SIMULATION OF NEW YORK CITY'S FELONY ADJUDICATORY SYSTEM

Lucius J. Riccio

Department of Industrial Engineering

Lehigh University

Abstract

A simulation model is described in this paper that was part of an analysis of New York City's Grand Juries and Supreme Courts. It has been used to test the relationship between additional court resources and the length of delay in the courts and the number of defendants in detention facilities.

Introduction

Like many other states, New York has been considering the implementation of court procedures generally known as "speedy-trial" rules. The rules contemplated in New York with respect to non-homicide felony cases are the following:

- A defendant shall be dismissed if his trial has not begun within 180 days of arr t, not including defendant-caused delay; and

- A defendant shall be released on his own recognizance if he is in custody and his trial has not begun within 90 days of arrest, not including defendant-caused delay.

The State Legislature, fearful that the courts could not possibly respond quickly enough to meet the demands such rules placed on them, opted for the rules advanced by the District Attorneys' Association. The rules were identical to those above with the exception that the milestone to be reached, rather than the beginning of trial, is the "ready for trial" declaration by the District Attorney. That is, the District Attorney must be ready to go to trial within 180 days of arrest not including defendant-caused delay.

The benefits of realizing speedy trials would be very great. The cause of justice is clearly vitiated by the lengthy delays presently characteristic of felony adjudication. The prosecution's case is weakened as delays drag out court proceedings; evidence can be lost, witnesses and victims may forget important facts or may die. On the other hand, oftentimes defendants will plead guilty simply to escape the detention facilities which in general are not very pleasant places (in Manhattan, the Detention Facility is appropriately called "The Tombs"). These two aspects of the problem do not cancel each other, but rather widen the possibilities for injustice.

Other aspects favoring speedy-trial legislation are that by insuring the speedy flow of defendants through the courts, detention populations would be minimized, and a greater degree of satisfaction would be imparted to other law enforcement agencies, such as Police, who have become increasingly critical of the courts.

New York City's Felony processing system can be simply described in the following way:

Felony arrests made by the Police are brought to the Criminal Court (Lower court)[1] for arraignment, at which time the defendant is informed of the charges against him. At this time, the case may be dismissed, the charge may be lowered to a misdemeanor, a plea may be offered[2] or the case may simply pass on to the next stage. If the defendant remains in the system he may be scheduled for Criminal Court appearances which generally lead to a hearing. However, many cases are taken out of the Criminal Court by the District Attorneys to avoid a hearing. In that way, they need not expose undercover agents. That occurs in many narcotics cases. Some defendants, for their own reasons, choose to waive the Criminal Court hearing.

The next step in the procedure is the presentation of the case to a Grand Jury. In addition to those cases described above, many cases start at this stage. District Attorneys often seek an indictment before they authorize an arrest. The Grand Jury can do one of the following things:

- Vote an indictment
- Dismiss the case
- Return the case to the Criminal Court for processing as a misdemeanor.

After an indictment is filed, the defendant is arraigned in the Supreme Court (Superior Court).[3] Pleas (which can be offered at almost any stage) can

be offered here. That generally depends on the presiding judge in the arraignment part and whether or not the county has pre-trial conference (PTC) parts. The pre-trial conference parts, which are the next step in the process (where they exist) have recently been established to improve and institutionalize the plea-bargaining procedure. If the county does not have PTC parts, reappearances may be scheduled in the arraignment part to try to get a disposition without going to trial.

Following the arraignment and PTC parts are the trial parts.[4] Trial parts are designated as either Legal Aid, regular (private counsel), homicide, or narcotics. Usually a case will require a number of appearances in the trial part before the case is either disposed or made ready for trial. The length of adjournment between appearances is a function of many things, often a function of the cause of adjournment. For example, the failure of a witness to be present at the proceedings may cause an adjournment. The length of the adjournment will then be a function of the availability of that witness. Some other reasons for the variation in the length of adjournment are counsel (or the arresting officer) being on vacation and the availability of an opening

on the court's calendar.

After all trial part proceedings have been completed and the defendant(s) has neither plead guilty nor had his case dismissed, the case is placed on the "ready and pass" queue. This queue contains all cases in the part ready to go to trial. A case gets placed in this queue when both parties declare readiness for trial or the judge decides there is no reason for further delay. A trial part will try only one case at a time.

Figure 1 is a diagram of the preceding description of the Felony Processing System. The status of the system at the beginning of this study can be summarized as the following:

As of the end of 1971, approximately 40% of all felony cases city-wide required more than 180 days from the time of arrest to disposition or first trial appearance. About 65% of the felony cases in detention need more than 90 days for disposition or first trial appearance. Detention populations in city prisons were in the 120-200% capacity range. The City was contemplating building a new "tombs" to accommodate the increased detention population. The cost of the facility would be in the neighborhood of $60 million, or about $48,000 a bed,

with the benefits, at this time, to the criminal justice process at best being questionable.[5]

Planning for the Speedy-Trial Rules

Although the installation of the speedy-trial rules had been expected for nearly a year, it was not until the winter of 1972 that an analysis of Supreme Court needs was undertaken. At that time two were performed; one by the New York City Budget Bureau and the other by The Committee on Court Delay (an Ad Hoc Group formed from the major city and state agencies dealing with the courts). The two analyses were similar both in approach and in conclusions. The Budget Bureau recommended the funding of 35 new trial parts and 1 additional Grand Jury while The Committee on Court Delay suggested 30 new parts and 4.2 Grand Juries. Most of the recommended trial parts would be temporary; their purpose would be to help eliminate the backlog.[6] Judges would be "borrowed" from civil case processing to man the trial parts. The cost of the new parts and Grand Juries was set at 3.7 million by the Budget Bureau while The Committee on Court Delay projected a $12 million expenditure to fund their recommendations.

Both studies employed reasonably similar input-output techniques for each county, the number of defendants disposed of in 1971 was divided by the number of trial parts to obtain a measure of trial part productivity. They divided that figure into the total number of defendants presently awaiting a trial part appearance to arrive at the additional parts needed to eliminate the backlog. Also, the productivity measure was divided into the expected increase in cases that will reach trial parts in the coming year to find how many parts will be needed to handle the greater case load.[7] The sum of these two calculations was the recommendation for additional trial parts. The Grand Jury figures were arrived at in a similar fashion.

Certainly a massive infusion of resources will reduce the backlog and speed up processing times. However, it is questionable whether such a dramatic increase is required. The need for an additional phalanx of trial parts to reduce the present backlog may not be real. The need would be a function of the amount of delay caused by resource constraints. The number of cases pending will approximately equal the average number of cases arraigned per day times the average number of days for case disposition. The average number of days for case disposition would approximately be equal to the number of appearances

337

times the average length of adjournment between appearances. With that in mind, it is clear that diminishing returns for additional trial parts may be reached rather quickly. The addition of parts will reduce the length of time between appearances. However, there are limits to that reduction because as discussed earlier, there are other reasons besides limited court time that affect the length of adjournment. Thus a point will be reached where the addition of more trial parts would only result in smaller calendar sizes and not a shortening in the time to disposition.

The trial part productivity measure is also somewhat misleading. Productivity, as it has been defined in the mentioned studies, is a function of many things, one being the size of the backlog. Productivity would increase as the backlog increased until the backlog is large enough to maintain full calendars. After that point, productivity would not increase as greatly and that increase would probably be the result of lower plea offerings.[8] Productivity is really a function of the number of defendant-appearances per day presided over by a Judge. Thus once calendars are full, and if all else remains the same, productivity can increase little.

The number of dispositions would also be quite sensitive to the amount of input. This is because although many cases take a long time for disposition, many cases "plead out" early in the process and also because as the backlog increases, greater pressure is placed on the system to dispose of cases. Essentially what I am saying is that "productivity" is a function of many things, and that additional parts may not be the entire answer.

Finally, the input-output models do not take into account the vicissitudes of some parameters. In the past, many parts would shut down during the summer and all parts would shut down for two weeks at Christmas. This is particularly debilitating in Manhattan.

The Committee on Court Delay recommended a list of improved court procedures such as standardized adjournment durations and calendaring procedures. If these recommendations could be implemented, it is quite possible only a few new parts (or maybe none at all) would be needed. However, they did not attempt to quantify and predict the effects of such improvements or whether they could even be implemented at all.

Neither study attempted to uncover the precise functional relationship between additional court resources and

the amount of delay in the system. Neither could say that their recommendations will reduce the delay in an optimal fashion, optimal in accordance with some well-defined criterion. Clearly that must be the objective of any detailed study dealing with court delay and speedy trial rules.

The Simulation Model

Recognizing the deficiencies of those studies, The Mayor's Criminal Justice Coordinating Council (CJCC) and The Budget Bureau authorized the development of a computer simulation model of the court system to analyze specifically the problem of delay as it relates to the speedy trial rules.

The vast majority of delay is due to the wait between appearances in one processing element or another. The present amount of delay can be described by frequency distributions generated from sampling studies. Such sampling studies are snapshots of the system at a given point in time, but are not of great value in trying to predict how the delay will change with changes in system resources. Waiting time is a function of many things, one of which being the size of the backlog. Thus it was projected that a simulation model that incorporated all of the vital aspects of the system would be able to generate

from within the delay distributions associated with various levels of resource allocation.

It was decided that the model would not simulate the Criminal Court proceedings, considering the time constraints on this study. The Criminal Court is replaced in the model by a probability distribution that describes the delay from arrest to held for Grand Jury.

The model is a descrete-event simulation updating itself on a daily basis.[9] All work is accomplished on weekdays. However, weekends are included since they count toward the speedy-trial rules. On each day it schedules new arrivals, "calls the calendars" of all grand juries and parts, and schedules trials. An "arrival" is a case placed in the held for Grand Jury queue. All of the processing units function in the following way:

When a case first arrives at a station, it is assigned a priori the number of appearances it will require for a disposition at that stage. It is then scheduled into its first appearance. The number of appearances remaining is retained and with each appearance, it is reduced by one. When all appearances have been completed, the model determines if the case goes on to the next processing stage or if it leaves the

system.

Probably the most critical aspect of the model and the real system is the calendaring procedures. It is critical in that it is very difficult to ascertain precisely what is done. The philosophy behind the model's scheduling algorithms can be broken into two parts. The first places limits on the amount of cases that will be scheduled for a working day. An "opening" is considered to exist on a day if on that day fewer cases than the limit have been calendared. Some scheduling is done by assigning a case to the first available opening. First appearances are generally assigned in this way. The second aspect to scheduling recognizes the multiplicity of causes that affect the lengths of adjournment. As such, the length of adjournment is found with the aid of a probability distribution. The date chosen for the next appearance is checked to be sure that the limits mentioned above have not been violated. Most re-appearance scheduling is performed in this fashion. All processing units give priority to jail cases over bail cases and re-appearances over first appearances. All criminal justice officials queried agreed with this formulation.

The model earmarks some Grand Juries and trial parts for homicide cases. Those units may handle other cases if there are openings on their calendars. Trial parts dispose of their daily calendars and then, if a case is ready, schedule a trial.

Many simulation models have the built-in assumption of independence between stages and uniformity within stages. Because each defendant's vital data is stored by the program and "carried" from stage to stage, this model is not restricted by the assumptions of independence and uniformity. Many variables were tested for functional dependency and categorical peculiarity. Jail and bail cases were handled differently throughout the model. Also, homicide cases generally required different parameters than non-homicide cases. However, most variable-pairs were found to have virtually no discernable amount of interdependency; e.g., the number of appearances prior to trial part seemed to have no affect on the number of appearances in a trial part. Those variable-pairs found to be in some way interdependent will be mentioned later in the text.

Input to the Model

As part of their study, The Committee on Court Delay commissioned a sampling study of felony cases that had

reached disposition in the year November, 1970 through October, 1971. That study provides a significant amount of reasonably good data. The following information was generated from their study.

As mentioned earlier the Criminal Court processing would be represented by a probability distribution. Figure 2 is a graph of the time from arrest to held for Grand Jury for all cases. The model used two distributions - one for jail cases and one for bacil cases, each having the same shape as Figure 2 but with averages of about 2 weeks and 7 weeks, respectively. The distribution in Figure 2 yields an average time of 4.73 weeks from arrest to held for Grand Jury with approximately 70% of the cases requiring one month or less to reach the held for Grand Jury stage. The distribution may have improved since the time that those cases sampled went through this segment of the system. However, this distribution is representative of system performance described as "current" which in court parlance is the proverbial goal of all court administrators. Current means the court is disposing of as many cases in a month as it receives and the time to disposition is within a certain specified criteria. Backlog can then be defined as the number of cases

that cause a violation of the above constraints. The definition of backlog used by this author has been and will remain synonymous with the number of cases pending. This is consistent with the usage of the term by the other studies referenced,

The sample also provided some information on the number of appearances required for disposition at arraignment and in trial parts. Figure 3 is the frequency distribution of the number of appearances that were required prior to trial part consideration. Since there were no PTC parts in Manhattan, all of these appearances were in the arraignment part. Figure 4 is the frequency distribution of the number of appearances in trial part required for disposition or readiness for trial. Very little is known about how fixed these distributions are, whether they will change with changes to the system.

As stated earlier, the number of appearances required for disposition at a stage is assigned to a case when it first arrives at that stage. When all appearances have been completed, the model determines if the case goes on to the next stage or if it leaves the system. Since this model is not restricted by the assumption of independence, the possibility of going

to trial was viewed as a function of the number of trial part appearances. This hypothesis was "tested" using the sample data. However, there were so few cases that went to trial (3.8% of the sample) that it was quite difficult to draw conclusions. Figure 5 shows both the fraction of cases going to trial for each number of appearances and the assumed probability distribution. The plot means that, reading on the assumed probability line, a case which required 10 appearances in a trial part has a 5% chance of going to trial when its tenth appearance has been completed.

Once a case has reached trial it is assigned a length of trial. Again, there is little data on this subject. Ms. Virginia Ambrozini, a consultant to CJCC, performed a study of Supreme Court operations in the summer of 1971. Figure 6 shows the results of her study with respect to the length of trial. Figure 6 also shows an assumed distribution of length of trial. This parameter could be quite sensitive to the speedy-trial rules. Presently, some defendants plead guilty as soon as the first juror is called. (That is the landmark signifying the beginning of trial.) They plead at that time knowing they have no case; they got as far as trial possibly hoping for a lower plea offer-

ing. Approximately 20% of the trials are disposed of on the day they were begun. With the implementation of speedy-trial rules, at least as they were originally constituted, this distribution might shift. More defendants might wait until trial before pleading. After the first juror is called there is less advantage to continuing.

A very important aspect of the model is that part dealing with the length of adjournment. This is certainly an under-researched area. However, the Ad Hoc Committee's sampling study does provide some information regarding that parameter. In that study, for each case reaching the trial part stage, they recorded the date of first trial part appearance and the date of the commencement of trial, plea, or dismissal. The difference between these two dates divided by the number of appearances minus one is the average time between appearances. Figure 7 is a plot of the frequency of those average times. This distribution does unfortunately include those cases that did go to trial. Those cases include the time between their last trial appearance and commencement of trial in their averages. That could account for the four data points having over 100 days between appearances (which I considered to be outliers). This plot

was used as an approximation for the lengths of adjournment.

The length of adjournment distribution is reflective of all the causes of adjournment and the causes of variation in the length of adjournment. Embedded in it is, of course, the resource restrictions placed on the calendaring procedures. To run the model with accuracy the part of the distribution caused by resource restrictions and manifesting itself in terms of calendaring interference should be subtracted out. This is important because the lengths of adjournment will fluctuate with the level of resources. The way the subtraction will be accomplished will be by running the model with a fixed calendaring procedure and comparing the distribution obtained with Figure 7. Admittedly, this is a crude technique.

It was thought that there was a possibility that the average time between appearances might increase as the number of trial part appearances increases, the reasoning being that the same reasons that cause many appearances might also cause longer adjournments. Figure 8 is a graph of the average time between appearances versus the number of trial part appearances, It appears that there is a slight upward trend in the averages. However, note the rapid drop

in observations just as the chart tends to move upward. Since the number of observations was small, it was assumed that the number of appearances has no effect on the length of adjournment. It was important to consider this point because if there was a correlation either positive or negative, it would have an impact on the spread of the distribution of time through the trial part phase of the system.

The other aspect of calendaring discussed earlier is the maximum number of defendants a trial part will schedule for a day. Since all parts give priority to jail cases over bail cases and re-appearances over first appearances, the model uses two limits for scheduling. The lower limit applies to first appearances for bail cases and the upper limit applies to jail cases and re-appearances. Preference is also shown by attempting to schedule jail case appearances with shorter average lengths of adjournment. Unfortunately, although this is an agreeable formulation, there is little data that provides an accurate assessment of those limits. Figure 9 is a frequency distribution of trial part calendar sizes obtained from a sample, compiled by the author, taken from the listings of trial part calendars printed in The New York Law Journal. It

is a very small sample; however, a pattern is clear. A further study would probably show that the calendar size of a trial part is a function of the judge. From Figure 9 I selected 15 as the lower limit and 20 as the upper limit for all non-homicide trial parts. Of course, judges get holidays and sick days and as a result judges sit on the average only 4 out of every 5 weekdays. The model compensates for the sick days and holidays by lowering the limits by 20% to 12 and 16, respectively.

Very little of the data needed for the model is known about Grand Jury presentations. There are no statistics available concerning the number of appearances or the length of time between appearances. It is assumed, before more information can be collected, that the distribution of appearances drop off sharply after one, much like arraignment.

When building the model, the question of work units arose. What constitutes a unit of work for a trial part? For a Grand Jury? It would seem a case would be the standard unit of measurement of judicial performance. However, virtually all court statistics are presently based on the number of defendants that pass through the various processing elements. Records are kept by defendant-count for several

reasons. One is that detention facilities house defendants, not cases. Another is that many defendants may be charged under one indictment or many indictments can be charged to one defendant. As the defendant traverses the system some indictments may be dropped, others consolidated. The ratio of defendants to cases will change several times during processing in a way presently unknown. Therefore, to bypass the difficulty of the defendant to caseload conversion, the model uses defendants as the operational unit of measure.

The model calculates a figure for the number of defendants in custody awaiting Grand Jury or Supreme Court appearances. Each defendant is assigned either a jail or bail status at the held for Grand Jury stage. A remand rate of about 75% is used in the jail-bail decision. (Of course, few defendants are actually remanded. Defendants in jail are for the most part people who cannot make bail.) The remand rate of 75% is held constant in the model not because in reality it actually remains constant but because it is subject to many forces and the function that governs the remand rate is not presently precisely known. The model is completely capable of incorporating a

remand rate function and one should be included when a reasonable formulation has been arrived at. One factor to consider is the percent occupancy of the jails. It may be the case that when detention populations reach the 150-200% capacity range, judges give lower bail decisions.

Validating the Model

Before obtaining projections of future needs, the model had to be "validated". Since total validity is virtually impossible, a better description of this procedure would be "building confidence" in the model. There are two ways this is done. The first is by having confidence in the structure and input to the model. The second is by demonstrating that operating results for a controlled run reflect past experience.

The structural assumptions of this model have been continuously tested in discussions with Criminal Justice officials from many parts of the system. Most of the input to the model was acquired from The Committee on Court Delay's sampling study, a sampling of statistical repute.

Once the model was operational, it was "fine-tuned" to adjust for errors in assumptions and input for those situations where there was not enough data to provide accurate estimation. A test run was made of the year 1971. Figure 10 compares statistics obtained from the model against actual data. The distributions of times to disposition were quite similar in form to those obtained from the Court Delay Committee's sampling study, however with slightly smaller variances. The spread of the model's distributions was less for a number of reasons. Some of the inputed data is based on averages, the use of which tends to tighten rather than spread distributions. Judge productivity and calendar size were determined by an averaging process. The percentage of defendants pleading out at arraignment is a function of the judge sitting in the arraignment part. This could fluctuate much more than the model permits. The model is capable of incorporating these more accurate aspects and should include them as better data is generated.

Figure 10 demonstrates the fluctuations in the number of defendants awaiting trial part appearances during the twelve months covered by the committee's sampling study. Also on that graph is the simulation's results for the test run.

In the Fall of 1971 Manhattan added a pre-trial conference part. All cases

go through this part between arraignment and trial part. Since there is little data about this part, it was assumed that almost all cases now have only one appearance in the arraignment part and the remaining appearances prior to trial part in the PTC part. That is the total number of appearances prior to trial part is still the same as before the addition of the PTC part. The capacity of the PTC part was established, in much the same way as the trial part limits, at 30 defendants per day. In February 1972, two trial parts, designated specifically for narcotics cases, were added. In the Spring of 1972, a grand jury was added. This last addition accounted, quite naturally, for a rise in indictments and a dramatic reduction in the number of defendants awaiting grand jury action. This met with widespread approval. However, the hidden effect of adding a grand jury was to send a sudden jolt to the rest of the system. Rather than reducing the total number of defendants in the system, it simply shifted the burden to the trial parts. The sudden rise in defendants awaiting a trial part appearance is really a transient effect rather than a steady state condition. Table 1 shows the total number of defendants in the system at three points in time. Admittedly,

this is not a complete analysis; however, I think the point is clear.

	Oct 30, 1970	Oct 30, 1971	May 18, 1972
Grand Jury	530	523	298
Arraignment	682	530	322
Trial Part	2165	2198	2527
TOTAL	3377	3251	3147

Table 1

The behavior indicated above was demonstrated in a test run of the first half of 1972. That run and all others following included a 20% increase in grand jury presentations over 1971.

Using the Model

As discussed earlier the purpose of the model is to determine what is required to reduce processing delays, the backlog of cases, and the detention population. The model was run simulating 400 days beginning with September 1972 using the system conditions monitored at that time.

The model was run without any further additions of resources other than those mentioned in the last section. By September the jolt given the system in the Spring by the additional grand jury had worn off. The model projected for the months September through the first half of December 1972 a decrease in the number of defendants awaiting trial part appearances from about 3200 to 2600. The two-week shutdown in December would

boost the backing log up to nearly 2800 but during the first six months of 1973 the model projects a steady decline to about 2300, at which time the total number of defendants in the system would be just under 3000 and the detention population would be about 1670.

Although the decline in the number of defendants looks encouraging, the percentage of cases taking less than 180 days for disposition is down only to 32%. Further analysis of the results indicated two things: (1) There would be an apparent bottleneck in the PTC part; and (2) the addition of trial parts would do little towards the reduction of delay unless tighter controls were placed on the lengths of adjournment. The average length of an adjournment in the trial parts was still over 18 days and the average calendar size in a trial part was under six defendants per day.

The model was run, starting again in September, with the addition of one PTC part. This expedited case processing dramatically and reduced the backlog to about 1725 defendants and a detention population of 1150 by June of 1973. However, the percentage of cases requiring more than 180 days for disposition was still high at 26%. The reduction of that percentage from 32% to 26% was almost entirely the result of the ability

of the additional PTC part to get defendants to plead out before going on to a trial part.

The addition of the PTC part shifted some of the burden to the trial parts. However, the average calendar size per trial part went up only to 6.3 defendants per day. Thus it is clear that it is not more parts but tighter controls on adjournment lengths that are needed for a reduction in delay. To compensate for the low calendar sizes two additions to the model were made: (1) A calendar size control mechanism, and (2) an emergency scheduling algorithm for cases approaching the 180 day limit. They are explained below.

The system to a certain extent is self-regulating with respect to adjournment lengths. As the backlog drops the average length of adjournment should drop. A simple control mechanism was incorporated whose purpose was to adjust adjournment lengths so that average calendar size would not drop much less than 6. Six was chosen because previous runs of the model showed six to be a consistent figure for that parameter, and assuming new trial parts will function, at worst, like old ones, six seemed reasonable. A lower limit was placed on adjournment lengths to account for uncontrolled variables.

347

No matter what preferences judges or DA's have for size of workload, it is certain that no one will want to be given blame for allowing someone to "escape" under the 180 day rule. Thus, as a case approaches that mark, both will accept slightly larger calendar sizes resulting from shorter adjournment lengths. To reflect this, the model was set-up so that when a defendant was in the system for more than 140 days he would be allowed adjournment lengths of 5 days on the average.

The model was run with these additions, again using 2 PTC parts and 17 trial parts. In that run, average calendar size rose only to 6.6 defendants per day. The number of defendants awaiting trial part appearances plummetted to less than 1500 in the first 4 months of the simulation. However, the percentage of defendants requiring more than 180 days for disposition dropped only to 22%. Thus it was clear that concentration solely on the trial parts was insufficient because many cases were coming to their first trial part appearance having already logged in nearly 180 days. Much of the pre-trial part delay was, from the model's point of view, caused by excessive delay of bail cases in the criminal court. This was the only

leverage point left in the system.

As stated earlier the distributions used by the model for criminal court delay were probably somewhat outdated due to improved administration within that court. The model was run as described in the paragraph above with all cases using the distribution previously used for jail cases only. That is, the model was run using an average criminal court delay of 2 weeks for all cases. It is quite possible that in reality this reduction has already been realized or it could be realized without an addition of resources.

The results of that run show only 18% of defendants requiring more than 180 days for disposition. This figure is an adequate system goal from the point of view of the model for four reasons. First, not all delay is caused by the state; some is caused by the defendant (request of an additional appearance or extra long adjournment) and as such is not chargeable to the 180 day rule. Second, all indicators point to the fact that the system has enough slack to allow additional special expediting of cases near the 180 day limit. Third, the 18% figure includes homicide cases (most of which take more than 180 days) which are not covered by the rule. Fourth, the 18% figure also includes some cases

arrested before the rules went into effect. That is, the true figure for cases covered by the rule is probably between 5-12%.[11] Taking into account defendant caused delay and possibilities for additional expediting, the system resources defined in this last computer run will be adequate to meet the requirements of the rule.

Conclusions

The results obtained from the model point to the following recommendations for Manhattan:

1. Add 1 PTC part.

2. Add trial part emergency expediting mechanism for defendants in the system for more than 120 days.

3. Reduce Criminal Court delay to an average of 14 days for felony cases.

Many aspects of the model could change, such as the number of appearances required for disposition, and as such would change the results and recommendations herein described. However, these predictions are based on rather conservative data. It seems unlikely, with the low utilization rates of most resources and the urgency of the situation, that any aspect of the system will relax and cause the model's results to be overly optimistic. The only parameter change that could possibly

cause major repercussions would be the number of grand jury presentations. However, even a large increase in that variable would not cause great damage since the constraining relationships in the system are not due to lack of resources.

This model has demonstrated its usefulness in the analysis of this court "crisis". However, it can, and should be, used to test system sensitivity to various parameters and to study other aspects of court system behavior. This simulation model is a powerful tool, useful to the process of understanding and improving court operations.

Notes

1. The Criminal Court disposes of all misdemeanor and violation cases in addition to being the first stage of the Felony Processing System.

2. Plea Bargaining is a mea of disposing of cases before they tie up valuable court resources. The vast majority of defendants adjudged guilty are disposed of via a guilty plea. The defendant is induced to plead guilty by being offered a shorter sentence than the expected sentence if found guilty as the result of a trial. The plea offering may be a plea of guilty to a lower charge.

3. The Supreme Court handles only

felony cases.

4. A trial part like an arraignment or PTC part is the term used for a courtroom staffed by a Judge, District Attorneys, and clerical staff.

5. In addition to the original cost there would be about $12 million in annual operating cost. Also, it will take a considerable amount of time to build the jail.

6. There are many definitions of the term backlog used to describe system status. The definition used in this paper is the number of cases pending.

7. Budget Analysts were more perceptive in this calculation. They took into account the fact that the present number of parts was insufficient to handle properly last year's caseload. Thus, before considering the expected increase in caseload, they added parts to allow the court to be a match for last year's input rate.

8. The backlog could be eliminated quickly if the court decided to lower significantly its plea offerings. That fact leads to difficulties in predicting system behavior; but more importantly it shows the imprecise nature of felony adjudication. Lower plea offerings would have the effect of lowering the average number of appearances for disposition.

9. New York City has five counties each being a Judicial District and each has their own Criminal and Supreme Court. The model is set-up to simulate the operations of each county separately. The majority of quantitative information in this paper refers to Manhattan.

10. See Jennings, "Quantitative Models of Criminal Courts," 39th National Meeting of ORSA, May 5, 1971.

11. The model was set-up so that those cases not affected by the rule would not have statistics collected about them. That run showed only 2.8% of defendants took more than 180 days for disposition. Another run was made setting the maximum trial part calendar size at 10 (had been 16). In that run 12% of defendants affected by the rule took more than 180 days. The backlog was reduced to a projected figure of 1700 defendants by June 1973.
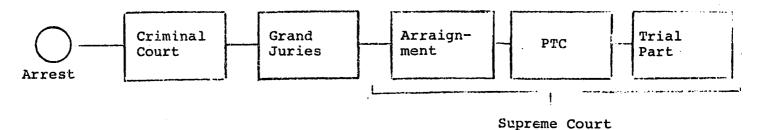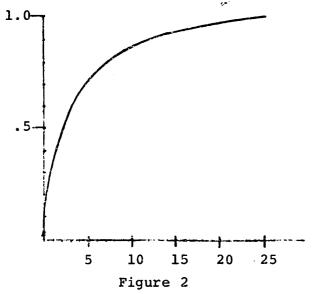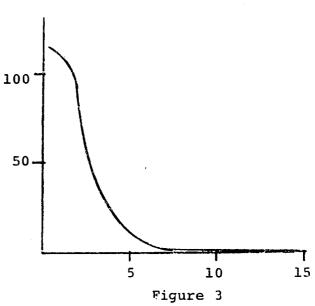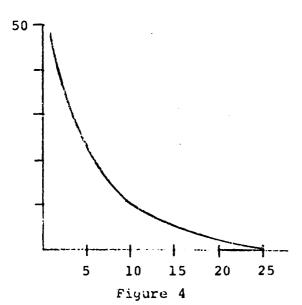
Figure 1
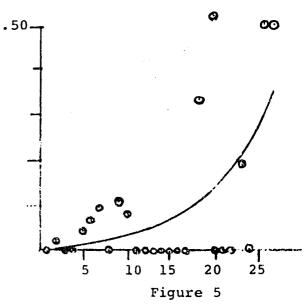Felony Processing System

Figure 2
Criminal Court Delay (Weeks)
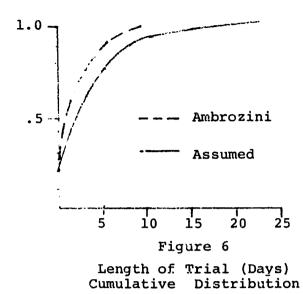Cumulative Distribution

Figure 3
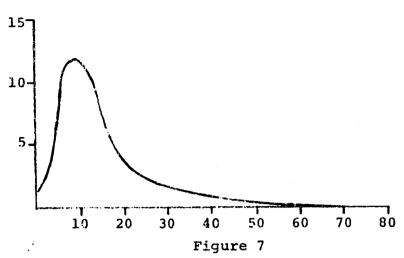No. of Appearances Prior to Trial
Part Frequency Distribution

Figure 4
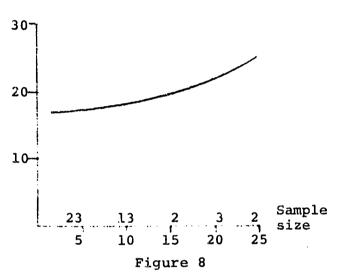No. of Appearances in Trial Part
Frequency Distribution
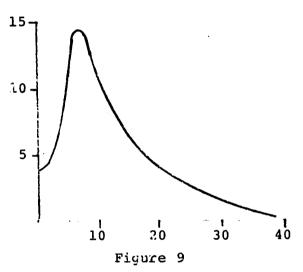

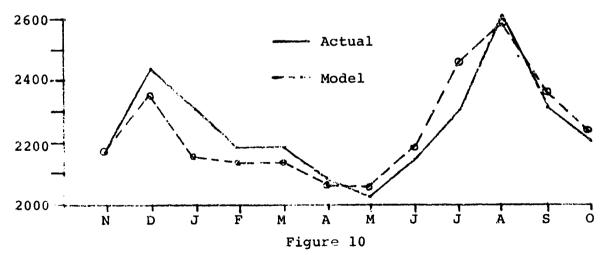Figure 5
Probability of Trial vs. No.
of Trial Part Appearances

Figure 6

Length of Trial (Days)
Cumulative Distribution

Figure 7

Length of Adjournment (Days)
Frequency Distribution

Figure 8

Length of Adjournment (Days)
vs. No. of Trial Part Appearances

Figure 9

Trial Part Calendar Size (Defend.)
Frequency Distribution

Figure 10

Actual vs. Model - Nov. '70 - Oct. '71