

99
Technical Report 1-202

049

COMPUTER-AIDED CRIME PREDICTION IN A METROPOLITAN AREA

prepared for the

Philadelphia Police Department

under a grant by the

Office of Law Enforcement Assistance

U. S. Department of Justice

 THE FRANKLIN INSTITUTE RESEARCH LABORATORIES

Technical Report 1-202

COMPUTER-AIDED CRIME PREDICTION
IN A METROPOLITAN AREA
FINAL REPORT

by

Donald P. Stein
Jay-Louise Crawshaw
Algird R. Barskis

December 1967

prepared for the

Philadelphia Police Department

under a grant by the

Office of Law Enforcement Assistance
U.S. Department of Justice



THE FRANKLIN INSTITUTE RESEARCH LABORATORIES
SYSTEMS SCIENCE DEPARTMENT

FOREWORD

The investigations described in this report were conducted by the Systems Science Department of The Franklin Institute Research Laboratories for the Philadelphia Police Department, under OLEA Grant No. 049.

The authors express their sincere gratitude to Police Commissioner Frank L. Rizzo; Mr. Phillip Carroll, Director of Central Services; Captain James Herron, Commanding Officer, Computer Unit; Lt. Joseph Krauss; and the entire staff of the Philadelphia Police Department for their wholehearted cooperation and support in this joint effort.

The authors also thank Dr. Robert Emrich of the Office of Law Enforcement Assistance, Department of Justice, for his continued guidance and encouragement; Dr. Marvin Wolfgang, University of Pennsylvania, for his invaluable advice and consultation; Dr. Samuel Messick, of Educational Testing Service, for his assistance with the finer points of multidimensional analysis; and Mrs. Sue Johnson, Dr. Alfred Blumstein, and the Institute for Defense Analyses (IDA) staff for their welcome criticisms and suggestions.

This effort would not have been possible without the substantial contributions of The Franklin Institute Research Laboratories staff, especially Dr. Daniel Landis for the development of the MDA techniques; and Miss Diane Reed and Mr. Bernard Epstein for editing and producing this report.

SUMMARY

RESULTS AND CONCLUSIONS

In September, 1966, the Philadelphia Police Department and the Franklin Institute Research Laboratories undertook an OLEA-sponsored project, "Operations Research for Crime Prediction," a \$76,400 grant for FY 1967 (OLEA Grant No. 049). The project's goal was to develop a crime-prediction model based on the conditions that surround specific types of crimes in Philadelphia; this model would be of great value in assisting the Police Department to deploy specifically trained tactical forces, such as stakeout teams and special-purpose vehicles, for prevention of the high-probability crimes.

The results of the past year are exceptionally encouraging. From a germ of an idea a year ago, the study has created a computer-based pilot model having great promise; the fundamental soundness of the underlying approach has been demonstrated, and there is every indication that a fully operational model of immediate utility to the police can be achieved with another two years of effort.

The conclusions are supported by the following results:

1. *A pilot model for burglaries has been developed and implemented.* The model can discriminate 20 to 40 percent of the burglaries from other crime types, based on a sample of 2800 crimes. This demonstrates the feasibility of the fundamental assumption that surrounding conditions differ for different crime types.
2. *The Philadelphia Police Department now has a program capable of testing the pilot model operationally.* The program will be used to test the ability of the model to predict burglaries (as distinct from no crime occurrences) in the "real world" of actual police operations. Based on these tests, the model can be refined to include a greater degree of normalization for no-crime situations, and incorporate improved cluster-analysis techniques.

3. *Once refined, this model will be used in police operations.* The computer program described in the preceding result has already been run through a remote console in the Police Communications Center, and can be used with the refined model. The operator will type in the current conditions for any police patrol sector; then, if the conditions in that sector match the conditions which commonly co-occur with a given type of crime, that type of crime will be "predicted" by the computer for the given sector. Officers on duty in that sector will then be alerted to watch for that crime type. This will be particularly valuable in combination with capability for data transmission to vehicles, now being tested.
4. *An initial model was developed for homicides.* Surrounding conditions for homicides apparently are indistinguishable from those for other crime types. This indicates that homicides may not be predictable and further investigation of homicides is probably not justified.
5. *Much of the data already collected to support the model is useful as a general-purpose data base.* If the presently available data concerning the characteristics of the crime were updated and augmented with data concerning the offender, it could be used as the nucleus of a general-purpose Police Department data base.

RECOMMENDATIONS

The project results strongly support the desirability of additional effort to derive maximal benefit from the model developed. To achieve this end, FIRL recommends that the following steps be taken:

1. *Develop data base further.* The project data base should be developed further, to support additional model refinement and to provide a quick-response data base for *ad hoc* use by the Philadelphia Police Department. Another year's crime data should be added; the 1960 census data should be updated, using projective techniques; and new variables should be incorporated.
2. *Refine model.* The multidimensional and regression analyses should be refined; and other techniques, such as multinomial discriminant analysis and adaptive pattern recognition, should be investigated.

3. *Conduct operational testing.* As the various crime-cluster predicting techniques are developed, they should be tested in an operational police environment, using the already-developed computer program. Results from the tests should be utilized to refine the model further.
4. *Conduct ad hoc studies for the Philadelphia Police Department.* The data base developed for the predictive model should be utilized for related *ad hoc* studies involving such areas as *modus operandi* and recidivism.
5. *Increase police participation.* The Philadelphia Police Department's active participation during this project has demonstrated its forward-looking approach to Police research and development, and its willingness to cooperate in long-range research efforts not having immediate payoff. It is recommended that the Department's technical participation be further augmented by the institution of a formal joint steering committee, and the inclusion of a full-time police officer as part of any future project team.

PROJECT ACTIVITIES

Technical Approach

Crime prevention, especially in metropolitan areas, depends greatly on effective allocation of tactical resources. If a scientifically accurate method for predicting specific crime occurrences were developed, then a powerful tool would be available to assist police commanders in their decision-making regarding the deployment of forces.

The technical approach was based on an analogy with weather forecasting. Weather prediction is made possible by the knowledge that certain combinations of factors, such as frontal systems and high- and low-pressure areas, tend to co-occur with particular types of weather. It was hypothesized that FIRL could identify specific combinations of factors which tend to co-occur with specific types of crime, and from these crime indicators could determine which crime types, if any, are likely in a given police patrol sector, on a given day, at a given time.

The present study has sought to develop an operations-research model to predict crime occurrence, hour by hour and sector by sector, based on combinations of factors which co-occur with crimes.

Data-Base Generation

From police commanders it was determined what factors are considered significant to the incidence of crime. Based on the results of a questionnaire sent to all Philadelphia Police commanders of lieutenant rank and above, and interviews with police officials, a list of potentially crime-related factors was developed.

The factors grouped naturally into three types: crime characteristics (weapon, crime type); temporary characteristics at the time of crime occurrence (weather, time of day); and characteristics of the neighborhood where the crime occurred (unemployment, economic level, housing).

Data Analysis

From the potentially crime-related factors, specific factor combinations were identified which tend to co-occur with specific types of crime. A sample of past crimes was selected and each crime was characterized by its surrounding factors. These data were subjected to mathematical analysis.

The crime sample was drawn from 1966 Philadelphia Police Department records (records for earlier years are not readily available). From the approximately 40,000 Part I and 110,000 Part II crimes which occurred in Philadelphia during 1966, approximately 2800 Part I, and 1800 Part II, crimes were selected for the sample. Each crime was characterized according to the crime type and by momentary and neighborhood conditions at the time and place of occurrence.

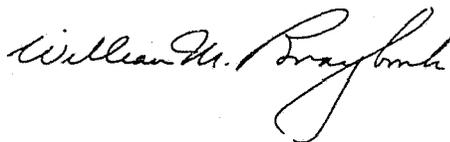
The mathematical analysis centered around a statistical technique called "multidimensional analysis" (MDA), which identifies clusters of crimes having similar surrounding characteristics. Much development work was required to adapt the MDA concept for use with crime data. Relevant existing techniques were modified for use with crime data; new techniques were developed as required, keeping in mind the objective of a useful end-product.

Two other techniques also were investigated. A conventional multiple-regression analysis was performed as a byproduct; its results were not directly usable, but directions for possible future development were identified. A multinomial analysis was perfected for testing specific hypotheses against the full crime sample.

Initial Implementation

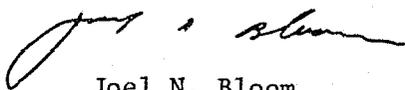
An operating model programmed for the Philadelphia Police Department's computer stores a list of previously identified crime clusters. The user would type the momentary conditions (weather, data, time of day, and so forth) and the district identification onto a remote console typewriter; the computer then would print out a list of patrol sectors for the specified district, and an analysis of how closely the conditions in each sector matched the conditions specified by the clusters.

Initially, this operating program will be used to evaluate and refine the model; then, the refined model can be used for allocation of Police resources. The allocation submodel can make use of sophisticated decision techniques to allocate patrol beats and shifts on a variable basis.



William M. Braybrook
Manager
Operations Research Laboratory

Approved:



Joel N. Bloom
Technical Director
Systems Science Department

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

TABLE OF CONTENTS

<i>Section</i>	<i>Title</i>	<i>Page</i>
SECTION 1.	BACKGROUND	
	A. Statement of the Problem	1
	B. Objective.	2
	C. Technical Approach	2
	D. Project Plan	3
SECTION 2.	DATA-BASE GENERATION	
	A. Crime-Factor Determination	5
	B. Data Collection	13
SECTION 3.	DATA ANALYSIS AND RESULTS	
	A. General Discussion	25
	B. Frequency Distribution	26
	C. Multiple-Regression Analysis	29
	D. Multinomial Analysis	46
	E. Multidimensional Analysis	49
SECTION 4.	INITIAL IMPLEMENTATION	
	A. Data Requirements	59
	B. Model Design	61
SECTION 5.	PROJECT HISTORY	67
SECTION 6.	CONCLUSIONS AND RECOMMENDATIONS	
	A. Conclusions	73
	B. Recommendations	73
APPENDIX A.	Questionnaire Development and Analysis	
APPENDIX B.	Development of Crime-Factor List	
APPENDIX C.	Data Base	
APPENDIX D.	Frequency-Distribution Analysis	
APPENDIX E.	Multiple-Regression Analysis	
APPENDIX F.	Multidimensional Analysis (MDA)	
APPENDIX G.	Programming	

} In separately
bound volume

LIST OF ILLUSTRATIONS

<i>Figure</i>	<i>Title</i>	<i>Page</i>
1.	Project Plan	4
2.	Initial Crime Factor List	12
3.	Random Sampling of 1966 Major Crimes	17
4.	Frequencies of Crime Types by Hour of Day	30
5.	Frequencies of Crime Types by Day of Week	31
6.	Normalized Frequencies of Crime Types by Phase of Moom	32
7.	Normalized Frequencies of Crime Types by Atmospheric Pressure	33
8.	Normalized Frequencies of Crime Types by Percent Nonwhite Population in Neighborhood	34
9.	Normalized Frequencies of Crime Types by Presence of PTC Transfer Point(s) in Neighborhood	35
10.	Normalized Frequencies of Crime Types by Presence of Senior High School in Neighborhood	36
11.	Data Matrices for Method I	40
12.	Data Matrix for Method II	41
13.	Results, Method I Comparative Distributions of Estimated Values for Burglaries and Aggravated Assaults	43
14.	Cumulative Distributions of Burglaries and Nonburglaries	45
15.	Sample Output for Multinomial Analysis Program	48
16.	Burglary Cluster "A" About Burglary No. 37 on Factor II	55
17.	Distribution of 200 Random Crimes with Respect to Burglary Cluster "A".	56
18.	Burglary Cluster "B" about Burglary No. 4 on Factor IX	57
19.	Distribution of 200 Random Crimes with Respect to Burglary Cluster "B"	58
20.	Flow Chart of Operational Model	62
21.	Original Data of Part II Crimes, 1966	70

LIST OF TABLES

<i>Table</i>	<i>Title</i>	<i>Page</i>
1.	Sources for Sociological Crime Factors	15
2.	Coding Procedure for Crime Data	19
3.	Property Codes	19
4.	Weapon Codes	21
5.	Premises Codes	21
6.	Coding Procedure for Census Data	22
7.	Scaling Quantities	24
8.	Crime Factors as Independent Variables	42
9.	Crime Factors Used by Operating Model	60

SECTION 1 BACKGROUND

A. STATEMENT OF THE PROBLEM

Crime prevention, especially in metropolitan areas, depends greatly on effective allocation of police resources. In some cities, these resources are allocated in response to or in anticipation of, "calls for service." In Philadelphia, on the other hand, resource allocation more closely parallels military force deployment; a "first line of defense," consisting of general vehicle and foot patrol, covers the entire city at all times. Superimposed upon the vehicle and foot patrol are *tactical* forces, which are deployed in varying patterns dictated by operational needs. Examples of tactical forces used in Philadelphia are stake-out teams, unmarked vehicles, and "tactical foot patrol" units.*

In the Philadelphia Police Department, as in a military organization, optimal deployment of these tactical resources greatly depends upon accurate intelligence as to what the 'enemy' (the criminal) is likely to do. Much of this intelligence is provided by informers, intelligence on known criminals, and the judgment of experienced police commanders. However, no *comprehensive* means now exists for supplementing this information with accurate correlative data-analysis. *If a scientifically accurate method for predicting specific crime occurrences were developed, then a powerful tool would be available to assist police commanders in their decision-making regarding the deployment of forces.*

*"Tactical foot patrol" units consist of a vehicle, and several policemen who "fan out" from the vehicle while maintaining communication with it. They can be recalled to the vehicle at any time and re-deployed to another location.

B. OBJECTIVE

For direction of police operations, the City of Philadelphia is divided into 22 districts, and further subdivided into approximately 300 vehicle patrol beats (called sectors). To be useful for deployment of forces, any scheme for crime "prediction" must at the very least discriminate among different patrol sectors, and different hours of the day. Accordingly, the project established the following overall *objective*:

To develop a model for predicting crime occurrence in the City of Philadelphia, hour-by-hour and sector-by-sector.

Such an operational tool, of course, should be designed so that it could be generalized to other metropolitan areas.

C. TECHNICAL APPROACH

The technical approach is based on the hypothesis that crime occurrence can be anticipated in a way similar to weather prediction. Underlying the total effort is the belief that if conditions which frequently co-occur with particular types of crime can be identified, then the conditions at a given location of the city at a particular time could be used to predict crime occurrence in the same way that temperature and barometric pressure are used to forecast a thunderstorm.

In order to develop a formal model, these conditions must be expressed in terms of certain [measurable] variables. The first step, then, is to investigate variables of all types - weather, time, neighborhood conditions - and to select those for which data is available and apparently relevant to crime occurrence. Then, by studying the values of these variables for some random sample of crime occurrences, groups (or "clusters") of crimes with similar surrounding conditions can be identified. Once identified, the relative presence or absence of a particular set of surrounding conditions can be used to indicate whether or not a crime of the associated type could be expected to occur.

The process of selecting appropriate variables (also called "crime factors") was envisioned as a combination of a state-of-the-art literature search and a survey of the opinions and suggestions of experienced police commanders. Multidimensional analysis, a sophisticated statistical technique, was selected to show meaningful combinations of crime factor values and identify clusters of crimes by determining the specific values of the factors that co-occur with specific types of crimes. Finally, an operational computer model, based on the identified crime clusters and programmed by the Philadelphia Police Department Computer staff, was proposed to relay hour by hour predictions for any city sector to police commanders.

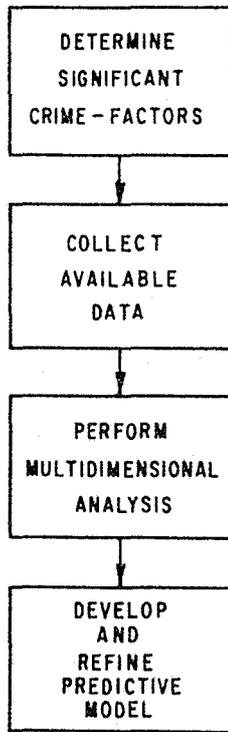
D. PROJECT PLAN

The original project plan envisioned two phases (see Figure 1). The first phase was to be developmental, its four main tasks comprising the bulk of the analytical part of the study. The crime factors had to be determined from past studies and other police data; data in support of these crime factors had to be amassed for use in the analyses to follow. The multidimensional analysis had to be programmed and run using the collected data, and finally, the framework for the predictive model had to be developed and refined.

Phase II would then serve to implement the theoretical model achieved in Phase I. A computer program was to be prepared to reflect the theoretical analysis; this program was to be used as part of an operational model, to be implemented on the Philadelphia Police Department's computer. Finally, an evaluation would be conducted so that the predictions of the model could be compared with real crime occurrences.

This original plan was followed quite closely during the project. The main points of departure are described in Section 5 of this report, Project History.

PHASE I



PHASE II

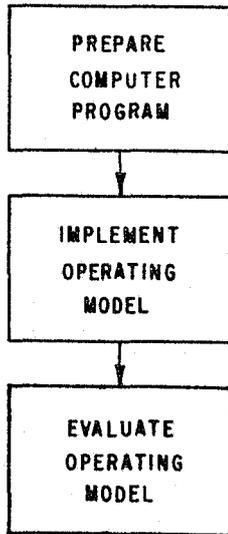


Figure 1. Project Plan

SECTION 2 DATA-BASE GENERATION

A. CRIME-FACTOR DETERMINATION

A first step in the project was the selection of a number of conditions or situations which have been found to co-occur with crimes, *i.e.*, crime factors, to be used as a common data base for all the analyses to follow. To begin this effort a literature search was made, guided by the suggestions of Dr. Marvin Wolfgang, Department of Criminology, University of Pennsylvania. The search yielded various studies linking many socio-economic characteristics and criminal statistics to locations where crimes occur most frequently or to persons who seem most likely to commit crimes. To summarize the search, a master list of suggested factors was compiled covering the areas of population, age, race and nativity, marital status, fertility, income, education, rent and property, police and crime statistics, occupation, employment, housing, physical characteristics, social factors, and other special factors such as business activity (see Appendix B, Table B-5).

The first rough factor list was drafted using the results of the literature search. The proposed model, however, was designed to rely on dynamic factors, such as time of day, season of the year, and weather conditions, as well as the more static socio-economic factors. Thus the rough draft was made up of "crime particulars" - time, date, location, type of premises, weapon, property stolen or injuries received, and age and sex of the offender; "environmental factors" - weather conditions, including phase of the moon, or special events co-occurring with the crime, such as parades or rallies; and "neighborhood socio-economic factors" - race, housing, rent, school enrollment, unemployment, occupations, population mobility, and income (see Appendix B, Table B-6).

In discussing the factors included in the rough draft of the crime factor list, two questions were raised: What data, if any, is available to support research on these factors? Would experienced police commanders agree with the crime factors selected? The project staff tried to answer these questions.

Data in support of the crime-particulars, or crime related factors, was sought in the Philadelphia Police Department report files. All crime-related factors mentioned except "weapon" and "extent of injuries" were found in Police Report form #49, routinely submitted for all complaints by the police officers in the field. Of those included in Report #49, all but "type of premises" and "age of offender" were available in punch-card form (see Appendix B, Tables B-1 and B-2).

The U.S. Weather Bureau and the Franklin Institute Weather Center were contacted to ascertain the types of weather data available (see Appendix B, Table B-3). The Franklin Institute had continuous trace plots of weather factors such as wind velocity, temperature, pressure, and humidity. The U.S. Weather Bureau had readings printed monthly for daily visibility, precipitation, wind speed, snow, temperature, pressure, and humidity. Some of these readings were daily averages while others were specific readings taken every three hours during the twenty-four hour daily cycle. A third possibility was the use of a weather tape, presently owned by the Philadelphia Electric Company, of hour-by-hour readings for visibility, precipitation, wind direction and velocity, temperature, humidity, and barometric pressure. An almanac provided a daily listing of the lunar cycle.

Contact with the Philadelphia Planning Commission established that all crime factors concerning socio-economic data could be found in the 1960 U.S. Census for Philadelphia and in a Public Information Bulletin, "Population Estimate for 1964" (see Appendix B, Table B-4). Because this crime factor data was not as recent as desired for the study, other sources, such as the City Economist's Office, the Urban Renewal Administration, and the City Department of Health, were investigated. Unfortunately, none of the data from these sources was in usable form for the study.

No one source could be found for data to support the special events category. Newspaper reports were considered as a possible source but the research necessary to provide special events data for each crime instance sampled was beyond the scope of this effort.

With the knowledge of available data sources, a preliminary list of crime factors was compiled. (see Appendix B, Table B-7). All crime related factors from the rough draft, except for "sex of offender," were retained in this new version. The time, date, and lunar cycle factors were grouped into a new category, "Time of Occurrence." The weather category was expanded to include visibility, wind speed, humidity, and pressure, and the sociological factors were expanded to include one or more measures in each of the following areas: age distribution, employment, housing, income, marital status, nativity, population, race, recidivism, rent, and school enrollment.

The second question was now tackled; How would police commanders in the field feel about the crime factors selected for the preliminary list? It was decided to solicit their suggestions and comments using a questionnaire (Appendix A).

The Crime Factor Questionnaire was designed to:

- (1) supplement the preliminary crime factor list;
- (2) verify the several factors included in the preliminary list;
- (3) investigate the relationships between the various factors and the different types of crime;
- (4) determine the degree to which police commanders rely on the various crime indicators for crime prediction and/or prevention, and;
- (5) solicit comments from experienced police commanders concerning the factors as well as the project in general.

For the preliminary factor list, measurable factors in the areas of crime particulars, time, weather, and sociological characteristics had been chosen without the benefit of field experience. It was recognized, however, that police commanders in the field might be aware of other factors, not included in the list, which their experience showed to be crime

indicators. Question 1-a was designed to elicit these additional factors. A crime factor matrix was set up with the preliminary factors preprinted. The commanders were then asked to add new factors in the space provided under the several category headings.

The commanders were asked in Question 4 to rank the five crime indicators they felt were most reliable, so that individual factors could be supported or rejected on the preliminary list, thus giving some indication of the value of these factors as crime indicators.

It was recognized that each factor on the list need not apply to every type of crime. Indeed, one type of crime might have none of the factors associated with it and thus be "unpredictable." To explore further the relationships between the crimes and the crime factors, the commanders were asked in Question 1-b to link factors with crimes on the crime/crime factor matrix. The Part I crimes were included in the matrix and space was provided for adding other crimes. Respondents were asked to double check the matrix position for a particular crime and factor to indicate a strong relationship; to single-check, indicating a relationship, or not to check the matrix position if crime and factor were unrelated.

As the factor list was being developed, it was hypothesized that police commanders rely to some extent on vague crime indicators, such as the "atmosphere" of the neighborhood, or even on abstract "hunches" rather than the defined and measurable factors included in the preliminary list. Three questions to test this hypothesis were included. Question 2 dealt with "circumstantial" indicators, those conditions the experienced commander feels may indicate trouble in his district. An unusual number of juveniles on the street or activity around a vacant house might alert a commander in the field to keep a sharp watch. In Question 3 the commanders were asked to discuss instances in which "hunches" or premonitions of trouble had proved reliable crime indicators for them. Question 5 was an attempt to discover other indicators which had proved reliable enough to become a part of routine observations.

Comments were encouraged on all of the first five questions, with a sixth question devoted entirely to comments, since it was recognized that

added comments on any topic in the questionnaire could provide further insights and valuable inputs not elicited by the structured questions.

An identification page for the respondent's name, rank, and district was included, but the information was kept confidential and restricted to the FIRL project staff to encourage maximum freedom of response. The preliminary factor list, and cover letters from the FIRL project leader and the Police Commissioner accompanied the questionnaire.

A briefing to acquaint police commanders with the project's purpose and scope was held at a weekly command meeting. The briefing illustrated the way factors would be used by the model to forecast crime and was followed by informal questions and answers. After the meeting, copies of the Crime Factor Questionnaire were distributed to all police commanders above the rank of sergeant.

Completed questionnaires were analyzed by tabulating responses question-by-question. These results were then scrutinized considering the five original objectives.

The police commanders added many factors to the preliminary list in response to Question 1-a. Each addition was recorded and all suggested factors were listed (see Appendix A, Table A-1). This augmented list was then presented for consideration at a subsequent meeting of police officials and FIRL staff. As a result of the commanders' suggestions, four new factors were added to the crime factor list: the number of schools in the area, business activity, land use, and transportation facilities.

The most important factors, (selected by the commanders in answer to Question 4) were tallied. The tally was repeated twice, separating the commanders first by rank and then by years of service (see Appendix A, Table A-4). Their selections verified all the preliminary factors except "phase of moon," with "location" and "day of week" receiving the most support. No significant differences were noted among responses from commanders of different ranks or various lengths of service.

Five types of tallies were made of the responses to Question 1-b.

- 1 and 2. All factors receiving checks (single and double) and then those receiving double checks alone were tallied to find out which factors were related to which crimes (Appendix A, Table A-2). Both tallies indicated the following relationships:

<u>Factor</u>	<u>Crime (s)</u>
Location	Burglary, Robbery
Time of day	Burglary, Robbery, Rape
Day of week	Burglary, Robbery, Rape
Phase of Moon	Aggravated Assault, Homicide
Precipitation	Burglary
Temperature	Aggravated Assault, Homicide
Special Events	Larceny
Race	Aggravated Assault

(Totals for the other factors were too small to confirm any relationship.)

3. A third tally was made of the factors respondents checked as being related to every crime (see Appendix A, Table A-3). While the percentage of commanders who responded in this manner was low, the factors selected were among or related to those on the preliminary list, further substantiating their importance.

- 4 and 5. Two final tallies were made - one of factors which were not checked at all and the other of crimes which were not checked (Appendix A, Table A-3). In the first instance "phase of moon" was rejected as a predictive factor by 39 percent of the commanders. However, since more than half of those responding had indicated some confidence in this factor, it was retained. The second tally indicated that less than 5 percent of those responding felt that a particular crime was *not* related to any factor. This response greatly supported the hypothesis that specific factors *are* related to specific crime instances.

The overall response to Question 1-b verified the preliminary factors and reinforced the hypothesis that there is indeed a relationship between the factors and crime occurrences.

Questions 2, 3, and 5 sought to determine what *other* crime indicators the police commanders relied on. Since an overwhelming number of respondents did not respond to Question 2, it was omitted in the analysis.

The "YES" and "NO" responses to the opening query in Questions 3 and 5 were tallied (see Appendix A, Table A-5). While only 31 percent of the respondents acknowledged that abstract "hunches" had proved reliable crime indicators, 60 percent indicated they relied on warning signals, which in most cases were described as some combination of more tangible information such as the previous day's crime report. This lent support to the hypothesis that crime prediction based on certain measurable crime indicators was a feasible and useful objective.

While comments as such were not formally analyzed, each questionnaire was read thoroughly by FIRL staff members. Important comments and suggestions were noted and designated for future reference.

Taking into account the results of the questionnaire analysis, a revised crime factor list was compiled (see Appendix B, Table B-9). The crime related factors remained basically the same, but "extent of injuries" was deleted (data were not available) and "type of item" and "dollar value" were combined as one factor, since "dollar value" could be used to measure the value of "type of item". The "time of occurrence" and weather factors remained unchanged. The sociological and special events categories remained intact, but "recidivism" was deleted since data were unavailable. A new category, "local characteristics," containing number of schools per sector, business activity, land use, and transportation facilities, was added as a result of the suggestions of the police commanders in the crime factor questionnaires.

Lack of readily available data necessitated a second revision of the crime factor list. "Age of offender" data were not readily available for every crime instance so that factor was dropped. The special events category and two of the local characteristic factors - business activity, and land use - also had to be deleted for lack of supporting data. An additional weather factor, "snow on ground" was added after it was decided that the precipitation measurement did not adequately separate snowy conditions from rain. Finally, as sources for the supporting data were determined, units of measurement for each type of factor data were established. The second revision of the crime factor list became the initial crime factor list (see Figure 2).

(Second Revision of Factor List)

Crime-Related Facts:

Type of crime
Weapon
Type of premises
Census tract and sector
in which crime occurred

Type of item

Time of Occurrence:

Hour
Day of week
Day of month
Month
Year
Phase of moon (full,
new, first quarter,
last quarter)

Weather

Snow (inches on
ground at 7 A.M.)
Visibility (in miles)
Precipitation (in
inches; daily total
water equivalent)
Wind Speed (in knots)
Temperature
Relative Humidity
Barometric Pressure

Sociological Factors:

Age Distribution % 15-34, total population
 % 60 and over

Employment % male unemployed
 % wage and salary workers

Housing % housing units owner-occupied
 % housing units in sound condition
 % housing units with 1.01 or more persons
 per room

Income Median family income, 1959

Marital Status % married, 14 years and older

Nativity % foreign-born

Population % growth or decline, 1960-1964
 % moved since 1955
 Median number persons per household
 % families, one or more under age 6

Race % non-white

Rent Median monthly gross rent

School Enrollment % enrolled in school, ages 5-34
 Median number of school years completed

Local Characteristics:

Number of important transportation transfer points
Number of schools - elementary, junior high, senior high

Figure 2. Initial Crime Factor List

B. DATA COLLECTION

Four main tasks comprised the data collection effort: determination of sources for data to support the selected crime factors, selection of crime samples, coding of data, and the scaling and merging of data into a final crime record for each sampled crime. Since complete information was not available for all crime occurrences prior to 1966, it was decided to select crime samples from that year and to use data compiled for every crime factor for each selected crime as a basis for the subsequent analyses.

1. Data Sources

Data collection began with the selection of sources from those previously investigated, to obtain data to support the crime factors on the initial crime factor list.

Listings of crime report punch cards containing data from the Philadelphia Police Department's Report Forms #48 and #49, were chosen as a source for the following crime-related information for each crime selected for the analysis samples: district/sector of occurrence; type of crime; type of property and property value for crimes against property; and time of occurrence (day of week, month, day of month, year and hour). The address of the crime occurrence was also available from the card listing as a means of determining the census tract in which the crime occurred. Address-census tract index listings were provided by the Philadelphia Police Department for this procedure. Weapon and premises information was available only from document crime reports on file at the Philadelphia Police Department. Thus it was decided to record the complaint number for each sample crime to facilitate a later report search for the weapon and premises associated with each crime sample.

Since the crime samples were to be drawn from crime occurrences in 1966, the Philadelphia Electric Company's weather tape, with readings only as far as 1965, had to be rejected as a source of weather data. In its place, printed monthly weather summaries were obtained from the U.S. Weather Bureau in Philadelphia for each month of 1966. Using the date and time data collected for each sample crime, the proper values for the weather factors associated with that date and time could be manually extracted from Weather Bureau summaries. The phase of the lunar cycle for each date and time were found in a 1966 almanac.

The several tables of the 1960 U.S. Census for Philadelphia and the Public Information Bulletin, "Population Estimate for 1964" (April 1966) were designated as sources for the sociological crime-factor data. Percentage data were used for these factors wherever possible. Table 1 lists the several sociological crime factors, the census table having the appropriate data and the computation necessary, if any, to form a percentage data figure where possible.

The Philadelphia Transportation Company and the Board of Education were selected as sources for the local characteristics data. From the PTC map of important transportation transfer points throughout the city, a binary indicator could be set up for each census tract, to show if a transfer points were located in that census tract. A list of city schools and addresses from the Board of Education was used to compile a list of the number and types of schools in each census tract. From this list three binary indicators were set up for each census tract to show whether or not one or more elementary, junior high, or senior high schools is located in that census tract.

2. Sampling

The Philadelphia Police Department supplied punch cards containing information on approximately 45,000 Part I crimes from 1966. (Part I crimes include homicide, rape, robbery, aggravated assault, burglary,

Table 1. Sources for Sociological Crime Factors

Crime Factor	Census Table	Computation necessary (if any)
% 15-34	P-2	sum 15-34 / total no. in tract
% 60 or over	P-2	sum 60 & over / total no. in tract
% male unemployed	P-3	no. unempl. male / male 14 & over
% wage & salary	P-3	private w. & s. / total employed
% owner-occupied	H-1	owner occ. / total
% sound housing	H-1	sound / total
% w/ 1.01 or more persons/room	H-1	1.01 or more / total
median family income	P-1	Direct
% married, 14 & over	P-2	married / total
% foreign-born	P-1	foreign-born / total
% growth	} *	difference / total
% decline		
% moved since 1955	P-1	100-(same house 1960/ residence in 1955)
persons per household	P-1	Direct
% families, 1 or more under 6 years	P-1	(Families w/ children under 6) / married couples
% non-white	P-1	100-(white / total)
median monthly rent	H-2	Direct
% enrolled in school	P-2	total 5-34 enrolled / total 5-34
school years completed	P-1	Direct

* Public Information Bulletin, City Planning Commission, April 1966.

larceny, and auto theft.) The cards were sorted by month of occurrence and crime type within month listed, and the listings bound in separate books for each month.

It was decided that a minimum of 5 percent random sample from each type of crime for each month would be taken (see Figure 3). Thus a count was made of each type of crime by month. If there were fewer than 20 particular crimes for a given month, the entire group was taken into the sample. Also, if the total count of a given type of crime for a given month was less than 400 (5% of 400 = 20), 20 crimes were chosen for the sample. For counts greater than 400, 5 percent of the total count was selected. Each selected crime was marked on the listing for future transfer of data to a coding form.

Using the above procedure a random sample of approximately 2800 Part I crimes during 1966, was obtained. The sample was divided into seven crime types and a random sample of 100 crimes was drawn from each type. The random samples for the seven crime types are designated 1 through 7. In addition, two random samples, designated as 8 and 9, were drawn, with replacement, from the entire population of 2800 crimes. In each instance the sampling procedure was the same as that used for obtaining the first large crime sample.

The entire Part I crime sample was intended for use by the multinomial analyses. The subsamples of 100 crimes for each crime type and the two 100-crime random samples were designated for the multidimensional analysis and the regression analyses.

Cards for the Part II crimes for 1966 were also received from the Philadelphia Police Department and a random sample consisting of 1800 crimes was drawn from two groups of cards, one representing Part II crimes for January through July and the other representing those for August through December.

PUNCHED CARDS OF
1966 MAJOR CRIMES

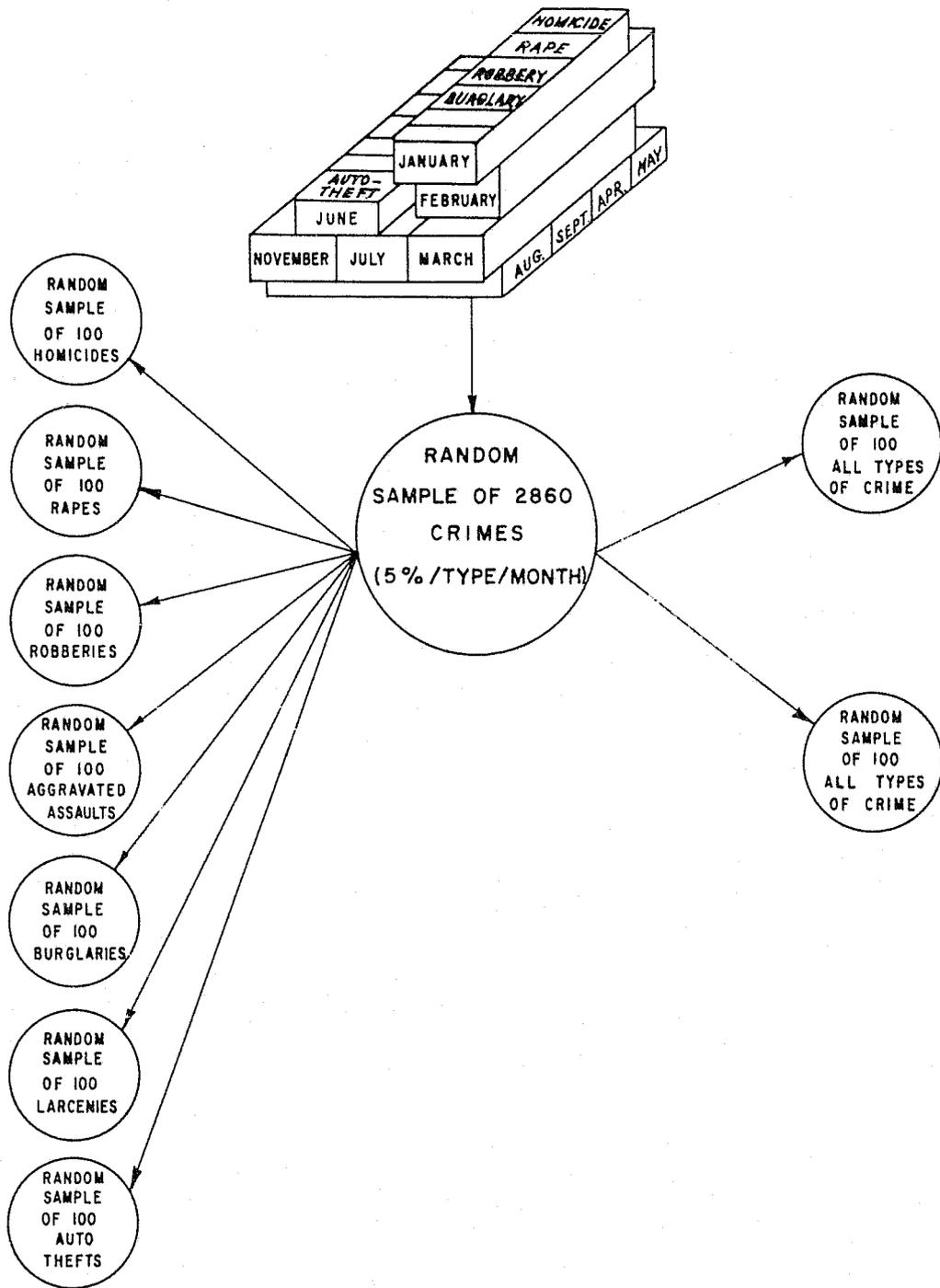


Figure 3. Random Sampling of 1966 Major Crimes

3. Coding

The basic analytic data unit was termed a "crime record" - the data values for each crime factor associated with a particular crime. A crime record for each of the sampled crime was found using three steps:

- (a) ascertain the crime-related facts and the date and time of occurrence for the crime;
- (b) using the date and time factors, select the proper weather data for the crime occurrence; and
- (c) using the census tract in which the crime occurred, select the proper sociological and local characteristic data.

A coding form was set up onto which crime-related information could be transferred directly from the crime card listings (see Table 2). The crime complaint number was copied to identify the crime record. Space was provided for a sample number, denoting the sample from which the crime was drawn. The district and sector identification consisted of a district number and a sector letter on the crime card listing but the letter was changed to a two digit numeric code, thus making the district/sector designation a four digit number on the coding form. For example, district 14, sector A is coded as district/sector 1401. Using the address given on the crime card listing and an address-census tract index, the proper census tract designation, again a number followed by a letter, was found for the crime. As with district/sector, the letter part of the census tract designation was coded as a number and the full four digit number entered onto the coding form. The crime code, a four digit number designating the type of crime, was copied directly from the crime card listing. (For interpretation of these codes see Appendix C, Section b.) For crimes against property, the property code representing type of property (see Table 3) and property value were copied directly from the crime card listing. For other crimes these were left blank. The day of the week, day of month, month, year, and hour for each crime occurrence were coded directly from the crime-card listings. Month, day of month, and year were represented by their usual two-digit numbers; for example January 1, 1966 was coded as 01 01 66. Day of the week

Table 2. Coding Procedure for Crime Data

Column(s)	Factor	Scaling, if necessary
1-6	Complaint Number	
7-10	Sample Number	
11-14	District/Sector	Sector letter is coded as number, i.e. A=1, B=2, etc.
15-18	Census Tract	Letter coded as above
19-22	Crime Code	(see Appendix C)
23	Property Code	(see Table 3)
24-28	Property Value	
29-31	Snow - inches on ground	(Use daily reading)
32	Day of week at 7 AM	1=Monday, 2=Tuesday, etc.
33	Phase of moon	1st Q = 1; Full = 2; 3rd Q = 3; New=4 (daily designation)
34-35	Month	
36-37	Day	
38-39	Year	
40-41	Hour	On 24 hour clock, $\frac{1}{2}$ & $\frac{3}{4}$ PM=13,
42-44	Visibility - nearest whole mile	(Nearest 3-hr reading)
45-47	Precipitation - in inches	(Daily total)
48-50	Wind Speed - knots	(Nearest 3-hr reading)
51-53	Temperature - degrees Fahrenheit	(Nearest 3-hr reading)
54-56	Relative Humidity	(Nearest 3-hr reading)
57-60	Barometric Pressure - inches of Mercury	(Daily average)
62-63	Premises Code	(See Table 5)
64-65	Weapon Code	(See Table 4)

Table 3. Property Codes

0	No property involved
1	Currency, bonds, etc.
2	Jewelry, precious metals
3	Furs
4	Clothing
5	Automobiles
6	Miscellaneous

was also coded numerically - Monday = 1, Tuesday = 2, and so on through Sunday = 7. Hour of the day was given by a two-digit number representing the hour on a twenty-four hour clock; for example, 1 AM = 01 but 1 PM = 13. Codes for type of weapon (Table 4) and type of premises (Table 5) were not available from the crime card listing. For these factors, the written reports corresponding to the complaint number for each of the sample crimes were sought in the Philadelphia Police Department central files and the proper information entered onto the coding form.

Once these crime-related facts were coded for each crime, the proper weather data could be added. The month and the day of the month were used to determine the proper phase of the moon, coded as 1 = first quarter, 2 = full, 3 = third quarter, and 4 = new; the daily total of precipitation, in hundredths of inches; and the daily average barometric pressure, in hundredths of inches of mercury. In addition the hour of occurrence was used to ascertain the nearest three-hour reading for visibility to the nearest whole mile; wind speed, in knots; temperature, in degrees Fahrenheit; and relative humidity, in percentage form. These items were all added to the crime-record coding form.

Rather than code a set of sociological and local characteristics data values for each crime, repeating the same values for crimes in the same census tracts, it was decided to set up the sociological and local characteristic data for each census tract in a separate data file. A computer program could then be written to select and add the sociological and local characteristics data for the proper census tract to each crime record. As a result, data values for the sociological crime factors for each census tract in the city were extracted manually from the census reports. Coding forms were set up for this data (see Table 6). For each census tract, the number-letter designation was coded as a four-digit number, as before, and followed by the nineteen sociological factors. The four binary-choice local-characteristic data values were also coded for each census tract after the sociological facts.

Table 4. Weapon Codes

CODE	WEAPON
1.	No weapon. (This category includes the following crime codes: all 500, 600, 700, and 311, 321, 323, 331, 341, 351, 361, 371, 381, 391.)
-1.	Information not available (If crime code is not listed above and no weapon is reported.)
2.	Hand-gun (Pistol or gun.)
3.	Shotgun or rifle (Any except hand-gun or gun.)
4.	Club (Pipe, brick, blackjack.)
5.	Knife (Any cutting instrument, bottle.)
6.	Strongarm (Fists, feet.)
7.	Chemicals (Acid, poison.)
8.	Automobile
9.	Other

Table 5. Premises Codes

CODE		TYPE OF PREMISES
Dwellings	1.	Apartment house
	2.	Other private residence
	3.	Public residence (hotel, rooming, house)
	4.	Unoccupied residence
Business	5.	Restaurant
	6.	Banks
	7.	Finance offices
	8.	Business offices
	9.	Drug stores (except chain stores)
	10.	Grocery stores or delicatessens (except chain stores)
	11.	Other stores
	12.	Chain stores
	13.	Repair shops
	14.	Warehouses
	15.	Alcohol sales: Taproom or bar
16.	State store	
Transportation	17.	Highway
	18.	Bus
	19.	Subway
	20.	Taxi
	21.	Car Lot - sales
	22.	Car Lot - parking
	23.	Gas station.
Public Services	24.	Hospital
	25.	School
	26.	Recreation facilities
	27.	Church
Other	28.	Vacant Lot
	29.	Other
	30.	Industry
	-1.	Information not available

Table 6. Coding Procedure for Census Data

Column(s)	Factor	Remarks
1-4	Census Tract	
5-6	% 15-34 years old	
7-8	% age 60 or over	
9-10	% male unemployed	
11-12	% wage & salary	
13-14	% owner-occupied	
15-16	% sound	
17-18	% with 1.01 or more persons/room	
19-20	median family income	$\times 10^2$, <i>i.e.</i> \$9000 = 90
21-22	% married, 14 & over	
23-24	% foreign-born	
25-26	% growth	
27-28	% decline	
29-30	% moved since 1955	
31-32	persons per household	expressed with one decimal place, <i>e.g.</i> 1.0, 3.1
33-34	% families, 1 or more children under 6 yrs	
35-36	% non-white	
37-39	median monthly rent	in whole dollars (0 through 999)
40-41	% enrolled in school	
42-44	school years completed	expressed with one decimal place, <i>e.g.</i> $8\frac{1}{2}$ years = 8.5
46	one or more PTC transfer points	1 = yes 0 = no
47	one or more elementary schools	1 = yes 0 = no
48	one or more junior high schools	1 = yes 0 = no
49	one or more senior high schools	1 = yes 0 = no

4. Data Preparation

After both groups of information were completely coded, the two groups - one of partial crime records and the other of sociological and local characteristic data by census tract - were transferred to punch cards. Two steps remained before the crime records could be considered complete.

- (1) Scaling of certain data values; and
- (2) Merger of all the crime factor data values into one crime record per sample crime.

For analysis, all crime factors were given a common range of values, zero through one hundred.* Of those crime factors to be used in the analysis, fifteen had ranges other than zero to one hundred and thus for these crime factors scaling was necessary. Five of the fifteen factors which required scaling were cyclic factors - day of week, hour, month, day of month, and phase of moon. For the crime record these factors were assigned raw values; scaling was left for later (see discussion of scaling under MDA section). The remaining ten crime factors - snow, visibility, pressure, income, persons per household, rent, property value, school years completed, precipitation, and temperature -- were scaled immediately and the scaled values placed in the crime record.

The first seven crime factors to be scaled were considered individually and a range of values obtained for each. In each case, the scaling quantities, called k and c , were determined to form the equation:

$$\text{scaled value} = k \times \text{raw value} + c$$

from the conditions:

$$0 = k \times \text{minimum value} + c$$

and

$$100 = k \times \text{maximum value} + c$$

* In practice, zeros act as "dominators" in matrix manipulations, and therefore were not used. The 0-to-100 range was adjusted to 1 to 101 during the MDA.

Table 7 gives the original range of each of the above mentioned crime factors and the "k" and "c" necessary to convert that range to the desired 0 through 100 range.

The last three variables were handled somewhat differently. Property value was set up so that the scaled value was equal to the raw value divided by fifty. If the scaled value then exceeded one hundred, one hundred was substituted. Precipitation and temperature had ranges which were approximately zero to one hundred, so the raw values which fell in this range were used as scaled values. Raw values which exceeded one hundred were replaced by one hundred and raw values less than zero were replaced by zero (see Table 7).

A computer program, called JCADATA (see Appendix G), was written to perform the above scaling and to merge all the crime factors into one crime record per crime. Each sample of Part I crimes was processed by the program and a two-card crime record produced for each crime in the sample. The development of the completed Part I crime records thus ended the data collection phase of the project. Processing and Analysis of the Part II crimes were postponed until completion of the Part I crime-cluster analysis (see Section 3).

Table 7. *Scaling Quantities*

CRIME FACTORS	k	c	OLD RANGE
snow	8.33	0.0	1-12 inches
visibility	6.67	0.0	1-15 miles
barometric pressure	16.67	-450.09	27.00 - 33.00 inches of mercury
income	1.39	-37.5	27 ($\times 10^2$) - 99 ($\times 10^2$) dollars
persons/household	32.3	-38.7	1.2 - 4.3 persons
rent	0.79	-22.4	28 - 154 dollars
school years completed	10.9	-73.9	6.8 - 16.0 years

SECTION 3 DATA ANALYSIS AND RESULTS

A. GENERAL DISCUSSION

The original project objective, as stated earlier in this report, was to *predict* crime occurrence, hour-by-hour and sector-by-sector. Mathematically true prediction (that is, the probability of a crime occurring, given a set of conditions) requires the analysis of not only past crimes, but also of past conditions where *no* crime occurred. But whereas past crime data is more or less readily available, past data on *no* crime events is not readily available except in statistically summarized form.

Therefore, it was decided that *as a first cut*, the analysis would be limited to a sample of individual past crimes. Under this scheme, the "prediction" of burglaries would compare past occurrences of burglaries against past occurrences of other crimes; rather than against all past situations where no burglary occurred. Mathematically, this corresponds to determining the probability of a *burglary* occurring, given a set of conditions, and given that *some crime* occurs. Technically, this amounts to crime-type *discrimination*, rather than *prediction*.

The *first cut* approach is justified by the difficulty of collecting specific past situations where no crime occurred. The rationale is two-fold. First, if a particular crime type, for example burglary, is to be "predicted," then the conditions surrounding non-burglary crimes could be used to provide an approximation to the no-burglary population; secondly, the amount of error introduced by this approximation would be exceeded by at least an order of magnitude, by the inaccuracies in the sample crime data.

Using this initial approach, one of two results would occur:

- (1) the rationale would be supported by the "real world" results; or
- (2) the theoretical deficiency would prove to be critical, necessitating the introduction of "no-crime" normalizations into the analytical model.

In either case, this initial approach seemed the most appropriate first step.

Accordingly, the analytical models would operate in the first instance on the crime samples described in the previous sections. The primary analytical technique would be multidimensional analysis (MDA). A series of computer programs were developed to apply this technique to the crime data and to set up an initial version of the prediction model.

Three additional analyses were performed tangent to the primary MDA. The length and complexity of the MDA made a multiple regression analysis desirable as an effort to ascertain quickly the extent to which certain combinations of crime factors co-occur with certain types of crime. The results of the MR, in turn, precipitated an analysis of the frequency-distribution of crimes over the range of crime factor values, an attempt to discern the complex interrelationships between crimes and crime factors. Concurrent with these two analyses, a multinomial analysis was designed. The latter was envisioned as a means of verifying, over the entire crime sample, hypothetical crime factor to crime relationships suggested by any of the other analyses.

The frequency-distribution, MR, multinomial, and MDA analyses are discussed in the following sections.

B. FREQUENCY DISTRIBUTION

1. Theory and Approach

A basic analysis technique used with sampled data on a single variable is frequency distribution analysis. The range of values of the variable under study is divided into equal intervals and for each interval a tally of the collected data points which fall within that interval,

called a frequency count, is made. The frequency counts for all the intervals make up what is called a frequency distribution, that is, the frequency of observed values distributed over the range of the variable. This measure can give a quick indication of the central tendency and dispersion of the sample and, from these, these same properties for the base population can be inferred.

In order to compare frequency distributions meaningfully, certain standardizing procedures are necessary. In comparing two frequency distributions in which the sample sizes for the two are not equal, it is helpful to use a percentage frequency distribution. For each interval the raw frequency count is divided by the total number of observations in the sample and multiplied by 100 to obtain the percentage of observed values which fall in that interval. Comparisons of these percentages are then meaningful regardless of differences in sample size.

Normalization, another type of generalization of frequency distributions, is useful for certain types of variables. Normalization adjusts the raw or percentage frequency counts to reflect properly the different population frequencies in different intervals. For example, a frequency count of burglaries against temperature can be "normalized" to reflect the fact that certain temperatures occur more frequently than others; this is done by dividing the number of burglaries in each temperature range by the number of times that temperature range occurred during the year.

Comparisons of frequency distributions can be visual. However, a more rigorous comparison can be made by testing for significant difference between distributions using a proportional t-test, a statistical technique to measure the likelihood that the difference between two distributions is due to chance. If this likelihood is small the differences between the two is termed significant.

Frequency distribution analysis was employed in the study at hand to investigate further the relationships between different crime types and the crime factors. Distributions for each type of crime over the

range of every crime factor were prepared and normalized where necessary.

2. Methods

For each crime factor, tallies were made by type of crime, for every value in the range of that crime factor. The entire crime sample was used as the base population. Then, intervals were set up to cover the range of values for each crime factor and the tallies were compiled to form raw frequency counts for each interval. Percentage frequency counts were then computed for each interval for every crime type.

The frequency distributions for some of the crime factors were then normalized. Weather factors, such as phase of moon, snow on the ground, and atmospheric pressure, were normalized by the number of days during the time span of the crime sample (the year 1966) that fell into each crime factor interval. On such a crime factor the raw frequency counts for each interval were divided by the associated number of days and the resulting normalized frequency expressed as a percentage. Socio-economic and local data crime factors, such as income, percent non-white, and number of PTC transfer points, were normalized by the population for census tracts having certain values for the crime factor. In other words, the population of all census tracts having between 0% and 10% non-white would be totaled and used to normalize the frequency count of a crime type for that same interval. Again, the normalized frequency counts were expressed as percentages.

A set of bar charts were then prepared for several of the more interesting distributions. Each set pertained to one particular crime factor and contained a chart over that crime factor for each type of crime. Cross-factor and cross-crime comparisons of these charts were then made.

3. Results

For each crime factor, the raw frequency counts and percentage frequency counts were recorded on a tabulation sheet (see Appendix D). The intervals of the crime factor's range were recorded across the top

of the sheet and the counts for each type of crime were entered in rows below. For crime factors which were represented by scaled values, a converted scale for the interval was also provided across the top of the tabulation sheet.

The raw frequency counts for crime factors which were to be normalized were also recorded on tally sheets, again by crime type. The normalization factor, either number of days or population, was also recorded for each interval. The normalized frequency counts and percentages were then computed and recorded below the raw frequency counts on the same sheet (see Appendix D).

Comparisons among the bar charts developed showed some interesting differences, but revealed little that was new to the experienced police officer. Figures 4 through 10 show the bar charts for several of the more interesting distributions. As might be expected, differences were noted between crimes against property, such as burglary, larceny, and auto theft, and crimes against person, such as homicide, rape, aggravated assault, and robbery.

C. MULTIPLE-REGRESSION ANALYSIS

The mathematical technique of multiple regression analysis is used to study the relationships between the value of one variable, called the dependent variable, and the values of two or more other variables, called independent variables. In each instance a hypothesis is made that the values of the independent variables in some way determine or "predict" the value of the dependent variable.

The data required to perform a multiple regression analysis are a set of observations, each containing values for the several independent variables and the associated value of the dependent variable. Using the method of least squares, a linear equation of the form

$$y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

is determined which best "fits" the observed data; that is, the equation

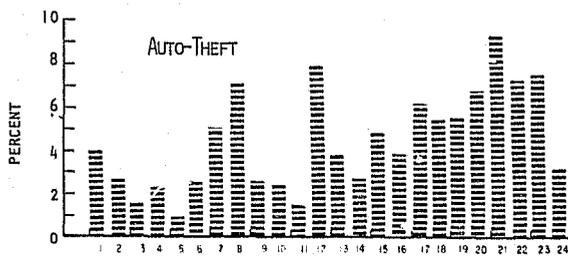
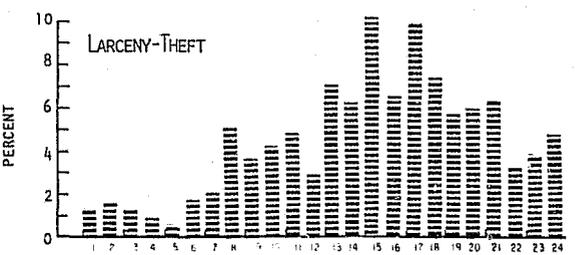
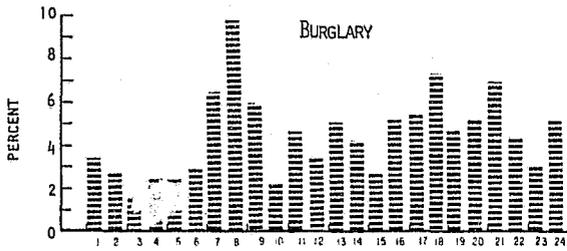
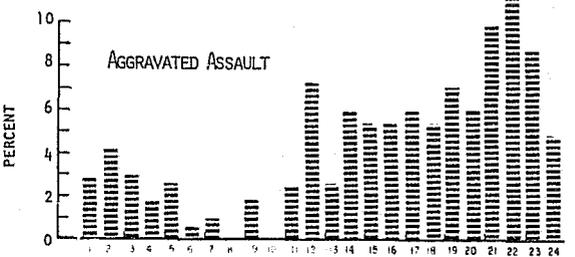
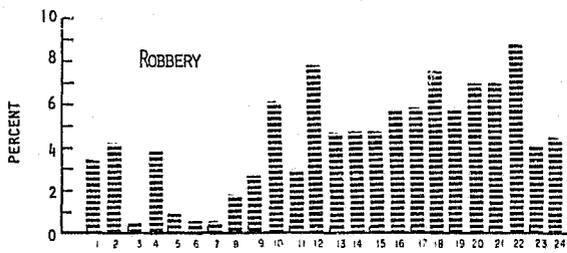
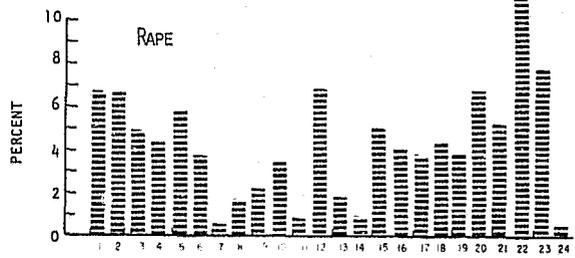
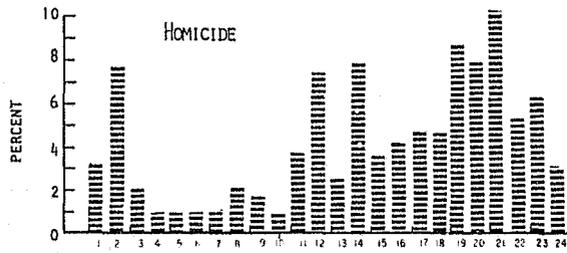


Figure 4. Frequencies of Crime Types by Hour of Day

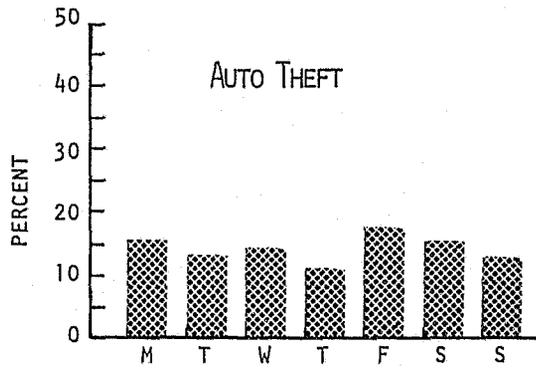
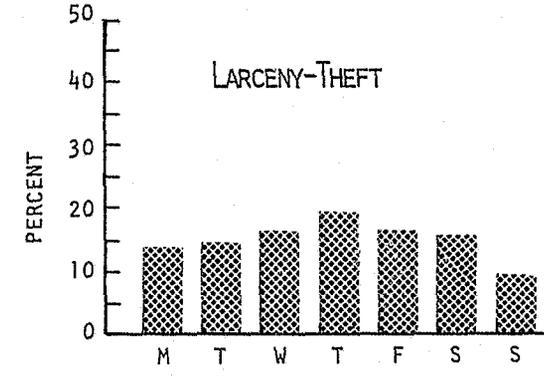
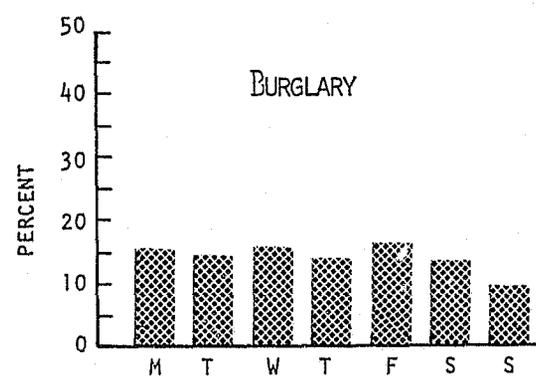
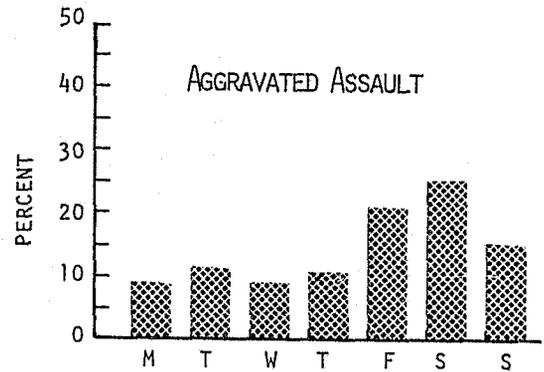
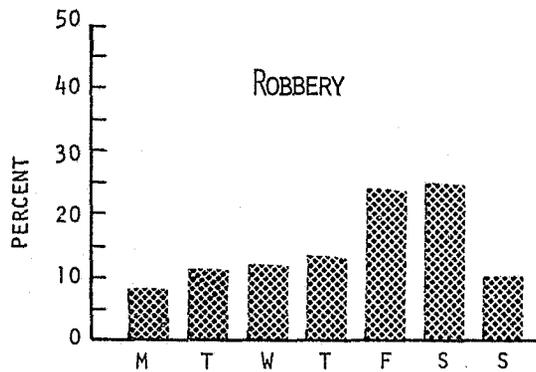
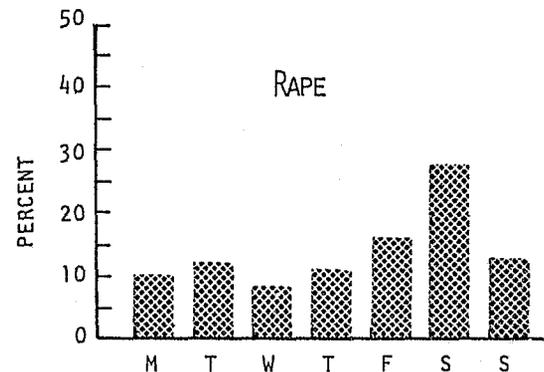
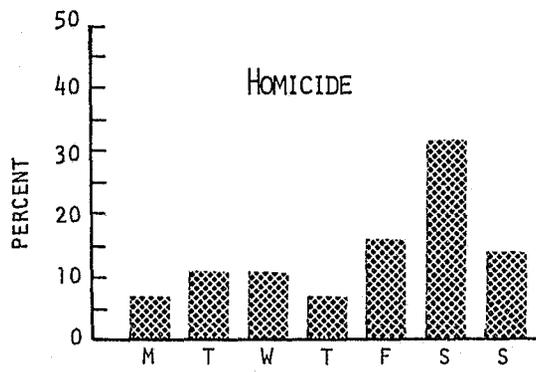


Figure 5. Frequencies of Crime Types by Day of Week

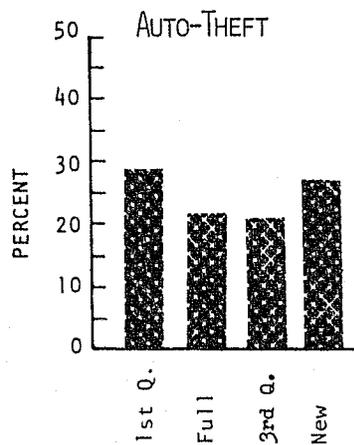
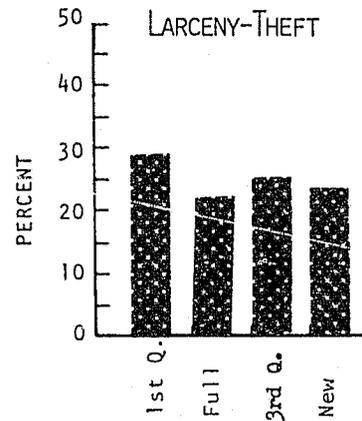
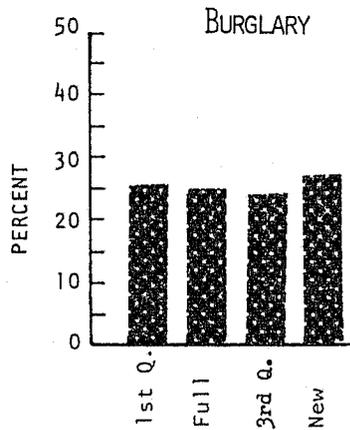
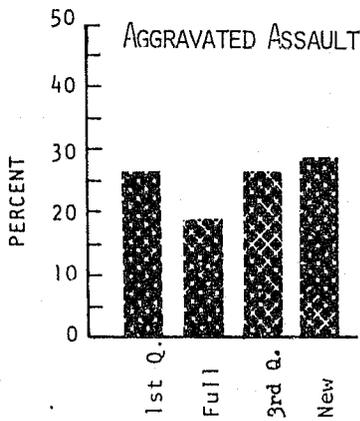
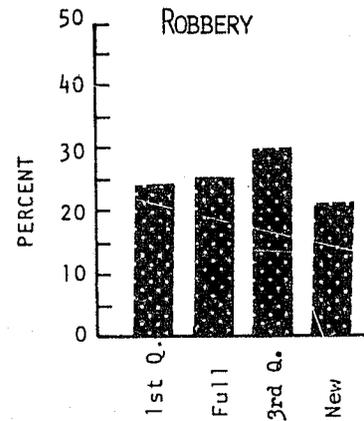
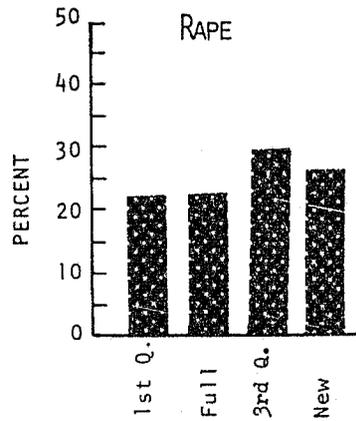
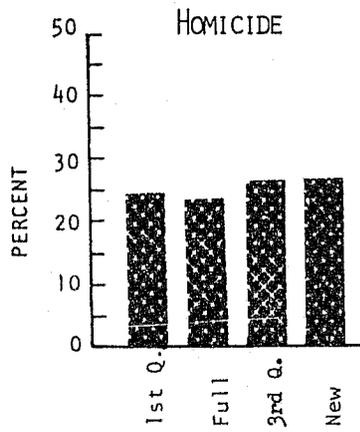


Figure 6. Normalized Frequencies of Crime Types by Phase of Moon

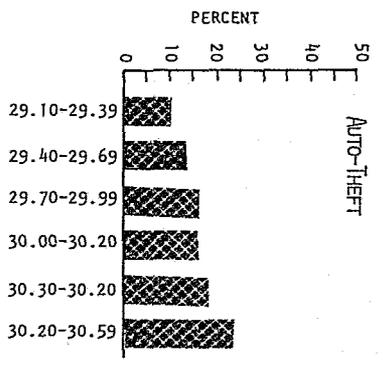
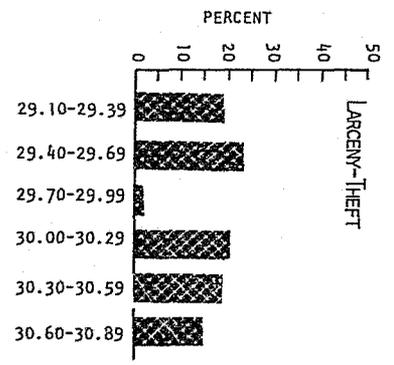
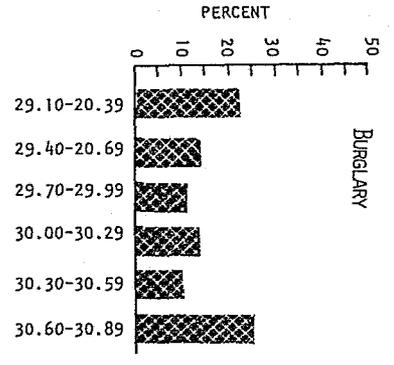
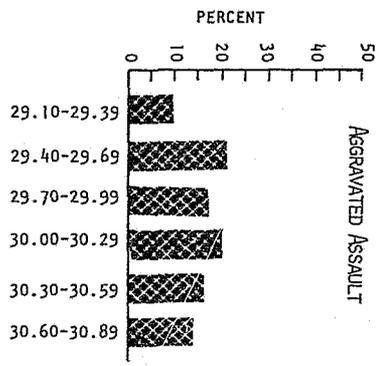
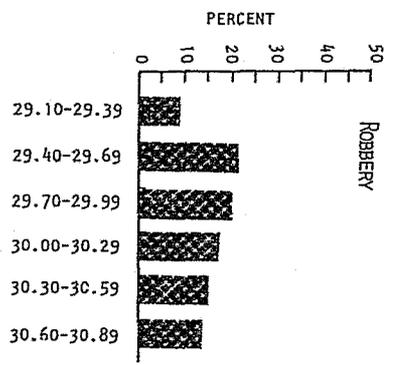
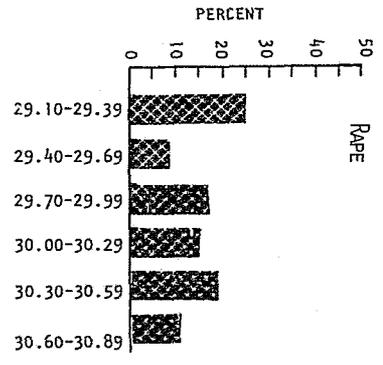
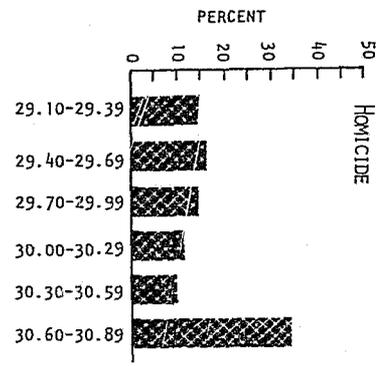


Figure 7. Normalized Frequencies of Crime Types by Atmospheric Pressure

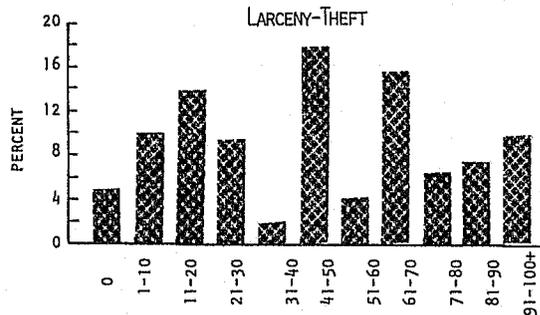
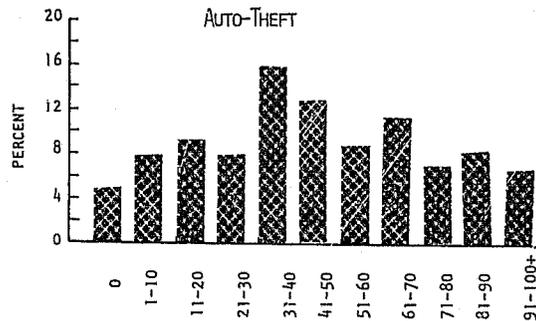
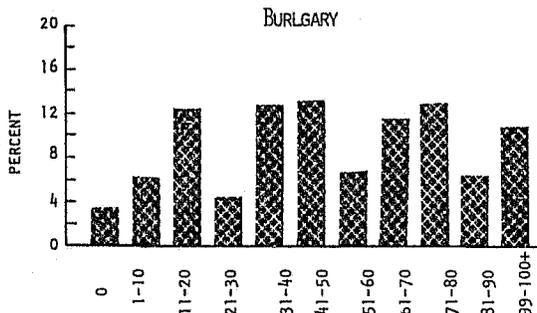
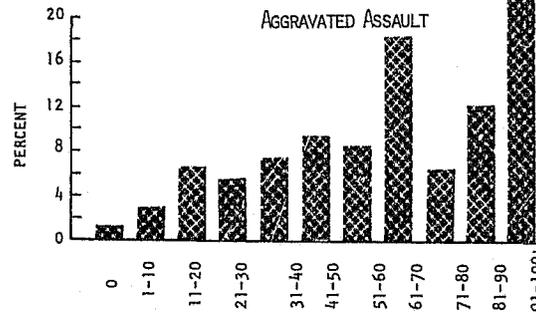
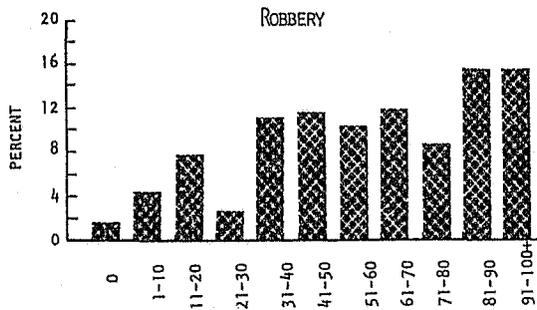
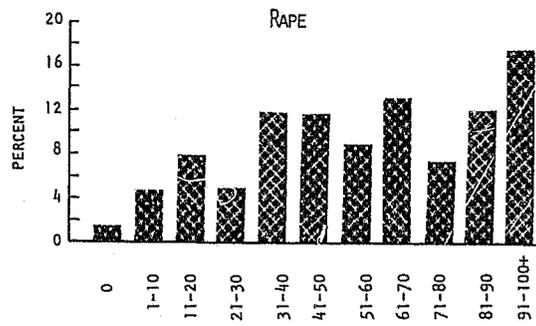
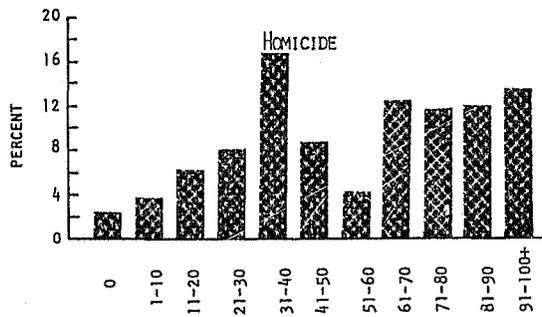


Figure 8. Normalized Frequencies of Crime Types by Percent Nonwhite Population in Neighborhood

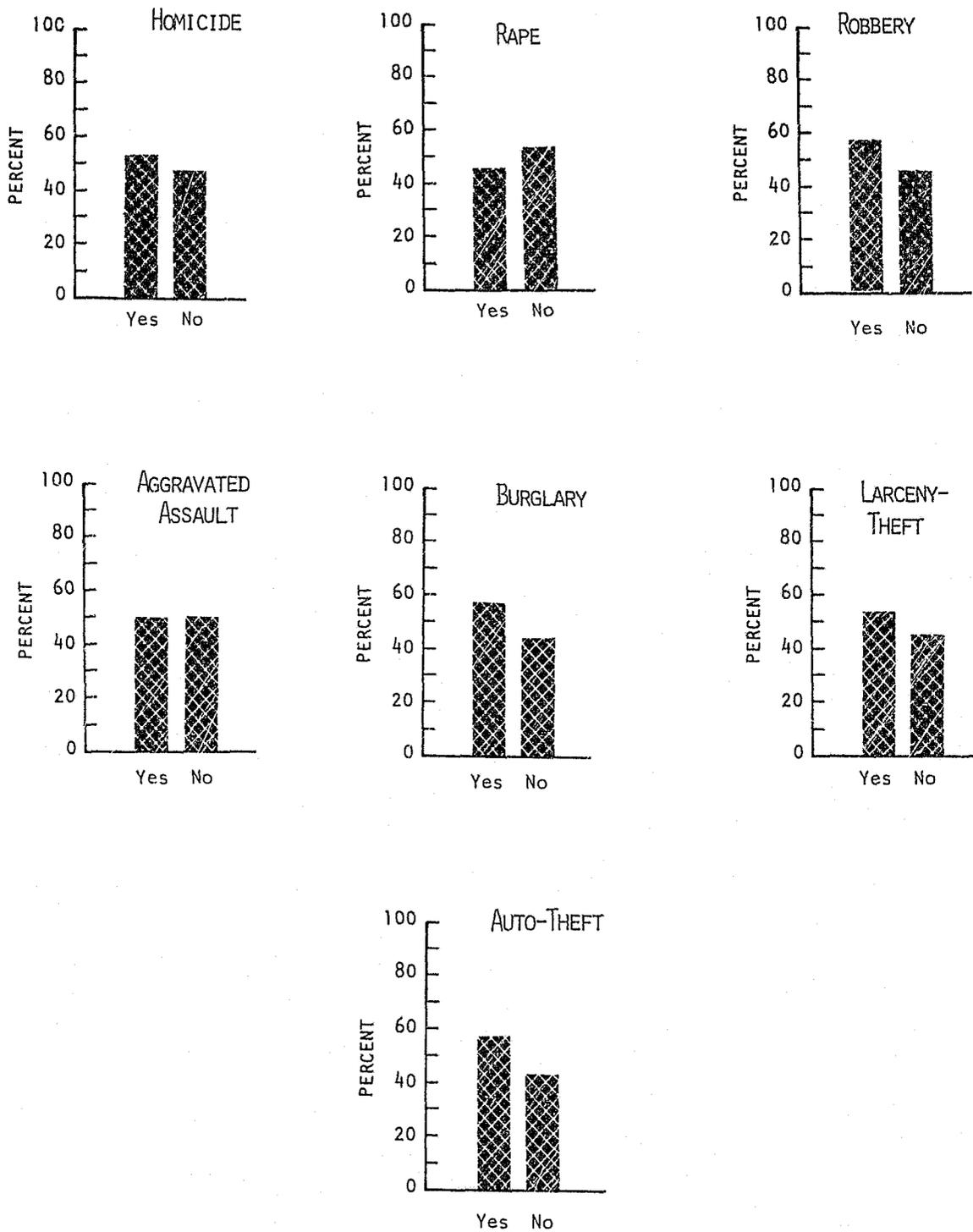


Figure 9. Normalized Frequencies of Crime Types by Presence of PTC Transfer Point(s) in Neighborhood

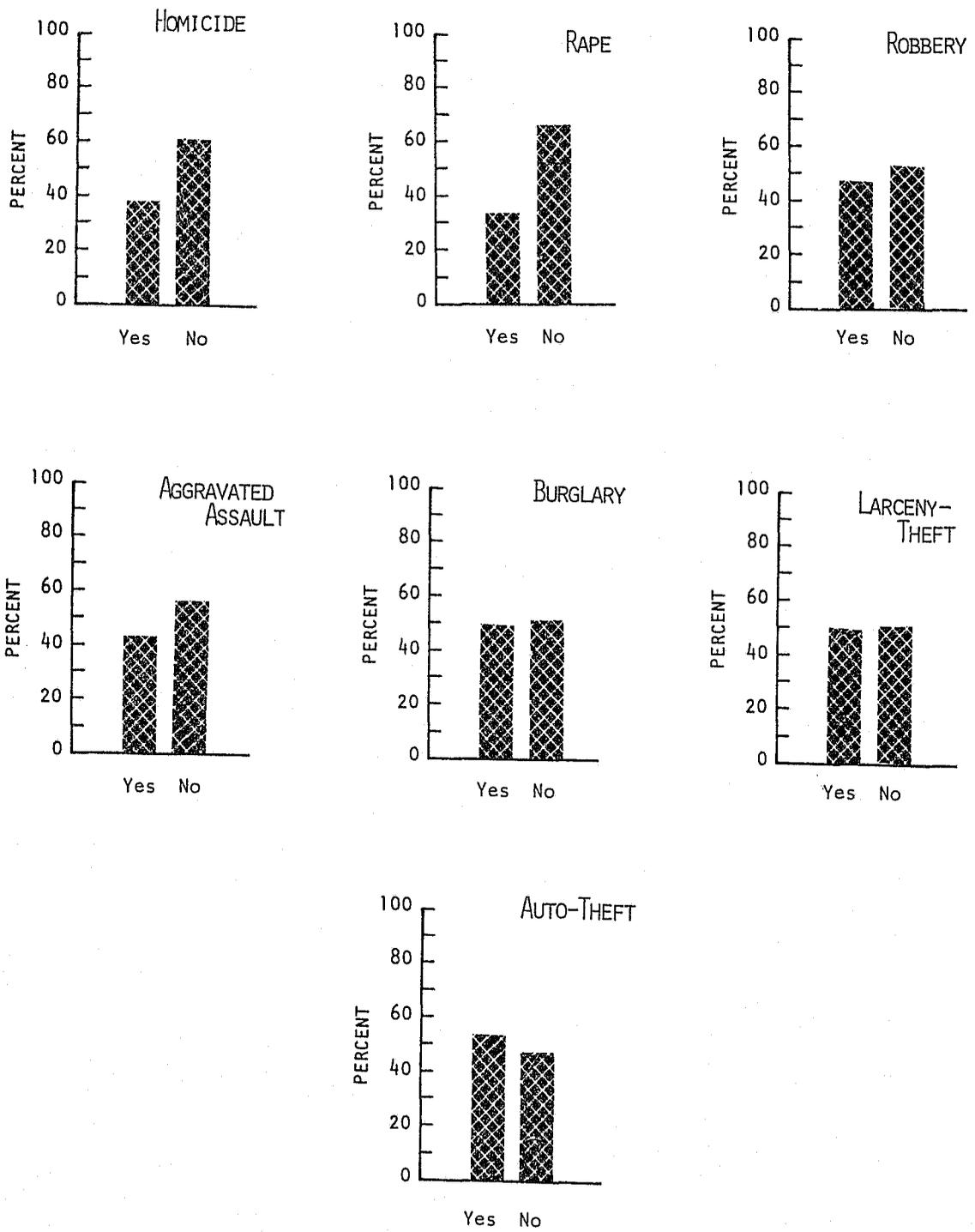


Figure 10. Normalized Frequencies of Crime Types by Presence of Senior High School in Neighborhood

for which, when the observed values of the n independent variables are substituted for the x_i 's, the value of y most nearly equals the observed value of the dependent variable.* The b_i ($i = 1, \dots, n$) in the above equation can be thought of as weights for the several independent variables because each b_i represents the relative influence the variable it modifies has on the value of the dependent variable; more specifically, b_i measures the change in y due to a unit change in x_i . The number a is a constant term which is used to preserve equality between the mean of the "predicted" values of the dependent variable and the mean of the observed values.

For best results, in the use of multiple regression analysis, two assumptions must be made about the data to be analyzed. First, it must be assumed that the selected independent variables are indeed related in a continuous linear fashion to the value of the dependent variable. If the relationship between the dependent variable and the independent variables is not linear or not continuous, the multiple regression will still produce the "best" straight line to fit the data and thus yield information as to the relative importance of the independent variables, but the equation obtained will not be a good "predictor" of the value of the dependent variable. That is, the differences between the predicted and observed values of the dependent variable may be quite large. Secondly, the assumption is made that the intercorrelations between the independent variables are at most negligible. If, in fact, high correlations do exist between two independent variables, the multiple regression will use one in the equation and delete the other under the assumption that the two variables will have the same effect on the value of the dependent variable and are thus redundant. For example, in a situation where n independent variables are being used to predict the value of y , and variables x_2 and x_3 are highly correlated, the equation produced may be of the form:

$$y = a + b_1x_1 + b_3x_3 + b_4x_4 \dots + b_nx_n.$$

(Note: An independent variable may also be dropped if the correlation between it and the dependent variable are found to be negligible.)

Whether or not these assumptions are met, the multiple regression analysis yields valuable information concerning the interrelationships among the

* see M.A. Efrøymesen, "Multiple Regression Analysis," in Mathematical Models for Digital Computers, Part V, (17), ed. by A. Ralston and H.S. Wilf (New York: Wiley, 1960).

variables involved and the "predictability" of the dependent variable and is thus a useful analytic tool.

It was decided to analyze the crime data using a multiple regression analysis to obtain a ~~quick~~ indication of the degree to which crime types might be discriminated from the associated crime factors. The crime data, however, were not ideally suited for multiple regression analysis. There was no indication that the crime factors were related linearly to the type of crime and indeed the dependent variable itself would have to be represented by a non-continuous variable, that is, "1" if the crime occurrence was of the desired type, and "0" otherwise. Correlations between the crime factors were expected to be high, especially among the weather factors. There was also the possibility of false correlations introduced by the restriction of the sample data to one year's time; for example, phase of moon and day of month might be correlated in a way peculiar to the year 1966. Ideally, for each type of crime, crime frequency counts could be made over the range of each crime factor and the proper transformation made to linearize these relationships. A sample over several years time could eliminate the chance of false correlations. However, the effort required for this type of data preparation was not consistent with the use of multiple regression as a *secondary* analysis tool; thus the multiple regression was run using readily available data instead.

The multiple regression analysis was performed twice; each time equations were derived for each of the seven crime types. The primary difference between the two applications, called Method I and Method II, was the type of data used for the independent variables.

In Method I, factor loadings, a by product of the multidimensional analysis, were used for the independent variables. These factor loadings were obtained for each crime in random samples 8 and 9 using the factor matrices for each of these samples which were produced during the course of the multidimensional analysis. The crime factor values for each of the sample crimes were weighted by the corresponding elements in each factor (or column) of the factor matrix and summed to form a loading on

each factor for each sample crime. Seven data matrices, one for each type of crime, were formed for each of the crime samples (see Figure 11). In each case the dependent variables were assigned according to the crime type under consideration; that is, "1" if the sample crime was of the type under consideration and "0" otherwise. The advantage of using factor loadings as independent variables was that the factors obtained from the factor analysis were calculated in such a way as to be independent of each other. Thus this method satisfied one of the necessary assumptions of multiple regression analysis. The method does have a disadvantage, however: since the factors are intricate combinations of the original crime factors, equations derived in this fashion are difficult to interpret in terms of the raw values of the crime factors.

Thus a second multiple regression was run, this time using the original crime factors as independent variables. The data matrix for Method II (see Figure 12) was constructed of the raw values for the crime factors for each of the large sample of 2800 crimes (see Table 8). The dependent variable, denoting type of crime, was constructed as before.

A basic result of each regression run was a simple correlation matrix showing the correlations among all the variables involved in the regression. Two correlation matrices were obtained for Method I, one based on Sample 8 as input and one based on Sample 9, while Method II yielded only one. Regression coefficients (b_i) and constant terms (a) were obtained for each equation for each crime type. There were twenty-one equations in all; fourteen derived by Method I and seven by Method II. For each of the derived equations, t-statistics and Beta coefficients were then calculated. The t-statistic, calculated for each regression coefficient, yields a confidence measure for that coefficient. For example, if the calculated t-value for a particular coefficient is greater than 2.0, the likelihood of that coefficient being determined by chance is only 5 out of 100. The Beta coefficients are a standardized form of the original regression coefficients. In other words, the regression coefficients are dependent on the unit of measure of the independent variables while the Beta coefficients are not. Thus Beta coefficients may be used to

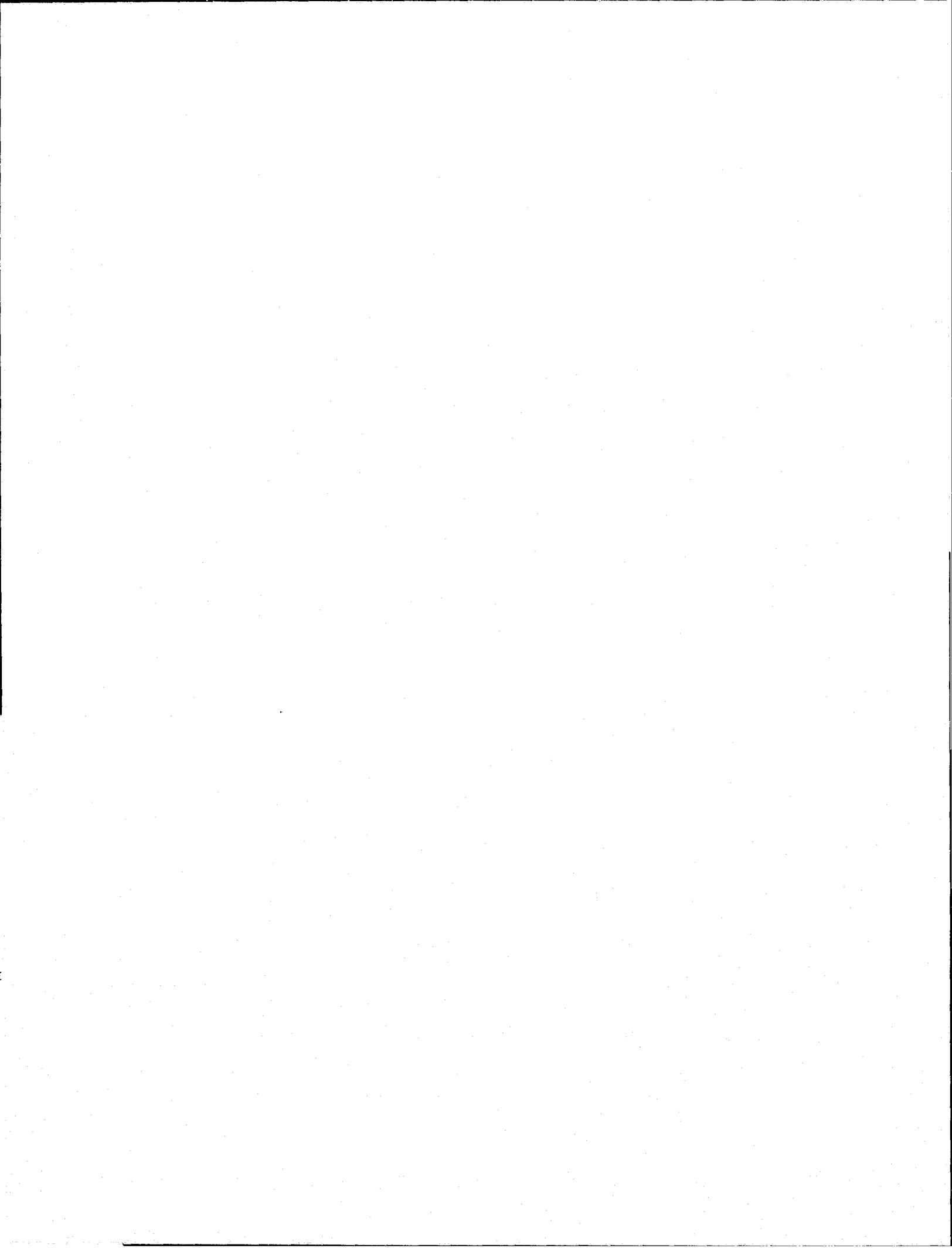
a. Sample 8

Crime	Homicide			Rape			Robbery			Aggravated Assault			Burglary			Larceny-Theft			Auto Theft			
	Factor 1	----	Factor 15	Factor 1	----	Factor 15	Factor 1	----	Factor 15	Factor 1	----	Factor 15	Factor 1	----	Factor 15	Factor 1	----	Factor 15	Factor 1	----	Factor 15	
1																						
2																						
3																						
⋮																						
100																						

b. Sample 9

Crime	Homicide			Rape			Robbery			Aggravated Assault			Burglary			Larceny-Theft			Auto Theft			
	Factor 1	----	Factor 16	Factor 1	----	Factor 16	Factor 1	----	Factor 16	Factor 1	----	Factor 16	Factor 1	----	Factor 16	Factor 1	----	Factor 16	Factor 1	----	Factor 16	
1																						
2																						
3																						
⋮																						
100																						

Figure 11. Data Matrices for Method 1



Crime	Crime Factors				
	1	2	3	...	35
1					
2					
3					
⋮					
2800					

Figure 12. Data Matrix for Method II

measure the relative influence of the several independent variables. The Beta coefficients may not, however, be used for comparison between variables from two different regression equations. Also for any coefficient, Beta and t must be interpreted simultaneously; if t is not significant, statements about Beta are of little value (see Appendix E for table listings of these results).

Each of the regression equations obtained was tested by substituting the observed data values for the independent variables and comparing the computed value of the dependent variable, called the estimated value, with the observed value. The estimated value was thus obtained for each of the 2800 sample crimes. Tallies of the range of estimated values for each crime type were made for each regression equation. Ideally, for any particular equation, the estimated values for crimes of the type to be discriminated by that equation should all be 1.0 while crimes not of that type should receive estimated values of 0.

Unfortunately, none of the regression equations obtained by Method I - using the factor analysis outputs - was able to discriminate crime types successfully over the sample of 2800 crimes. For example, one of the two regression equations obtained for homicide was able to identify

Table 8. Crime Factors as Independent Variables

REGRESSION ANALYSIS	INDEPENDENT VARIABLES			RANGE
1	Snow	Inches	Scaled	0-100
2	Visibility	Miles	Scaled	0-100
3	Precipitation	Inches	Scaled	0-100
4	Wind Speed	Knots	-	0-100
5	Temperature	Degrees	Scaled	0-100
6	Relative Humidity	Percent	-	0-100
7	Pressure	Inches of Mercury	Scaled	0-100
8	Income	Dollars	Scaled	0-100
9	Persons/House	Number of Persons	Scaled	0-100
10	Rent	Number	Scaled	0-100
11	School years completed	Number of Years	Scaled	0-100
12	PTC			0-100
13	Elementary School(s)			0-100
14	Junior School(s)			0-100
15	Senior High School(s)			0-100
16	Month			1-12
17	Day			1-31
18	Hour			1-24
19	Age Percent 15-34	Percent		0-99
20	Age Percent 60 and over	Percent		0-99
21	Percent Males Unemployed	Percent		0-99
22	Percent Wage and Salary	Percent		0-99
23	Percent Owner-Occupied	Percent		0-99
24	Percent Sound Housing	Percent		0-99
25	Percent with 1.01 or more per room	Percent		0-99
26	Percent Married	Percent		0-99
27	Percent Foreign-Born	Percent		0-99
28	Percent Growth	Percent		0-99
29	Percent Decline	Percent		0-99
30	Percent Moved	Percent		0-99
31	Percent Families, 1 or more under six years	Percent		0-99
32	Percent non-white	Percent		0-99
33	Percent Enrolled in school	Percent		0-99
34	Day of Week			1-7
35	Phase of Moon			1-4

91% of the homicides, but only 20% of all the crimes identified as homicides were actually homicides. That is,

$$\frac{\text{number of homicides identified correctly}}{\text{total number of homicides}} = 91\%$$

$$\frac{\text{number of homicides identified correctly}}{\text{total number of crimes identified as homicides}} = 20\%$$

Similar results were observed for the other regression equations.

A graph comparing the distribution of two different types of crime, burglary and aggravated assault (see Figure 13), further illustrates the behavior of the regression equations obtained using Method I. These distributions were obtained by plotting the percentage of crimes, burglaries or aggravated assaults, that received estimated values of 0, 0.1, 0.2, ..., 1.0 using the burglary regression equation. This graph shows that while the estimated values for burglary deviated widely from the observed

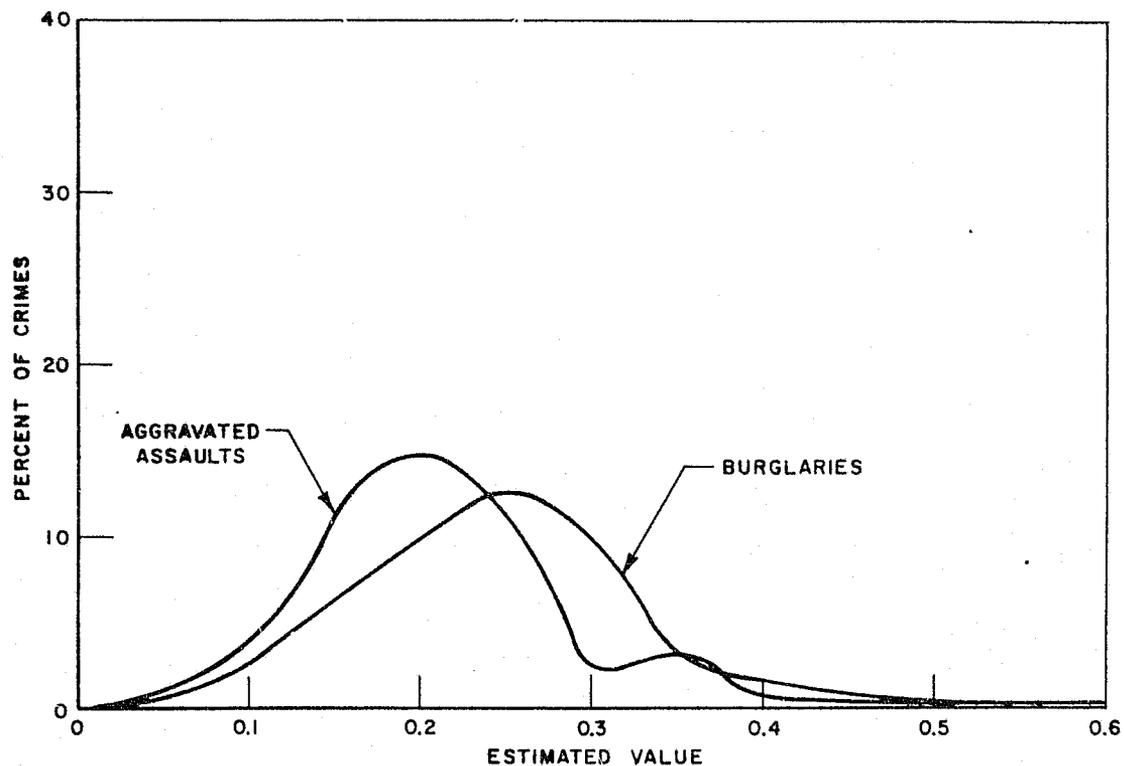
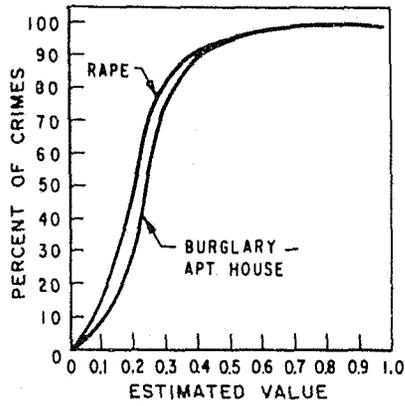


Figure 13. Results, Method I Comparative Distributions of Estimated Values for Burglaries and Aggravated Assaults

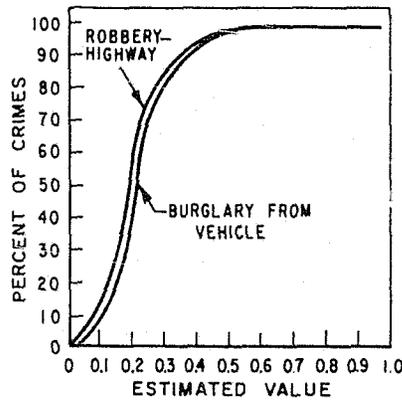
value of 1.0 and the estimated values for aggravated assault ranged far above the observed value of 0, there was a noticeable separation in the two distributions. The burglary distribution is definitely shifted toward the 1.0 end of the scale, the end that denotes burglary. This slight but definite separation suggests that while some of the Factors contribute to both aggravated assault and burglary there are some which do discriminate between these types of crime. The appearance of such a discrimination in the results of such a restricted analysis supports the hypothesis that a further refined multiple regression analysis might be able to discriminate more accurately.

The results of Method II were again slight but promising. Of the seven regression equations none was successful at discriminating the crime type it represented. Again, however, the distributions of estimated values for the several crime types were distinguishable. Figure 14 shows the cumulative distributions of crime types within the burglary classification as compared with cumulative distributions for non-burglaries (all were obtained using the burglary equation). As might be expected, the smallest separation occurred between the distributions for highway robbery and burglary from a vehicle, while the largest occurred between willful killing and burglary from a non-residence. This suggests the reasonable hypothesis that the crime factor values co-occurring with the former two types of crime are quite similar, while those co-occurring with the latter pair are not. These situations typify the results of the other six equations.

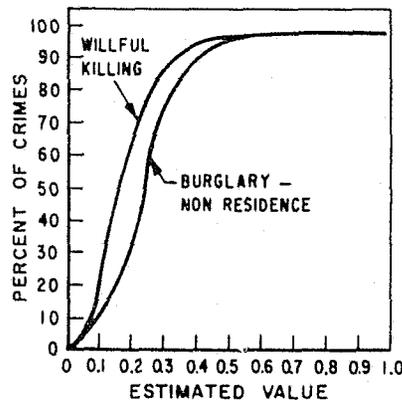
Given the several limitations under which the multiple regression analysis was run, results showed the multiple regression technique to have definite promise. With the proper normalization of crime factors and the reordering of regression-variable ranges according to increasing frequencies of crime occurrence, the chances of obtaining good discrimination would be substantially increased. Then, inclusion of "no crime" situations in the data would allow actual crime *prediction*.



a. Rape and Apartment - House Burglary



b. Highway Robbery and Burglary from Vehicle



c. Willful Killing and Nonresidence Burglary

Figure 14. Cumulative Distributions of Burglaries and Nonburglaries

D. MULTINOMIAL ANALYSIS

1. Theory and Approach

Theoretically, it is a simple process to determine specific combinations of crime factors which co-occur with certain types of crimes. It would suffice to enumerate all possible combinations of values for the crime factors and then count the number of observations in the sample which have each combination of values. Then, given any combination of crime factor values, one could check this combination against the tallied observations and determine which type of crime to "predict."

This exhaustive analysis, while ideal in theory, rapidly becomes unmanageable in practice. For instance, if one were considering 15 crime factors, each of which could have ten different values, the total number of combinations to be investigated would be 10^{15} . Thus, for this study of *thirty-eight* crime factors, each of which could have possibly *one hundred* different values, the multinomial technique was discarded as impractical as a primary analytic tool.

On a smaller scale, however, the multinomial technique retains its usefulness. To test a specific hypothesis, that is, a hypothesis that a certain combination of crime factors co-occurs more frequently than not with a certain crime type, it would be possible to use the multinomial approach. The crime sample could be searched to ascertain how many crimes co-occurred with the required combination of crime factor values. This group could then be divided into crimes of the desired type and crimes not of the desired type. A comparison of the number of crimes in these two groups could then serve to support or reject the hypothesis. Thus it was decided to use multinomial analysis to verify hypotheses produced by the other analytical techniques.

2. Methods

A computer program, called JCSRCH (see Appendix G), was designed to perform the multinomial analysis to test specific hypotheses. Given a set of crime factor values and a specific crime type, the program

searches through the entire crime sample and tallies two sums:

S_1 : the number of crime occurrences for which the given crime factor combination exists but which are not of the given crime type; and

S_2 : the number of crime occurrences for which the given crime factor combination exists and which are of the given crime type.

After these sums are found a ratio of occurrence is computed.

$$R = \frac{S_2}{S_1 + S_2}$$

This ratio gives an indicative measure of the degree to which the given combination of crime factors co-occurs with the given crime type and accordingly reinforces or weakens the hypothesis being tested.

3. Results

The multinomial analysis was written to test specific hypotheses; it was subsequently debugged and tested. Figure 15 shows the results for one test hypothesis which was run. The hypothesis to be tested was "a larceny (crime-code 616) is the most likely crime type in sector 23C and census tract 47D on Tuesday, November 1, 1966 when the moon is full." The printout gives the crime type being tested followed by the given conditions; note that the values for some of the crime factors are coded representations, such as 2303 for 23C and weekday 2 for Tuesday. Next follows the ratio of occurrence, in this case 0.333333. The true numerator and denominator are also given to indicate the relative importance if the ratio; that is, the ratio of 0.333333 would assume more importance of its numerator and denominator were 33 and 99, respectively, rather than the 1 and 3 indicated in this test.

The range and confidence level of the ratio then can be determined with the aid of appropriate binomial tables.

In general the output from a run of the multinomial analysis program includes, for each hypothesis given, the computed ratio of occurrence and the values of S_2 and $S_1 + S_2$. Also, if the given combination of

PROBABILITY OF LARCENY 616.00

GIVEN SECTOR 2303.00

CENSUS TRACT 4704.00

YEAR 66.00

WEEKDAY 2.00

MOON 2.00

MONTH 11.00

DAY 1.00

EQUALS 0.333333

NUMERATOR EQUALS 1.0

DENOMINATOR EQUALS 3.0

Figure 15. Sample Output from Multinomial Analysis Program

crime factor values contains more than one value, then a series of ratios, S_2 values, and $S_1 + S_2$ values are computed, one set for the full list and one for each sublist, subtracting one of the given values each time. For example, if the hypothesis to be tested was "burglary given knife, non-residence, and 10 PM," the R , S_2 , and $S_1 + S_2$ values would be computed for

- a. burglary given knife, non-residence, 10 PM;
- b. burglary given knife, non-residence; and
- c. burglary given knife.

This capability for multiple results was added mainly for the practical purposes of investigating several different hypotheses during one computer run.

As crime factor combinations are identified with crime clusters by the multidimensional analysis, these hypotheses can be tested over the entire crime sample using the multinomial technique.

E. MULTIDIMENSIONAL ANALYSIS

1. Theory and Approach

Multidimensional analysis comprises any of several statistical techniques used to study the interrelationships among a group of events. The events, each described by a given set of measures, are arrayed as points in a multidimensional space. Then, the several dimensions and the projection of each event on each dimension are identified. Events which have similar loadings, or projections, on the several dimensions would tend to "cluster" in the space; the cluster, then, denotes a particular type, or class, of event. Once the space is established, a new event can be arrayed in the space and identified merely by noting within which, or near which, cluster, if any, the new point falls.

In the study at hand, MDA was selected as an appropriate technique to study the relationships among crime occurrences, where each crime is described by specific values of the crime factors. A crime space, or crime distance space, could then be formed, its dimensions representing complex combinations of the original crime factors. Once a multidimensional crime space is achieved, each crime occurrence could be arrayed in the space and crime clusters could be identified. Finally, the ability to identify a new set of crime factor values with one or more of the existing crime clusters would add the "predictive" capability to a model based on the MDA technique.

In order to use standard MDA techniques, however, the measures involved, in this case the crime factors, are subject to two restrictions. First, each measure must be quantifiable on a monotonically increasing scale; that is, if v_i is the value on a certain crime factor for crime i and if the distance between two crimes on any crime factor is given by

$$d_{AB} = |v_A - v_B|$$

then if A, B, and C are three crimes such that $v_A > v_B > v_C$, d_{AC} should be greater than d_{AB} on the crime factor in question. Secondly, the

distance between crimes on each crime factor must obey the properties of a metric space; that is, for any crimes A, B, and C

$$d_{AB} + d_{BC} \geq d_{AC}$$

Most of the crime factors satisfy these requirements. However, a certain few, such as "day of the week," are cyclic and thus violate the first restriction above. For example, if crime A occurs on a Tuesday, denoted by day 2, and crime B occurs on a Sunday, day 7, the distance between crime A and crime B on the "day of the week" factor is not 5 days, but 2 days; that is

$$d_{AB} = 2 \neq |v_A - v_B| = |2 - 7| = 5$$

In such a case, the actual differences do obey the metric property, but the actual differences are not the same as the numeric differences between the crime factor values.

These data limitations led to the selection of an MDA technique which was unconventional in that it did not require the measures used to be quantifiable on a monotonically increasing scale. This technique, developed by Tucker and Messick, relies upon the differences between each pair of events on each measure used rather than the actual values of the measures.* In deriving these differences special techniques were developed and used to assure the computation of the actual, rather than the numeric, differences for each cyclic measure.

The Tucker-Messick technique also differs from conventional MDA in that it provides not one multidimensional space to represent the group of events being studied, but several spaces, each one representing a different but consistent classification of these events. In this way, clusters of events which may have been overlapping and confounded in one multidimensional space may be seen clearly in several multidimensional spaces each presenting these events from a different perspective.

*L. R. Tucker and S. Messick, "An Individual Differences Model for Multidimensional Scaling," Psychometrika XXVIII (December, 1963), pp. 333-367.

The Tucker-Messick method was applied to data samples for each type of crime. Thus several multidimensional spaces were derived for each crime type, each describing crimes of that type from a different point of view.

The next step was to set up a framework for projecting a new set of crime factor values ("current conditions") into one or more of the existing crime spaces and determining whether they cluster with the sampled crime. Theoretically, new sets of crime factor values could be arrayed into new spaces corresponding to each of the derived spaces.* The matrices, which represent the new spaces and the established spaces, could then be compared using a method derived by S. Messick.† The results of this comparison would then indicate the likelihood that the new event comes from the same population as the old events. This method does not require the actual identification of clusters within the several crime spaces.

Computer programs were written and designed to carry out this analysis. However, it became apparent that the physical size of the crime data arrays and the complexity of this analysis rendered this mode of comparison unmanageable. The computing procedures were lengthy and the data storage and manipulation requirements were unwieldy and time consuming. Accordingly, several alternate techniques of cluster analysis were developed. As clusters were identified they were tested against the conditions surrounding a sample of known crimes.

2. Methods

As described in Section 2-B-2, random samples of 100 crimes were selected for each crime type, seven in all, and two additional random samples were also selected without regard to crime type. For each of these samples the analytical steps below were followed:

- a. A crime by crime factor matrix, called C, was constructed, where each row represented a crime and each column a crime

*Ibid., p.341.

†S. Messick, "Within-Group Covariance Factor Analyses: Notes on a Model due to L. Tucker" (unpublished manuscript), pp. 1-11.

factor (see Appendix G, Program JCADATA). The scaled values of the crime factors for each sample crime are recorded in this matrix.

- b. Differences were then formed between each possible pair of crimes on every crime factor, giving special consideration to the cyclic crime factors. These differences were arrayed in a matrix X, in which each row represents a crime pair and each column, a crime factor. From X a matrix called P was constructed by the rule $P = X^T X$. P is a symmetric matrix similar to a correlation matrix and thus can be factor-analyzed (see Appendix G, Program JCPREP).
- c. A factor analysis was then performed on matrix P, using the method of Principal Components (see Appendix G, Program JCFACT) to obtain the independent factors, or viewpoints, on which the multidimensional spaces were to be based. The factor analysis yielded a factor matrix A, which contained the independent factors as rows, and a diagonal matrix Γ^2 containing the eigenvalues of the matrix P.*
- d. The factor matrix A was then rotated obliquely (see Appendix G, Program JCROTA) to obtain a more meaningful representation of the derived factors in terms of the original crime factors. This rotation yielded a transformation matrix T and the rotated factor matrix B, where $B = TA$.
- e. The original difference matrix X was then converted into a matrix Z of crime pair projections in the several multidimensional spaces by the process

$$Z = (XA^T (\Gamma^2)^{-1}) T^{-1}$$

(See Appendix G, Program JCMATM.)

Each column in Z now represents a separate multidimensional space and each entry within a column represents the distance in that space between two crimes, corresponding to the crime pair rows of X.

Once multidimensional spaces, represented by the columns of Z, were formed for each crime type, the problem of cluster analysis was tackled. As previously mentioned, the theoretical approach was explored but eventually discarded as unmanageable. Consequently, several cluster analysis techniques of a similar format were tried. Each point in the space was considered a potential cluster center. Using the known inter-point distances, points were then added one by one to the cluster starting

*Tuoker and Messiook, op. cit., p.338.

with the point nearest the chosen cluster center. In this manner a set of clusters was constructed around every point in the space. Now, two things were needed to determine the actual clusters. First, an algorithm was necessary to determine at what point the cluster ended, that is, where the process of adding a new point to the cluster should cease and the cluster be termed complete. The distance from the last point added to the cluster center in this way could then be called the radius, or perimeter, of the cluster. Second, a figure of merit, or measure of "goodness", was needed so that the best clusters in the space could be selected for use in the identification and prediction process.

The clustering techniques differed in how the figure of merit and radius criterion were computed and in what use was made of non-burglary data. In several techniques, the figure of merit was some combination of a measure of dispersion for the points in the cluster and the number of points itself.

A simple example would be

$$FM = \frac{N}{\sigma^2}$$

where N is the number of points in the cluster and σ^2 is the variance of the points from the cluster center. Other techniques used a measure of the actual versus the expected distribution of points in a cluster, considering the distances from the center to the points in the cluster as a probability distribution. The non-burglary data was used in some instances to normalize the figure of merit. The radius criteria used were just as varied. Sometimes the FM was calculated for each step in the cluster formation process and the step with the maximum FM selected as the cut off. Another approach was to divide the distances into intervals between the center and the mean distance from the center and then select the interval with the least number of points between the maximum FM and the mean. Programs were written to accomplish these several clustering techniques but no documentation is included since no firm and final technique was selected as best.

3. Results

The several clustering techniques were applied successively to the homicide and burglary samples; and each set of clusters obtained was tested against the two random samples of mixed crime types (Samples 8 and 9). A computer program, not finalized and thus not documented, was set up to perform this testing in the following manner.

Each set of crime factor values for a random crime was treated as a current situation, CS; that is, as if the crime factor values were known and the crime type were to be predicted. Differences were formed between this CS and all the crimes in the burglary sample, thus forming a modified version of the X matrix, called X_{CS} . Using the method by which the original X matrix was converted into the Z matrix (see previous discussion of method, part e.), an extension of the Z matrix, called Z_{CS} , can be formed from X_{CS} . Each column of Z_{CS} now gives the distances between the CS and the crimes within each multidimensional space. Prediction can then proceed as follows; if crime 6 in multidimensional space 1 has been determined the center of a cluster of radius r_{61} , since entry z_{61} of matrix Z_{CS} represents the distance between the CS and crime 6 in multidimensional space 1, $z_{61} \leq r_{61}$ means that CS falls within the determined burglary cluster. This prediction can then be tested against the true crime type of the CS. The discriminating ability of any clustering technique can then be assessed by ascertaining the number of correct identifications which can be made using the clusters so defined. No clusters were found which discriminated homicides. Figures 16 through 19 show some initial results of the clustering and testing techniques for burglaries. Figures 16 and 18 show two clusters developed in the burglary sample. The distribution of the entire burglary sample about the cluster-center is shown in each with an arrow marking the radius which was selected as the boundary of the cluster. Figures 17 and 19 show the test results over the two random samples of mixed crime types, 200 crimes in all, for each of the clusters shown. For cluster "A", Figure 16 shows that the non-burglaries which fell within the cluster

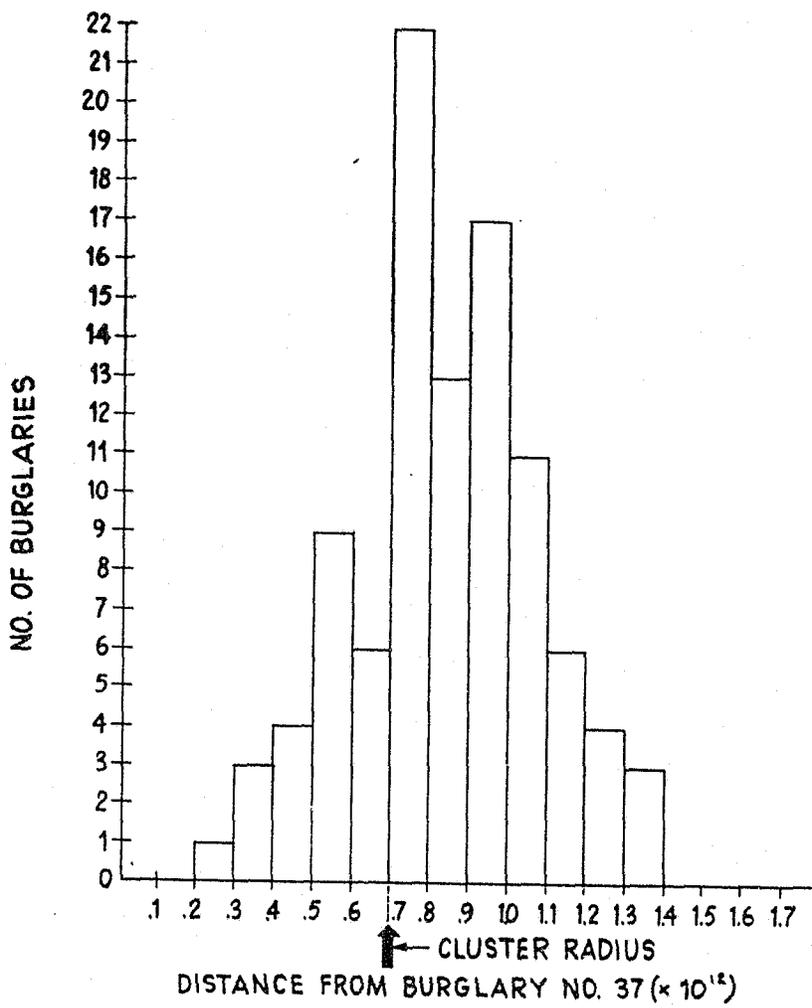


Figure 16. Burglary Cluster "A" About Burglary No. 37 on Factor II

radius consisted of robberies and larcenies, both crimes similar to burglary in nature. The test results for cluster "B", Figure 19, are similar but here larcenies were the only non-burglaries identified as burglaries. In both cases crimes against the person, such as homicide, rape, and aggravated assault, never fell within the limits of the burglary cluster.

These results support the present model's discriminative capabilities. A determination of the model's *predictive* capabilities awaits completion of implementation of the operating model (See Section 4 of this report).

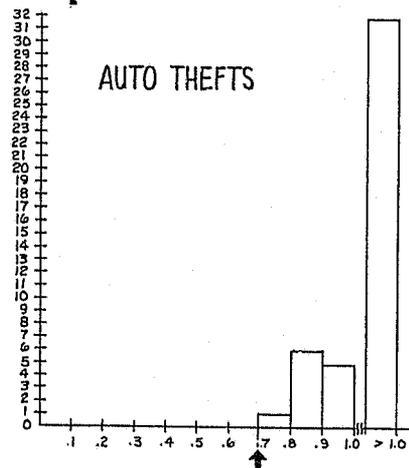
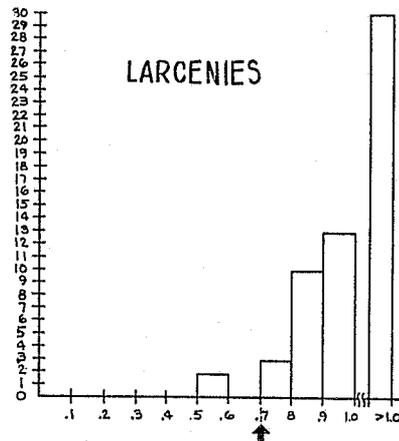
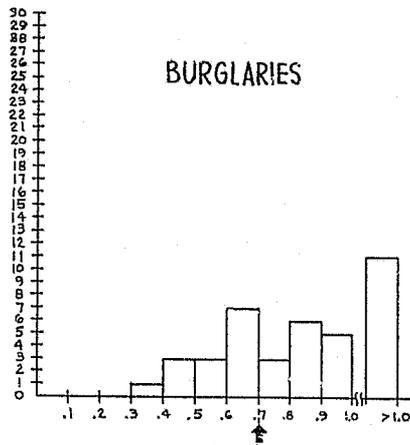
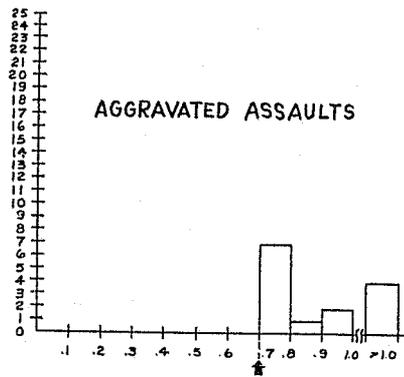
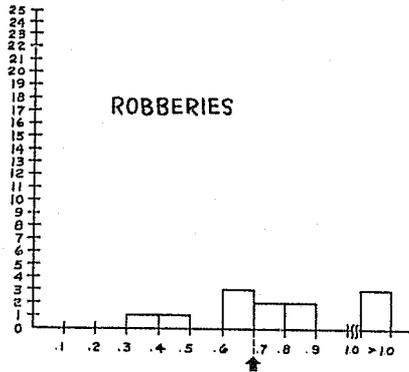
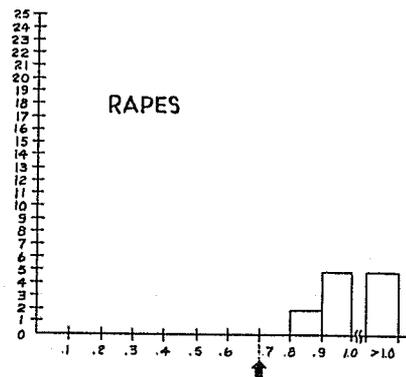
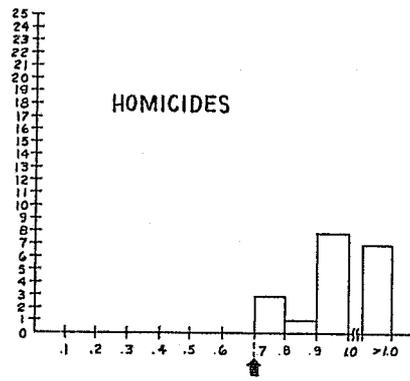


Figure 17. Distribution of 200 Random Crimes with Respect to Burglary Cluster "A"

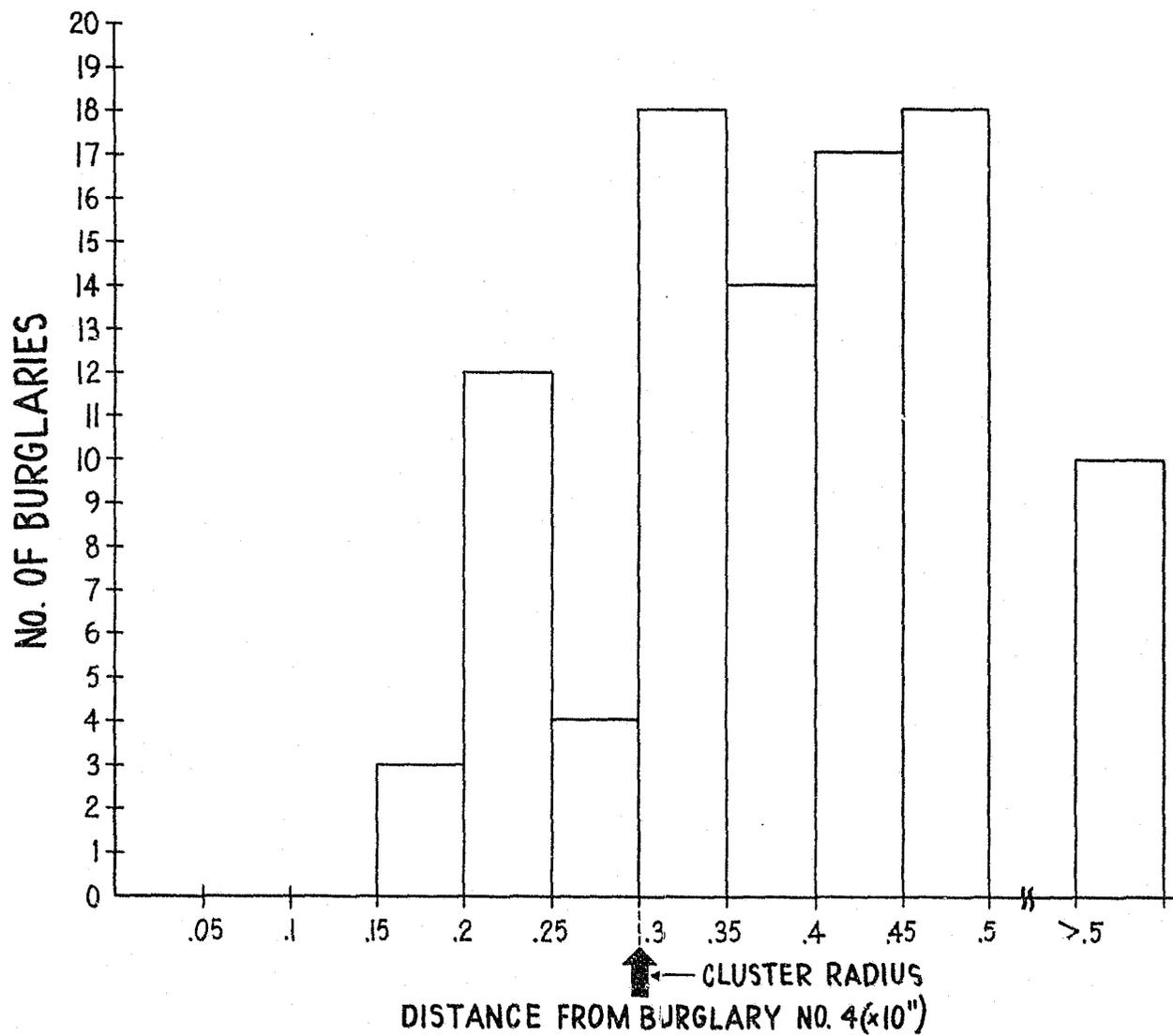


Figure 18. Burglary Cluster "B" About Burglary No.4 on Factor IX

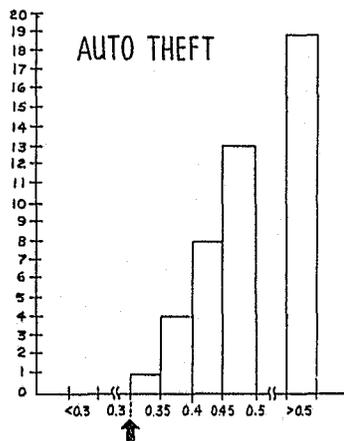
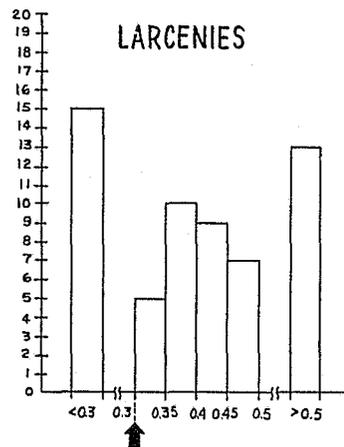
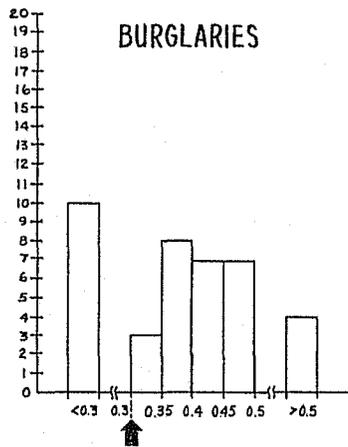
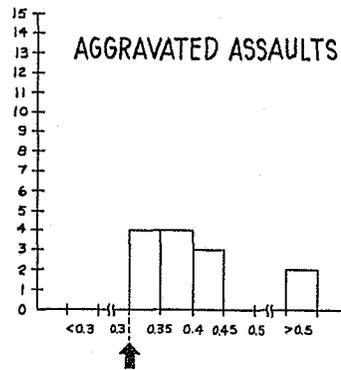
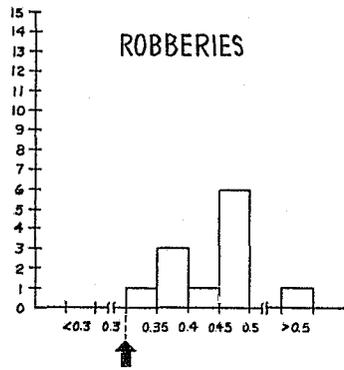
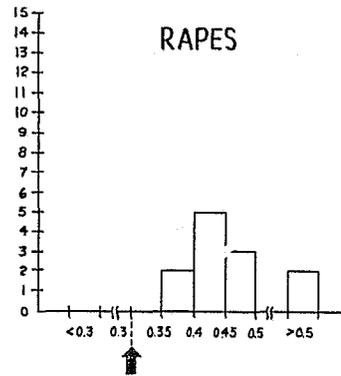
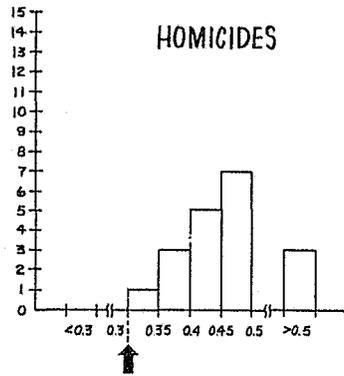


Figure 19. Distribution of 200 Random Crimes with Respect to Burglary Cluster "B"

SECTION 4 INITIAL IMPLEMENTATION

To learn more about the everyday operation of the system and obtain some indication of the strengths and weaknesses of the analysis, an initial version of the operational model was prepared. The initial version, designed for implementation on the Philadelphia Police Department computer system, uses magnetic disks for external storage and a remote terminal for input/output. The actual programming of the model, as well as the creation of the disk data bank, was done by members of the PPD staff, with the assistance of the FIRL project staff.

A. DATA REQUIREMENTS

For this first attempt, the model's scope was limited to a consideration of burglary conditions. Data pertaining to those clusters selected by the multidimensional analysis as "burglary clusters" were therefore needed to operate the model. These data and the other two data arrays necessary were stored on a magnetic disk external to the machine, a storage mode which prevents the program model from being limited by a computer with a small internal memory. Four disk storage files were set up.

The first file contains the values of the twenty-three socio-economic crime factors (see Table 9) associated with each sector. The file was organized by police districts, twenty-two in all, and by police sectors within districts. Each sector record contains the sector identification, the district number and an identifying letter, and the twenty-three variable values associated with that sector.

The second and third files contain data which pertain to the selected "burglary clusters." The first contains records of cluster-center data

Table 9. Crime Factors Used by Operating Model

CRIME FACTOR	UNIT OF MEASUREMENT
Day of Week	
Month	
Day	
Hour	
Phase of Moon	
Snow	Inches
Visibility	Miles
Precipitation	Inches
Wind Speed	Knots
Temperature	Degrees
Relative Humidity	Percent
Pressure	Inches of Mercury

CRIME FACTOR	UNIT OF MEASUREMENT
Age Percent 15-34	Percent
Age Percent 60 and over	Percent
Percent Males Unemployed	Percent
Percent Wage and Salary	Percent
Percent Owner-Occupied	Percent
Percent Sound Housing	Percent
Percent with 1.01 or more per room	Percent
Percent Married	Percent
Percent Foreign-Born	Percent
Percent Growth	Percent
Percent Decline	Percent
Percent Moved	Percent
Percent Families, 1 or more under six years	Percent
Percent non-white	Percent
Percent Enrolled in School	Percent
Income	Dollars
Persons/House	Number of Persons
Rent	Number
School years completed	Number of Years
PTC	
Elementary School (s)	
Junior School (s)	
Senior High School (s)	

for each of the clusters. Each cluster-center record contains the scaled values for all of the thirty-five crime factors associated with that center and cluster. Storage space was provided for up to one hundred cluster-center records.

Next, records of factor-weights, each record associated with a particular burglary cluster, were stored. Each factor record contains thirty-five weights, corresponding to the thirty-five crime factors. These weights were determined during the MDA cluster analysis (see Section 3) and are necessary to transform the difference between two crime situations into a distance figure in the multidimensional space of a particular factor. Up to sixteen factor records may be stored in the third disk file.

The final disk file contains fifteen numbers necessary for scaling the five cyclic crime factors. For each of the cyclic crime factors three quantities are stored: the maximum difference allowable for that crime factor; a code number designating whether the maximum range of the crime factor is even or odd; and the scale factor necessary to convert the proper raw difference into a scaled value.

B. MODEL DESIGN

The general flow of the program model for any given run is described in Figure 20. At the start of the run a police commander, working at the remote terminal, will enter the current values for the first twelve of the thirty-five crime factors (see Table 9). These twelve represent the values of the dynamic factors of time, date, and weather conditions in the city at the time of the run. Next the commander will query the model concerning burglary conditions in some or all of the city's police sectors. If he does not wish a report for the entire city, the commander may specify only those police districts (groups of ten to twenty police sectors) which he wants the model to process.

The model will begin by scaling four of the twelve input crime factor values - snow on ground, visibility, precipitation, and barometric pressure - into the range zero to one hundred. The scaling is performed as follows:

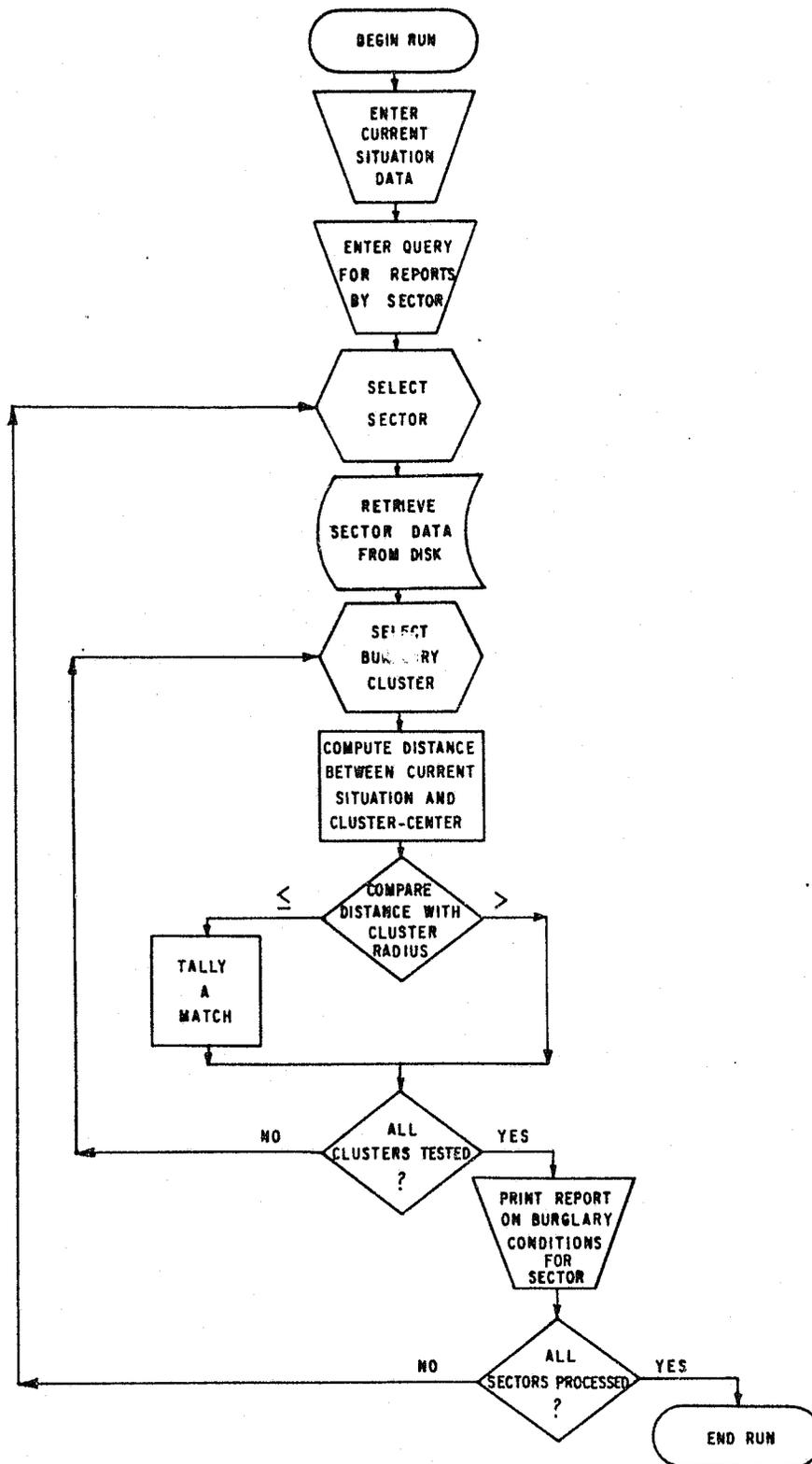


Figure 20. Flow Chart of Operational Model

snow	scaled value = 8.33 * raw value
visibility	scaled value = 6.67 * raw value
precipitation	scaled value = 100. * raw value; any value greater than 100 is set equal to 100.
pressure	scaled value = 16.67 * raw value - 450.09

Next, each of the specified sectors will be considered. In each case the remaining twenty-three crime factor values for the sector under consideration will be retrieved from disk storage. These values are for the more static variables, such as age distribution, housing, unemployment, and schools, which were determined sector-by-sector and have previously been scaled into the proper range. The values from storage will be merged with the input values to form a current-situation record, a representation of the conditions in the sector for which the report is desired.

In order to determine whether the current-situation conditions are associated with burglary-cluster conditions, a technique similar to that used in the MDA cluster analysis (See Section 3) must be used. The current-situation record is compared with the cluster-center record of each "burglary cluster." A distance figure, the distance in the multi-dimensional space between the current-situation and the cluster center, is calculated by the following method:

Differences are taken between the current-situation and the cluster-center on each of the thirty-five crime factors. Differences for the five cyclic crime factors - day of week, month, day of month, hour, and phase of moon - must then be scaled. In each case, if the raw difference is less than or equal to the maximum allowable difference for that crime factor it need only be multiplied by the scale factor for that crime factor to obtain the scaled difference. When the raw difference is greater than the maximum allowable difference, however, the proper difference must be obtained before it can be scaled. The parity of the code number for the crime factor is checked and the proper difference is computed by:

code number even

proper difference = twice maximum difference -
raw difference,

or

code number odd

proper difference = twice maximum difference
plus one - raw difference.

The proper difference is then multiplied by the scale factor to obtain the scaled difference. When all the cyclic differences have been scaled the entire group of differences is combined to form a difference of thirty-five values.

The difference record must now be transformed into a distance figure in the multidimensional space in which the burglary cluster exists. This is accomplished by multiplying each value of the difference record by the corresponding value in the factor record associated with the cluster under consideration and summing the products. The distance figure thus obtained is called Z.

The distance Z is then compared with the radius of the burglary cluster. If Z is less than or equal to the radius, the current situation is said to fall "within" the cluster. A tally of the number of "matches," instances in which Z falls within a burglary cluster, is kept for the sector under consideration. This process is repeated until the current-situation has been compared with all the burglary clusters.

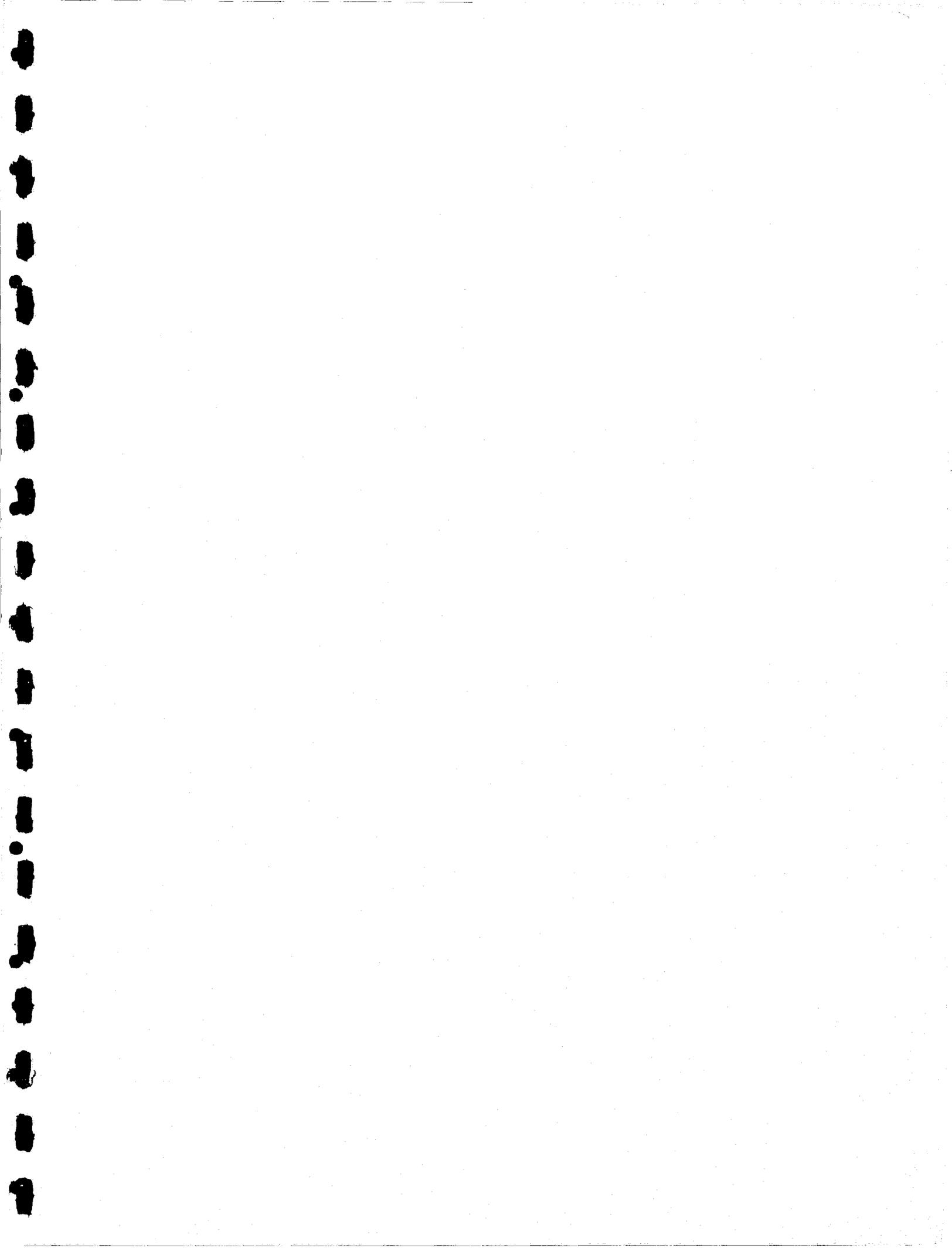
After the current-situation has been tested against all burglary clusters, a report code to reflect the number of matches is generated for the sector as follows:

<u>Code</u>	<u>Meaning</u>
GOOD	Current-situation falls within 75-100% of all burglary clusters.
FAIR	Current-situation falls within more than 25% but less than 75% of all burglary clusters.
POOR	Current-situation falls within 25% or less of all burglary clusters.

The report code is then printed (with the identification code for the sector being considered) at the remote terminal, thus relaying the result directly to the police commander who initiated the run. The entire process is then repeated until reports have been generated for all the sectors which were indicated by the commander.

Initially, this operating model will be used to evaluate and refine the crime-cluster techniques. When the clusters are sufficiently reliable, operational testing and use will be undertaken.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100



At this meeting the crime factor list was approved for the initial analysis. It was also decided that the crime sample to be used would have to be restricted to the current year, 1966, since the Philadelphia Police Department does not retain crime records for previous years in their entirety. No firm decision had yet been made as to which types of crimes would be used in the analysis.

While these data requirements were being completed certain problems appeared, necessitating some refinements in the proposed analytical technique. The introduction of cyclical data, such as "time of day," made it necessary to select and modify an unconventional multidimensional analysis technique. Also, nonnumeric data, such as "type of premises," required the development of a scaling technique to quantify these data for analytical purposes. These necessary revisions of the proposed analytical procedures made it advisable to initiate two supplemental analysis efforts: a multiple regression analysis and a multinomial analysis. Accordingly, development of these techniques was begun and continued parallel with that of the multidimensional analysis.

In January 1967, the decision was made to draw the initial crime sample from all crime types, that is, from both Part I and Part II crimes. The data collection process was begun, but two problems became apparent at the outset. First, the Philadelphia Police Department crime cards containing the required data, were still in use at the time of the data collection effort, making it impossible to use them as input to the analysis. Second, certain crime factors, in particular "type of premises" and "type of weapon" were no longer available in punched-card form. The large sample size along with the necessity for data transfer and record-searching, increased the data collection effort substantially.

The Philadelphia Police Department delivered crime cards to FIRL for data transfer intermittently throughout the months of January and February, 1967. By the beginning of March, Part I crimes had been completely sampled and coded and the data collection effort for the Part II crimes was begun. Upon completion of the Part II sample in April, however, a preliminary analysis revealed that the Part II cards

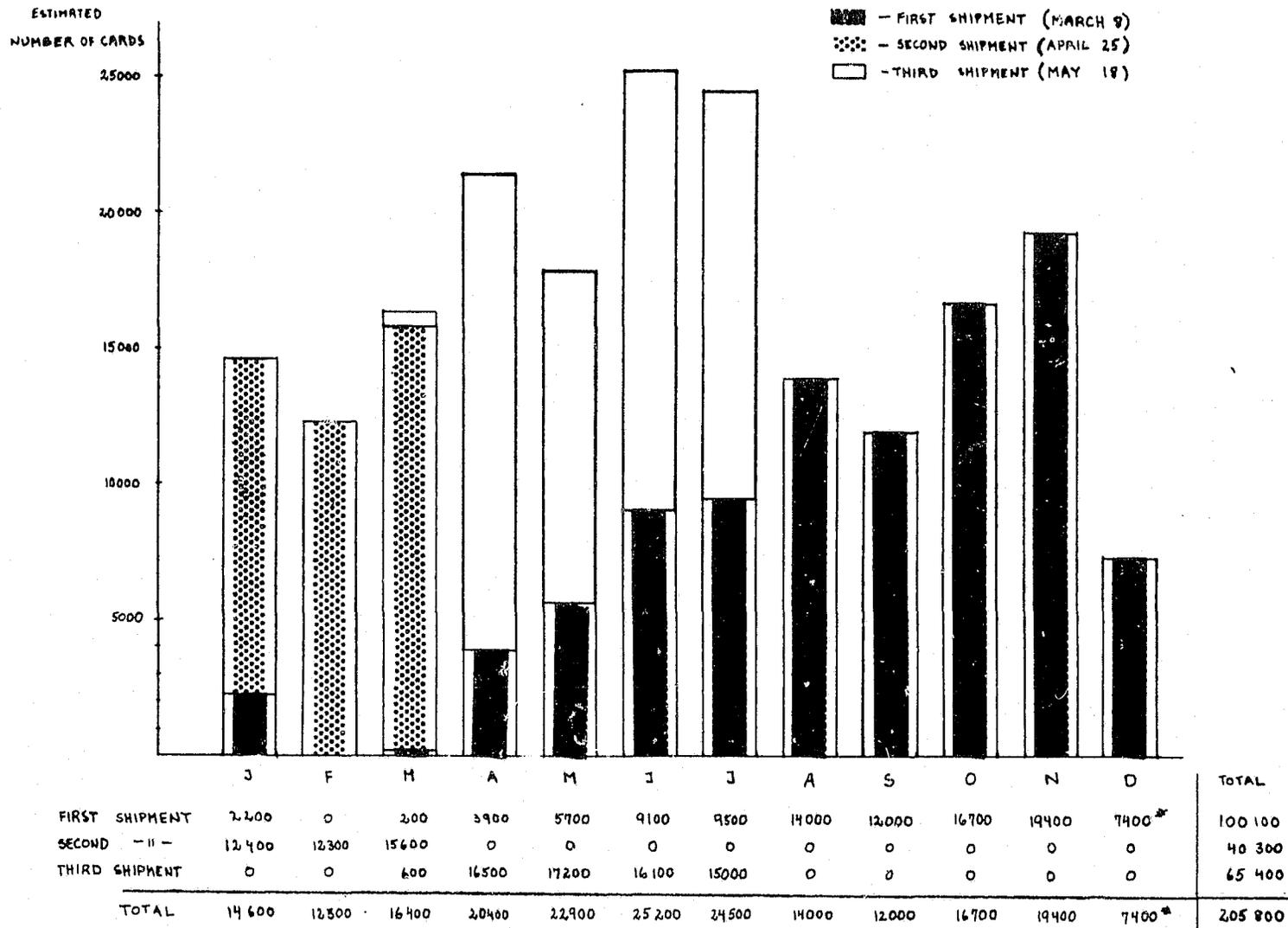
provided for sampling did not in fact constitute the complete set for 1966. Thus it was necessary to take a stratified sample of Part II crimes from the original set and from two additional sets provided by the Philadelphia Police Department. Figure 21 shows the composition of the three Part II card shipments. The December data for both Part I and Part II crimes was incomplete since cards representing these crimes were not available at the time of data collection. The data collection and coding for this sample was completed by early June 1967.

As the data collection proceeded for the Part II crime sample, analyses were begun on the completed Part I sample. Two runs of the multiple regression were made and the results documented for mention in the final report. The multinomial analysis was programmed and tested and thus readied for use in testing specific hypotheses generated by the multidimensional analysis. The MDA itself was performed on each crime type and thus multidimensional spaces representing each crime type were constructed.

The theoretical technique for identifying new sets of crime factor values in terms of the developed multidimensional spaces was found to be unmanageable. Complex manipulations involving large data matrices caused this technique to be discarded as impractical.* As a result, new clustering and identification techniques were developed. Burglary was used as the "test" crime type to which these techniques were applied and several "burglary clusters" were identified and subsequently tested against two random samples of crimes for comprehensiveness and accuracy of discrimination.

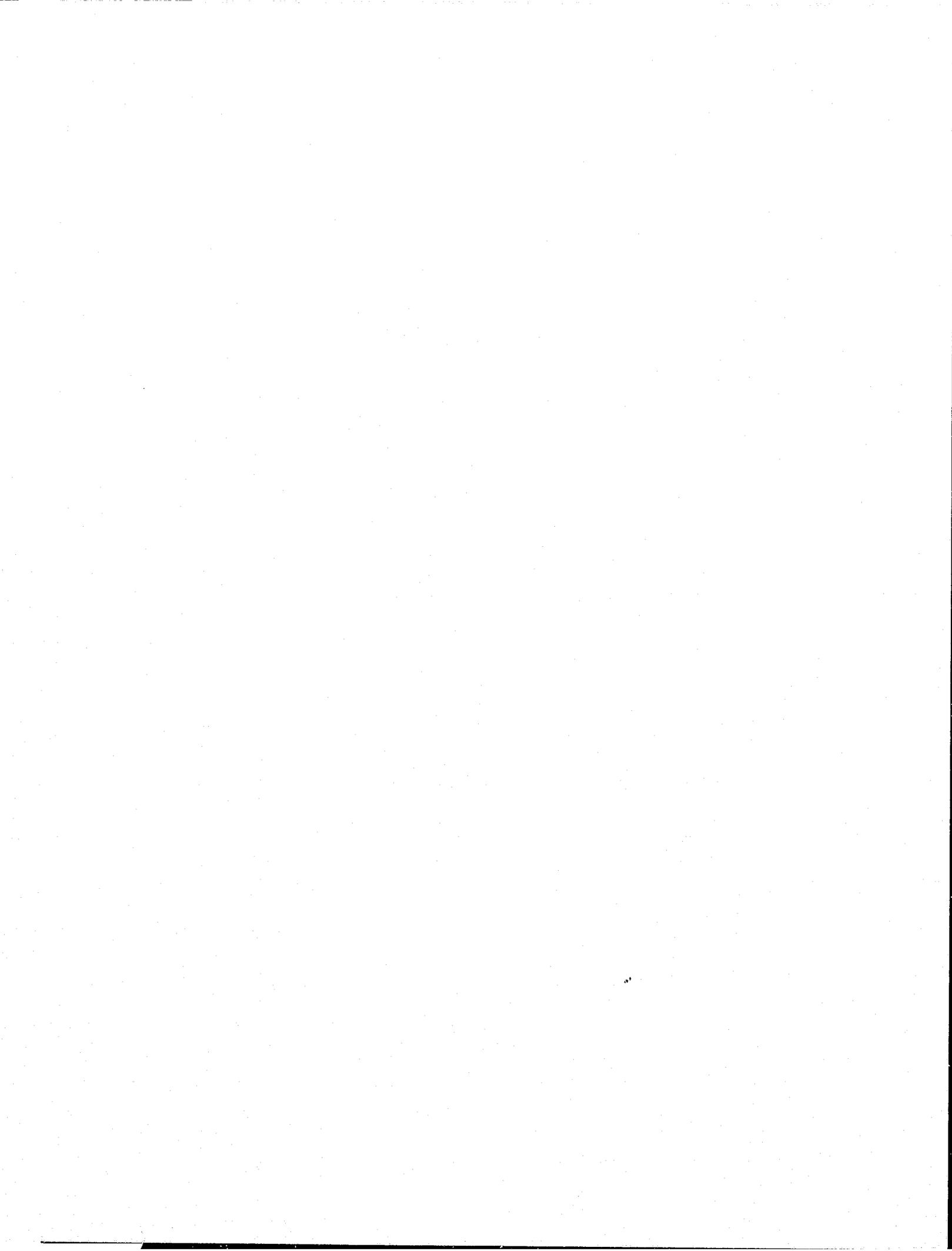
With the aid of FIRL personnel, the Philadelphia Police Department computer staff then flow-charted and programmed operational version of the analytical model, which is now in the process of being debugged. FIRL is furnishing data inputs and an initial set of burglary clusters

* For example, one such matrix contained 10,000 elements and another, 38,000 elements.



* - INCOMPLETE

Figure 21. Original Data of Part II Crimes, 1966



which will be used by the model to predict burglary occurrences on an operational basis, hour-by-hour and sector-by-sector. The actual performance of the operational model will then be evaluated to establish its strengths and weaknesses. Subsequent refinements will be made in the analytical model.

As the implementation of the operational model continued, a sentence outline of the final report was submitted to the Office of Law Enforcement Assistance. Subsequently, the final report, documenting the one-year effort to develop an operations research model for crime prediction, was prepared.

Technical work on the project was concluded in September 1967.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

SECTION 6

CONCLUSIONS AND RECOMMENDATIONS

A. CONCLUSIONS

The present study has achieved two important goals: first, it has demonstrated the fundamental soundness of the original concept. The results of the several analyses demonstrate the existence of differences in surrounding conditions for different crime types, and give every indication that crime *prediction* using a computer-based model can be achieved.

Secondly, a great quantity of necessary data has been gathered and refined, and a solid mathematical foundation has been established. An extensive data base has been created, several analytical problems have been met and overcome, and computer programs have been developed. Through these efforts, a better understanding of the problem has been gained, and a firm basis has been established from which the model can be refined into a useful police tool.

As in any other research effort, all of the problems could not have been, and indeed were not, anticipated. Unforeseen difficulties have hindered the full achievement of the ambitious work plan presented in the original proposal. However, the accomplishment of the above goals strongly supports the desirability of completing the effort. The following section describes what is required to bring the past year's effort to completion.

B. RECOMMENDATIONS

It is recommended that this effort be continued to completion by adding another two years of effort. The four objectives for the proposed additional effort should be:

- a. Continue developing a police data base;
- b. Complete refinement of the prediction model;
- c. Test the crime prediction systems in actual operation; and
- d. As a data-base byproduct, perform *ad hoc* studies for the Philadelphia Police Department.

The specific recommendations are stated below.

Recommendation 1. Develop data base further.

The project data base should be developed further, to support additional model refinement, and to provide a quick-response data base for *ad hoc* use by the Philadelphia Police Department. Specifically, the following steps of data base development are recommended:

- a. *Add crime data for an additional year.* The data base population for the FY67 effort consisted of 1966 crimes from Philadelphia Police Department records. The sample's short time span introduces many false correlations into the analyses because of the particular conditions and relationships peculiar to 1966 (for example, the relationship between phase of moon and day of month; or the particular weather for 1966). For this reason, it is recommended that an additional year's crime data be added to the data base.
- b. *Update census data.* The sociological (neighborhood) conditions surrounding each crime were taken from 1960 census data and are thus seven years out of date. It is recommended that projective techniques now being explored by the U.S. Census Bureau be used to update this census data to reflect current conditions in Philadelphia more accurately.
- c. *Add new variables.* The initial set of variables was a representative list of those for which data were readily available. Many potentially useful variables were not included because usable data were not available. It is recommended that new variables be added to the data base, such as additional population-age variables and land-use data, permitting experimentation with different mixtures of variables.

In addition, variables should be added to the data base to support its use for *ad hoc* police studies.

For example, additional characteristics of individual offenses should be added to permit action strategies based on *modus operandi* (m.o.), and to identify crimes with similar characteristics.

Recommendation 2. Refine Model.

During the previous year, two mathematical techniques were explored: multiple-regression analysis (MR), and multidimensional analysis (MDA). The MR results were not satisfactory as a predictive tool, but showed promise deserving further development. The main thrust of the analysis, MDA, resulted in some initial clusters for burglary. It is recommended that both techniques be further refined, and additional techniques be investigated. Specifically, the following steps are recommended:

- a. *Continue to develop and refine the MDA.* The MDA should be refined in four respects. First, and most important, the cluster analysis techniques should be further refined to incorporate a greater degree of normalization for 'no-crime' conditions. Second, various mixtures of variables should be tried, which would be made possible by the expansion of the data base (Recommendation 1, above). Third, sub-clusters should be explored; the initial analysis dealt only with clusters for each Part I crime type (for example, robbery and burglary clusters). Sub-clusters within each crime type, such as drugstore robberies and apartment house burglaries, should be investigated separately to determine their tendency to cluster. This technique, if successful, will permit model outputs not only by sector and time of day, but also by such factors as type of premises. Fourth, macro-clusters of 'preventable' crimes should be explored. All crime types which appear 'preventable' by similar strategic action should be grouped for analysis, resulting in fewer data cells, which should yield greater precision.
- b. *Develop and refine MR techniques.* The major short-coming of conventional multiple-regression analysis is that the regression variables (predictors) do not in general correlate positively over their entire range, nor negatively over their entire range, with crime occurrence. Instead, a variable is likely to correlate positively with crime occurrence over part of its range, and negatively over part of its range.

For example, auto theft peaks on weekends, dropping off toward midweek, so that the variable 'day of week' on a scale ranging from Sunday to Saturday correlates negatively with auto theft over the first part of its range (that is, Sunday through Wednesday), and correlates positively with auto theft over the second part of its range (Wednesday through Saturday). Recommended modifications to the MR to eliminate this shortcoming include three steps:

- (1) Perform a complete frequency-distribution analysis, showing normalized frequency-of-occurrence of each crime type over the entire range of each variable. (This analysis was begun during the current project);
 - (2) Re-order the regression-variable ranges according to increasing frequencies of crime occurrence for each crime type. This will insure that each variable can have only a positive correlation with crime occurrence, over its entire range; and
 - (3) Run both linear and nonlinear regression analyses on the resulting variables.
- c. *Investigate other discriminant-analysis techniques.* Project time and funding did not permit the investigation of other potentially useful approaches to crime-occurrence discrimination, such as multinomial discriminant analysis and adaptive pattern-recognition techniques. Such techniques should be explored to determine whether one of them might yield better discrimination and prediction than either MDA or MR.

Recommendation 3. Conduct operational testing.

As the various crime-cluster predicting techniques are developed, they should be tested against "real-life" data. Testing should be a continuing process, involving several iterations of the *test-refine-test-refine loop*.

The operational testing should center around the PPD's computer, using the computer program described in Section 5 of this report.

The program has the capacity to

- a. Implement any crime-clustering techniques;
- b. Accept current conditions input from the console in the PPD Communications Center; and
- c. Output a crime-cluster analysis.

The crime-cluster analysis output should be compared with actual crime occurrences to determine the predictive efficiency of the technique under test. To refine the models, specific instances of failures should be analyzed to determine why the failure occurred; these individual analyses should be used to make necessary improvements to the models. As the model refinement continues, the judgment of experienced police commanders on the scene should be incorporated where possible.

Recommendation 4. Conduct ad hoc studies for the Philadelphia Police Department.

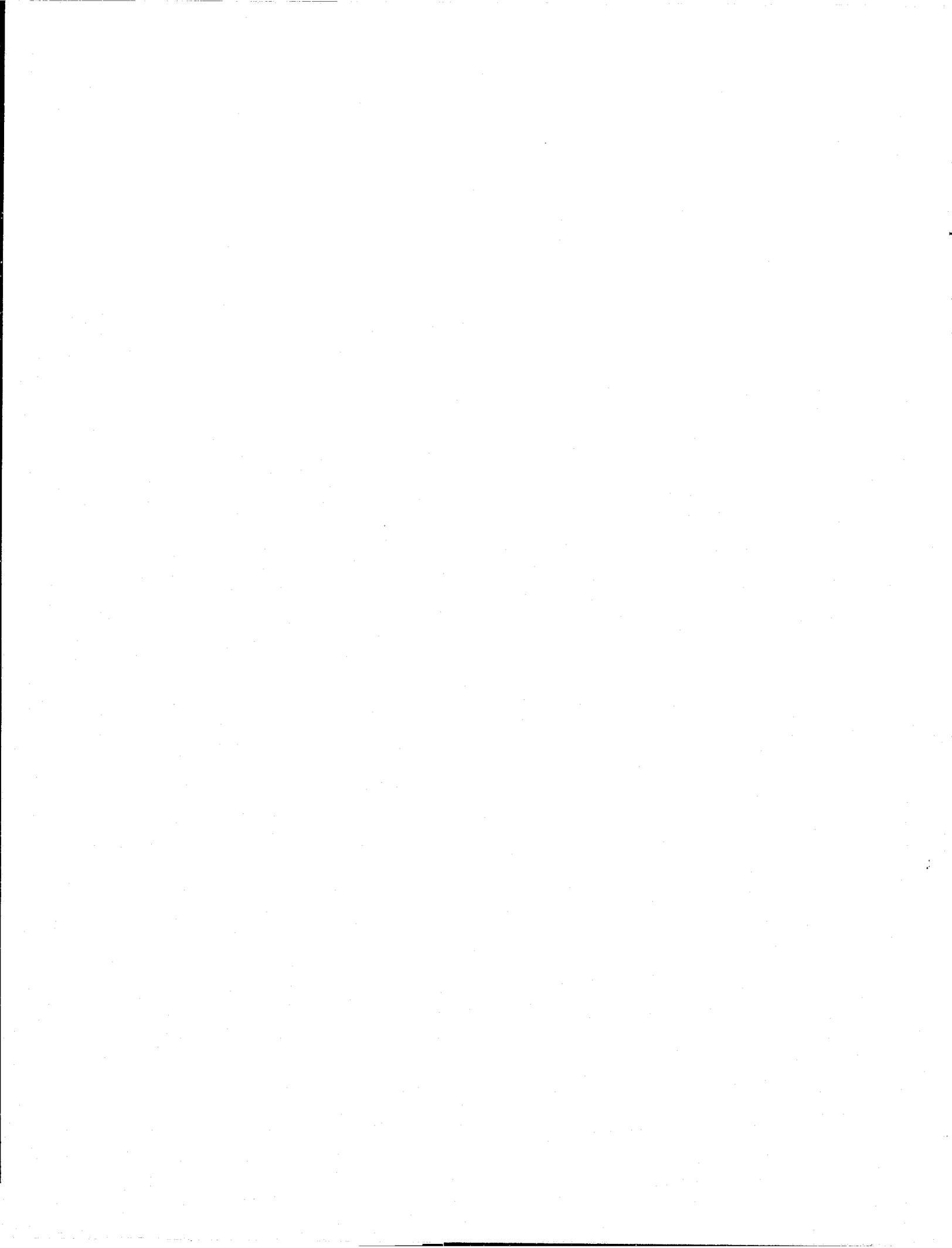
Much of the data already collected to support the model has been found to be useful as a general-purpose data base. As the work continues, several new applications of the project data base to Police Department operations will probably become apparent. These additional applications should be studied on an *ad hoc* basis, as they arise, utilizing the project data base.

For example, two such applications have already been identified. The first is to generate a profile of the offender. This requires adding characteristics of the offender (in particular, his previous police record) to the data base. Then, an analysis of offenders' characteristics (age, previous arrest record, and so forth), by crime type, could be generated. This would permit the use of action strategies based on the type of person committing certain offenses; for example, if it should turn out that certain crimes of violence are frequently committed by a certain type of convicted felon who had been granted an early parole, then an appropriate suggestion might be made to the parole officials.

The second application already identified is use of m.o. (*modus operandi*) data to facilitate apprehension or prevention (via stakeout or other tactical deployment in instances of reported similar crimes committed by the same offender (so-called "crime waves")). The m.o. characteristics of recent crimes should be entered into the data base; these could then be analyzed to identify groups of crimes having common m.o. characteristics. The techniques described above (MR and MDA) could also be used here.

Recommendation 5. Increase police participation.

Throughout all future work, police participation must be the keynote. It is recommended that police officials play an increased role in any future project effort through the institution of a joint steering committee and that a full-time police officer be a part of any future project team. These recommendations follow from the belief that constant communication between the researchers who are developing the model and the police commanders who will use the resulting tool is vital if the effort is to reach a successful and relevant conclusion.



END