorney, it should increase hiring by nineteen percent.[42]

The falling dominoes approach of Markov chain analysis can be helpful even when probabilities are unavailable, provided that we

42. The preceding analysis assumed that the only percentages or probabilities undergoing change are the percentages of cases involving release on recognizance, release on bond, and jail detention. In some situations a change in the percentages on one tier can affect the percentages on another tier as well as the quantity of cases. For example, after bail reform only the relatively poor risks would remain in jail. They may be less likely to plead guilty at a .70 rate because they may be more accustomed to jail and thus are *less* vulnerable to the prosecutor's offers. On the other hand, they may be more likely to be convicted if their cases go to trial, suffering less stigma than from pleading guilty, and thus are *more* vulnerable to the prosecutor's offers. These two considerations allow us to reasonably assume the .70 guilty pleading rate for defendants in jail will remain after bail reform.

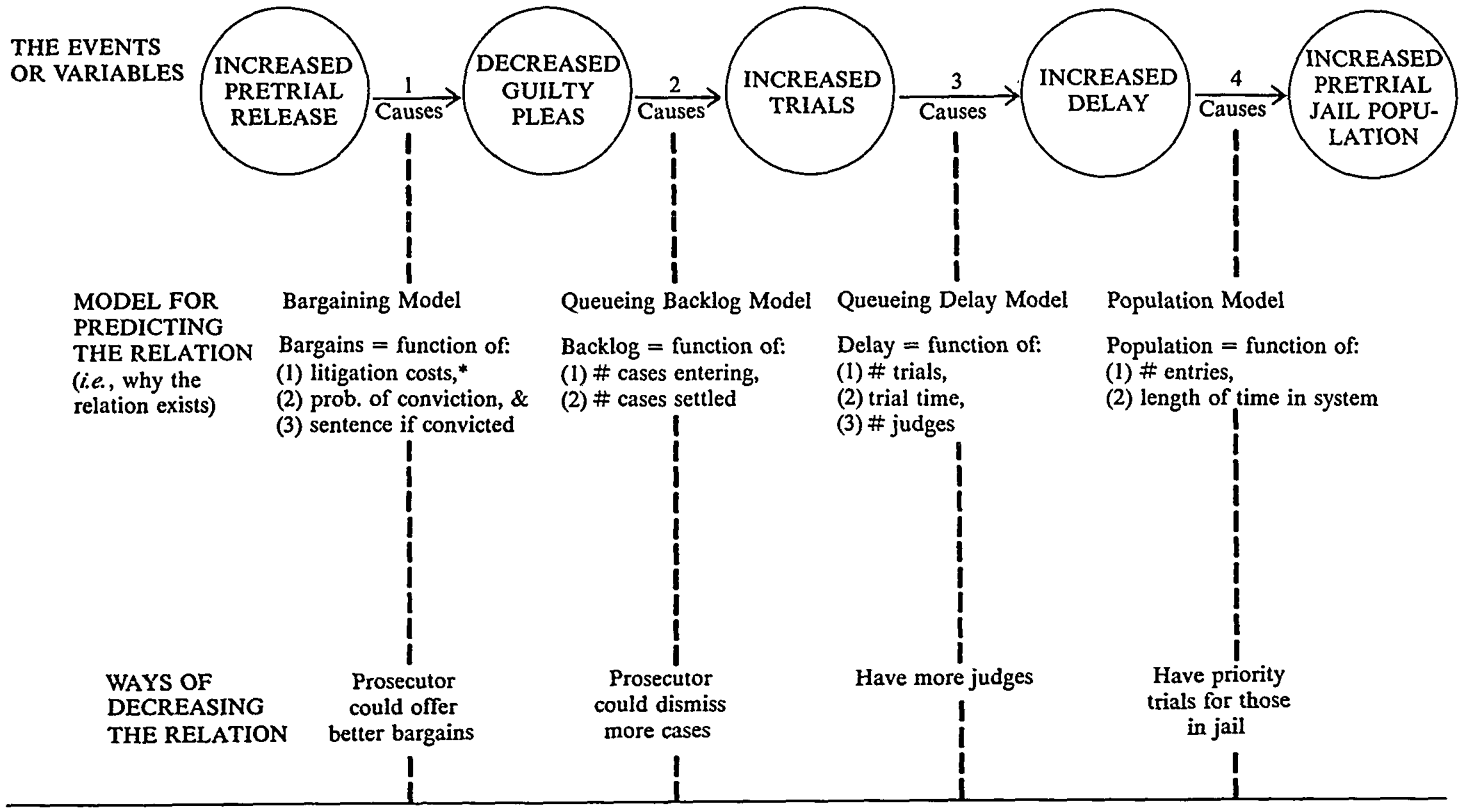
know the effect on a subsequent event of a change in a previous event. The relation between increased pretrial release rates and the size of the jail population is an excellent example of how the analysis of a chain of events can provide insights that might otherwise be missed. Using a simple causal analysis, one might logically conclude that if pretrial release rates are increased, pretrial detention population will decrease. Chain analysis reveals, however, that this conclusion is not necessarily correct. Figure 6 represents the relevant events that might occur as a result of increasing the pretrial release rate. An increase in the pretrial release rate will likely result in a decrease in guilty pleas because many defendants who plead guilty do so in return for a prosecutor's promise to recommend a sentence equal to time served or to recommend probation. Jailed defendants who demand a trial may have to be incarcerated for a longer period before trial; they may still lose their case and have to serve additional time. If many defendants who were formerly incarcerated pending trial are now released, they will be less vulnerable to prosecutorial pressure to plead guilty. A decrease in guilty pleas will likely result in a corresponding increase in trials because pleas of not guilty are essentially pleas for a trial.[43] An increase in trials will likely result in increased delay. Queueing models and common sense confirm these relationships. If, however, jailed defendants must wait longer for a trial, the initial benefits of a decreased jail population will, of course, be somewhat neutralized.

Deductive models explain the four relations or causal arrows in Figure 6. Bargaining models, which explain the first relation, show that plea bargains are a function of litigation costs and probability of conviction and sentence as perceived by both the defendant and the prosecutor. One of the defendant's most important litigation costs is time spent in jail awaiting trial.[44] Increasing the pretrial release rate lowers that litigation cost. This likely lowers the defendant's upper limit of amenability to plea bargains, making it less likely that his upper limit will be above the prosecutor's lower limit and thereby reducing the probability of a settlement. The queueing models for determining

43. *See* Lenihan, *Telephones and Raising Bail: Some Lessons in Evaluation Research*, 1 EVALUATION Q. 569, 579 (1977); Wald, *Pretrial Detention and Ultimate Freedom: A Statistical Study*, 39 N.Y.U.L. REV. 631, 633 (1964).

44. *See* C. FOOTE, STUDIES ON BAIL 722-30 (1966); D. FREED & P. WALD, BAIL IN THE UNITED STATES 39-48 (1964); W. THOMAS, BAIL REFORM IN AMERICA 110-18 (1976).

FIGURE 6. THE IMPACT OF PRETRIAL RELEASE ON THE PRETRIAL JAIL POPULATION



*including the cost of sitting in jail awaiting trial backlog sizes and the amount of delay or time consumption per case respectively explain the second and third relations. The fourth relation is a population model using as its inputs the lower rate of entry and the higher average length of time a detained defendant remains in jail.

In addition to predicting and explaining the chain of relations from pretrial release to the size of the jail population, each model suggests meaningful ways of reducing or stabilizing the occurrence of undesired relations. For example, the bargaining model suggests that the prosecutor could stabilize the rate of guilty pleas in the face of an increased pretrial release rate by making more attractive offers to released defendants. Even if the rate of guilty pleas declines, the prosecutor can still stabilize the volume of trials by increasing the dismissal rate; he is not obligated to prosecute merely because the defendant pleads not guilty. If the volume of cases increases, the amount of delay can be stabilized by providing more judges and other court personnel. If delay increases, there does not have to be a corresponding increase in the average pretrial detention time provided the prosecutor gives priority to the trials of incarcerated defendants. If data similar to the data in Figure 5 were inserted in Figure 6, we could be more precise in determining the effects of prior events on subsequent events and could suggest effective changes in the system. Even without such data, however, the chain analysis reasoning process can produce useful descriptive and prescriptive insights.[45]

B. *Time Series Analysis*

A common social science method of predicting the future from the past is time series analysis.[46] This analysis—also called longitudinal or

45. In the legal process literature there have now appeared a number of examples of Markov chain reasoning applied to predicting the probability that a given convicted defendant will reengage in criminal behavior within a certain number of time periods after being released from prison. *See, e.g.*, Belkin, Blumstein, & Glass, *Recidivism as Feedback Process: An Analytical Model and Empirical Validation*, 1 J. CRIM. JUST. 7 (1973); Rardin & Gray, *Analysis of Crime Control Strategies*, 1 J. CRIM. JUST. 339 (1973); Slivka & Cannavale, *An Analytical Model of the Passage of Defendants Through a Court System*, J. RESEARCH IN CRIME & DELINQUENCY 132 (1973); S. Deutsch, J. Jarvis, & R. Parker, A Network Flow Model for Predicting Criminal Displacement and Deterrence (1977) (unpublished paper); D. Greenberg, Recidivism as Radioactive Decay (1975) (unpublished paper); T. Schelling & R. Zeckhauser, Law and Public Policy: Policy Analysis (1975) (unpublished course materials); D. White & S. Hung Uh, Juvenile Court Records and Markov Chains (1976) (unpublished paper).

46. *See* S. KIRKPATRICK, QUANTITATIVE ANALYSIS OF POLITICAL DATA 385-509 (1974); D. LEEGE & W. FRANCIS, POLITICAL RESEARCH: DESIGN, MEASUREMENT, AND ANALYSIS 383-96 (1974); R. PINDYCK & D. RUBINFELD, ECONOMETRIC MODELS AND ECONOMIC FORECASTING

over-time analysis—involves obtaining and processing data on one or more variables for many points in time. It is generally contrasted with cross-sectional or over-space analysis where data is obtained and processed over one or more variables for many places at a single point in time. The processing of the data in either of the analyses tends to emphasize the development of linear regression equations of the form, $Y = a + BX$, or nonlinear regression equations of the form, $Y = aX^b$, where Y is the variable being predicted to, and X is the variable being predicted from. The a and b coefficients are determined by computing the Y and X scores or their logarithms for each time point or for each place point along with a regression analysis program.[47]

There are many ways to classify time series analyses. One useful method is to categorize the analysis as univariate, bivariate, or multivariate. Univariate analysis uses information only on the Y variable; the X variable simply consists of consecutive numbers corresponding to the time periods. For example, if we program nationwide crime scores for the preceding ten years as variable Y, the numbers one through ten as variable T, and a regression analysis program, the computer will provide numerical values for the a and b parameters of the equation: Crime $= a + b$ (time period). We can use this equation to predict the amount of crime in the eleventh or twentieth time period by simply inserting these figures into the parentheses, multiplying by the value of b, and adding the value of a. This, in effect, extends a trend line to the ten time data points. If we suspect that crime is not increasing at a constant rate, but rather at an increasing or decreasing rate, we can program the logarithms of crime scores for the preceding ten years, the logarithms of the numbers one through ten, and the same regression analysis program. The computer will then provide numerical values for the a and b parameters of the equation: Crime $= a$ (time period)^b. A numerical value of b greater than 0 but less than 1 indicates crime is

(1976); S. WHEELWRIGHT & S. MAKRIDAKIS, FORECASTING METHODS FOR MANAGEMENT (2d ed. 1977); Russett, *Some Decisions in the Regression Analysis of Time-Series Data*, in MATHEMATICAL APPLICATIONS IN POLITICAL SCIENCE (J. Herndon & J. Bernd eds. 1971).

47. In cross-sectional analysis, the units analyzed are *places* that have a variety of characteristics or variables. In time series analysis, the units analyzed are *time points* that have a variety of characteristics or variables. One kind of cross-sectional analysis has a time element and uses places as the units of analysis. Some variables, however, are expressed in terms of change over time. For example, the places might be cities that have NAACP chapters and the variables might be changes in housing discrimination. *See* S. NAGEL & M. NEEF, THE APPLICATION OF MIXED STRATEGIES: CIVIL RIGHTS AND OTHER MULTI-POLICY ACTIVITIES (1976); Bohrnstedt, *Observations on the Measurement of Change*, in SOCIOLOGICAL METHODOLOGY 113-33 (E. Borgatta ed. 1970).

increasing at a diminishing rate; a numerical value of b greater than 1 indicates crime is increasing at an increasing rate. The computer, using both the linear and nonlinear approach, will also yield the percentage of the Y variation that is explained by the T variation, indicating which approach has the greater explanatory power.

Bivariate analysis uses information on the Y variable and an X variable that is not a time period counter. For example, the Y variable might be murders for each year since 1900, and the X variable might be executions for each of the same years. The computer analysis would then yield equations of the form: murders $= a_1 + b^1$ (executions), if we are predicting murders from executions; or executions $= a_2 + b_2$ (murders), if we are predicting the converse. We might achieve a more precise result (*i.e.*, account for more of the variance on the dependent variable) by using a lagged independent variable. To predict murders from executions using this kind of variable might mean that for each data pair the Y score is at t and the X score is at t-1. This would yield numerical values for the parameters of the equation, murders$_t = a + b$ (executions$_{t-1}$), on the assumption that it takes about one year for executions to influence the subsequent murder rate. One can experiment with a two-year or longer lag to determine which lag accounts for the greatest change on the murder variable.

If, for example, we suspect that the unemployment rate might influence the relation between executions and murders, then we can split the seventy-seven data pairs into two subsets. The subsets consist of murder-years in which unemployment was relatively low and in which unemployment was relatively high. We should then obtain two separate regression equations of the form murders $= a + b$ (executions) for each subset of data. The multivariate alternative approach enters scores for each year on murder, execution, and unemployment. The computer then gives numerical values to the parameters in an equation of the form: murders $= a + b_1$ (executions) $+ b_2$ (unemployment rate). A future value for executions (based on zero if executions are abolished) and a future unemployment rate can then be inserted. Solution of the formula yields a prediction of the future quantity of murders. Multivariate analysis can also process similar additional variables and those that are lagged to show delayed relations or logged to show nonlinear relations.

Another classification of time series analyses relevant to legal process research is based on whether the independent variable—often a policy variable—undergoes continuous change or a single interruption. An example of a policy variable undergoing continuous change is the number of executions over the last seventy-seven years. The Connecticut speed crackdown of the 1960's and the British adoption of compulsory breathalyzer testing, both of which used highway accidents as the dependent variable in the time series analysis, are examples of policy variables undergoing single interruptions.[48] Interrupted time series may show a strong noncausal relation between the *X* policy and the *Y* suspected effect. The noncausal relation may exist because: (1) *Y* changed due to a third variable; (2) *Y* recently increased or decreased before the policy change and was merely regressing to its average position; (3) *Y* was already moving upward or downward and did not change its rate or direction when *X* changed; and (4) *Y* fluctuates over time and its fluctuation when *X* changed was part of its normal pattern. These alternative explanations must be eliminated before one can conclude that the change in *X* caused the change in *Y*.[49]

Time series analysis is also classified according to whether its purpose is: (1) to describe graphically how a variable changes over time; (2) to predict future scores on a variable from knowledge of the time period (*T*), its prior *Y* score, or a score on another *X* variable; or (3) to determine the causes of change over time on the *Y* variable. The causal analysis may use a variety of methods and alternative explanations. For example, one might try to determine which of two relations is valid or stronger: the extent that easing divorce laws causes increases in divorce rates, or the extent that increases in divorce rates cause divorce laws to be eased. This determination may require holding a third variable (*e.g.*, urbanism) statistically constant in the multivariate relation and may also require looking for joint causation relations. Special forms of bivariate regression analysis, multivariate regression analysis

48. *See generally* Ross, Campbell, & Glass, *Determining the Social Effects of a Legal Reform: The British "Breathalyser" Crackdown of 1967*, in LAW AND SOCIAL CHANGE (S. Nagel ed. 1970); Campbell & Ross, *The Connecticut Crackdown on Speeding: Time-Series Data in Quasi-Experimental Analysis*, 3 LAW & SOC. REV. 33 (1968).

49. These alternative explanations are less likely to be viable if *X* changes frequently and *Y* consistently undergoes a corresponding change. For materials specifically dealing with interrupted time series, see G. GLASS, V. WILLSON, & J. GOTTMAN, DESIGN AND ANALYSIS OF TIME-SERIES EXPERIMENTS (1975); QUASI-EXPERIMENTAL APPROACHES (J. Caporaso & L. Ross eds. 1973); Campbell, *Reforms as Experiments*, 24 AM. PSYCHOLOGIST 409 (1969). Some legal policies undergo neither continuous change nor periodic interruptions. *E.g.*, the effect on oil productivity of changes in the oil depletion allowance. *See* J. Bond, A Longitudinal Analysis of the Effects of the Oil Depletion Allowance: Empirical Evidence to Resolve Conflicting Evaluations (unpublished paper presented at the meeting of the Midwest Political Science Association, at Chicago, Ill., 1976).

(two-stage least-squares analysis), and interrupted time series may be used to separate the effects of divorce laws on divorce rates from the effects of divorce rates on divorce laws.[50]

A computerized time series analysis not only gives a and b coefficients and percentages of variance accounted for (R^2), but also numbers that can be compared to probability tables to determine the probability that a, b, and R^2 can be as large as they are, given the number of time points, when they might actually be zero due to a distorted sample. This kind of testing for statistical significance is virtually the same in regression analysis over time points or over many places at one time point. Time series differs, however, because it is more likely to involve autocorrelation, which can disrupt the probability calculations.[51] To determine if autocorrelation is present, each time point must be given a residual score (Y'_t), which equals its actual score (Y_t) minus its predicted score ($\hat{Y}_t$), and a lagged residual score (Y'_{t-1}) equal to $Y_{t-1} - \hat{Y}_{t-1}$. One then inputs these Y'_t and Y'_{t-1} scores into the computerized regression analysis for each time point to determine whether a predictive relationship exists. If the residual scores correlate highly, adjustments must be made in the original regression equation ($Y_t = a + bX_{t-1}$) by lagging, logging, using other transformations, or adding additional variables to reduce the autocorrelation so that the chance probability calculations will be more meaningful.

Various research examples from the legal process field illustrate how time series analysis can clarify predictive relations that would be unclear under an analysis of many places at one point in time. It would, for example, be more meaningful to compare judicial behavior in Missouri before and after the 1940 transition from an elected to an ap-

50. On causal analysis with time series or cross-sectional data, see D. HEISE, CAUSAL ANALYSIS (1975); Nagel & Neef, *Causal Analysis and the Legal Process*, in RESEARCH IN LAW AND SOCIOLOGY (R. Simon ed. 1977); note 45 *supra*.

51. Autocorrelation refers to the extent that the lagged residual scores of the persons, places, or other units of analysis correlate with each other. A residual score is the difference between one's actual score and one's predicted score. For example, the actual crime score for a given city in 1970 might be 60 on a scale from 0 to 100. The predicted crime score based on city size might be 80. Thus the city has a residual score of -20, meaning the actual score is 20 below the predicted score. A lagged residual score is a residual score for the city at an earlier point in time. For example, using a two-year lag, the lagged residual score for 1970 would be the residual score for 1968, which might be -15. If each residual score for the city and its two-year lagged residual score differs by 5, perfect autocorrelation is present. If, however, those relations are highly random, no autocorrelation is present. The presence of autocorrelation reduces the certainty that the slope of the relation between crime and city size (or other relations) is not due to chance sampling probability. On autocorrelation and other aspects of time series analysis, see M. BRENNAN, *supra* note 3.

pointed judiciary than it would be to compare the appointed judiciary of Missouri with the elected judiciary of Illinois at the same time point. The former is more meaningful because other variables that might affect judicial behavior remained constant, whereas the effects of these variables across state lines were unrelated.[52] Similarly, while it is probably not meaningful to compare the murder rates of Michigan, which has abolished capital punishment, with Mississippi, which has not, it might be meaningful to compare Michigan's murder rates before and after it abolished capital punishment. Instead of an interrupted time series, one could undertake a continuous time series relating homicides to executions for many or all states.[53]

The relation between anticrime expenditures and crime occurrence provides an excellent illustration of the value of time series in clarifying legal process relations. If crime and expenditure figures for many cities at one point in time are entered into a regression analysis, the resulting regression equation: crime = $a + b$ (expenditures) will have a positive b coefficient. This implies, contrary to common sense notions, that as we increase anticrime expenditures the crime rate increases. This spurious positive relation tends to remain despite attempts to lag or log the expenditure variable or to hold other variables constant. If, however, the crime and expenditure scores for many points in time for each city are entered into a computer, a negative b coeffficient is somewhat more likely to occur. This result implies that as we increase anticrime expenditures the crime rate decreases.

The corresponding coefficient or slope is also more likely to be negative if those variables that may simultaneously influence crime and expenditures are held constant over time. One meaningful way to do this

52. *See* S. NAGEL, COMPARING ELECTED AND APPOINTED JUDICIAL SYSTEMS (1973); R. WATSON & R. DOWNING, THE POLITICS OF THE BENCH AND BAR: JUDICIAL SELECTION UNDER THE MISSOURI NON-PARTISAN COURT PLAN (1969).

53. *See generally* H. MATTICK, THE UNEXAMINED DEATH: AN ANALYSIS OF CAPITAL PUNISHMENT (1965); THE DEATH PENALTY IN AMERICA (A. Bedau ed. 1967); Ehrlich, *The Deterrent Effect of Capital Punishment: A Question of Life and Death*, 65 AM. ECON. REV. 314 (1975).

In determining the effects of reapportionment, a before and after analysis is more meaningful than an analysis between states that are malapportioned and those that are properly apportioned. The across-states or cross-sectional approach finds that apportionment has no effect on the legislative output, probably because many other causal variables cannot be meaningfully held constant when comparing two sets of states at the same point in time. The over-time approach, however, reveals changes in legislative output before and after reapportionment when reapportioned states are compared with other states over the same time period. *See* Bicker, *The Effects of Malapportionment in the States: A Mistrial*, in REAPPORTIONMENT IN THE 1970'S (N. Polsby ed. 1971); Cantrall & Nagel, *The Effects of Reapportionment on the Passage of Nonexpenditure Legislation*, in DEMOCRATIC REPRESENTATION AND APPORTIONMENT (L. Papayanopoulos ed. 1973).

is through cross-lagged panel analysis. By entering into a computer for each city studied a crime score for each year (Y_t), an expenditure score for each prior year (X_{t-1}), and a crime score for each prior year (Y_{t-1}), we can obtain numerical values for the coefficients in the equation: $Y_t = a + b_1(X_{t-1}) + b_2(Y_{t-1})$. Crime at time t is related to expenditures at time t-1 because crime rates tend to reflect anticrime expenditures in the previous year rather than the present year. Crime for the previous year is, in effect, held constant in this equation because it tends indirectly to hold constant the variables, other than anticrime expenditures, that influence previous crime and thus probably influence present crime.[54] Logarithms of these three variables in the regression analysis can also be used to consider whether anticrime expenditures and the other variables have a diminishing returns relation with crime rates. This technique produces the numerical values for the coefficients or parameters in the equation $Y_t = a(X_{t-1})^{b1}(Y_{t-1})^{b2}$.[55] In this non-

54. For a discussion of the relative predictive power of equations of the form, $Y_t = f(Y_{t-n})$, as compared to $Y = f(X_1, X_2, X_n)$, see Goldman, Hooper, & Mahaffey, *Caseload Forecasting Models for Federal District Courts*, 5 J. LEGAL STUD. 201 (1976). One usually can predict a variable more accurately from the same variable at a prior point in time than from other variables. Predicting from other variables, however, may provide a better understanding of the cause of fluctuations in the variable and an understanding of the methods that will favorably influence those fluctuations. By developing a regression equation to predict Y_t from Y_{t-1}, Y_{t-2}, and so on down to Y_1 (or from X_{t-1} down to X_1), one performs a kind of prediction similar to Markov chain analysis, especially if the Y variable that one seeks to predict is scored zero or one for absent or present and can be treated like a probability. A form of regression analysis called probit analysis explicitly works with Y variables that are probabilities. This procedure is called cross-lagged panel analysis because it examines the relationship between crime and expenditures now and at an earlier point in time (lagged analysis), with over-time data (panel analysis), and across two variables. If one compares: (1) the relationship between crime and both prior expenditures and prior crime with (2) the relationship between expenditures and both prior crime and prior expenditures, one can find the degree of reciprocal causation between crime and anticrime expenditures. The first relationship indicates the extent to which crime is influenced by anticrime expenditures, holding prior crime constant; the second relationship indicates the extent to which anticrime expenditures are influenced by crime, holding prior expenditures constant. On the use of time series to analyze reciprocal causation, see Heise, *Causal Inference from Panel Data*, in SOCIOLOGICAL METHODOLOGY (G. Bornstedt ed. 1970); Nagel & Neef, *Causal Analysis and the Legal Process*, in RESEARCH IN LAW AND SOCIOLOGY (R. Simon ed. 1978).

55. Deserialization is an alternative method of obtaining numerical values for the coefficients in the equation: $\text{Crime}_t = a + b(\text{Crime}_{t-1})$. This equation yields a predicted crime score ($\hat{Y}$) for each time point and then a residual or difference score (Y'), where $Y' = Y - \hat{Y}$. This score represents the amount of crime that occurred in the city in a given year that could not be explained or predicted from the previous year's crime or the social variables determining the previous year's crime. A Y' score and an expenditure score (X for each year can be entered into a computer to obtain the numerical values for the equation: $Y' = a + bX$. For this equation, the value of b is the slope or marginal rate of return of an extra dollar spent on reducing the unexplained crime

linear multivariate time series equation, the marginal rate of return in crime reduction per additional monetary unit spent is $b_1 \; A(X)^{b1-1}$, where A is the product of $(Y_{t-1})^{b2}$, Y_{t-1} is the prior year's crime score, and the coefficients are provided by the computerized regression analysis. This marginal rate of return follows from the optimum level and optimum mix analysis rule that if $Y = aX^b$, the slope of Y to X is baX^{b-1}.[56] With this information, we can allocate funds in at least three ways: (1) in proportion to the multipliers and the exponents in each city's marginal rate of return; (2) by simultaneously solving a series of equations where the marginal rates of return are set equal to each other, as in optimum mix analysis; or (3) by working with one of the newly developed nonlinear programming routines available at many computer centers. This time series analysis is thus capable of both describing the relation between crime occurrence and anticrime expenditures within each city and of being used prescriptively for more efficient allocation of scarce resources in the legal system.[57]

C. *Difference and Differential Equations*

A difference equation uses a dependent variable or variable to be predicted as of a given point in time (Y_t) expressed as a function of itself at an earlier point in time (Y_{t-n}). Each point in time is an integer, and at each time point the value of Y moves upward or downward. It can be contrasted with a differential equation in which changes in Y

---

occurrence. The deserializing approach may in some ways sound simpler than the cross-lagged panel analysis. It, however, has a defect: prior crime (Y_{t-1}) explains much of the variance in crime (Y_t) and thus virtually nothing is explained by anticrime expenditures (X). This result does not mean expenditures have no influence, but simply that the relation of crime to prior crime covers the influence of expenditures because prior crime includes the influence of prior expenditures as well as demographic and other variables.

56. *See* text accompanying note 34 *supra*.

57. Nagel & Neef, *Allocating Resources Geographically for Optimum Results*, 3 POLITICAL METHODOLOGY 383 (1976); Nagel & Neef, Optimally Allocating Anti-Crime Dollars Across U.S. Cities and Anti-Crime Activities (paper presented at the annual meeting of the Midwest Political Science Association, 1978). For additional examples of time series analysis applied to the legal process, see H. ZEISEL, *supra* note 1, at 251-62; Grossman & Sarat, *Litigation in the Federal Courts: A Comparative Perspective*, 9 LAW & SOC. REV. 321 (1975); Rossell, *School Desegregation and White Flight*, 90 POLITICAL SCI. Q. 675 (1975-1976); R. Kagan, The Business of State Supreme Courts (unpublished paper presented at the annual meeting of the Midwest Political Science Association, 1977). The Rossell article asserts that the white flight phenomenon from central cities to suburbs has been occurring over time without substantial change as a result of court desegregation orders. An analysis merely comparing cities ordered to desegregate with cities that have not been so ordered gives the appearance that white flight is greater in the former cities because of the desegregation orders.

are continuous rather than abrupt. Both difference and differential equations are distinguishable from time series regression equations; the former are deduced from the subject matter, while the latter are induced by fitting a curve to many time data points. Although they do not necessarily refer to time, this section focuses only on time-oriented difference and differential equations.[58]

An example of difference equations in the legal process is the set of equations describing the relations in plea bargaining between the first counter offer and the initial offer, any counter offer and the immediately prior counter offer, and any counter offer and the initial offer. From the nature of the subject matter, it seems logical that the first counter offer for the defendant (D_1) will be equal to the defendant's initial offer (D_0) plus an increment. It also seems logical that the increment would represent a percentage of the distance from the defendant's upper limit (L) to his initial offer (*i.e.*, the distance L-D_0). In short, the relation between the first counter offer and the initial offer of the defendant can be symbolized $D_1 = D_0 + \% (L\text{-}D_0)$. Similarly, the first counter offer for the prosecutor (P_1) is equal to the prosecutor's initial offer (P_0) minus a decrement which represents a percentage of the distance from the prosecutor's lower limit (L) to his initial offer (*i.e.*, the distance P_0-L). The relation between the first counter offer and the initial offer of the prosecutor thus can be symbolized $P_1 = P_0 - \% (P_0\text{-}L)$, where % and L for the prosecutor are unlikely to be the same as % and L for the defendant.

It follows from the preceding reasoning that the relation between any counter offer and the immediately prior counter offer for either the defendant or the prosecutor can be symbolized $F_t = F_{t-1} + \% (L - F_{t-1})$, where F represents an offer by either the defendant or the prosecutor to settle the case by reducing the jail term. The t represents the time period or the round in the series of paired offers and counter offers. The equation is sensible for the defendant because his limit is always higher than or equal to his previous offer. Thus, $\% (L\text{-}F_{t-1})$ represents a positive number of years to be added to his previous offer. The equation also is sensible for the prosecutor because his limit is always lower than or equal to his previous offer. Thus, $\% (L - F_{t-1})$ represents a negative number of years to be subtracted from his previous offer.

58. On difference and differential equations, see M. BRENNAN, *supra* note 3, at 226-45; F. CORTES, A. PRZEWORSKI, & J. SPRAGUE, SYSTEMS ANALYSIS FOR SOCIAL SCIENTISTS (1974); C. DINWIDDY, *supra* note 3, at 116-26, 133-49, 199-216.

Solving a difference equation generally means expressing Y_t in terms of Y_0 rather than in terms of Y_{t-1}. If Y can only be expressed in terms of Y_{t-1}, then to determine the value of Y_6 would entail first determining the values of Y_0 through Y_5. In the plea bargaining example, expressing F_t in terms of F_0 requires experimenting with the preceding equations by expressing F_1 in terms of F_0, expressing F_2 in terms of F_1, which is then expressed in terms of F_0, and so on until one observes a relation between F_t and F_0. This relation, $F_t = L + [(1 - \%)^t (F_0 - L)]$, can be solved by subtracting the decrement, which is in brackets, from the defendant's upper limit and adding the increment, which is in brackets, to the prosecutor's lower limit. The bracketed material will be a decrement for the defendant because his initial offer (F_0) will be lower than his limit (L) and an increment for the prosecutor because his initial offer (F_0) will be higher than his limit (L). The $(1-\%)^t$ indicates that a smaller portion of that F_0-L distance is added to or subtracted from L at each successive t stage. For example, $(1-\%)^3$ is a smaller portion than $(1-\%)^2$, assuming the splitting rate (or %) remains roughly constant from stage to stage.

This kind of analysis can be useful in a variety of ways. It can enable one to predict whether the prosecutor and defendant are likely to converge or settle. If the defendant's upper L limit is higher than the prosecutor's lower L limit, a settlement is likely. It can also enable one who knows the respective L, %, and F_0 figures to predict the number of stages or time periods that are likely to converge. That, in turn, indicates that if we want faster and surer convergence, we should seek to: (1) increase the defendant's upper L limit and initial F_0 offer, (2) decrease the prosecutor's lower L limit and initial F_0 offer, and (3) increase the splitting rate for both the defendant and the prosecutor. The analysis also leads to speculation about how L, %, and F_0 are determined. These components include the predictability of conviction and sentence and the litigation costs of the defendant and prosecutor. In addition, the analysis suggests that by increasing predictability (to avoid misperceptions) and decreasing litigation costs (to avoid coerced settlements), plea bargains will more accurately reflect trial sentences without the time and expense of trials.[59]

59. This same kind of analysis can also be applied to analyzing and improving civil out-of-court settlements. The equation $Y_t = Y_0 (1+r)^t$, where Y_t is the future value of an amount of money, Y_0 is the initial value, r is the interest rate per year, and t is the number of years, represents a good example of a differential equation that can be applied to the legal process. Suppose a plaintiff in a civil case tells a defendant that he will withdraw his lawsuit if the defendant will pay

III. SUMMARY

This article has presented a series of time-oriented models which we hope will prove useful in reducing delay or predicting future events in the legal process. Queueing theory is especially useful for reducing backlog and delay by deducing the implications of arrival and servicing rates. It also indicates the importance of decreasing the arrival rate and the number of stages and increasing the service rate and the number of processors to reduce delay and backlog. Dynamic and sequential programming contributes to the reduction of delay and backlog by determining the optimum sequence of cases and case stages. Critical path analysis and flow chart models identify stages in the legal process that are most in need of delay reduction and indicate the effects of input and parameter changes on those stages. Optimum level analysis in the form of a time-oriented model focuses on minimizing the sum of the delay costs and the speed-up costs. Optimum mix analysis enables allocation of scarce resources among different programs or groups of personnel to maximize time reduction for a given budget or to minimize expenditures under a maximum time constraint. Optimum choice analysis is helpful in developing incentives for judges, prosecutors, and defense counsel to act in a manner that promotes delay reduction.

A number of models designed to systematically predict future events

---

$3,000. The defendant surmises that if the case goes to trial five years hence and he loses (assume a 2/3 chance of losing), the plaintiff will be awarded $6,000. The defendant thus projects the case will have an expected value of $4,000. The defendant must decide whether he should pay the plaintiff $3,000 now or pay the plaintiff $4,000 in five years. Using the preceding differential equation and the current interest rate (which we assume is .06), the equation thus becomes $Y_t = 3000(1.06)^5$, and Y_t thus equals $4,015. The defendant should therefore put his $3,000 into a savings account at 6 percent, pay $4,000 to the plaintiff five years hence, and have $15 left over. This conclusion, however, is shortsighted. It fails to consider litigation costs as well as the possibility that a single defendant, by not accepting the plaintiff's offer to settle, stands to lose $3,000 if he loses the case (*i.e.*, $6,000 damages minus the $3,000 rejected offer) or to save $3,000 if he wins the case. Further, the equation does not fully consider the variable of inflation. The $6,000 damage award in five years may implicitly take inflation into account. But with greater inflation than anticipated, the $6,000 would be less valuable.

This differential equation example differs from the previous example involving the choice of a federal or state court. *See* note 25 *supra*. In the critical path example, the plaintiff needed to know the present value of a future monetary amount determined by the formula $Y_0 = Y_t(1+r)^t$. In contrast, in the present example, the defendant needs to know the future value of a present monetary amount. Both examples, however, used differential equations because: (1) they express Y at one point in time in terms of Y at another point in time; (2) they are deductively determined from the subject matter rather than inductively determined from time data points; and (3) t can be any decimal part of a year, unlike a difference equation which involves integer jumps. On difference equations applied to plea bargaining, see Nagel & Neef, *Plea Bargaining, Decision Theory, and Equilibrium Models: Part I*, 51 IND. L.J. 987 (1976); *Id. Part II*, 52 IND. L.J. 1 (1976).

have been explored. Markov chain analysis enables examination of events that cause a domino-like chain reaction through a series of branching or successive probabilities and relations. Time series analysis is a valuable tool for: (1) describing graphically how an occurrence changes over time; (2) predicting future occurrences from trends, cycles, or relations with other occurrences; (3) determining causes of fluctuations in occurrences; and (4) providing input into prescriptive or optimizing models. Difference and differential equations are also useful in prediction, in suggesting variables that are subject to manipulation for reducing delay or producing other desirable results, and in reaching cost-efficient decisions in civil and criminal settlements.

Unfortunately, the frequency of the application of these models has not been proportionate to their potentiality and has been limited to those skilled in the methods of modeling but not necessarily knowledgeable of the legal process. What may be particularly needed is an awareness of these modeling theories among practicing attorneys, judges, judicial administrators, legal researchers, and others involved in the legal process. They need not become professional modelers, but they should develop a better understanding of the potentialities and limitations of modeling approaches in order to more constructively apply them in their work. The authors hope that this article has contributed to that awareness.[60]

60. This awareness is expected to increase as a result of sophisticated delay reduction research projects being conducted by the American Judicature Society, the Federal Judicial Center, and the National Center for State Courts. An additional stimulus might be the new Court Delay Reduction Program of the Law Enforcement Assistance Administration which is making generous grants available for delay reduction research. As this article goes to press, the *American Bar Association Journal* reports: "The chairman of one of the country's largest corporations thinks that business management techniques can be used to solve some of the problems of the courts. Irving S. Shapiro, chairman of the Du Pont Company, told the Economic Club of Pittsburgh recently that he's convinced such management techniques would be practical as well as effective." *Lawscope*, 64 A.B.A.J. 1212 (1978).

The authors of this article and Nancy Munshaw of the Department of Urban and Regional Planning at the University of Illinois are contemplating the application of most of the models discussed in this article to data available on the Washington, D.C. court system. More empirical data has been compiled on that court system than on any other in the United States, but the data has been used only in flow chart analysis. *See* DISTRICT OF COLUMBIA COURTS: 1976 ANNUAL REPORT (1977); PROMIS RESEARCH PROJECT: HIGHLIGHTS OF INTERIM FINDINGS AND IMPLICATIONS (1977); REPORT OF THE PRESIDENT'S COMMISSION ON CRIME IN THE DISTRICT OF COLUMBIA (1966).

APPENDIX

GLOSSARY OF TERMS

Symbol	*Represents*
QUEUEING THEORY	
1. Basic Rates	
A	Arrival rate of cases into the system per unit of time
S	Service rate of cases completed by the system per unit of time
A/S	Ratio of cases arrived to cases serviced per unit of time
2. Time Spent	
T	Total time spent in the system by an average case
T_W	Time spent waiting by an average case before service begins
T_S	Time spent servicing an average case
3. Number in Backlog	
N	Total number of cases backed up in the system
N_W	Number of cases in backlog awaiting servicing
N_S	Number of cases in backlog being serviced
4. Other Symbols	
P	Probability of a certain amount of time spent by a case (total, waiting, or being serviced), or of having a certain number of cases in the backlog (total, waiting, or being serviced)
C	Number of processing channels

OPTIMUM SEQUENCING

P_1, P_2	Pleading time for cases 1 and 2
T_1, T_2	Trial time for cases 1 and 2
X	A characteristic of a case used to predict the amount of time cases will consume, *e.g.*, the plaintiff's or defendant's settlement offer and the kind of personal injury or crime involved

CRITICAL PATH METHOD

1. Estimated Time for Each Processing Stage

T_O	Optimistic time, *i.e.*, estimated time if things go well
T_P	Pessimistic time, *i.e.*, estimated time if things go poorly
T_L	Likely time, *i.e.*, estimated time in light of what usually happens
T_E	Expected time calculated from T_O, T_P, and T_L

2. Present Value of Future Payoff

t	A given time period or the quantity of time periods
r	The rate of return that can be obtained by depositing money in a savings account or other investment for t years

OPTIMUM LEVEL ANALYSIS

1. Relating Time and Judges

T	Time in days per average case
J	Judges, number of
a	Either: (1) time in days consumed by the average case when there is only one judge available to process cases, as in $T = a/J$, or (2) number of judges needed to reduce the average time consumed to one day, as in $J = a/T$

2. Costs

Y_1	Actual or predicted number of dollars wasted over a given number of days prior to trial for a case or an average case (delay costs)
Y_2	Actual or predicted number of dollars expended on judges in order to achieve a given number of days awaiting trial by an average case (speed-up costs)
Y	Total costs, *i.e.*, delay costs plus speed-up costs for a given number of days

3. The Parameters or Constants for Predicting the Costs from Time Consumed

A_1	Predicted number of dollars in delay costs if only one day is consumed, as in $Y_1 = A_1(T)^{b_1}$
A_2	Predicted number of dollars in speed-up costs if only one day is consumed, as in $Y_2 = A_2(T)^{b_2}$
b_1	Ratio between a percentage change in delay costs and a one percent change in the number of days consumed per average case
b_2	Ratio between a percentage change in speed-up costs and a one percent change in the number of days consumed per average case

OPTIMUM MIX ANALYSIS

P	Prosecutors, number of
D	Defenders, number of
$\$J$	Judges, expenditures for
$\$P$	Prosecution, expenditures for
$\$D$	Defenders, expenditures for
G	Total of dollars available to be allocated

M_1, M_2	Minimum number of dollars to be allocated to activities 1 and 2

OPTIMUM CHOICE ANALYSIS

$+a$	Benefits from making time-saving decisions
$-b$	Costs of making time-saving decisions
$-c$	Costs incurred in making time-lengthening decisions
$+d$	Benefits from making time-lengthening decisions

TIME SERIES ANALYSIS

Y_t	Actual score at time t
Y_{t-1}	Actual score at time $t-1$ or the immediate prior time point
$\hat{Y}_t$	Predicted score at time t
Y'_t	Residual score at time t or difference between actual and predicted score
R^2	Percentage of variance accounted for
Log	Logarithm to the base 10, *i.e.*, an exponent (x), such that 10^x equals a given number (N)

DIFFERENCE EQUATIONS

D_0, P_0	Initial offer of defendant or prosecutor
D_1, D_2	Counter offers of defendant at times 1 and 2
P_1, P_2	Counter offers of prosecutor at times 1 and 2
L	Defendant's upper limit or prosecutor's lower limit
F	Offer of either defendant or prosecutor of the sentence he is willing to settle for without a trial

%	Portion or percent of the distance from the defendant's upper limit down to his initial offer or from the prosecutor's lower limit to his initial offer

DIFFERENTIAL EQUATIONS

r	The rate of return that can be obtained by depositing money in a savings account or other investment for t years
t	A given time period or the quantity of time periods
Y_0	The initial or present value of an investment or a damage award

BASIC FORMULAS

Formula	*Represents*
QUEUEING THEORY	
1. Time Spent	
$T = 1/(S\text{-}A)$	Waiting time
$T_W = T(A/S)$	Total time
$T_S = T - T_W$	Servicing time
2. Number of Cases in Backlog	
$N = (A/S)/(1\text{-}A/S)$	Total number of cases in backlog
$N_W = N(A/S)$	Number in waiting line
$N_S = N - N_W$	Number being serviced
OPTIMUM SEQUENCING	
Avg. $T = (T_1 + \ldots + T_n)/N$, where $T = T_w + T_s$	Average time
$T = a + b_1X_1 + \ldots + b_nX_n$	Prediction of time consumed from case characteristics assuming linear or constant relations
$T = aX_1^{b1} \ldots X_n^{bn}$	Prediction of time consumed from case characteristics assuming nonlinear relations, diminishing effects, or increasing effects
CRITICAL PATH METHOD	
$T_E = (T_0 + 4T_L + T_P)/6$	Relation between expected, optimistic, likely, and pessimistic times
$P = A/(1+r)^t$	Present value of a future amount considering the interest rate and the number of time periods

OPTIMUM LEVEL ANALYSIS

1. Basic Relations

$T = a/J$	Relation between time in days per average case and number of judges, *e.g.*, $T = 1500/J$
$J = a/T$	Relation between number of judges and time in days per average case, *e.g.*, $J = 1500/T$

2. Costs

$Y_1 = A_1(T)^{b_1}$	Delay costs, *e.g.*, $Y_1 = \$5(T)^2$
$Y_2 = A_2(T)^{b_2}$	Speed-up costs, *e.g.*, $Y_2 = \$165{,}000(T)^{-1}$
$Y = A_1(T)^{b_1} + A_2(T)^{b_2}$	Total costs, *e.g.*, $\$5(T)^2 + \$165{,}000(T)^{-1}$

A_1 = (Wasted cost per day per jailed defendant) × (Percent of defendants who are jailed) + (Wasted cost per day per released defendant) × (Percent of defendants released)
E.g., $A_1 = (\$7)(.50) + (\$3)(.50) = \$5$

A_2 = (a) × (Salary per year)/(Number of days per year)
E.g., $A_2 = (1500)(\$40{,}000)/(365) = \$165{,}000$

3. Relation Between a Change in Total Costs and a Change in Time per Case
$\Delta Y/\Delta T = (b_1)(A_1)(T)^{b_1-1} + (b_2)(A_2)(T)^{b_2-1}$
E.g., $\Delta Y/\Delta T = (2)(\$5)(T)^{2-1} + (-1)(\$165{,}000)(T)^{-1-1} = 10(T) - 165{,}000/(T)^2$

OPTIMUM MIX ANALYSIS

Numerical values are generally used in the formulae below. They are based on the hypothetical data from the text, rather than using symbols for the parameters or constants. The numerical values for prosecutors and public defenders are calculated analogously to the way they were above for judges.

1. Relation Between Time and Number of Judges, Prosecutors, and Defenders

$T = 1500/J$ $\quad$ $T = 1200/P$ $\quad$ $T = 1000/D$

2. Salary per day (annual salary/365)

   \$40,000/365 = \$110   \$30,000/365 = \$82   \$20,000/365 = \$55

3. Relation Between Time and Dollars for Judges, Prosecutors, and Defenders

   $T = 165{,}000/\$J$ $T = 98{,}400/\$P$ $T = 55{,}000/\$D$

4. Relation Between a Change in Time and a Change in Dollars for Judges, Prosecutors, and Defenders

   $\Delta T/\Delta \$J = -165{,}000(\$J)^{-2}$ $\Delta T/\Delta \$P = -98{,}400(\$P)^{-2}$ $\Delta T/\Delta \$D = -55{,}000(\$D)^{-2}$

5. Equation for Relating T to \$$J$, \$$P$, and \$$D$ to Consider Overlapping Effects
   $T = a(\$J)^{b_1}(\$P)^{b_2}(\$D)^{b_3}$

OPTIMUM CHOICE ANALYSIS

$EV_S = (+a) + (-b)$	Expected net value (benefits minus costs) from making a time-saving decision
$EV_L = (P)(-c) + (1-P)(+d)$	Expected net value from making a time-lengthening decision

TIME SERIES ANALYSIS

1. Univariate and Bivariate Relations

Crime = $a+b$(Time Period)	Predicting crimes for various time periods
Murders = a_1+b_1 (Executions)	Predicting murders from executions
Executions = a_2+b_2(Murders)	Predicting executions from murders

2. Cross-Lagged Panel Analysis to Relate Crime to Expenditures

$Y_t = a+b_1(X_{t-1}) + b_2(Y_{t-1})$	Linear relation with b_1 as the marginal rate of return

$Y_t = a(X_{t-1})^{b_1}(Y_{t-1})^{b_2}$	Nonlinear relation with $b_1 a(Y_{t-1})^{b_2}$ $(X)^{b_1-1}$ as the marginal rate of return

3. Deserializing to Relate Crime to Expenditures

$\hat{Y}_t = a+bY_{t-1}$	Predicting crime at time t from crime at time $t-1$
$Y' = Y_t - \hat{Y}_t$	Defining residual or unexplained crime as the difference between actual crime and predicted crime
$Y' = a+bX$	Predicting residual crime from anticrime dollars with b as the marginal rate of return

DIFFERENCE EQUATIONS

$D_1 = D_0 + \%(L-D_0)$	Relation between first counter offer and initial offer of defendant considering the defendant's upper limit
$P_1 = P_0 - \%(P_0 - L)$	Relation between first counter offer and initial offer of prosecutor considering the prosecutor's lower limit
$F_t = F_{t-1} + \%(L-F_{t-1})$	Relation between any counter offer at time t and the immediate prior counter offer
$F_t = L+[(1-\%)^t(F_0-L)]$	Relation between any counter offer and the initial offer considering the percentage rate at which the defendant or prosecutor splits the difference between his last offer and his outer limit

DIFFERENTIAL EQUATIONS

$Y_t = Y_0(1+r)^t$	The future value at time t of an amount of money offered to settle a case now, given the interest rate per time period and the number of time periods until time t is reached