Application of Time Series Methodology to Crime Analysis

by

Clifford W. Marshall

June 15, 1977

Department of Mathematics
The Polytechnic Institute of New York
333 Jay Street, Brooklyn, N. Y.
11201

This project was supported by Grant Number 76-TA-99-0028 awarded by the Law Enforcement Assistance Administration, U. S. Department of Justice, under the Omnibus Crime Control and Safe Streets Act of 1968, as amended. Points of view or opinions stated in this document are those of the Author and do not necessarily represent the official opinion or policies of the U. S. Department of Justice.

ABSTRACT

Various types of crime data, such as reported crime, are normally collected over time by police departments. These data are used in a variety of forms for various purposes. The most significant thing about such data is that they form a series of values developing in time rather than a collection of values specified at a particular time. The time varying nature of crime data is often overlooked or minimized with the result that erroneous or misleading observations may be made. This report describes various levels of mathematical detail for addressing the study of time series, particularly reported crime. The methodology is described and illustrated. Illustrations are based on real crime data as well as synthetic time series specifically generated for the purpose of this study.

The report includes discussion of one of the most complete methodological approaches to time series analysis: the Autoregressive Integrated Moving Average (ARIMA) method. It also discusses the widely used Census X-11 program. Technical problems related to the relatively complex programs (software) required for some of the methodology are discussed.

This report is not intended as a text on the mathematical statistics of time series but rather as a bridge from that body of knowledge to its actual implementation in crime analysis.

ACKNOWLEDGMENTS

Most of the background analysis leading to this report was done as part of a project carried out at The Urban Institute for The Police Foundation. The entire effort owes its existance to the direction and encouragement of Joseph H. Lewis of The Police Foundation. The Author's colleagues at The Urban Institute: Alfred I. Schwartz and Sumner N. Clarren were sources of continual help of both a conceptual and technical nature. Complicated problems of software implementation were solved by Robert Teitel of The Urban Institute Computer Services Group.

The use of actual reported crime data was made possible by the kind approval and encouragement of The Cincinnati Police Division in the person of Carl Lind.

A number of people reviewed parts of this report and their efforts are greatly appreciated. In particular Alfred I. Schwartz, Phil McGuire (Head of Crime Analysis Section, New York City Police Department), and Ralph Swisher and his colleagues at The Law Enforcement Assistance Administration have contributed a number of improvements to the report. Any shortcomings that remain are the sole responsibility of the Author. The idea of using a specific physical example (the Sump) in addition to actual crime data is due to Phil McGuire.

Thanks are certainly due to the efforts of Lucy Prescia, Carol Devlin, and Helen Charest at The Polytechnic Institute of New York for typing the text together with its complicated tables and formulas.

Introduction

In the field of crime analysis one encounters a number of records consisting to data that occur at different times. Such data records are called time series and in one form or another they form a part of many descriptions or analyses of crime activity. The study of time series can range from a direct presentation of the data to highly sophisticated statistical analysis.

The purpose of this report is to provide practical guidance to planners and analysts who have sufficient mathematical and statistical background to read and understand clearly written materials concerning time series analysis techniques, even though they are not specifically acquainted in or experienced with the particular techniques being discussed.

There is increasing availability and visibility of these methods, and an LEAA research program is now testing the use of stochastic analysis of reported crime data in a number of cities. Thus it is likely that the techniques will be used with increasing frequency. Therefore it is desirable for LEAA to have materials available to minimize the misuse of the more advanced expensive methods and the possibility of misinterpreting results of time series analysis.

This report is not meant to be a basic textbook, but rather a supplement to such readily available texts. It is especially meant for use by those likely to employ available canned computer programs without adequate study of basic texts or formal training in the use of the method. It is also intended to have utility for monitors of studies employing time series analysis, to aid them in judgments about the appropriate use of such sophisticated methods.

The range of uses remains a primary purpose of LEAA regarding this report. To that end it is intended to acquaint many potential users with the report through courses developed by LEAA for advanced analysis of crime data as well as to distribute it to a general user audience.

The report consists of eleven sections covering a wide range of statistical approaches to the analysis of time series.

Sections 1 through 4 cover material of an elementary and basic nature stopping short of considerations leading to the development of mathematical models representing time series. In these four sections some of the possible statistical tests for comparison of time periods within one time series or across two time series are given. Methods of presentation and interpretation are also discussed. The statistical methods are of the "parametric" variety though in many cases appropriate non-parametric techniques could be employed as well. Since a major goal of the report is to deal with the less widely known, more advanced, techniques, the sections on elementary methods do not attempt to give broad coverage of many possibilities but rather present and illustrate some methods of demonstrated utility.

Sections 5 through 11 discuss the development of mathematical models that may be used to represent time series. There are a number of reasons for wishing to create such models, one of the most important being the possibility of forecasting subsequent values in the series. Because of the relatively advanced nature of the mathemematical model building technique it is presented and illustrated in detail. In addition the implementation of the technique, utilizing computer programs is described in such a way as to identify a number of potential problem areas.

TABLE OF CONTENTS

Section 1 Basic Consideration In Using Reported Crime Data.

Crimes reported to the police are recorded as a standard procedure. The type of crime, locale, date and time of occurrence are among the information recorded. Various other information may also be noted, depending on the standard police procedures used in a particular jurisdiction. However, the type, location, and time (including date) are universally recorded providing a kind of data common to all police jurisdictions. The form in which these data are kept or might be available differs, ranging from reports filed by officers to computerized data bases. Reported crime data are discussed in this section to identify the basic features of what such data are like, how they can be obtained for study, and what types of methodologies may be effective for operational data analyses.

From the point of view of data analysis the most fundamental properties of reported crime data are that they are found data and that they occur over time.

Properties of Found Data

Found data are distinct from data obtained as the result of special surveys or data gathering activity. They are data collected as part of the regular operations and are not specifically designed for a specific analysis. There are advantages and disadvantages in using found data as distinct from survey data.

Major advantages follow from the relative availability of found data resulting in a great savings in cost over survey data. Found data are also likely to have good reliability because of the ongoing nature of the collection process. However, a major caution in employing found data is to understand and check the data collection and processing system. Such systems are subject to change over time and also to errors that can enter and be undetected for some time unless an explicit effort is made at checking.

Another disadvantage of found data is that they may not be available in the form desired for analysis. In such a case the possibility of conversion to appropriate form must be determined and the cost of such activity weighed against the benefits expected from the intended study.

Found data must also be appropriate to the objectives of an investigation. In fact the major argument for using survey or developed data, with the costs and problems involved, is that no found data exist that are appropriate for a particular study. Arguments close to this point of view are presented relative to the use of reported crime data. These data may be contrasted with victimization data obtained by survey methods at considerable effort (cost). It is argued that victimization data give a much truer picture of actual crime activity than does reported crime. There seems to be considerable evidence for this view and it need not be considered further here. The point in the present context is that found data, such as reported crime, may have shortcomings for some studies while being entirely suitable for others. A prime example of this is the use of reported crime data to compare different situations involving police activity that responds to reports of crime, rather than as an accurate description of all crime activity in a single situation.

Care must be exercised in adapting found data to specific analysis but there is no reason to believe a _priori_ that such data cannot lead to reliable and useful results. Further discussion of the found data aspects of reported crime data are given in (1) and will not be pursued here except as they arise within the context of related matters.

Time Series Nature of Reported Crime Data

The time series nature of reported crimes is a theoretically more important aspect than the found data quality. While the latter imposes various cautions and potential limitations on the use of the data it is their time series nature that governs the kind of analyses that should be employed. There is a very real difference between a collection of data values that are independent of time and a collection in which the values are obtained at various times.

To appreciate the significance of time series data, it is necessary to consider the concept of a random sample as used in mathematical statistics.

Random Sample

Many of the widely used statistical tests depend on the concept of a random sample. In such a sample each data point is supposed to assume a value independently of all other data points and the probability law governing the values assumed is the same for each data point. If either of these attributes are not satisified the sample does not properly fit into many of the commonly used statistical procedures. However, it is often difficult to determine the extent to which a sample is, in fact, random.

Statistical procedures can be devised for testing a sample but such testing increases the complexity (and cost) of a statistical study.

In addition it may be that the sample is not random. What should be done then? The widely used practice is to employ standard tests even though the assumption of a random sample is not satisfied (or has not been checked). In many cases this procedure leads to meaningful results because of the strength of the tests as indicators of statistical variation.

Dependence of Data

If the sample data are in fact dependent to some extent so that the sample is not truly random incorrect results can be obtained. Moreover, the extent of dependence may not be known so that the confidence one might want to have in making statistical tests cannot be established.

This situation can be present in any statistical sample but is particularly likely in time series data where a value at one time has an opportunity to affect subsequent values.

The concept of dependence may be illustrated by the number of sales recorded at a store. Two kinds of samples can be considered: the sales by item classes in a single day for comparison with a similar sample at another store, or the number of sales of a particular item recorded each day for several weeks. The former is not dependent on time and it may be very reasonable to assume that the sales level of items do not affect each other (of course they may). In that case the sales picture at the two stores can be studied by normal statistical tests assuming random samples.

Time Series

In the second case however, a time series is involved and the sales level of a particular item, such as raincoats, may depend on previous sales levels. The sale of raincoats depends on random effects such as present weather conditions and possibly recent past values as well. Heavy rain a few days past may cause raincoat sales to have been high with the result that many people have raincoats when, a few days later, it rains again. The sales level will go up due to the rain but not as much as if the earlier period had recorded lower sales levels.

Reported crime, being a time series, may not provide random sample data for use in standard statistical tests. On the other hand, such tests are relatively simple to apply and understand. It is worthwhile to consider a range of possible levels for the statistical analysis of time series ranging from the most elementary and direct to advanced considerations in which the possibility of interaction effects between values is measured and described.

Four Classes of Statistical Procedures

The full range of methodological possibilities can be covered by dividing it into four distinct, though related, procedural classes: Elementary Statistical Methods, Preliminary Time Series Analysis, Detailed Statistical Analysis, and Stochastic Models. This report will describe each of these approaches as they apply to the study of reported crime data. Concepts will be illustrated using actual reported crime data and other numerical examples as described later. The general ideas of each class of analysis is presented here to orient the reader regarding the following sections of the report.

1. Elementary Statistical Methods treat a collection of time series data values as though it were a random sample. Simple statistical quantities related to the sample mean and variance are calculated and used to test for differences in crime activity. Such methods are simple to apply and are widely known for ease in communication of ideas. They may indicate valid effects or they may not if the sample points are not independent. There is no possibility to make forecasts of future values so that in any evaluation application one must wait until a reasonable size sample of data points has been obtained. This kind of situation can be thought of as static evaluation as opposed to dynamic evaluation where, after an initial data generation period, effects can be evaluated in an on-going way by using forecast values.

2. Preliminary Time Series Analysis consists of qualitative and simple numerical study of the reported crime data for some time period. The data values are plotted as functions of time and the general character of the series is observed. The sample mean and standard deviation values, for various lengths of time, are computed and may be shown on the time series plot to assist in making observations about the nature of the data. Trends and cyclic variation can often be observed directly from the plot. Some of the variation in data values can be represented by specific graphs (such as linear trend or some periodic variation). The fitting of such specific variations with time yielding deterministic curves (by judgment alone or using techniques of regression analysis) can be undertaken as part of preliminary analysis and gives a rough forecasting ability. The main features of preliminary analysis are to see if the data seem to have special trends and to qualitatively estimate the statistical variation present.

3. Detailed Statistical Analysis uses the data values to compute quantities called autocorrelations (and partial autocorrelations) which fully characterize a time series. These values can be used directly to represent and study time series or they may be used as a basis for constructing a mathematical model representation of the time series. The autocorrelations are more difficult to compute than the simple elementary statistics but do not require nearly the effort that model construction does. However, the detailed statistics alone can not forecast, and their use is limited to static type comparisons between time series.

4. Stochastic Models are mathematical formula representations of a time series. There are several approaches to such mathematical models, one giving a high level of generality is the so called autoregressive integrated moving average model (ARIMA) described in a later section. This method was developed by G. E. P. Box and G. M. Jenkins (4) and is alternatively referred to as the Box/Jenkins technique. These models are difficult to form because the numerical parameters must be estimated from the data by means of an extensive numerical calculation procedure. It is completely out of the question to attempt such calculations without the use of a computer and appropriate software programs. The value of such models is their ability to forecast which holds the potential for dynamic evaluation.

Selection of the Proper Method

The important thing in applying any of the above types of analysis is to understand the level of effort required and what can be expected from the expenditure of that effort. In all cases the need to communicate useful information to others should be emphasized. Simple plots and tables are most often useful for this purpose as described in subsequent sections. Table 1.1 gives a summary of the four types of analysis methodology.

Methodology	Level of Effort (Cost)	Amount of Information Used from the data	Level of Theoretical Ability to represent the series data	Ability to ?
Elementary Statistical Methods	Low	Minimal	Variable	None
Preliminary Analysis	Low-Moderate (depending on detail)	Minimal	Variable	Possible. Reliability Unknown
Detailed Analysis	Moderate	All	Good	None
Stochastic Models	Great	All	Good	Possible. Reliability measurable

Table 1.1
Comparison of Methodologies

No matter what type of methodology is to be applied the time series data, must be formed into an operational data base. This means that usable raw data must be available and it must be possible to put those data into the form desired for use. There are many aspects of this requirement as described in (1). One major set of considerations is how to select various aggregations of data. Time location, and crime type (classification) are the three quantities to be aggregated in the analysis of reported crime.

Aggregation by Time

The most likely time units are days, weeks, and months. Though reported crime is recorded by day it is most often only available by month as historical data. In deciding upon a level of aggregation one must balance between having too little data and having too long a historical record. For most statistical analyses one needs a fair number of data values and for stochastic models the complex estimation requires a relatively large sample for meaningful results.

Time Period

Months therefore lead to long historical records which tend to mask out effects one wishes to study. Months have different compositions of type of day that can also mislead analysis. On the other hand days tend to vary too much and have little statistical regularity. Thus, weeks seem to be desirable as the level of time aggregation in reported crime analysis with some use of monthly or daily data in supporting studies.

Geographic Area

The level of geographic aggregation depends on the specific analysis provided there are enough data values to yield meaningful statistical results. However, too large a region, such as an entire city, may combine too many different effects and fail to indicate useful information. Thus one selects some region, large enough to provide data appropriate to a particular study, but not so large as to allow a washing out of statistical effects.

Classification of Crime Types

Reported crime can be classified in a number of ways. Even following the uniform crime reporting system (FBI's UCR) different aggregation of crime types can be made. The same general rules apply: aggregation should be at a level appropriate to the analysis with large aggregates avoided on the basis of their lack of meaning because of interaction effects.

In applications of the various time series models to crime data one might do well to consider the sensitivity of the analysis to variations in two parameters that are usually not recognized as parameters, geographic bounds and crime type classification. Problems may arise from the preliminary specification of a very fuzzy system. The preliminary specification accepts the existing political boundaries and the existing crime categories without giving much thought to what real effect this may have on the models derived. If the models are sensitive to these parameters; the conclusions, evaluations, and recommendations resulting from these studies will also be sensitive to these parameters.

Studies to date yield no empirical proof of the sensitivity of the statistica techniques to boundaries or aggregation of crime types but experience produces some feeling that the systems under examination can be specified in alternate ways which seem at least as reasonable as accepting the existing boundaries or classifications. A sociologically defined boundary such as "neighborhood" may be more appropriate than precinct boundary; even bounding by land use may be more effective a control than precinct. There is also some indication that criminals are essentially opportunistic and switch easily between crime categories possibly making a time series model composed of certain types of Larceny plus Burglary a more reasonable series for modeling than either Larceny or Burglary along.* Use of modified regional aggragation may be implemented by the Census Bureau geocoding procedures being established throughout the country. A recent reference to current developments related to this methodology is given in (12).

An Analogous Physical Example

This section will conclude with an example of actual reported crime data used to illustrate the methodological procedures in the following sections; but first another example will be described. This example illustrates a physical situation rather than a social one such as reported crime. In this way the procedures to be described are illustrated by two distinct kinds of examples, one based on a physical system and the other based on actual social type data.

The physical system used is a water supply for a steam driven turbine. The purpose of the water supply, called a sump, is to maintain a fixed level of water to the steam system. During the operation, some water evaporates, some condenses, and some is added or drawn off in a deterministic way at fixed times each day. The water level in the sump can be denoted by $L(t)$ where t denotes the time at which L is recorded. For the example used in this report, L will be recorded on the hour so that each day provides a time series containing 24 points.

Numerical values used in the example are generated from a mathematical model of the sump level. This procedure is similar to that described for generating synthetic time series in Section 11.

This model expresses $L(t)$ as a combination of a specific function of time $F(t)$ and a random change $R(t)$ that accounts for the net change in water level due to evaporation, condensation, and steam system demand. The expression $F(t)$ shows how much water is drawn off or added at specified times. Figure 1.1 shows a typical time plot for the example series $L(t)$.

*These cautions in forming aggragations are due to Philip McGuire, Director of Crime Analysis, New York City Police Department.

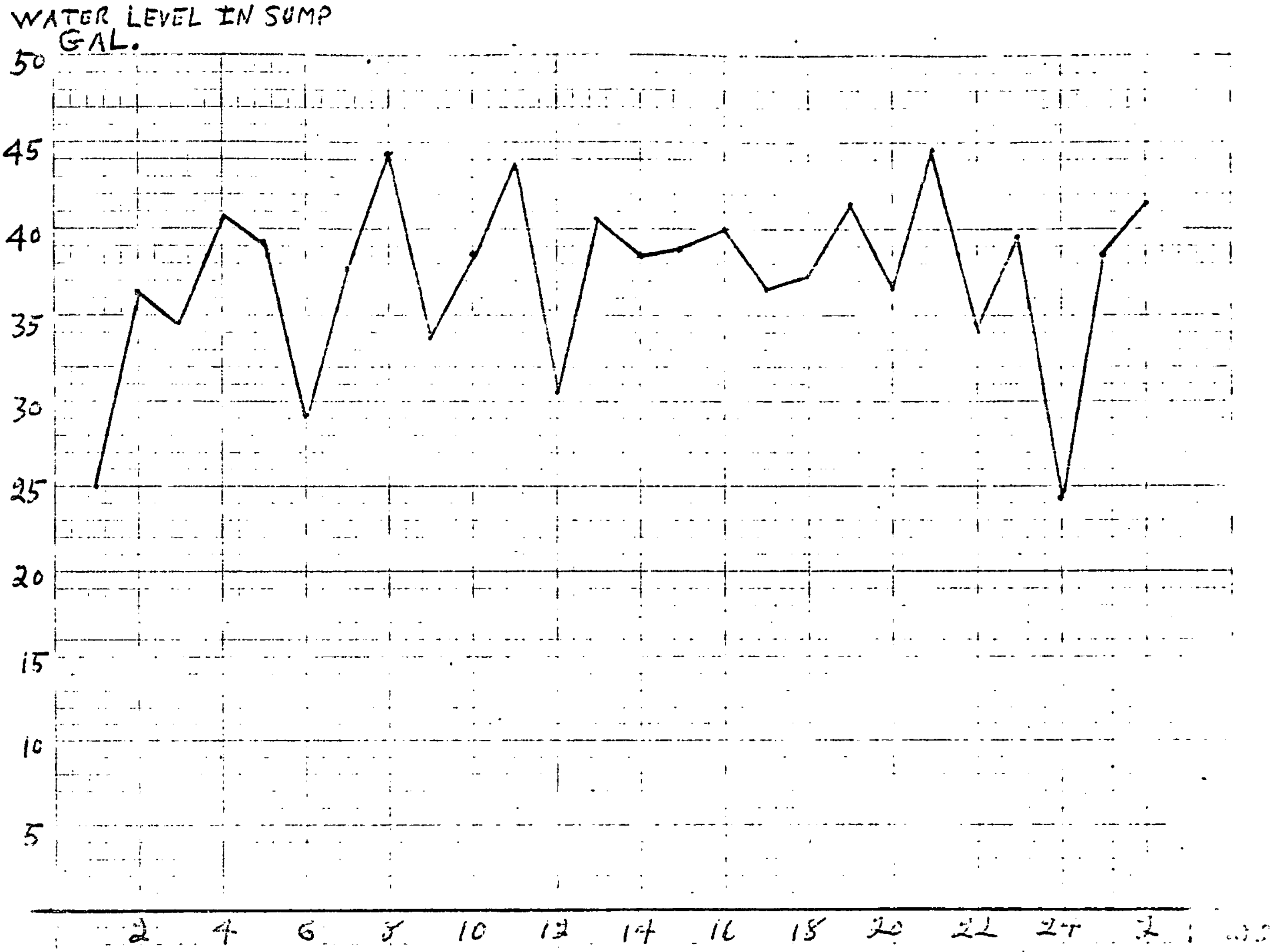


Figure 1.1 Typical Example Time Series (Physical Data)

Crime Data Examples

In the following sections similar numerical examples will be used to illustrate the respective procedures described in each section of the Report. Procedures will be illustrated using reported crime data from Cincinnati that were available in operational data bases generated for other studies (as discussed in (1), (2), and (3)).

For illustrative purposes, certain crime types were selected on the basis of good data, representation of various types of statistical data, and relative interest for crime analysis. The cases used are: Rape (as an example of almost purely random type series), Burglary (both complicated statistically and important for analysis), Robbery, Auto Larceny, and Aggravated Assault. The geographic region for aggregation is the Police District level of which Cincinnati has six. Time is aggregated by week in most cases as this is felt to provide the best combination of statistical variability, length of historical record, and number of data points. Monthly data are also used for some examples because of their general interest and wide availability. Data used in the examples are taken from a data base in which each time series is 217 weeks long (starting in 27th week of 1971 and going through the 35th week of 1975). The collection and management of these data is a significant part of any time series study and is discussed in detail in (1). The present report will simply use the time series to illustrate methodology and procedures.

Section 2 Elementary Statistical Techniques

Elementary statistical techniques utilize collections of data for which numerical indices are calculated. These indices are called statistics and may be used either directly or in combination to measure differences between collections of data. In applications to studies of time series such as reported crime, a collection of data is defined by the crime type and reporting region (location) to which the data values belong. The number of data values in the collection depends on the length of time covered and the interval of time by which the reported crime data are grouped. For example, with weekly data a six-month (26 week) period might be used. The techniques in this section will be illustrated by examples using particular crime types (e.g., robbery) and locations (e.g., District 1) for several 26-week long time series.

Let the time series under study consist of n data values (e.g., n may be 26 weeks or 36 months). The actual data values can be specified by means of a subscripted variable X_i where the subscript i takes values 1 through n to denote each of the n data values contained in the time series. It is difficult to work directly with all the time series values, each of which represents a distinct measurement of the same kind of quantity, such as the number of robberies in a police district in a distinct week. The values differ for a variety of reasons, and it is difficult to tell very much about the nature of the quantity under study (e.g., robbery) from the numbers alone. The most elementary statistic one can consider is the numerical average value. This is called the (sample) mean for a collection and is denoted by $\overline{X}$. The formula for numerical calculation of $\overline{X}$ is given below.

Computation of the Mean Value, Variance, and Standard Deviation

The mean value, of a collection of data, by itself may say very little about the quantity under study. The mean must be viewed in terms of the variability that is present in the data. If the variation is great, the mean will not say much about the individual values. However, if the variation is small, the mean value is indicative of (close to) actual sample values that may be expected.

Variation is measured about the mean value as the average sum of squares of differences from the mean. This measure of variation, called the variance, and denoted by S^2, is computed by the formula given below. Since values are squared to compute the variance, it does not have the same dimension as the data points themselves. Thus, the variation is measured by taking the square root of the variance yielding a quantity called the standard deviation.

Elementary sample statistics are computed as follows:

mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ where X_i are the sample values and n is the sample size.

variance $S^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$

standard deviation $S = \sqrt{S^2}$.

The sample mean and standard deviation values are computed by using the above formulas, illustrated below for a typical time series arising from the sump example described in Section 1.

A typical time series consisting of 96 data values, which shall be called SUMP 1 has the values: 25.0, 36.5, 34.6, 40.2, 39.2, 29.1, 38.0, 44.4, 33.8, 38.6, 43.5, 30.6, 40.5, 38.5, 38.9, 40.1, 36.7, 37.2, 41.6, 36.9, 44.7, 34.3, 39.7, 24.2, 38.6, 41.2, 41.7, 43.0, 45.9, 29.6, 42.7, 34.3, 37.6, 38.3, 34.5, 37.0, 39.5, 31.1, 45.0, 32.3, 34.3, 38.5, 32.8, 47.6, 37.4, 30.3, 40.8, 34.2, 36.4, 40.3, 38.4, 33.4, 27.2, 30.7, 40.8, 39.5, 41.5, 40.4, 40.5, 42.5, 29.5, 41.3, 40.9, 40.3, 30.1, 33.8, 43.7, 30.0, 37.8, 44.9, 35.7, 32.2, 44.2, 30.7, 45.6, 36.6, 39.8, 31.5, 38.1, 33.5, 44.0, 36.9, 42.3, 36.7, 37.9, 26.6, 41.2, 44.7, 35.4, 32.4, 26.5, 42.9, 40.2, 36.8, 29.0, 36.2. For this series the sample mean is 37.2, the sample standard deviation is 5.18, and the range is 24.2 to 47.6.

Reported crime data can also be used to illustrate the calculation procedure for mean and standard deviation values. Weekly data for a typical 26 week period showing robberies in a police district are given by the following time series.

Reported Crime Example

4, 8, 8, 7, 12, 10, 6, 10, 8, 10, 4, 8, 7, 9, 7, 15, 12, 7, 11, 8, 14, 7, 13, 9, 11. For this series the sample mean and standard deviation are 8.9 and 2.8 respectively. The range is 4 to 15.

Comparison of Time Series

For evaluation or other operational purposes, it is often desirable to compare two different time series. The series may represent different time periods, different geographic regions, or other kinds of difference. Because of the variability present in data it is confusing to attempt to compare collections in terms of their component data points. Such comparisons can be attempted, e.g., by plotting individual points for each collection as a graph over time and making visual comparison. Though time series plots can suggest a variety of results and serve to guide various considerations, their interpretation is too subjective and qualitative for direct use in characterizing time series data.

Thus one is led to compare the average values of the two series. This idea leads to the use of concepts from elementary mathematical statistics which derive from the concept of a random sample as described in Section 1. Each data point is assumed to be subject to the same probabilistic process and to be unaffected by the other data values. In time series the data points are associated with distinct time periods in sequence. In every time series, the time points may encompass one or more time periods. Differences between time periods can be studied within one series or between two time series. To study indications of change between time periods, one assumes the n sample points in both time periods as representing a random sample from a population comprising a theoretical set of events characteristic of both time periods. If the time series values are not unrelated, this assumption is not satisfied, and the elementary statistical techniques may not give reliable results.

On the assumption of random samples from each time period, one may compare the periods for indications of difference. To do so one uses a test of the hypothesis that each period has the same (theoretical) mean value, so that sample means are, within statistical variation, the same. In testing such an hypothesis one must not compare only the sample mean values. To do so ignores the random fluctuations present in the data. It is necessary to compare values in terms of units specified by the amount of variation present, as measured by sample variance values. For this reason a sample statistic z is employed. In this section z, defined by the formula given below, is considered to have a standard normal distribution. Such a procedure is based on the view that samples are large enough so that the sample mean values are approximately normally distributed.

In comparison of two populations, subscripts or change of letter denote the populations, X and Y are used here.

The normal statistic z on the hypothesis that $\mu_x = \mu_y$ (theoretical means are equal) is given by the following formula:

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

n_x and n_y denote the number of cases in the X and Y populations respectively

In order to provide an example of basic statistical analyses, five six-month time period employing reported crime data are given. Two of these are immediately prior to initiation of a new team policing activity called COMSEC and the other three follow its implementation on March 4, 1973. The periods are designated as I, II, III, IV, and V; and cover the following time intervals:

I. 26 weeks: 197210-197235
March 5, 1972, (Sunday) through September 2, 1972 (Saturday).

II. 26 weeks: 197236-197309
September 3, 1972, (Sunday) through March 3, 1973, (Saturday).

III. 26 Weeks: 197310-197335
March 4, 1973, (Sunday) through September 1, 1973, (Saturday).

IV. 26 weeks: 197336-197409
September 2, 1973, (Sunday) through March 2, 1974, (Saturday).

V. 26 weeks: 197410-197435
March 3, 1974, (Sunday) through August 31, 1974, (Saturday).

Statistical analyses are presented here for all six police districts, denoted by D1, D3, D4, D5, D6 and D7. They are also presented for the city as a whole and for the city except for District 1, denoted by CITY and CD1 (complement of District 1), respectively. There are, therefore, a considerable number of statistics to contemplate. As a help in such considerations, statistical significance is indicated on the data tables.

Basic statistical data are presented for each crime type and each location by period. The sample mean and sample standard deviation are shown in each case, and are denoted by $\overline{X}$ and S respectively. Sample statistics are based on a 26-week time period in every case. Table 2.1 presents the mean and standard deviation values.

It is possible to make a total of ten comparisons between the five time periods, each comparison being made between two related periods. The total of ten comparisons gives such extensive material that it is extremely difficult to appreciate. Three of the possible comparisons are selected as examples here. These are:

- Comparison 1--Period I and Period II. Both periods are prior to COMSEC. Period I covers the same range of dates as does the post-COMSEC Period III, one year later. Period II is just prior to COMSEC. Comparison of these two periods helps to indicate trends and seasonal variations that might have an effect on other comparisons involving before-and after-COMSEC periods.

- Comparison 2--Period I and Period III. These give a comparison of two periods falling in exactly the same calendar range, one before and one after COMSEC.

- Comparison 3--Period II and Period III. These give a comparison of two periods falling immediately before and after COMSEC's initiation.

These three comparisons illustrate some of the more meaningful comparisons for evaluation purposes between periods having similar seasonal positions. Such comparisons do not incorporate seasonal variation and are more likely to indicate actual changes.

In multiple comparisons there is a possible loss of statistical significance due to the use of the same data in more than one comparison. A number of statisticians have addressed this issue. However, it is felt to be of minor value here and is not considered in this report.

Analysis of comparisons during Period I and Period III must consider trends and other gross changes over a year separation. Analysis of comparisons in Period II and Period III must consider activities taking place at the time in addition to the COMSEC activity. Such activity as changes in the reporting system must be considered in all comparisons.

In each of the comparison situations, the later time period mean is subtracted from the earlier period mean. Thus positive values of z indicate a decrease, shown as an arrow down in Tables 2.2, 2.3 and 2.4. A negative value for z indicates an increase, shown as an arrow up. Since values may increase or decrease, the significance levels of statistical values must be based on a two-tail (test) significance number. For the normal statistic z, the significant value at the (two tail) 5 percent level is 1.96 and at the 1 percent level is 2.58. In the table, arrows are shown for significant values, i.e., for values equal to or greater than 1.96. Highly significant values are indicated by an H when the statistic exceeds the 2.58 value, i.e. at the 1 percent level.

Of course, significant and highly significant relate only to the numbers and the true significance is hard to say because of "other" factors, such as trends, seasonal effects, poor data, etc. The statistical results suggest situations but do not prove them.

A number of observations and interpretations may be drawn based on the statistics given in the Tables. These are intended primarily to guide and stimulate readers in developing their own analyses. It must be stressed that these remarks provide, at most, indications of possible situations and do not carry the force of statistical significance due to the time series nature of the data.

Because of the large number of cases, it is helpful to summarize the significant situations in order to recognize more easily indications of change or other possible results. Comparison 3 was not covered in this way because of the likelihood of seasonal effects contributing to these cases.

Table 2.5 gives a summary of the significant and highly significant cases for the two comparisons involving pre- and post-COMSEC periods.

Table 2.2 shows very few significant increases. Thus there may be a trend or seasonal effect that results in less reported crime in Period II than in Period 1. In comparison between Period III and these periods, this should be kept in mind.

Table 2.5 shows that many indications of significant difference are the same for the two comparisons. These cases, therefore, suggest a certain strength of effect that may indeed signal a post-COMSEC change.

REPORTED CRIME BY TYPE AND REGION SHOWING SAMPLE MEANS AND STANDARD DEVIATIONS

Period	Crime Type	D1 $\bar{X}$	D1 S	CD1 $\bar{X}$	CD1 S	City $\bar{X}$	City S	D3 $\bar{X}$	D3 S	D4 $\bar{X}$	D4 S	D5 $\bar{X}$	D5 S	D6 $\bar{X}$	D6 S	D7 $\bar{X}$	D7 S
Period I	Rape	1.2	1.1	4.8	2.2	6.0	2.5	.73	.98	.77	.93	1.3	1.0	.77	.93	1.2	.97
	Robbery	8.9	2.8	22.9	7.1	31.8	7.6	2.0	1.6	6.2	2.8	5.4	3.2	3.3	2.1	6.0	2.3
	Aggravated Assault	5.0	2.8	10.2	3.9	15.2	5.5	1.6	1.6	2.2	1.5	2.2	2.2	1.4	1.4	2.8	1.6
	Burglary	35.3	8.0	164.7	16.8	200.0	19.7	32.6	7.2	32.8	5.9	34.8	9.0	31.7	6.1	32.7	7.6
	Auto Theft	12.1	4.8	51.8	12.1	63.9	13.5	9.1	3.5	9.6	4.2	13.5	6.2	8.3	3.9	11.4	4.4
Period II	Rape	.96	.85	2.6	1.7	3.5	2.0	.27	.44	.58	.93	.42	.74	.27	.52	1.0	1.1
	Robbery	9.7	3.1	23.9	7.0	33.7	7.8	2.7	1.5	5.1	2.8	6.5	2.4	3.1	1.9	6.5	2.8
	Aggravated Assault	4.0	2.2	8.8	3.5	12.8	4.7	1.3	1.3	2.0	1.3	2.0	1.4	1.2	.89	2.3	1.5
	Burglary	30.0	8.2	156.0	22.7	186.0	27.3	33.8	6.3	30.1	6.4	36.4	8.9	26.5	8.3	29.2	6.3
	Auto Theft	12.2	3.8	42.8	9.3	55.0	9.2	7.7	2.9	7.3	3.1	11.6	4.9	7.5	2.9	8.5	3.6
Period III	Rape	.88	.85	3.7	2.0	4.6	2.5	.38	.62	.65	.87	1.0	1.1	.58	.74	1.1	.96
	Robbery	8.6	3.9	16.8	4.4	25.5	6.2	1.9	1.4	4.5	1.9	3.5	1.4	2.3	1.4	4.7	2.6
	Aggravated Assault	4.8	2.8	11.5	5.1	16.4	6.4	2.2	1.4	2.9	1.9	2.0	1.2	1.5	1.6	3.0	2.0
	Burglary	24.7	5.0	175.6	18.5	200.3	19.3	38.0	6.5	39.2	9.6	39.4	8.5	30.2	6.3	28.8	6.2
	Auto Theft	10.2	3.9	43.5	6.9	53.6	8.5	9.3	2.6	9.9	3.5	10.7	3.1	6.3	3.3	7.3	3.2

Table 2.1

TABLE 2.2 REPORTED CRIME BY TYPE AND REGION SHOWING TEST STATISTIC z

(Comparison Time Periods I-II)

Crime Type	Region							
	D1	CD1	City	D3	D4	D5	D6	D7
Rape	.72	H 4.0↓	H 4.0↓	S 2.0↓	.80	H 3.8↓	S 2.5↓	.69
Robbery	-.98	-.50	-.89	-1.7	1.4	-1.4	.36	-.70
Aggravated Assault	1.4	1.4	1.7	.69	.51	.39	.61	1.2
Burglary	S 2.4↓	1.6	S 2.1↓	-.64	1.6	-.64	H 2.6↓	1.8
Auto Theft	-.08	H 3.0↓	H 2.8↓	1.5	S 2.2↓	1.2	.84	H 2.6↓

TABLE 2.3 REPORTED CRIME BY TYPE AND REGION SHOWING TEST STATISTIC z

(Comparison Time Periods I-III)

Crime Type	Region							
	D1	CD1	City	D3	D4	D5	D6	D7
Rape	1.1	S 1.9↓	S 2.0↓	1.3	.40	1.0	.89	.37
Robbery	.32	H 3.7↓	H 3.3↓	.14	H 2.6↓	H 2.8↓	S 2.0↓	S 1.9↓
Aggravated Assault	.26	-1.0	-.73	-1.5	-1.5	.41	-.24	-.40
Burglary	H 5.7↓	S -2.2↑	-.06	H -2.8↑	H -2.9↑	S -1.9↑	.87	S 2.0↓
Auto Theft	1.6	H 3.0↓	H 3.3↓	-.35	-.28	S 2.1↓	S 2.0↓	H 3.8↓

TABLE 2.4 REPORTED CRIME BY TYPE AND REGION
SHOWING TEST STATISTIC z

(Comparison Time Periods II-III)

Crime Type	Region							
	D1	CD1	City	D3	D4	D5	D6	D7
Rape	.40	S -2.1↑	-1.8	-.71	-.40	S -2.3↑	-1.8	-.34
Robbery	1.1	H 4.4↓	H 4.2↓	S 2.0↓	.90	H 5.5↓	1.8	S 2.4↓
Aggravated Assault	-1.1	S -2.2↑	S -2.3↑	S -2.4↑	S 2.0↑	0.0	-.83	-1.4
Burglary	H 2.8↓	H -3.4↑	S -2.2↑	S -2.4↑	H -4.0↑	-1.2	-1.8	.23
Auto Theft	S 1.9↓	-.31	.57	S -2.1↑	H -2.8↑	.79	1.4	1.3

TABLE 2.5 SUMMARY OF SIGNIFICANT CASES

(Comparison 1 of Period I-III)
(Comparison 2 of Period II-III)

Crime Type[1]	Comparisons 1 and 2 by Region															
	D1		CD1		City		D3		D4		D5		D6		D7	
	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Rape			S↓	S↑	S↓							S↑				
Robbery			H↓	H↓	H↓	H↓		S↓	H↓		H↓	H↓	S↓		S↓	S↓
Aggravated Assault				S↑		S↑		S↑		S↑						
Burglary	↓H	H↓	S↑	H↑		S↑	H↑	S↑	H↑	H↑	S↓				S↓	
Auto Theft		↓S	H↓		H↓			S↑		H↑	S↓		S↓		H↓	

SYMBOLS
S = Significant value.
H = Highly significant value.
↑ = Increase in crime.
↓ = Decrease in crime.
Blank Cell = No significant difference.

In the applications of statistical analysis, there is widespread use of the so-called t-test. In many cases there may be no reason for using t-tests, and in fact their use may have little theoretical foundations. For these reasons, the normal z statistic is often the most appropriate one to use. Due to the interest many people have in the t statistic, it is discussed and illustrated after the following general discussion of elementary statistical procedures.

There are three elementary procedures that may be considered for comparing two collections of data to test the hypothesis discussed above. These are described below and considered in terms of the basic assumptions operating when each is used.

--Large sample theory was used with the Z-statistic above. In this approach one assumes that the sample values constitute a random sample (independent and from the same distribution) and that the sample size is large enough so that the sample mean has an approximately normal distribution.

In comparisons of two populations, e.g. time series, there is no need to assume equal variance of populations when using the Normal statistic. Variances are estimated by sample variance values.

When the population variances are known, the statistics are very accurate. They become somewhat less accurate when variances must be estimated.

When the sample population is normal, the method is exact. For nonnormal populations, the method becomes inexact.

The sample size required for necessary accuracy depends upon the characteristics of the sampled population. As these deviate from normality larger size samples are required. For skewed and biased data, fairly large samples may be required, but in many cases more than ten items will result in accurate enough measures of the characteristics for the sample mean distribution.

In the above examples with 26 sample points and no special skew properties, the normal approximation is considered to be very good.

--Small sample theory using the t-statistic. In order to overcome problems with using a sample variance when the true variance is not known and the sample is too small for a good sample estimate, one may use the t-static. In this method, one also assumes a random sample, but there are two other assumptions required: the population is assumed to be normal and in comparing two populations, both are assumed to have the same variance. Neither of these assumptions are needed in the large sample theory which is one strong reason for using it.

A widespread practice is to employ the t-statistic because of concern with "small samples" and to have little concern with the three assumptions underlying the use of this statistic. In many cases the samples are, in fact, large enough so that the central limit theory is in effect and the sample means are acting as though they are normal. Thus, use of the t-statistic does no harm since it yields essentially the same results as the large sample procedure in such situations. However, the use of the t-statistic is not improving on large sample theory in such cases and it should be avoided on the grounds of misleading sophistication. Numerical examples of the t-statistic are given and compared with some Z statistic results.

--The modified t-statistic when two populations do not have the same variance. It was pointed out above that in using the t-statistic to compare two populations, one had to assume equal variance. In considering the modified t approach one must first determine when the population variances are not the same. The F statistic given below is used for this purpose. It is illustrated for some cases of data in this study. In such cases where the hypothesis of equal variance is rejected (significant F value) the modified t-statistic is appropriate. That statistic is called τ here and its formula is given below.

To use the τ statistic, a complicated formula must be employed to compute the appropriate degrees of freedom (DF). This formula is given below and some illustrative DF values have been calculated. One obtains significant values by using t values with the calculated DF.

It has been argued above that the z statistic is often the most appropriate one in the study of reported crime data using 26-weeks or longer time periods. Since many people are interested in the utility of the t statistic, it was introduced above and the associated formulas for both the t and modified t statistic (τ) are given below. The F statistic, defined by the formulas below, is used to test the hypothesis that two collections (time series) represent populations with the same theoretical variance (standard deviation). It must be employed if one wishes to test for the necessity of using a modified t-statistic. The following formulas are given in general form for two sample populations X and Y. The special forms used with sample sizes both equal to 26 are also given since these are used in the illustrative examples.

The t-statistic on the hypothesis $\mu_x = \mu_y$

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{n_x s_x^2 + n_y s_y^2}} \sqrt{\frac{n_x n_y (n_x + n_y - 2)}{n_x + n_y}}$$

, with $n_x + n_y - 2$ degrees of freedom.

In the 26 week time period case $n_x = n_y = n$ so that

$$t = \frac{\overline{X} - \overline{Y}}{\sqrt{s_x^2 + s_y^2}} \sqrt{n-1} = \frac{\overline{X} - \overline{Y}}{\sqrt{s_x^2 + s_y^2}} \quad (5)$$

The F statistic on the hypothesis $\sigma_x = \sigma_y$

$$F = \frac{n_x s_x^2 / (n_x - 1)}{n_y s_y^2 / (n_y - 1)}$$

. When $n_x = n_y$ the formula simplifies

to $F = \frac{s_x^2}{s_y^2}$ with $n_x - 1$ and $n_y - 1$ degrees of freedom, which are both 25 in the illustrations.

The modified t-statistic under the hypothesis $u_x = u_y$

$$\tau = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_x^2}{n_x - 1} + \frac{s_y^2}{n_y - 1}}},$$

and when $n_x = n_y = 26$ yields

$$\tau = \frac{\overline{X} - \overline{Y}}{\sqrt{s_x^2 + s_y^2}} (5) = t.$$

For the modified t-statistic the formula for degrees of freedom is

$$2 + DF = \frac{\left(\frac{s_x^2}{n_x - 1} + \frac{s_y^2}{n_y - 1}\right)^2}{\frac{\left(\frac{s_x^2}{n_x - 1}\right)^2}{n + 1} + \frac{\left(\frac{s_y^2}{n_y - 1}\right)^2}{n_y + 1}},$$

With $n_x = n_y$ this becomes

$$2 + DF = \frac{\left(s_x^2 + s_y^2\right)^2}{s_x^4 + s_y^4} (n + 1) \text{ where } n + 1 = 27$$

in the case of 26 week long series. Note that the degrees of freedom are equal to the right side of the equation minus 2. In the reported crime study from which the illustrations are taken (3) most of the cases had a DF value about 50 and there was no particular need for considering the modified t-statistic. In crime analysis one would most often employ Normal statistics or go to more advanced methodologies such as are presented in subsequent sections.

The use of the t-statistic is illustrated in Tables 2.6, 2.7, and 2.8 which present t-statistic values for some cases in the reported crime study previously used to illustrate the z-statistic. It can be seen thay they have essentially the same numerical values as the z-statistic. In fact the formulas show that one formula uses $\sqrt{25}$ and the other $\sqrt{26}$ and this is the only numerical difference between them. Of course, one must refer to different numbers for identifying the levels of significance values. For the t-statistic the significance level values are given by the t random variable with 24 degrees of freedom. For two-tailed tests at the five percent and one percent levels, the significance values are 2.06 and 2.80 respectively.

The modified t-statistic may also be illustrated by considering the associated F test which is required to determine when the modification is called for. This test provides an indication of different variances between two samples (time periods). Table 2.9 gives some examples of the F statistic values. Significant cases corresponding to F values of 1.95 or 2.61 are indicated in these tables and it can be observed that relatively few such cases occurred.

Because of the similarity of the t and z statistics in values and the lack of any theoretical reason for using any form of t-test, it is often desirable to use the "large sample theory" z statistic approach.

All formulas in this Section are taken from Reference (6).

TABLE 2.6 REPORTED CRIME BY TYPE AND REGION SHOWING TEST STATISTIC t

(Comparison Time Periods I-II)

Crime Type	Region							
	D1	CD1	City	D3	D4	D5	D6	D7
Rape	.70	4.0	3.9	1.9	.79	3.7	2.4	.67
Robbery	-.96	-.50	-.87	-1.7	1.4	-1.4	.35	-.69
Aggravated Assault	1.4	1.3	1.7	.68	.50	.38	.60	1.1
Burglary	2.3	1.5	2.1	-.63	1.6	-.63	2.5	1.8
Auto Theft	-.08	2.9	2.7	1.4	2.2	1.2	.82	2.6

TABLE 2.7 REPORTED CRIME BY TYPE AND REGION SHOWING TEST STATISTIC t

(Comparison Time Periods I-III)

Crime Type	Region							
	D1	CD1	City	D3	D4	D5	D6	D7
Rape	1.1	1.8	2.0	1.3	.39	1.0	.88	.35
Robbery	.31	3.7	3.2	.14	2.5	2.7	2.0	1.9
Aggravated Assault	.25	-1.0	-.71	-1.5	-1.4	.40	-.24	-.39
Burglary	5.6	-2.2	-.05	-2.8	-2.8	-1.9	.86	2.0
Auto Theft	1.6	3.0	3.2	-.34	-.27	2.0	2.0	3.8

TABLE 2.8 REPORTED CRIME BY TYPE AND REGION SHOWING TEST STATISTIC t

(Comparison Time Periods II-III)

Crime Type	Region							
	D1	CD1	City	D3	D4	D5	D6	D7
Rape	.39	-2.1	-1.7	-.69	-.39	-2.3	-1.7	-.34
Robbery	1.1	4.3	4.1	1.9	.89	5.4	1.7	2.4
Aggravated Assault	-1.1	-2.2	-2.3	-2.4	-2.0	0.0	-.82	-1.4
Burglary	2.8	-3.3	-2.1	-2.3	-3.9	-1.2	-1.8	.23
Auto Theft	1.8	-.30	.56	-2.1	-2.8	.78	1.4	1.2

TABLE 2.9 REPORTED CRIME BY TYPE AND REGION SHOWING TEST STATISTIC F

(Comparison Time Periods I-II)

Crime Type	Region							
	D1	CD1	City	D3	D4	D5	D6	D7
Rape	1.5	1.7	1.6	6.3**	1.0	2.0*	3.2**	.83
Robbery	.82	1.0	.95	1.1	1.0	1.8	1.2	.67
Aggravated Assault	1.6	1.2	1.4	1.5	1.3	2.5*	2.4*	1.1
Burglary	.95	.55	.52	1.3	.85	1.0	.54	1.5
Auto Theft	1.6	1.7	2.2*	1.5	1.8	1.6	1.8	1.5

* Significant value.
** Highly significant value.

In the elementary procedures discussed and illustrated above a major assumption is that the sample is random. This means that each value is independent and drawn from the same probability distribution as every other value. However, the general approach to time series analysis takes a completely opposite point of view from the one reflected by an assumption of random sampling. The time series viewpoint is that the data may be related rather than independent (that they are, in fact, correlated) and that each time point has its own particular probability distribution. Thus arguments based on a random sample may not apply to time series.

The assumptions upon which general time series models are constructed are very basic. Therefore, those are the most general models, and analyses of any other kind must be compatible with them when applied to time series data. If that is not the case, the basic assumption, random sample, upon which the hypothesis testing is based (for normal or t-statistic procedures) is not valid. Time series models often assume stationary data which means each point in time has the same distribution, however, in general, the points are not independent. It is the general lack of independence that makes time series data samples nonrandom samples.

Time series models are discussed in other sections of this report. They require larger samples than the 26-week series used for the numerical illustrations in this section. A major objective of time series studies is to employ such models (e.g., autoregressive integrated moving average) to give theoretically justified statistical analyses of data that might allow one to discover indications of changes. Such models do not only address levels, but may provide an insight into the nature of the time series.

It should be noted that the simple statistical analyses described in this Section may be particularly useful in two ways:

--Effects can be so strong that they will be truly indicated by the elementary analyses even though there are theoretical objections to the procedures. Thus, if one looks upon such analyses as indications and views them in connection with other information, they can be useful. They should not be viewed as having the same kind of statistical significance as analyses in which the theoretical assumptions are shown to be satisfied.

--The nature of some crime types, such as rape, is such that the time points are likely to constitute a random sample. In such cases the elementary analyses are valid in the full statistical sense. One value of more general time series analysis is that it helps to establish which crime types are of purely random character.

An alternative to using statistical comparison of sample statistics such as those discussed above is to present the variation in graph form. For time series data such as reported crime, each time period is assigned a position on the horizontal axis. For example if there are seven time periods available, then seven, equally spaced positions are assigned on the time (horizontal) axis and given the labels of the corresponding periods. Above each period, the mean value of the sample for that period is indicated by a point. This assigns the dimension of "number of reported crimes" to the veritical axis. The variability that is present in the data at each time period is indicated by showing a short line one standard deviation above the mean value and another the same amount below the mean. Such a diagram gives direct visual appreciation of the upward and downward changes in the mean values for different time periods. In addition, it shows how meaningful the changes in mean are by indicating the strength of variability present. If the one standard deviation above and below the mean results in large intervals, the means do not well represent the data and their movement up and down does not necessarily show time changes in data levels. On the other hand if the variation spread is relatively small the mean does represent the data and changes in the mean value are likely to be statistically significant. Actual significance for crime analysis in such cases would depend on the magnitude of the difference.

Figure 2.1 shows the alternative method for observing statistical variation for selected crime types. The five time periods (each six months long) are indicated as points on the horizontal axis. For a selected crime type, such as robbery in D1, a line segment two standard deviations long, centered at the sample mean value is drawn above each period.

By considering plots, such as those shown in Figure 2.1 one can observe the changes in mean value directly. Since the standard deviations are also shown, one can see how representative the mean is likely to be. One can also get an impression of how the random occurrence of events is operating. A large value indicates a "loss of control". That is, some kind of extreme change in the situation. Smaller values of the standard deviation indicate some level of statistical stability, allowing more reliable interpretations of data and derived statistical analyses.

Figure 2.1 indicates that the variability in robbery increased in Period V resulting in a lower degree of statistical control for this crime type. This was not the case for aggravated assault which kept about the same variability in all periods. Burglary shows a marked decrease in mean values which is particularly meaningful because of the relatively constant level of variability. Of course the variability is in fact fairly large in all these cases as one must expect from social as distinct from engineering type data.

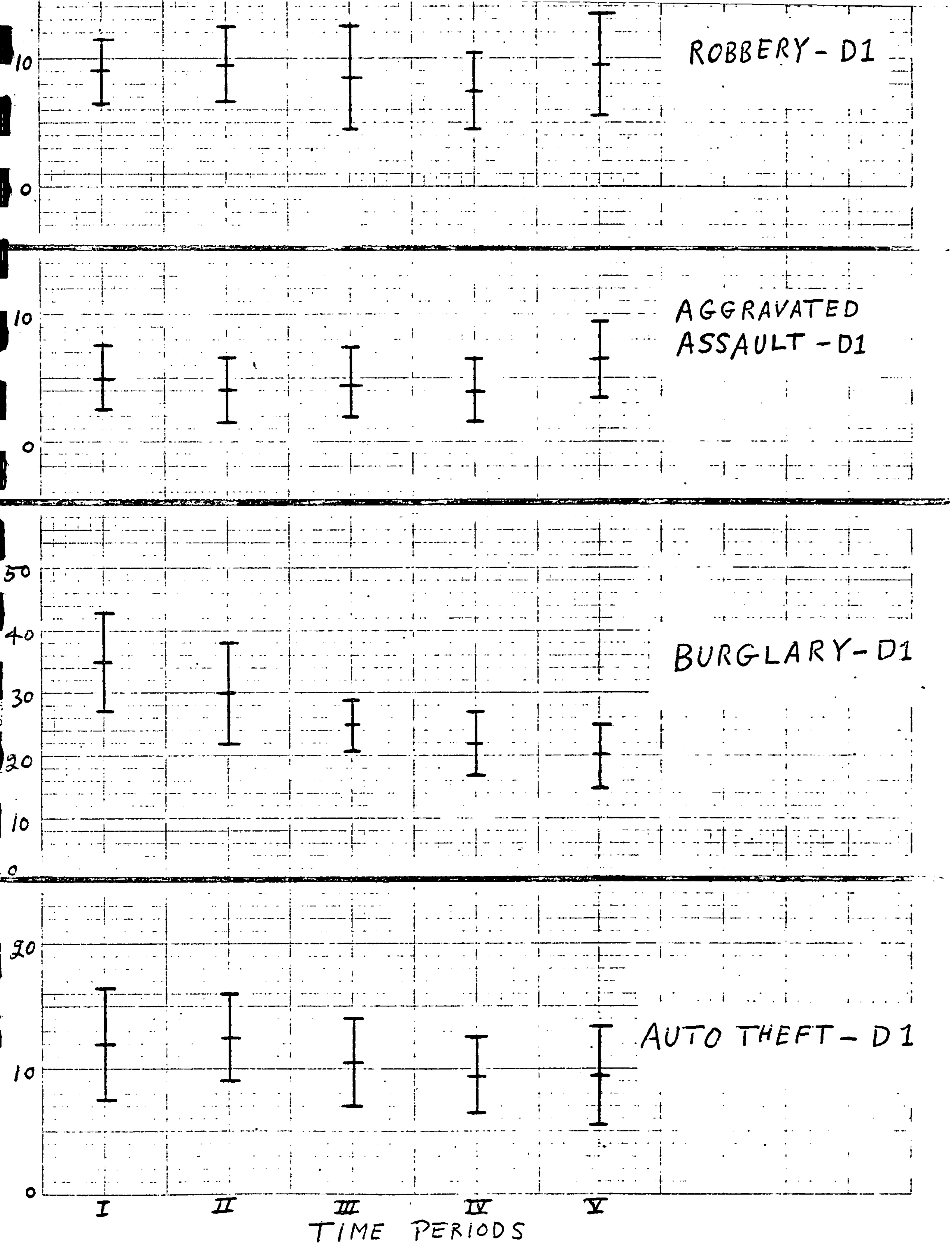


Figure 2.1. Pictorial Presentation of Elementary Statistical Variation

Section 3 Preliminary Analysis of Time Series

Before any kind of quantitative statistical analysis of time series data is undertaken it is advisable to carry out some amount of preliminary analysis. The details and extent of such analysis depends on the kind of subsequent investigations that are intended and the magnitude of the entire statistical effort. In the case of a small effort the preliminary type analysis may form essentially the complete study or at least represent a major part of it. When extended investigations are carried out including detailed statistical calculations and stochastic modeling, the preliminary analysis will represent only a small fraction of the overall effort.

The purpose of preliminary anlysis is to understand the time series data under investigation. During this phase of study one becomes familiar with the kind of reliability and underlying statistical variability that are present in the data. This understanding is qualitative in nature, sometimes assisted by numberical measurements. It may suggest subsequent quantitative analysis and lead to useful conclusions about the data.

In this Section the following features of preliminary analysis of time series will be described and illustrated:

o Qualitative nature of time series data relative to its reliability, and utility for various levels of statistical analysis. Component models.

o Investigation of non-stationary character and cyclic trends.

o Investigation of homogenity of the data and advisability of data transformation.

Qualitative Analysis

The most direct approach to begin any study of time series data is to make a plot of the data values against time. Even such a basic tool of analysis requires several decisions for its implementation. The kind of data values must be selected. In analysis of reported crime this means selecting the geographic region and crime type; for example robberies in District 1 might be selected. The time scale must also be selected, for example the plot might be by day, or week, or month. Aggregation of data by region, crime type, and time period can affect the nature of the plot and also of subsequent statistical analysis.

Another problem with time series plots is simply the effort required to produce them. To be useful time series should contain a fairly large number of data values leading to rather long plots. Many different cases of series data may be desired; for example one may wish to study some ten crime types for six police districts. One can very soon be faced with plotting between 70 and 100 graphs, each with over 70 data points. This is a substantial task if undertaken by hand.

Fortunately if the data are developed into a computerized data base the plots can be done by most computer centers on a digital plotter. It is also possible to form plots with the printer but these are not nearly as satisfying as plots produced on a plotter. If one has the option of using a plotter it should certainly be used. Many plots of high quality can be produced at low cost (plotters operate "off line" from tape produced by the central computer). The computer produced time series plots provide a powerful tool for preliminary analysis, particularly when many plots are desired.

Figures 3.1 and 3.2 illustrate computer produced plots of reported crime data. In the plot of Figure 3.2 the overall mean value and sample standard deviation values are shown by three horizontal lines. The lines on either side of the mean show an amount equal to one standard deviation. Though these values are for the entire series and thus do not represent information about short segments of the series, they do indicate something about the overall level and extent of statistical variation present in the data.

Even when more detailed studies intend to use weekly data it may be of value to consider time series plots of monthly data as a step in preliminary analysis. Reported crime in particular suggests such an approach. Such data are more likely to be available as monthly time series than by week. Therefore analysis of data can start sooner if monthly data are used and the value of going after weekly data can be assessed. If the monthly data seems to be in good shape and further analyses are contmeplated weekly data can be requested. The details of data selection are described in Reference (1).

Monthly data are more likely to indicate certain seasonal trends and overall properties of some time series. In some ways their underlying structure is simpler than for weekly series and can be more easily developed into a preliminary component type of model. By considering a simple components model one can gain some feeling for the statistical variation present in the data. This is illustrated by Figure 3.3 showing robbery in District 1 for a 54 month time series.

A simple analysis of this kind of data is shown in Figure 3.4 which divides the full time series into three distinct periods. The first is a rather stable series from January 1968 to March 1970 then a transition region exists during which the level of the series increases to a new stable series with a mean about 12 robberies a month greater than for the earlier period. The transition region can be represented by a varying mean curve as shown. These simple analyses may employ regression to fit underlying trend curves or combine the use of visual fits and calculated mean and standard deviation values.

The lines for the transition region were found as regression lines for regions determined visually as being distinct in a qualitative way (by subjective judgment). The upper and lower lines represent one standard deviation difference for each region being considered. The December values show a strong seasonal effect for robbery. They distort the other data and should not be included in calculations of the mean and standard deviations. Such points are out-lyers for which a specific justification should be sought rather than mixing them in with the nominal statistical data.

By considering data plots, imposing sample mean and standard deviation values, identifying changing regions and special characteristics one gains an important familarity with the time series under consideration. Making some simple model forms, with or without regression techniques serves to increase and sharpen this familarity to a point where the analyst can embark on detailed statistical studies with some initial feeling about the data. This plays the

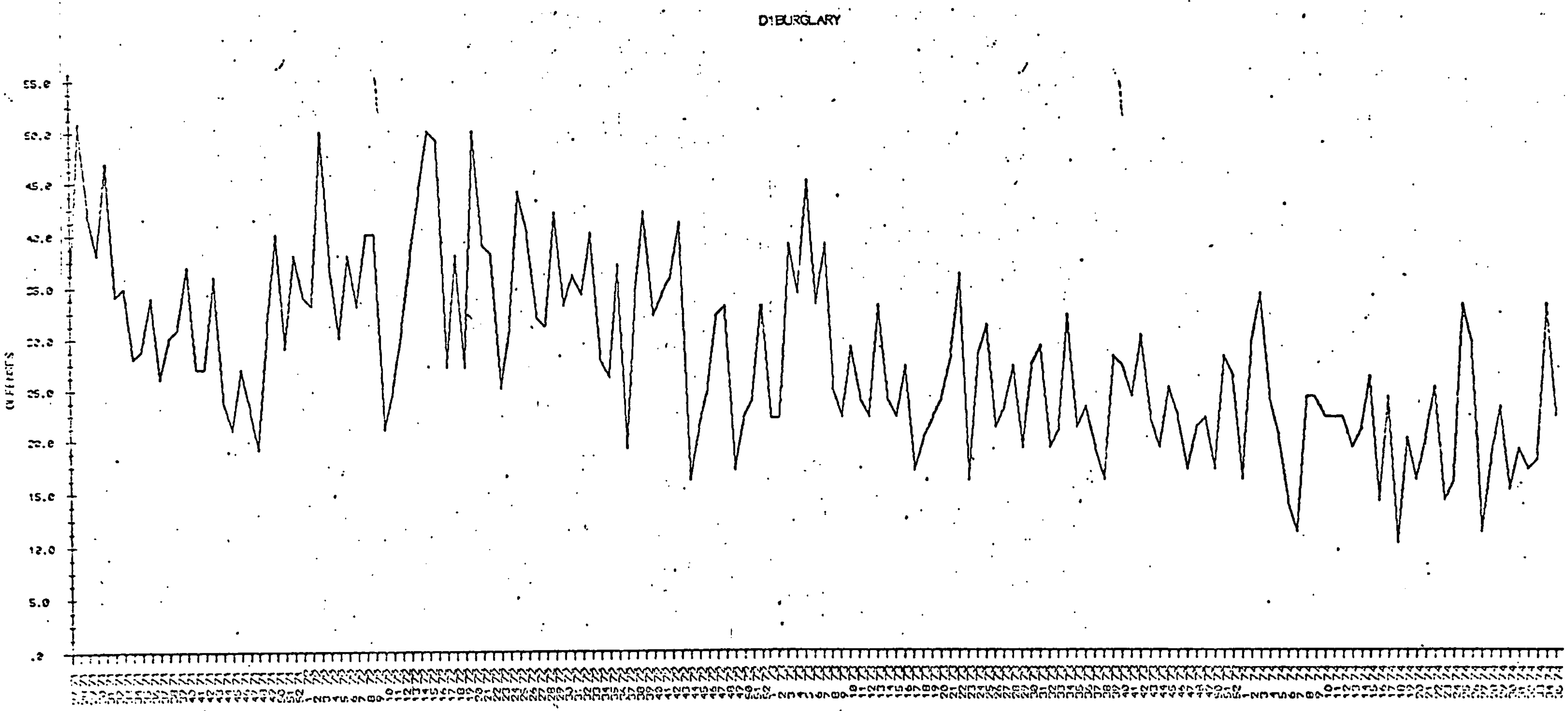


FIGURE 3.1
BURGLARY IN DISTRICT 1 BY WEEK

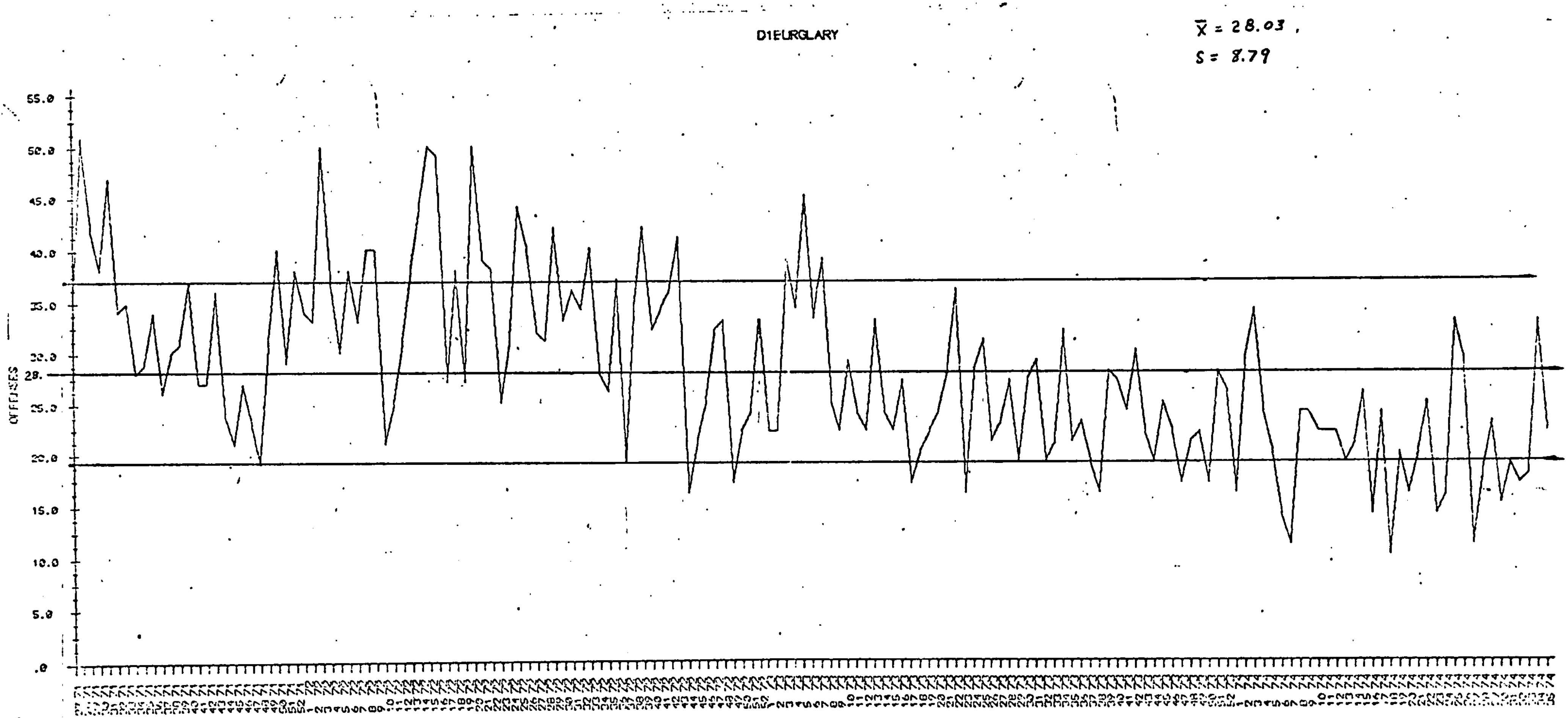


FIGURE 3.2

BURGLARY IN DISTRICT 1
SHOWING MEAN AND STANDARD DEVIATION

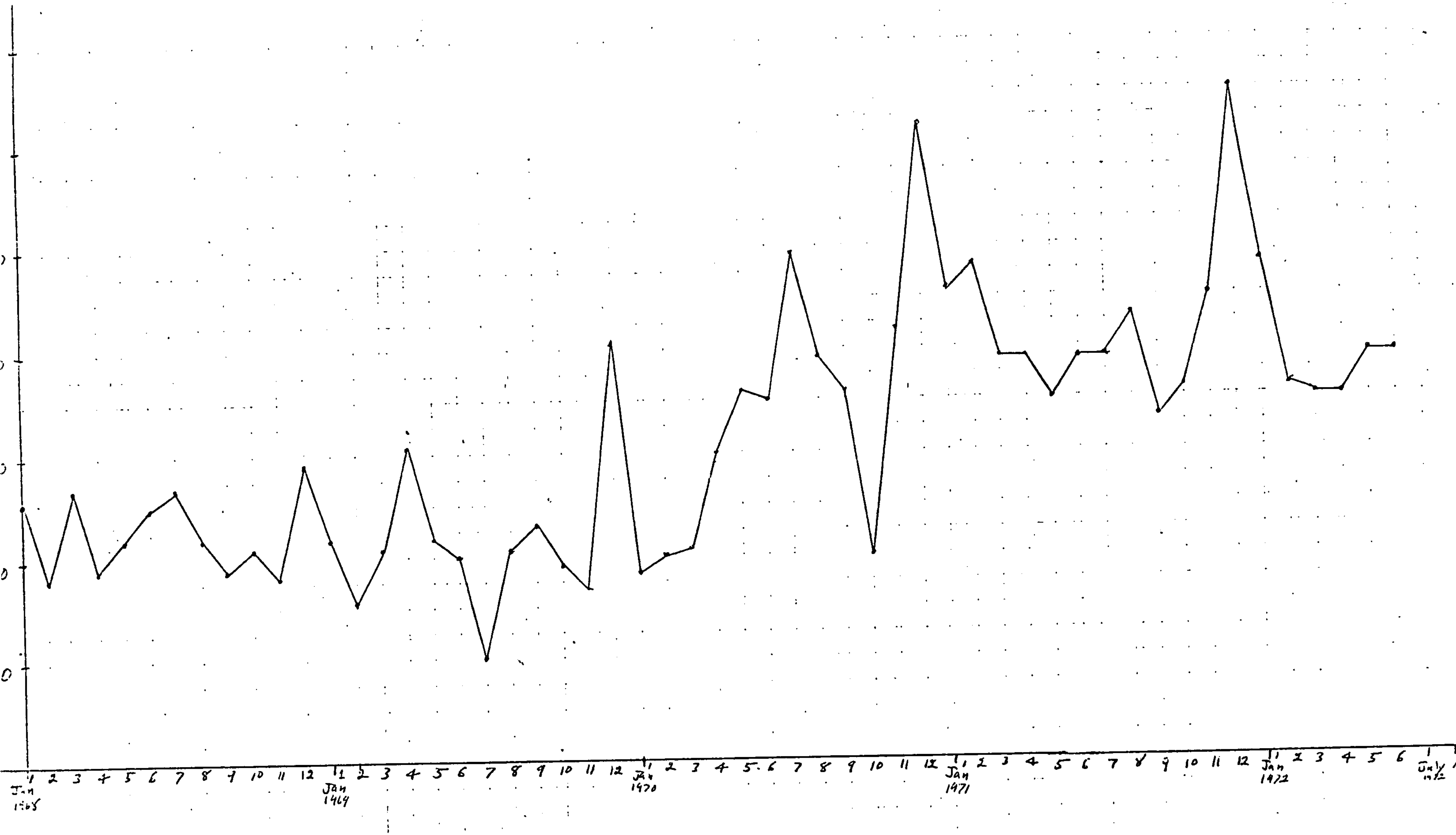


FIGURE 3.3

ROBBERY IN DISTRICT 1 BY MONTH

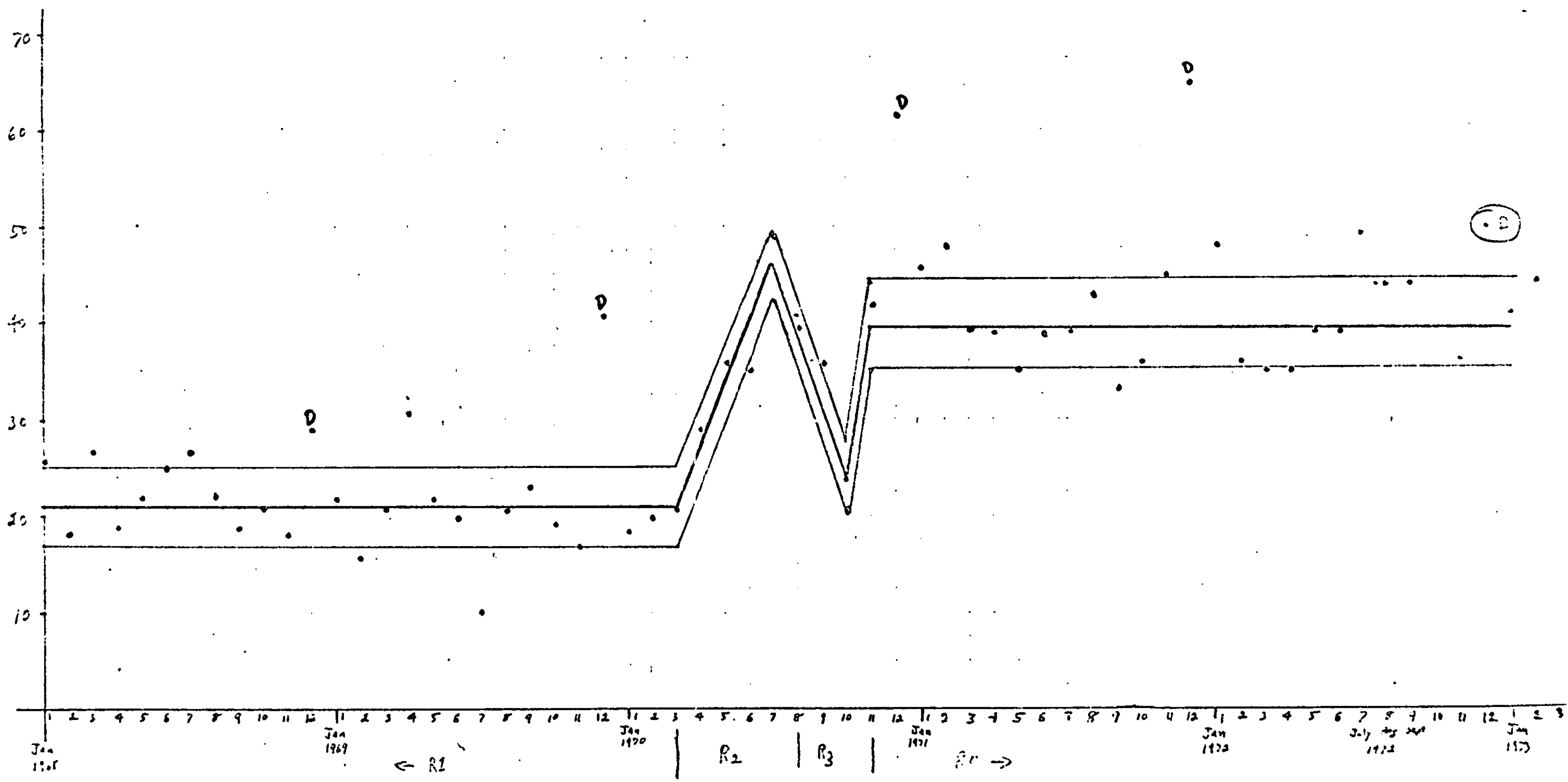


FIGURE 3.4

ROBBERY IN DISTRICT 1
SIMPLE PRELIMINARY ANALYSIS

role, for analysis, that real would experience with the phenomena under study plays for the potential customer of the analysis. Therefore preliminary analysis of the kind indicated above helps in communication between the analyst and the customer as well as providing a basis upon which to continue into more detailed studies.

Investigation of Trends

There are two major kinds of trends that may be present in time series data. One is a change in level resulting in a non-stationarity of the underlying statistical nature of the data. The other is cyclic variations which occur regularly in a series. These are stationary in nature but can distort interpretations of the data if not identified and measured.

Stochastic models require stationarity of the data. They can accommodate cyclic trends if these are properly identified. Both kinds of trends can be identified by means of detailed statistical analysis as discussed in Section 4. However it is often an advantage to study time series plots with particular attention paid to the existence of trends. By making qualitative identification of trends in the preliminary analysis stage one can consider appropriate modifications of the data (such as differencing to reduce non-stationarity) and more easily understand the nature of detailed statistical calculations (autocorrelation values).

Figure 3.5 shows a time series plot for the Sump water level example.

In Figure 3.5 the existence of a trend is clearly indicated. Since the data for this example were generated by a synthetic procedure it is known that the trend line has slop .2. Such a line is shown in the figure with a starting value equal to the mean of the same series without the trend (Sump 1 example) plus a shift of 2 units which was also introduced into the trend line. The deterministic removal and addition of water to the sump can also be observed though there is enough random variation to somewhat mask the various deterministic factors.

Investigation of Homogenity

If one wishes to develop a mathematical model of a time series it is desirable to have about the same amount of variation over the length of the series. This property is called homogenity. By observing a plot of the data one can often judge the degree to which the series is homogeneous. When the series is not homogeneous the data can be transformed by raising it to some power k. If the best power to use seems to be k=0 a natural log transformation is used. The particular power to use is based on judgment and tested by checking the transformed data for homogenity. One can gain some initial experience by forming two or three transformed series and noting the effect of the powers used. It is often useful to use one fractional and one integer power as a start.

Sometimes a semi-quantitative approach can be used to test time series data for homogenity. This is known as the range-mean plot. Such a plot is based on a set of time series values $\{z_t\}$. A spacing value S is selected and the terms $\{z_{t-ks}\}$ are collected into a set S_0 the terms $\{z_{t-1-ks}\}$ into a set S_1 and so forth, where $k = 1, \ldots, m$ and m is determined by the amount of available data (length of the time series z_t).

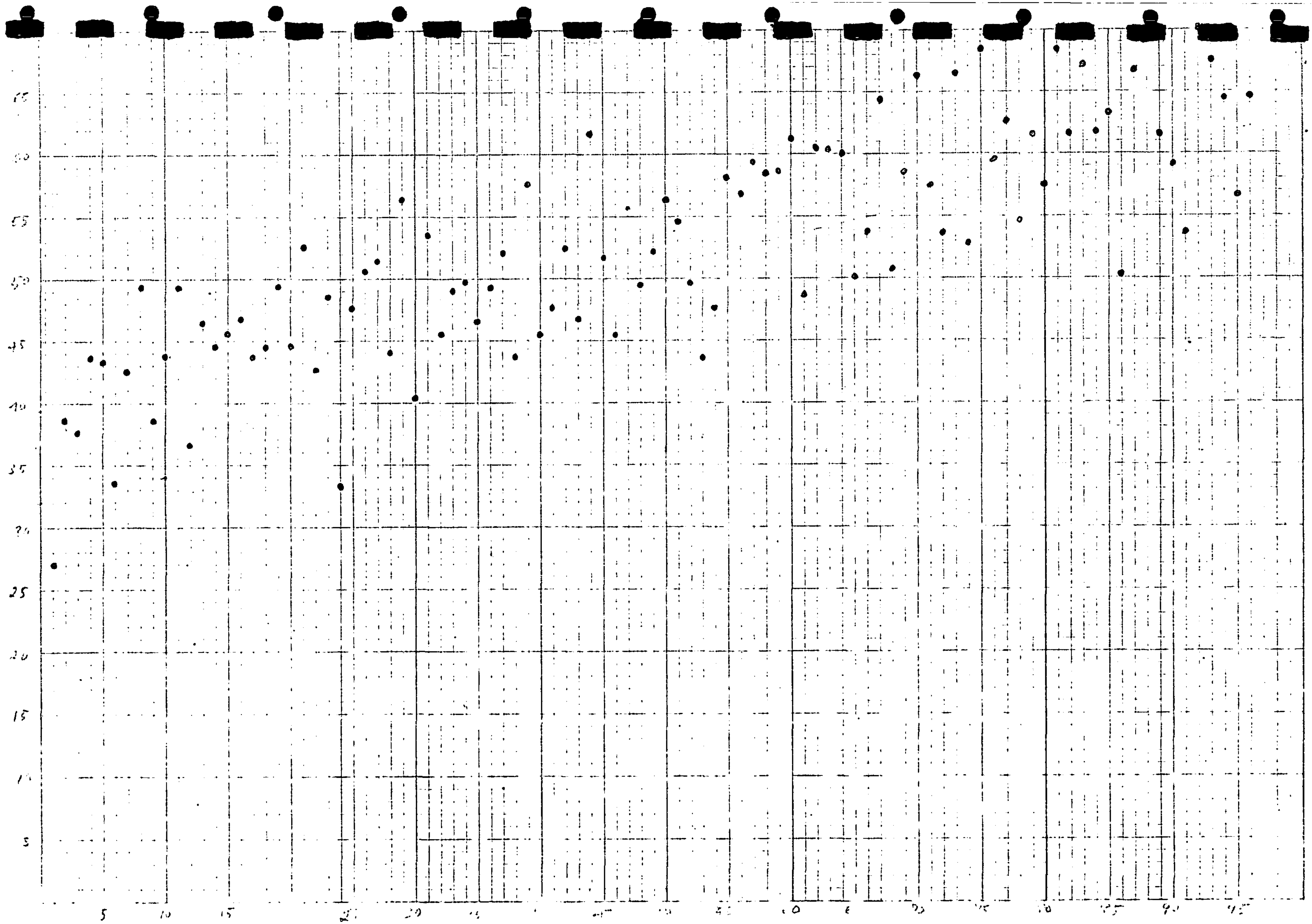


FIGURE 3.5 SUMP 2 DATA

Suppose that p such sets of data are obtained in this way, where p is determined by the amount of available data. The values m and p are selected so as to use most of the data values, give the sets S_j, j = 1, . . . , p reasonable content (each contain m values), and have enough sets so as to get a reasonable number of points on the scatter plot. For each of the sets S_j, one obtains the range R_j of data values (largest value minus smallest value) and the mean value M_j of the data points. Each such pair (R_j, M_j) forms a point on the scatter plot of range R vs. mean M. If the range does not depend on the mean, then no transformation of data is indicated. Otherwise, the dependencey observed may suggest a useful transformation.

There are two ways in which the scatter plot can indicate a lack of functional dependence between the range and the mean: One way is to have the p points fall in very general positions. The other is to have them fall in an almost horizontal pattern. It may be appreciated that this analysis of data transformation is highly subjective in most cases; but it is felt to be of value as a general guide and might be useful in analysis of reported crime.

The range-mean analysis for homogenity can be illustrated for robbery data as shown in Figure 3.6 for police District 1 in Cincinnati. The data series was divided into seven subsets each covering six months of weekly data. Though robbery data have been modeled directly the plot in Figure 3.6 indicates that a transformation might yield more easily analyzed data. A log transformation is suggested by the 45^{o} line through the data. For lines having less slope a positive power between zero and unity is suggested; for greater slopes negative powers are more apprpriate. This may be explained by considering the case of a slope of 20^{o} (less than the 45^{o} slope corresponding to 0) which might suggest a power like .5. For this transformation relatively small values are not changed too much while larger values are changed by a greater amount thus tending to bring the variation of data values to the same level.

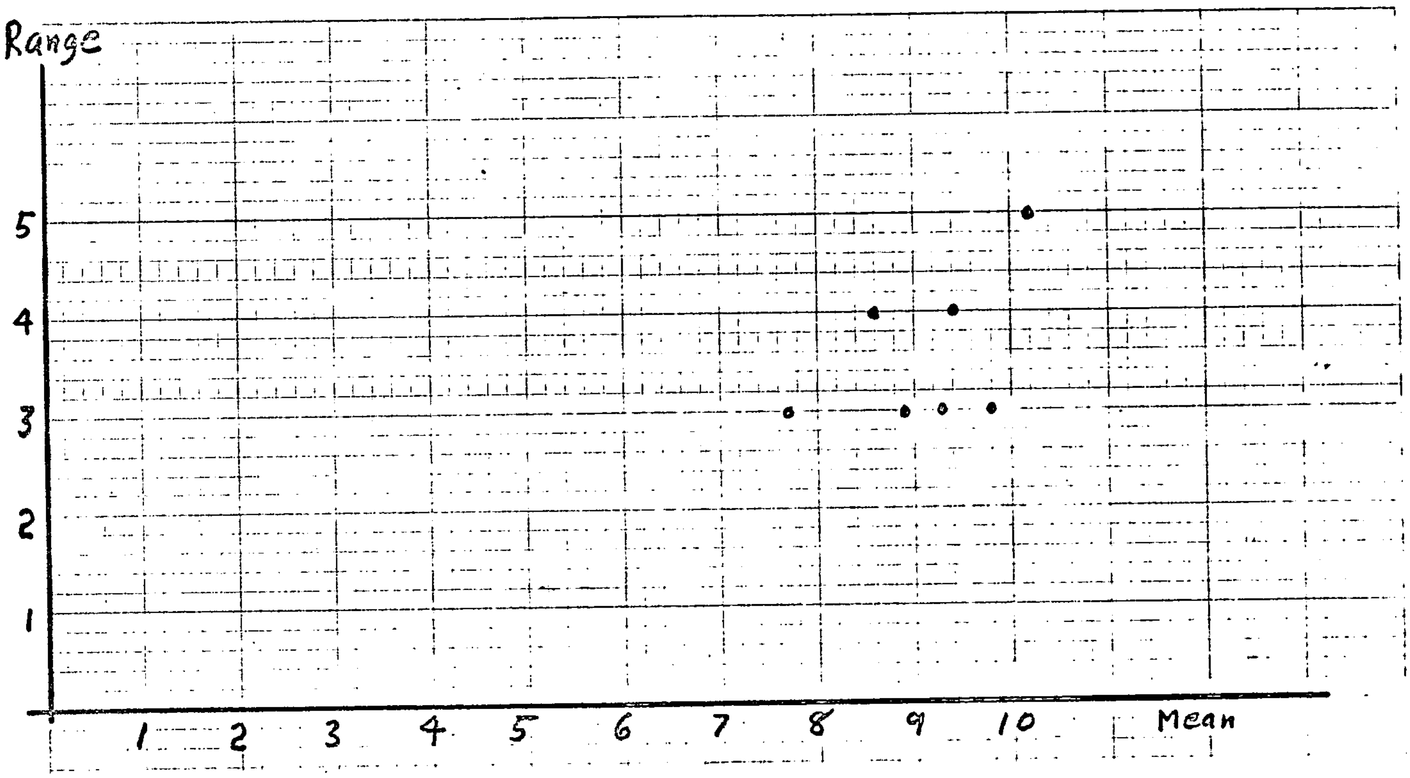


FIGURE 3.6
RANGE - MEAN PLOT FOR ROBBERY

Section 4. Detailed Statistical Techniques

The sample mean and sample standard deviation of a time series are two numerical characteristics of a series but fall far short of complete characterization of the time series as represented by data.

In the study of time series the mean and standard deviations should be considered. However, there are two major reasons for developing a methodology for constructing detailed representational statistics in addition. When there is a strong effect that is reflected by changing mean values more detailed statistics will provide greater appreciation of the kinds of differences present. Alternatively the change in time series samples may occur in fundamental aspects of the time series rather than in the levels indicated by the sample mean. The mean values may fail to indicate change when, in fact, the two time series are different. In such a case simple statistical procedures would indicate no change, possibly missing important changes that are, in fact, present.

The case in which two time series can be completely different without an indication of that difference being shown by the mean value or standard deviation can be illustrated by easily constructed examples designed for that purpose. However, Figure 4.1 shows an illustration using actual time series data. In Figure 4.1 both time series have a mean value of 14 and standard deviation of five (to the nearest integer). The plots certainly look quite different (the peaks and valleys generally having very different magnitudes and falling in different positions) and in fact one is for auto theft and the other is for total nonindex crimes by week in police District 1 of Cincinnati (from mid-1971 through 1972). Thus the use of mean and standard deviation values can completely miss significant differences between time series.

Because of the random nature of time series such as crime data, the difference in appearance of the two plots does not necessarily imply a true difference between these time series. For example, they might look like Figure 4.1 and both graphs represent different time series of the same crime type. A more critical test of difference, not based on the appearance of the graphs, is to look at a number of detailed statistical values computed for each series. These autocorrelation values are discussed below. Figure 4.2 shows 20 such values specified as lag values for the autocorrelation of each crime type. These values do indeed indicate a considerable difference between auto larceny and nonindex crime.

Sample correlation values give a more complete characterization of a time series by providing additonal numerical quantities derived from the data. These values express the degree to which data at various time intervals (e.g. weeks) called lags, are correlated. A lag of zero corresponds to the sample variance. Increasing the lag corresponds to measuring the extent, over time periods, to which series values are similar, i.e., correlated. Since the sample values are all for one time series, correlation is between different time data values of the same series. Thus, the sample statistics are called autocorrelations, denoted by acf (autocorrelation function). Figure 4.2 shows 20 autocorrelation values for the reported crime series given in Figure 4.1.

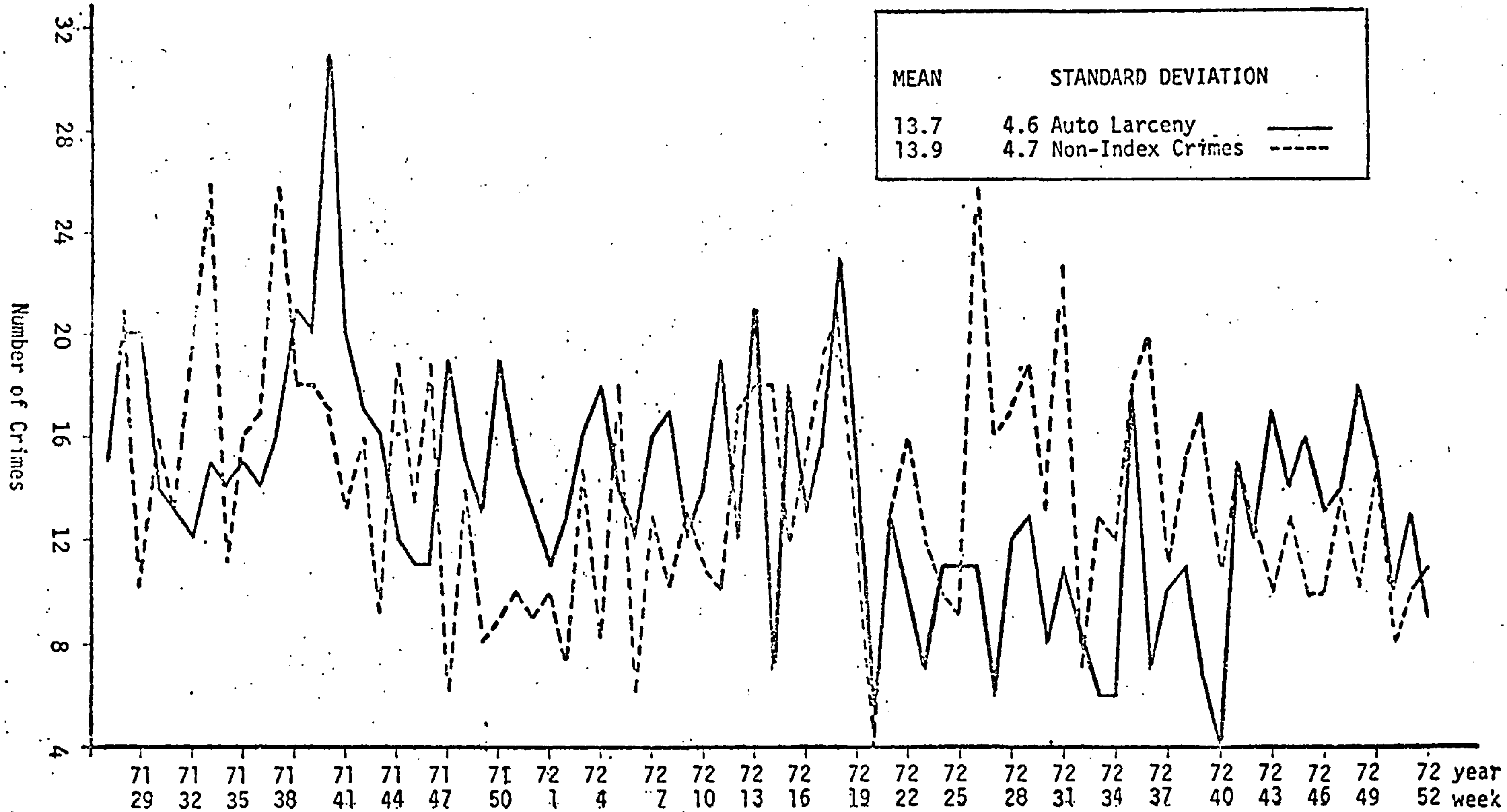


FIGURE 4.1 WEEKLY PLOT OF AUTO LARCENY AND NONINDEX CRIMES FOR DISTRICT 1

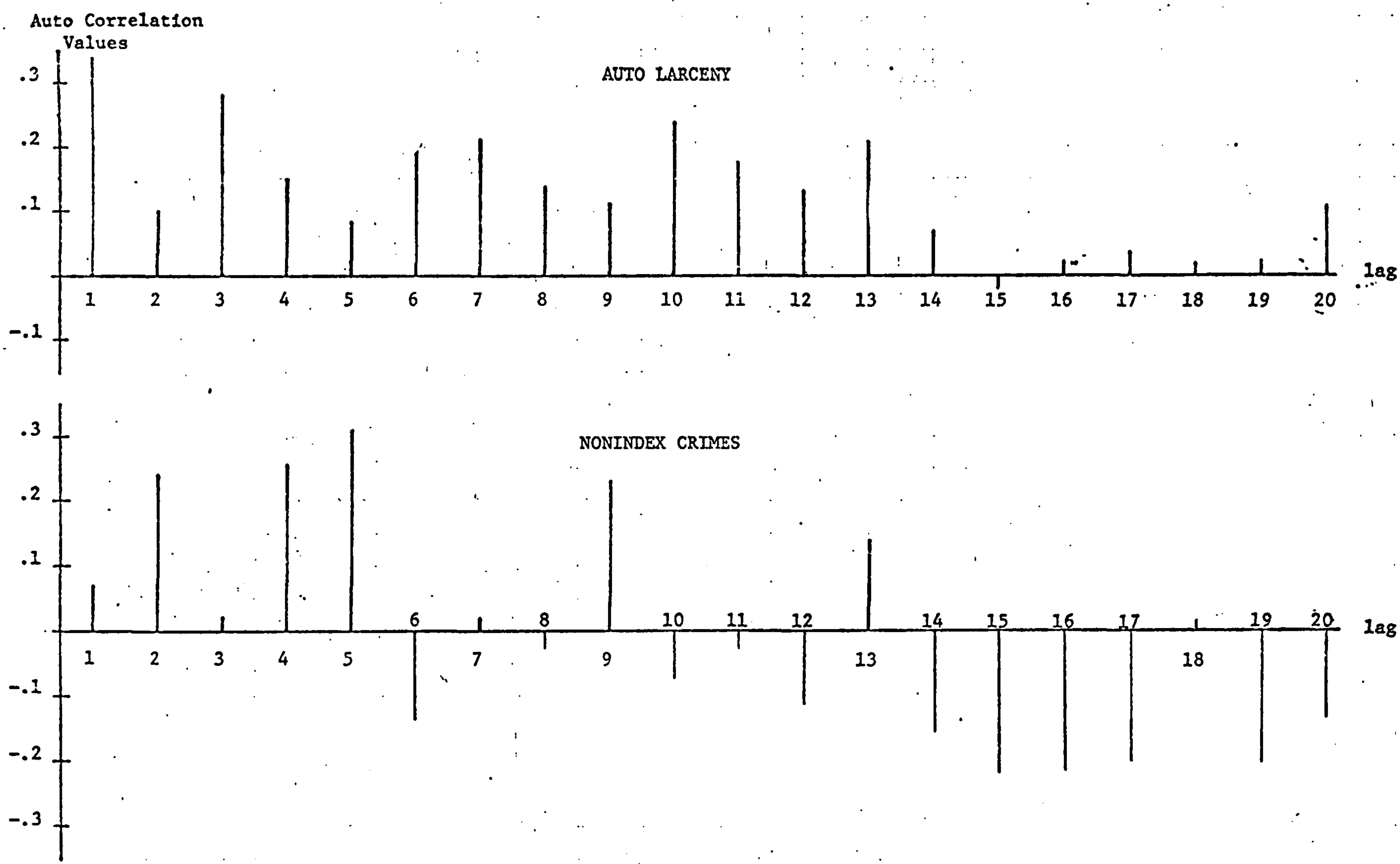


FIGURE 4.2 AUTOCORRELATION VALUES FOR 20 LAGS IN DISTRICT 1

Autocorrelation values are computed from the time series data by using formulas from mathematical statistics. These formulas produce values of desirable form which also have an associated statistical thory so that in appropriate cases statistical tests of significance can be applied. The method for computing autocorrelation values follows:

Denote the time series by $\{w_i\}$ and its mean by $\bar{w}$ where $\bar{w} = \frac{1}{n} \sum_{i=}^{n} w_i$, n is the number of data points in the time series. Then the autocorrelation of lag k is $r_k = c_k / c_0$ for k = 0, 1, ..., K where K is the maximum lag considered and the autocovariance values c_k are defined by the formula:

$$c_k = \frac{1}{n} \sum_{i=1}^{n-k} (w_i - \bar{w})(w_{i+k} - \bar{w})$$

for k = 0, 1, ..., K. Note that c_0 is the sample variance and r_0 is always unity (by definition).

Autocorrelations can be presented in either numerical form as tables or in graphical form as shown in Figure 4.2. The numerical values are useful when the exact values are needed whereas the graphical form is most useful in giving a picture of the general nature of the autocorrelations associated with a time series. For many purposes of analysis, such as the formulation of mathematical models of the time series, the pattern of autocorrelation values plays a major role. Such patterns are best seen by means of the graphical presentation. In the examples both tables and plots of autocorrelations will be illustrated.

An artificial time series representing the water level at 96 time units (hour) for the sump example yields the autocorrelations listed in Table 4.1 and shows in the plot of Figure 4.3.

Lag	1	2	3	4	5	6	7	8	9	10
Autocorrelation	-.26	-.03	-.11	-.01	.03	.14	-.17	.18	-.26	0.0
Lag	11	12	13	14	15	16	17	18	19	20
Autocorrelation	.04	.06	.12	-.14	.01	-.03	-.09	.17	-.01	-.14

Table 4.1. Autocorrelations for Sump 1

Since the Sump 1 time series has 96 data points and $1/\sqrt{96} = .1$, only autocorrelations greater than .1 in magnitude have any possible significance. These correspond to lag values of 1, 3, 6, 7, 8, 9, 13, 14, 18, 20. Of these the values at 3, 13, 14, and 20 are near .1 and likely to have little meaning. The values at 6 and 18 seem to reflect the deterministic in phase changes being made to the water level at six hour separations. This is also true of the relatively strong autocorrelations at 1 and 9 due to the combination of 3 hour and 6 hour deterministic changes in water level.

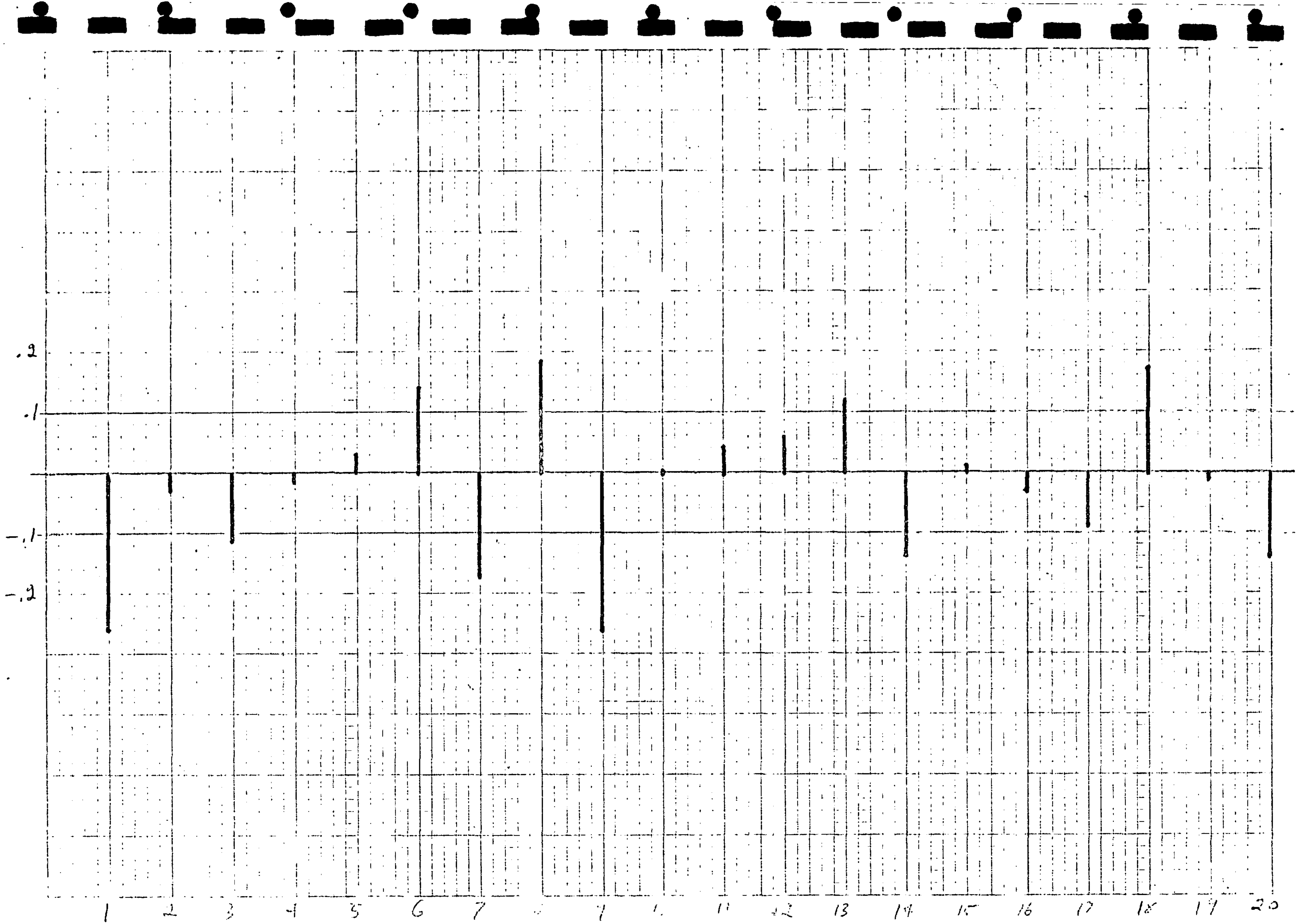


FIGURE 4.3 SUMP 1 AUTOCORRELATIONS

If the time interval for the data is weeks, then the lag is measured in weeks. The autocovariance values C_k measure the variability about the full series mean $\overline{W}$ due to terms in the series separated by k weeks (lag k). When k=0, this is the sample variance resulting from the full series of n data values. Autocorrelations are measured in units of the variance C_o and must take on numerical values between -1 and +1. A high positive (e.g., .8) value for an autocorrelation of lag k indicates that values separated by k weeks are similar in that if one data point is high the data point k weeks away will be high and if one is low both will be low. Large negative values (e.g., -.75) indicates the opposite effect. The weekly values k weeks apart are related but in an inverse way so that when one value is high the other is low. Of course, these are statistical indications only and represent a likely general behavior of the series data. Particular values may not actually conform to the likely behavior indicated by the series autocorrelations. When the autocorrelation C_k is near zero (positively or negatively), data values separated by k weeks are not likely to have a significant relationship to each other.

Autocorrelation values may indicate a number of things about a time series. If the sample data points do not reflect changing levels of crime or other major changes in the data collection system, the series is said to be stationary. In such cases, the autocorrelation values would be expected to decrease in magnitude (disregarding the positive or negative nature of the data). This is because values should become less related as they become separated in time except for three kinds of situations. One situation in which the autocorrelation values do not decrease quickly as lag value increases is the non-stationary series mentioned above. In such cases, there is some real effect operating to produce data that are correlated. A trend line is a typical indication of such an effect. For example, if burglary increases due to the pressures of a bad economy, there can be a general increasing trend rendering the time series of burglaries non-stationary. Autocorrelations can be used to indicate the presence of such trends.

Another situation that causes large autocorrelation values at various lag amounts is a so-called seasonal or periodic effect. If robberies are high in the third week of every month, then some indication of this will be shown by rather large autocorrelation values for lag value 4 (the approximate weekly separation). Autocorrelation values can, therefore, be used to indicate the presence of periodic effects. For reported crime data, periodic effects are not likely to occur in weekly time series. They are more likely to show up, if they exist at all, in monthly data. Some aspects of a periodic effects study are given in Section 9.

The third situation which may yield relatively large autocorrelation values in various patterns is when the time series data have some degree of statistical structure leading to the possibility of representation as a mathematical model. The autocorrelation values can be used to indicate the most likely such model form and show how it should be developed. This is the approach to stochastic models discussed in Section 6.

When there are no autocorrelation effects, the time series is said to be purely random in character. The term random may be used in a number of ways, usually implying the opposite of deterministic. A random occurrence is one that depends on some kind of probabilistic framework which may be supposed to be acting, which governs the occurrence. In this report, random is used in this sense and more particularly to denote a lack of correlation effects between joint events which taken together are subject to probabilistic laws, such as the time series values of reported crime. These examples present autocorrelations for Police District 1 Cincinnati (D1), all of Cincinnati except D1 (CD1), and for the city. A maximum lag value of 20 is used. Autocorrelations were computed for the 87-week pre-COMSEC and a 78-week post-COMSEC period in each of the regions D1, CD1 and city.

The autocorrelations may be presented as tabulated data or as graphs. In either case, there must be some measure of purly random difference from zero autocorrelation. One wishes to use the acf values to characterize and interpret the time series. In such applications, "significant" numerical acf values are employed. However, it may be that some nonzero values (at various lags) have no real significance, but occur only from purely random data effects, "small" sample sizes and numerical round off. This situation is dealt with by stipulating a value measuring the sample vaiability of the sample acf quantities. If n is the number of values in the time series (size of the time series sample), then $1/\sqrt{n}$ is considered to represent a single standard deviation of randomness in acf values. In many studies of time series $2/\sqrt{n}$ is used as the critical value, only greater values being taken to have nonrandom significance. In the examples $1/\sqrt{n}$ was used to prevent loss of information. It may be noted that for the 78 and 87 week series the quantity $1/\sqrt{n}$ is .11 to the nearest hundredth and this value is used in all cases.

The graph presentation of acf values is extremely useful and can serve as a guide in series comparisons and in the basic structure stage of stochastic modeling. Examples of autocorrelation plots will be given following the numerical table.

Detailed comparison of time series requires consideration of the numerical acf values. These are illustrated in Table 4.2 which shows all acf values greater than .11 magnitude, up to lags of 20. For reference the table also gives sample mean, $\bar{x}$, and sample standard deviations, s, for all cases.

Figures 4.4, 4.5, and 4.6 give autocorrelation plots for Rape, Robbery, and Burglary corresponding to the numerical data of Table 4.2.

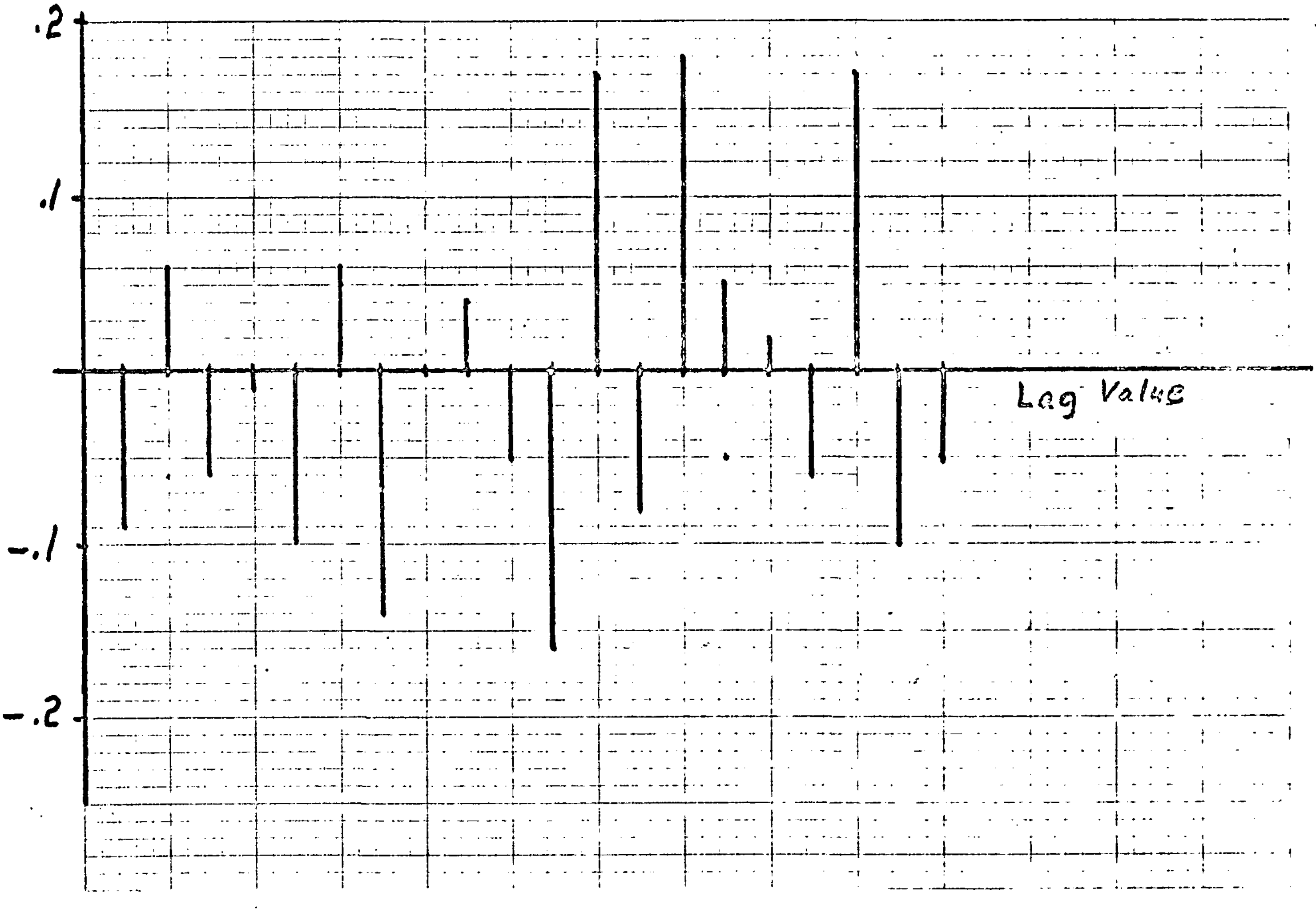


FIGURE 4.4 RAPE

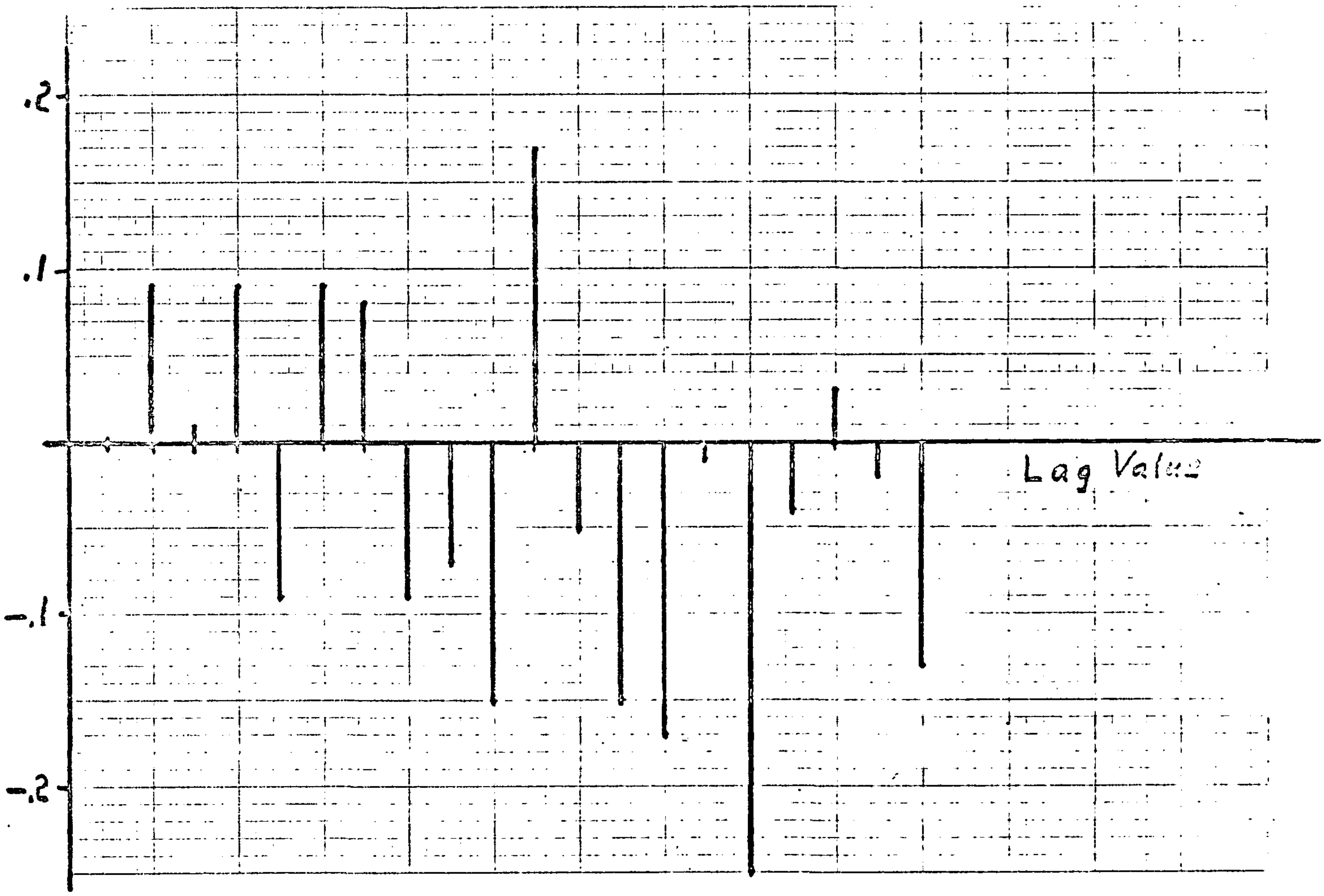


FIGURE 4.5 ROBBERY

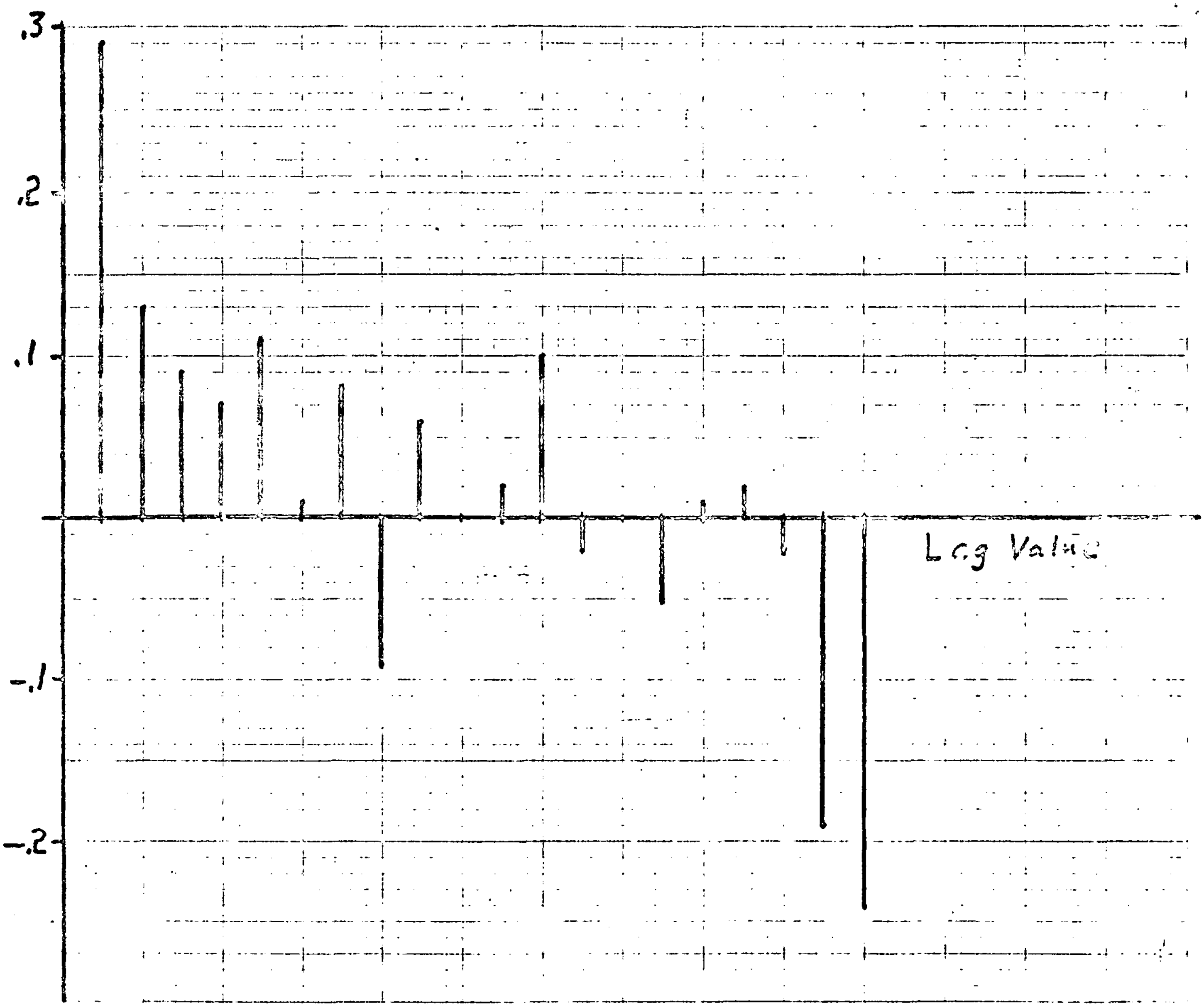


FIGURE 4.6 BURGLARY

TABLE 4[redacted] SIGNIFICA[redacted]T AUTOCOR[redacted]ELATION V[redacted]LUES FOR R[redacted], R[redacted]ERY, [redacted]GGR[redacted]ATE[redacted]SSA[redacted] AN[redacted] BURGLARY

CRIME TYPE	RAPE						ROBBERY						AGGRAVATED ASSAULT						BURGLARY					
REGION	D1		CD1		City		D1		CD1		City		D1		CD1		City		D1		CD1		City	
PERIOD	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post	Pre	Post
$\overline{X}$	1.0	1.1	3.5	3.6	4.5	4.7	9.4	8.6	24.2	18.9	33.6	27.5	4.9	5.3	9.8	13.4	14.7	18.7	33.0	22.5	163.9	175.9	196.9	193.4
S	1.0	1.1	2.0	2.2	2.3	2.6	3.1	3.8	6.6	5.0	7.8	6.7	2.6	2.9	3.5	5.9	4.7	7.5	7.9	5.8	20.6	21.7	23.8	22.2
ACF LAG 1				.12		.21		.21		.17						.34		.36	.29		.31	.33	.28	.32
2			.23		.21	.14		.21								.25		.29	.13	-.12	.29	-.22	.25	-.25
3		.36	.22		.16	.14		.22	.20		.21					.42	.15	.39		.23	.17		.13	
4				-.23									.12	.20		.17	.13	.30			.12		.12	
5			.20		.22					.16		.16			.12	.44		.44			.23		.24	
6				-.13	.17	-.14					.16		.20	.12		.38	.13	.31		.21				
7	-.14			.21	.14	.13			-.14					.17	.15	.12		.19		-.18		-.14		-.12
8													.17	.25		.24		.39						
9			.14	-.14	.24								.16			.22		.25		.36				
10							-.15					-.12		.12		.19		.26						
11	-.16						.17							-.12		.24		.19						-.12
12	.17	.23												.15	-.16	.26		.23		.33		-.13		-.12
13							-.15		.12	.16						.20		.25				-.14		
14	.18		.17		.15		-.17	-.19		.16			-.12			.16	-.12	.19						
15				.17		.18										.15				.15	-.20		-.22	
16							-.25		-.21		-.24									-.18				.12
17				-.16		-.17								.14		.15		.13			-.15			.13
18	.17	-.16	-.13	-.15		-.13			-.16		-.17		-.23			.12	-.23	.14						
19									-.12	-.16									-.19	-.18		-.23		-.23
20			-.12				-.13			-.30		-.22		.14										

* Pre ≡ 87 week period before altered operation
** Post ≡ 78 week period after " "

SYMBOLS

$\overline{X}$ - Sample Mean
S - Sample Standard Deviation
Blank Cell - ACF Smaller than $1/\sqrt{n}$ - .11

By considering the acf values in Table 4.2 and Figures 4.4, 4.5, and 4.6 a number of observations can be made indicating representational and evaluative information. Such observations are illustrated by examples given below by crime types. The interpretation of acf values requires some familiarity with autocorrelational analysis; however, some indication of how such interpretations are made is given in each case below, details may be found in Reference 4.

RAPE

Rape is considered to be essentially random with possibly some seasonal effects. The autocorrelations for rape do indicate a rather random type of time series. They do not indicate long lag seasonal effects. Weekly data are not likely to show such effects. There is some lack of pure randomness at lags that may relate to monthly (three to five weeks) effects. In Dl there seems to be an effect after COMSEC at lags three and 12 that were not present before. No clear differences are indicated in CDl or city.

ROBBERY

Robbery seems largely random in DCl and city for both time periods. In Dl it has gone from largely random to some indication of autocorrelation after COMSEC, indicated by the first three lag values. There are no very strong indications of change or effects for robbery.

AGGRAVATED ASSULT

In this case there are striking changes in the autocorrelation patterns in CDl and city. In each case aggravated assault has gone from an essentially random type time series to a series showing highly autocorrelated characteristics. By contrast, there was no such change in Dl, both the before and after series are essentially random. Because of the difference in Dl and CDl behavior, changes in classification or reporting procedures are not likely to have produced changes of the type shown for aggravated assault. Following are the most likely causes of such changes:

- Instability of the time series due to trends developing which may destroy the so-called stationarity of a time series and produce extensive autocorrelation values (for many lags). Which is the situation in the example data.

- Some pattern of aggravated assault that is unusual, related to changes in the reasons for such crimes. An extreme example of this situation would be the emergence of guerrilla warfare out of what was previously (random) civil crime of the aggravated assault type. This would be likely to also produce a trend effect.

BURGLARY

In CDl and city, burglary seems to be similar before and after COMSEC. It seems to be an autocorrelated process as discussed in Section 6. However, it is more strongly autocorrelated, indicated by more lag values, before COMSEC. The autocorrelated nature of burglary in Dl before COMSEC is diminished, for small lags, after COMSEC. However, some monthly variation (lags 9 and 12) may have come in after COMSEC in D1. Such effects may be due to reporting procedures, non-stationarities in the time series, or actual effects in the crime pattern.

A problem with detailed statistical comparison of time series using autocorrelations (or partial autocorrelations) is the essentially qualitative and subjective nature of the comparison. This is compounded by having to deal with a large number (in the example, 20) of statistical values. Stochastic models provide one method of characterizing a time series by means of a structure type and a few parameters as discussed in Section 5. The structure type is selected on the basis of autocorrelation (and partial autocorrelation) patterns, a subjective procedure. Thus the autocorrelations form an important set of statistical data to be used by themselves for analysis or as part of the more complex procedure of stochastic model building.

Another set of detailed statistical values that may be computed from the data is the partial autocorrelations. These are more complicated than autocorrelations, both to compute and to understand. They do not add much information to direct detailed statistical analysis but may be helpful in formulating stochastic models. Their use in such applications depends on the fact that, like autocorrelations, partial autocorrelations have patterns characteristic of particular forms of time series structure. By employing both autocorrelation and partial autocorrelation patterns, the underlying form of a time series may be indicated.

The partial autocorrelation is an involved quantity to compute. It is numerically equal to the estimated value of the last term in an autoregressive model of order k (as defined in Section 6). Thus, for each k one considers an autoregressive model of order k for the series and computes the last (highest power) coefficient of the autoregressive operator polynomial. For such a calculation, it is necessary to sequentially compute some other terms of the autoregressive model as well. If L is the highest lag value to be considered in forming the partial autocorrelation function (PACF), then $L < K$ and the formulas for computing the PACF value Φ_{mm} are:

$$\Phi_{mm} = \begin{cases} r_1 & \text{if } m = 1 \\ \dfrac{r_m - \sum_{j=1}^{m-1} \Phi_{m-1,j}\, r_{m-1}}{1 - \sum_{j=1}^{m-1} \Phi_{m-1,j}\, r_j} & \text{if } m = 2, 3, \ldots, L \end{cases}$$

where $\Phi_{mj} = \Phi_{m-1,j} - \Phi_{mm}\Phi_{m-1,m-j}$ for $j = 1, 2, \ldots, m-1$

As in the case of autocorrelations derived from actual time series, the PACF will contain a number of values that are not really significant for model structure considerations. It is shown in Reference 4 that one can omit consideration of values lying less than $1/\sqrt{n}$ from the zero value, where n is the number of data points. It is common practice to use the less critical value $2/\sqrt{n}$ in this way. However, in many cases of reported crime it has been found that the stricter condition was able to be met in model formulation.

Figure 4.7 shows a typical partial autocorrelation plot. It gives values to lag 15 for Robbery in a police district based on 87 weeks of data.

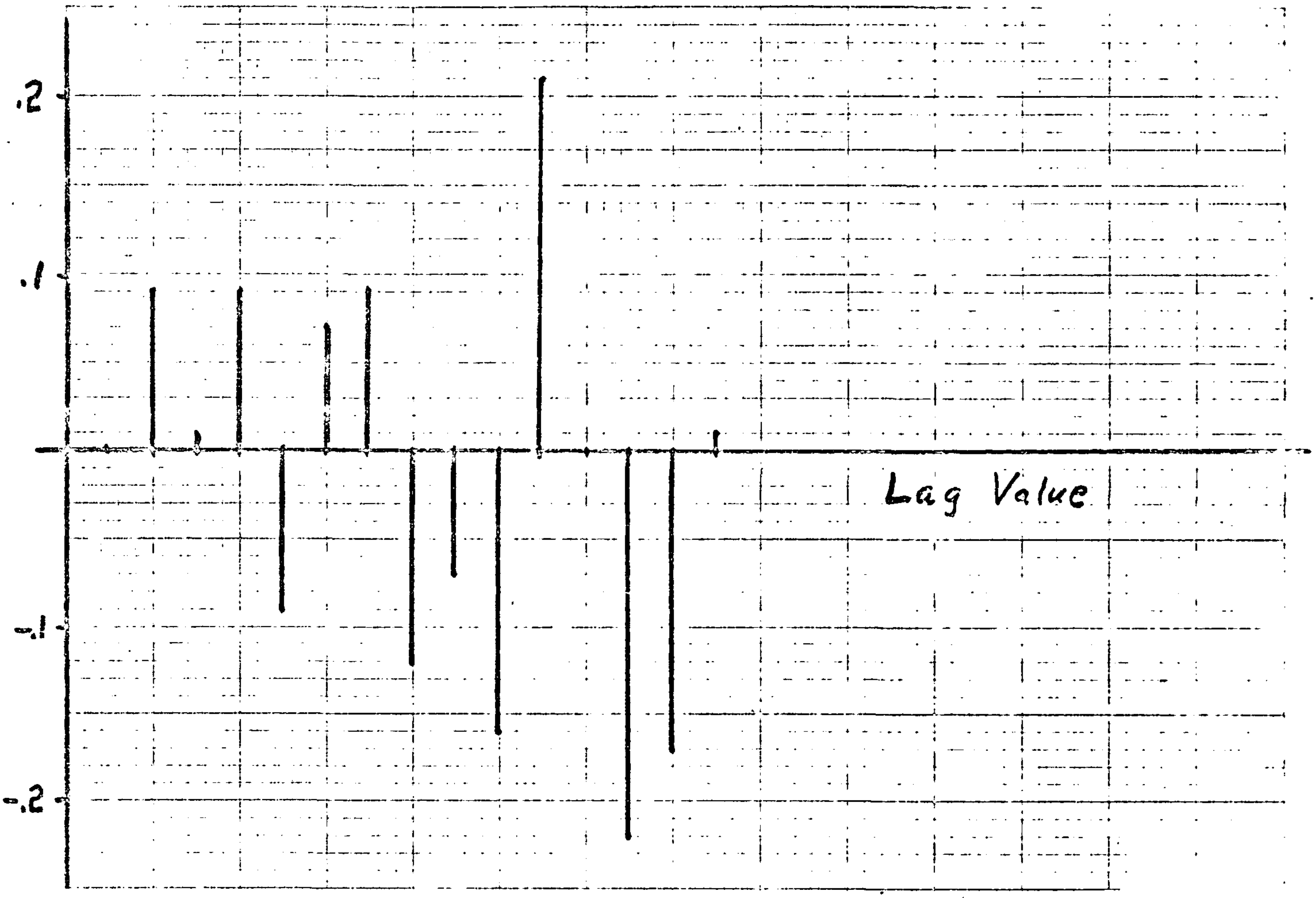


FIGURE 4.7 PARTIAL AUTOCORRELATIONS
ROBBERY

In practice seasonal variations will also disturb the ACF and PACF patterns by producing values greater than $1/\sqrt{n}$ that are not part of the major model pattern indicators. This can be dealt with by incorporating seasonal effects in the model, or by ignoring such features in the basic model and dealing with them by other methods (e.g., components models).

When considering the detailed statistics methodology itself, as distinct from its role in the stochastic modeling process, the autocorrelation values are used to represent the time series and indicate information about it as illustrated by the examples given above. Such information can relate to trend effects, periodic (seasonal) effects, or underlying correlations between data values separated by various lag (weekly) amounts. By studying sets of autocorrelation values for different time series, one can test the series for differences in any of these characteristics. In the context of an evaluation, comparisons of interest may be made between pre- and post- series and between test regions and other regions in which changes under evaluation were not in operation.

Section 5. Basic Ideas of Time Series Models

A mathematical model of a time series consists of a combination of formulas and procedures for describing the time series. In an ideal situation the model would produce, for each time value, the numerical data point in the time series. Because of the non-deterministic nature of statistical time series actual models do not give the same values as a particular series. However if the model is a good representation of the time series it will yield values that are as likely as the actual values within the degree of statistical variation (randomness) present.

This property of the model allows its use in forecasting and also gives a specific mathematical expression characteristic of a time series. For some applications this is more desirable than collections of numerical indices (statistics). Shortcomings of mathematical models lie in the difficulties in their construction, the role of subjective reasoning required in some approaches, and the distinct possibility that good representations may not be possible for certain time series.

When one wishes to develop a mathematical model for a time series, the approach to dealing with the random character or stochastic aspects of the model must be selected. There are essentially three basic approaches to the development of stochastic models for time series. They have features in common, but are held by their major proponents to differ widely in philosophy. They certainly differ in details.

1. COMPONENTS MODEL

The components model is the most direct and is widely used in varying degrees of sophistication. Careful consideration of such models is supported by a number of statistical analysts. Components models play a role in many analyses and should be considered to some extent in any study (e.g., they are basic to the Census Two method for seasonal variation analysis, discussed in Section 9). Models of this type represent a time series in terms of several parts including a trend component, a seasonal variation and a residual random part. In product form the parts are considered to be multiplied together. In additive form the parts are added to represent the time series. This basic model approach can be carried to considerable detail but involves a high degree of subjective reasoning, guided by impressions of the data more than by quantitative analysis of the data (though this is by no means absent). In a sense very simple approaches to time series fall into this general category. When a trend or other deterministic effect is strongly indicated by an initial study of time series plots, such effects should be removed before carrying out other types of model formulation. To this extent, some form of component analysis should play a role in any time series study, often as part of the preliminary analysis as already discussed and illustrated in Section 3.

Components models are developed in terms of three (or more) parts: a trend component denoted by $H(t)$, a periodic (cyclic) component $C(t)$ and a random variation $R(t)$. In the product type of model the series $z(t)$ has the form:

$$z(t) = H(t)\ C(t)\ R(t),$$

and in the additive model it has the form:

$$z(t) = H(t) + C(t) + R(t).$$

It may be observed that one need only consider additive models, since a log transformation of product model data will put those data into additive model form.

The components H, C and R can be developed with various levels of sophistication. A simple direct procedure, that often yields a useful model, is to fit H and C by subjective study of a plot of z(t). The residuals, resulting from the selected H and C, comprise sample points of the process R(t). One can obtain more sophisticated models of this type by fitting some postulated forms of H and C to the data by regression analysis methods (that minimize the square error or residual value). This technique is employed in order to reduce the variance of the R(t) component.

Figure 5.1 shows data for a synthetic time series called SUMP 3. There is a strong degree of randomness indicated but it is not too variable. Most data points lie within one sample standard deviation of the mean line as indicated by the parallel lines above and below the mean. There does seem to be a tendency for particularly low values at intervals of six time units as indicated by the circled data points. This can be thought of as a cyclic component C with period of six. The presence of this term tends to lower the mean so far as the remaining, more nearly random points, are concerned. By removing the extra low points (placing them in the term C) the mean will be shifted upwards and the random part will account very well for the data within one standard deviation. There seems to be no trend component present (series is stationary). To find more detailed model forms the autocorrelations must be considered, these are shown in Figure 5.2 at the end of this Section.

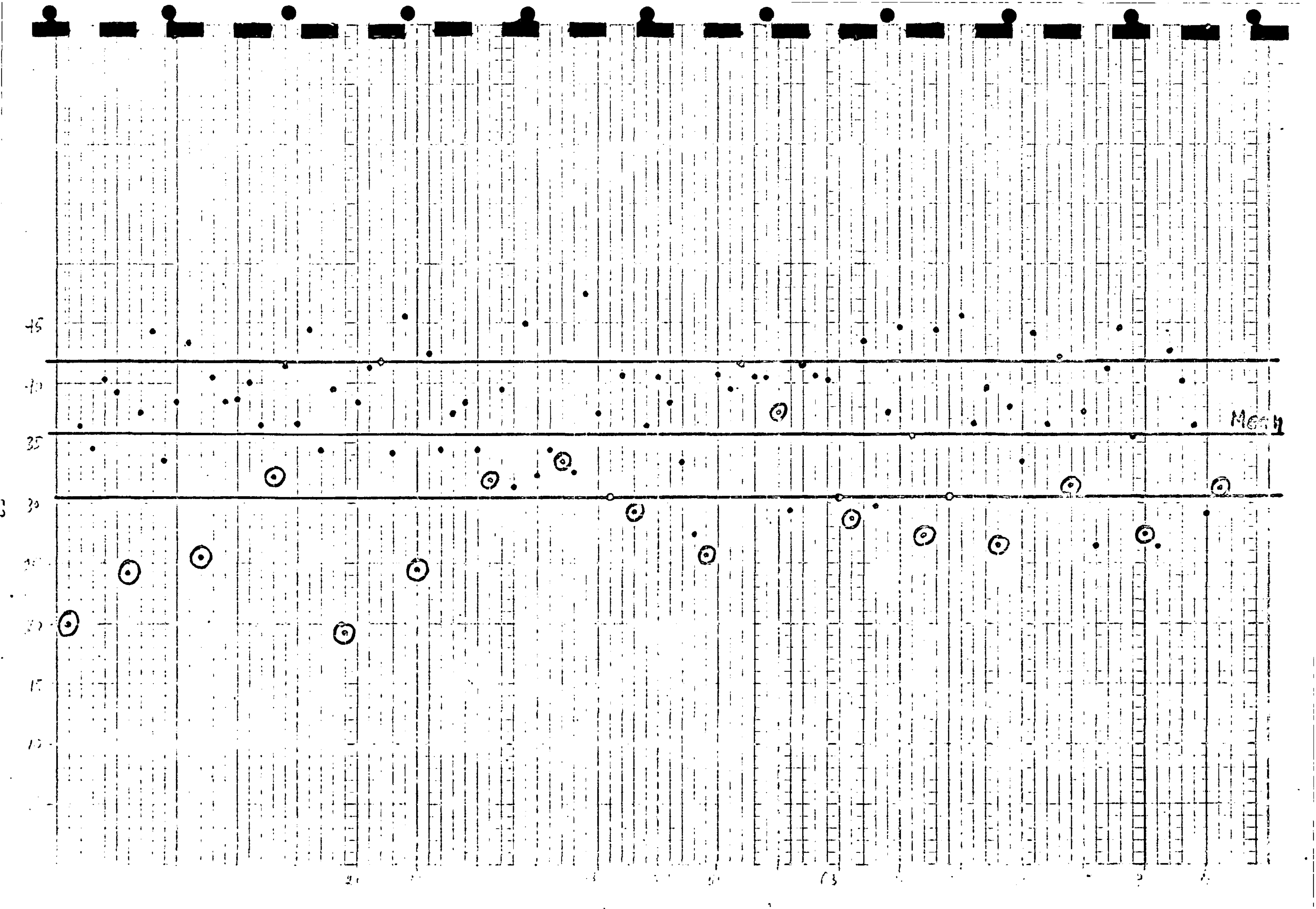


FIGURE 5.1 SUMP 3

2. EXPONENTIAL SMOOTHING

The second major approach to time series is commonly called "exponential smoothing" and is presented in detail in Reference (5). This approach is logically developed by applying methodology of systems engineering to time series. It draws on the mathematical techniques of transform and spectral analysis as they apply to the assumed stochastic structure postulated as a model form. Exponential smoothing assigns weights to probabilisticly occurring past data and uses a combination of such values to forecast time series values. It has had notable success as a forecasting method and is widely used.

Advocates of the method point to its relative directness of approach, utilization of the data to assist in model formulation and decrease in subjectivity over the components approach. On the other hand, detractors from the method indicate that it is very much of an ad hoc methodology. The model having little or nothing to say about the stochastic process being modeled, however good it may be at forecasting. Moreover, for some of its more widely used forms, it can be considered as a subcase of the general stochastic model and "falls out" of that more comprehensive approach in appropriate cases.

Its strength lies in its relative simplicity where it works (i.e., forecasts well). In the use of reported crime data it may be desirable to have good representational models for baseline and subsequent time series so that significant changes in the character of the series, implying possible changes in operational effectiveness, could be detected. This consideration, together with the observation that the general method will often yield an exponential smoothing type model where appropriate, directs consideration toward a third, more comprehensive methodology for time series analysis.

3. ARIMA METHOD

The third approach is called the Autoregressive Integrated Moving Average (ARIMA) method. It is also known as the Box/Jenkins technique after its major proponents, G.E.P. Box and Gwilym Jenkins. (Reference 4). This method begins with a preliminary study of the time series and possibly some number of differenced (or transformed) series. This phase of study considers the possible utility of transforming the series data (or data from some differenced series) to make the resulting data more amenable to stochastic model formulation. Typical transformations are the logarithm (base e) and square root (or other exponential).

After one selects a form for expressing the data, it is desirable to have the series as "stationary" as possible. Stationarity means that the random behavior of the series does not depend on the particular time origin so that the model will apply to the series independently of special aspects of the times at which data are recorded. In practical terms a stationary series does not have trend effects. If the original series is not stationary, the first or second difference series may become more stationary as may higher degree difference series. Thus it is the desire for stationarity that causes one to study differenced series.

The original series may be denoted by (z_t) where t is the time at which data value z_t was recorded (or indicates an aggregate time interval in which z_t falls, such as a week). Then the first difference series is $\nabla z_t = z_t - z_{t-1}$.

Similarly the second difference is $\nabla^2 z_t = \nabla(z_t - z_{t-1}) = z_t - 2z_{t-1} + z_{t-2}$. Though higher degrees of differencing can be used in producing stationarity, reasonable models based on the ARIMA technique should not require (or employ) more than second-order differencing. Use of higher orders implies a complicated time series for which a reliable sophisticated model will be very hard to formulate. Should one succeed in such a formulation, it would be difficult to interpret and in particular to use as a basis for indicating change. Stationarity is desirable for the development of particular stochastic models (e.g., in parameter estimation). The properties of the series and its differenced series can be used to indicate stationarity or its absence.

After preliminary study, the ARIMA technique distinguishes three distinct phases of stochastic model construction: the identification phase, parameter estimation and diagnostic checking.

Identification

In the identification phase, the general form of model is selected. The identification phase is carried out by studying autocorrelation and partial autocorrelation data in either graphical or numerical form. As discussed in Section 4, these statistical values indicate the nature of a time series and point out special features that may be present (such as seasonal effects or correlations between series values). Theoretical studies of autocorrelation and partial autocorrelation values show distinct patterns corresponding to different kinds of ARIMA model forms. Thus by comparison of actual autocorrelation and partial autocorrelation data with standard patterns, some indication of the most appropriate model form can be obtained. Of course, the real data are often complicated and truly typical patterns are seldom found though there is usually enough effect of one form or another to allow a reasonable identification.

The model's structural form (in terms of mathematical expressions) combines autoregressive terms representing previous values of the series and moving average terms representing weighted contributions of purely random past inputs (random shocks) to the series. In most cases the level of sophistication for models will be up to, at most, second order in both the autoregressive and moving average parts. Higher order models seem to result in overly complicated models for most purposes. Some details of ARIMA model form are given below and in Section 6 and are fully exponded in Reference (4).

When the model form has been determined, the series data are used to estimate parameter values specifying a particular stochastic model. Such estimation is based on minimizing the expected sum of squares of the residuals that represent the difference between the actual series and the series resulting from the model. This procedure can result in very good estimates for the parameters, particularly when appropriate statistical hypotheses are satisfied. However, the minimization procedure is rather complicated and requires the development of considerable computer software for its general implementation.

The last major step in the ARIMA technique is called diagnostic checking. In this step the residual values, not accounted for by the mathematical expressions of the model, are generated and studied in various ways to determine how well the model represents the actual time series. The ideal situation is when the residuals form what is known as a white noise process of small variability (the major feature being no autocorrelation) about a zero mean. A white noise process is a time series where autocorrelation values are essentially zero (except for the nominal or zero lag value which is always unity). This means that there is no correlation effect between the time series data values. Technically white noise also implies that each data value is a sample point governed by the same probability form, called the Normal probability. However, the significant feature is the lack of correlation rather than the assumed form of the underlying probability law. Of course, the ideal situation is seldom achieved in practice and one judges how satisfactory the model is from how closely the residual series approximates the white noise characteristics.

The statistical significance of stochastic models depends to some extent on the number of data values available for testing model structure and estimating model parameter values. For simple processes, relatively few points may give satisfactory results. In complicated cases or when seasonal variations make large contributions to the series, many data points are needed for satisfactory model building.

The autoregressive Integrated Moving Average (ARIMA) model stipulates a detailed mathematical expression for the time series $z(t)$. It is supposed that $z_t = z(t)$ depends on previous values of the time series and also on previous (and present) values of a random residual denoted by a_t at time t. The general form of the model is

$$\Phi(B)\ (1-B)^d z_t = \theta(B)a_t + \theta_o \qquad (1)$$

where B is the shift operator, defined by $Bz_t = z_{t-1}$, $\Phi(B)$ and $\theta(B)$ are polynomial operators, $(1-B) = \nabla$ is the differencing operator, defined by $\nabla z_t = z_t - z_{t-1}$, and d is the degree of differencing applied to the data. The polynomial operators are of the form $\Phi(B) = 1 - \Phi_1 B - \Phi_2 B^2 - \ldots - \Phi_p B^p$, and $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q$. In this form seasonal effects are not explicitly considered.

The model structure may account for such effects satisfactorilly: if not, additional formulation is required. The general ARIMA model, as developed by Box and Jenkins (see Reference 4), deals with seasonal models by introducing a seasonal shift operator s into the general form shown in equation 1 above. Seasonal operators are denoted by $H(B) = 1 - h_1 B - \ldots - h_n B^n$ and $R(B) = 1 - r_1 B - \ldots - r_m B^m$. If s is the period of seasonality and D is the degree of seasonal differencing, the general ARIMA form is:

$$\Phi(B)\ (1-B)^d\ H(B^s)\ (1-B^s)^D z_t = \theta(B)\ R(B^s)a_t + \theta_o .$$

Details will only be given for the non-seasonal form in this Section.

Considering the differenced series $w_t = (1-B)^d z_t$, one finds that it has two parts: an autoregressive part $\Phi(B)w_t$ depending only on previous values of w_t and a moving average part $\Phi(B)a_t$ depending only on previous shocks of purely random process a_t. In theory the a_t process is described as white noise; that is, a series of normal random variables with zero mean, all having the same variance, uncorrelated overtime. These a_t random variables have been referred to as random shocks, indicating that they account for random variations in the time series not attributed to (accounted for by) the previous terms in the w_t series itself.

The term $\Phi(B)w_t$ is called the autoregressive part of the series since it expresses the present value z_t in terms of a weighed sum of previous values. (The weights are the coefficients in the polynomial $\Phi(B)$.) A purely autoregressive process has the model from $\Phi(B)w_t = a_t$.

The term $\theta(B)a_t$ is called the moving average part of the series. It expresses a sum of weighed contributions due to past random shocks. (The weights are the coefficients in the polynomial $\theta(B)$.) A pure moving average process has model form $w_t = \theta(B)a_t$. Box and Jenkins call this an integrated moving average since it includes a contribution of many individual moving averages.

Some appreciation of what the moving average model means can be gained by considering the simple case

$$w_t = (1 - rB)a_t$$

which may be written as

$$(1 - rB)^{-1} w_t = a_t$$

or

$$w_t = a_t - rw_{t-1} - r^2 w_{t-2} - r^3 w_{t-3} - \dots ,$$

This formulation expresses the present value w_t as the present random shock minus contributions from past values of the w_t series. When $0 < r < 1$, the contributions from the past are reduced by weights that vary exponentially. Hence this type of model is often called exponential smoothing (see Brown in Reference 5). Some additional aspects of the modeling process including considerations of stationarity and transformation of data are discussed in Section 6.

Parameter Estimation

Once a model form has been selected from the considerations of the identification phase, the parameters of the model must be estimated from the time-series data. These parameters are the coefficients in the two operators $\Phi(B)$ and $\theta(B)$ and when seasonal variation is considered one also includes the parameters of the seasonal operators.

Diagnostic Checking

The third phase of the ARIMA technique is called diagnostic checking. In this phase the mathematical model form is used to compute a residual stochastic process from the numerical time-series values. The residual process is theoretically the same as the series of the past shocks $\{a_t\}$. If the model is an exact representation of the series, the residual series will be a white noise process. That is, an uncorrelated normal process with zero mean value function and constant variance. The diagnostic check considers how closely the actual residuals are to forming a white noise process. The autocorrelation function of $\{a_t\}$ is computed. Small values indicate that the residual series is uncorrelated. If, in addition, the mean is close to zero and the standard deviation is of reasonable size, the model is accepted as a satisfactory representational model. It should be noted that the size of the residual standard deviation has no connection with the accuracy of the model. It only relates to potential use of the model as discussed elsewhere in this report.

An additional feature of the diagnostic checking phase is calculation of forecasted values for the time series under study. Forecasted values can be compared with series values that occur subsequently. In addition, the general character of the predicted values can be considered. Reasonable results from such considerations will indicate operational utility of the model as a forecasting device. They also contribute to a general feeling of reliability of the model as a representation of the actual time series. However, it may be that a good representational model will not yield particularly good forecasts In such a case the model may still prove useful for evaluative study.

In all cases it should be observed that the forecasts are expected to become less reliable as they move further forward in time from the point of evaluation. In making the forecasts, past values of the series and the random shocks are used in the model form. When new values are required, computed values (forecasts) are used for the time-series values and zero is used for the unknown future shocks (as in Reference 4, one can show that this gives the best estimate for future shocks in the sense of minimum mean square error estimation). Forecasting is discussed and illustrated in Section 8.

Seasonal variation or other cyclic type trends can be important in time series models. These are discussed in Sections 6, 7, and 9.

The SUMP 3 example data is shown in Figure 5.1 and given a preliminary type of analysis in the beginning of this Section. Figure 5.2 shows the autocorrelation plot for SUMP 3 data. It clearly indicates the strong deterministic correlation separated at intervals of six time units (note strong autocorrelation at lag of 6, 12, and 18). The correlation in opposite sense is indicated by the values separated by 6 time intervals, starting at lag 3. This synthetic model was constructed as an ARIMA type (1, 1) with two deterministic terms added so that its form was

$$(1 - \varphi_1 B)z_t = (1 - \theta_1)a_t + \theta_0 + \text{Deterministic Terms}$$

The values used were $\varphi_1 = .25$, $\theta_1 = 1.5$, $\theta_0 = 28$. The deterministic terms

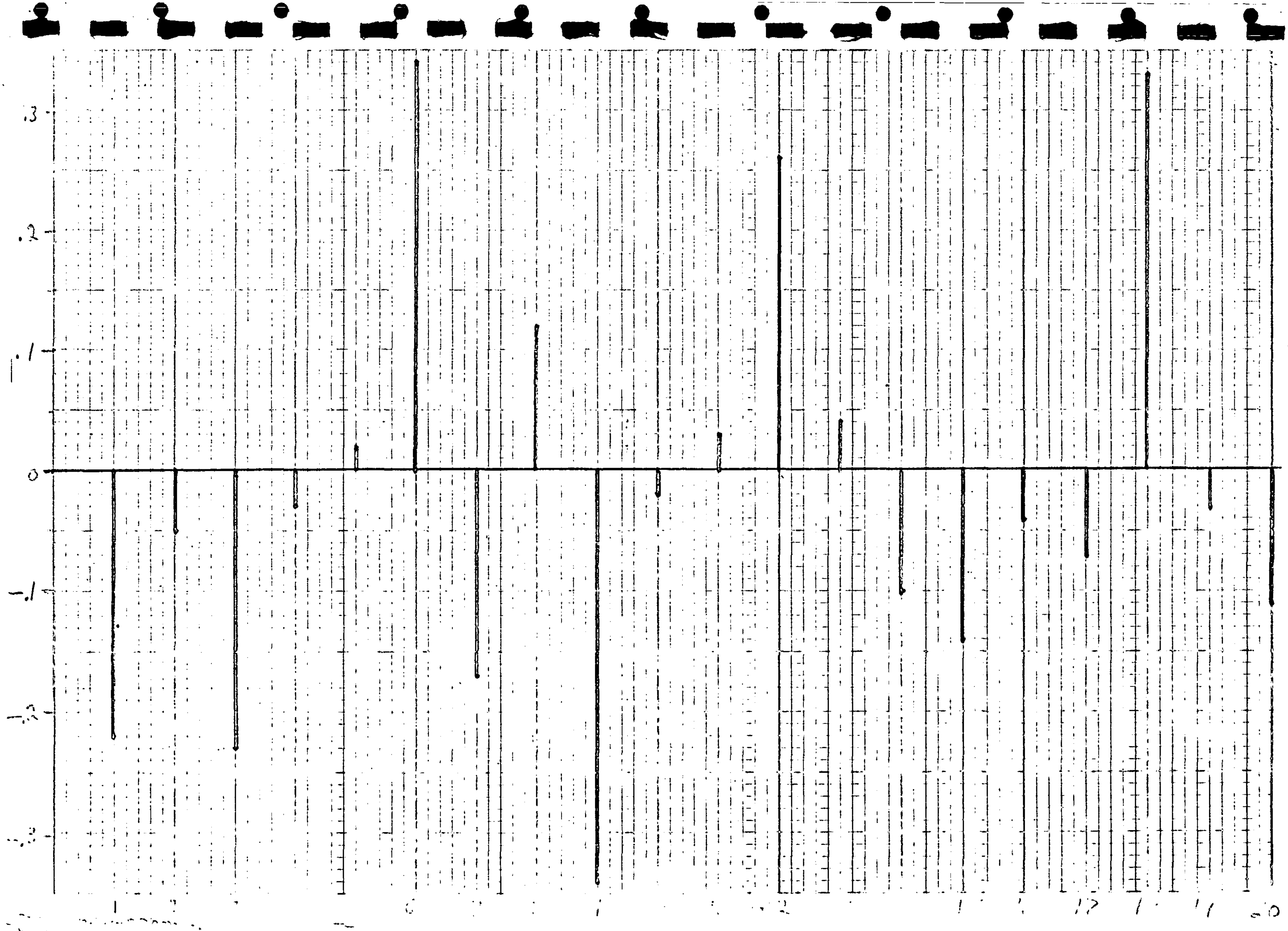


FIGURE 5.2 SUMP 3
ACF

Section 6. Methodology of ARIMA Models

The methodology employed for the study of time series data by formulating ARIMA type models consists of several distinct tasks and requires computer software for carrying them out. These tasks, or modeling steps, include:

--Plot time series and plots of transformed or differenced series as required.

--Compute sample mean and standard deviation for a series of data values.

--Compute sample autocorrelations and partial autocorrelations to specified number of lag values.

--Make a set of estimates for the parameters of a selected mathematical model of the time series. Estimates are most often based on a minimum mean square estimation program as discussed below.

--Calculate residual time series not accounted for by the model.

--Test the model to see how well it represents the time series. This involves testing the residual series for white noise characteristics, observing smoothing ability of the model and making short-range forecasts.

Much of the above process is commonly known as the ARIMA technique as previously discussed. Various alternatives are possible for carrying out the ARIMA model formulation of a time series such as with reported crime data. In particular, the usual ARIMA procedure employs an identification and decision step in which the autocorrelations and partial autocorrelations for various lag values are studied. From these, one attempts to gain characteristic patterns that suggest an appropriate model form. Then the form is selected and model parameters are estimated.

However, the overall point of view is to get a good model, which is defined as one having certain residual characteristics. How such a model is obtained is not important. Thus, in cases where a large number of time series are to be considered, a more direct approach is desirable. To proceed in such cases it may be possible to assume that relatively simple models will be satisfactory. Several simple model types can then be chosen for study and an attempt made to produce each of these models for every series. This involves estimating the desired parameters.

In some cases useful models may not be obtained because the data do not lead to useful estimates of the parameters. This situation enforces the viewpoint that only simple models could be obtained on the basis of the assumption of elementary series models. By attempting to produce candidate models for each series, a large number of models can be constructed. To carry out the more usual ARIMA procedure would take a great amount of time, and computer use. Once the candidates are produced, a study of each allows selection of the best model for each time series. Best is determined from properties of the residual. (Some commercial procedures are becoming available that seem to follow similar pragmatic methods to obtain useful ARIMA models.)

For careful development of a time series model an ideal procedure is to conduct preliminary analysis of the kind discussed in Section 3 together with the basic identification phase of ARIMA modeling. The autocorrelation can suggest things like lack of stationarity and seasonal variations which can also be considered by means of the plots and simple statistical studies employed in preliminary analysis. Transformation of data or differencing of the series prior to detailed model formulation are often suggested on the basis of preliminary analysis or part of the ARIMA identification phase. These preliminary considerations are discussed in Section 3.

The procedure described above will be illustrated in detail for two cases: one showing a good model and one showing a poor model. Properties determining a good model will be indicated as part of the illustration rather than being given in more abstract formulation. They are, to some extent, subjective, but do provide operational guidelines for selecting mathematical models of time series.

Before presenting the illustrations, the model form will be discussed somewhat further to provide a background for discussion of the illustrations and also to allow definition of a possible new level of crime measure based on the stochastic model concept. The measure may be called the residual crime level and may prove to be more meaningful than the sample means as a scalar (i.e., single numerical) indicator of crime level. Detailed study of the residual crime level is beyond the scope of work described in this report.

The time series may be denoted by a collection of values z_t which equals the value of the series at time t. Thus, in the study of reported crime if the time series for burglary in District 1 is being considered, z_5 would denote the number of burglaries reported in District 1 during week number 5 (which would correspond to some calendar week in the time series). The nonseasonal mathematical models express the value z_t in tems of previous series values such as z_{t-1}, purely random amounts a_t (called random shocks) that cannot be accounted for by the previous series values, and a nonrandom constant term. If the model uses p previous series values and q previous shocks, then the value of z_t is expressed in the form:

$$z_t = \sum_{i=1}^{p} \Phi_i z_{t-1} + a_t - \sum_{i=1}^{q} \theta_i a_{t-i} + \theta_o$$

To specify such a model the p+q+1 parameters Φ_i, θ_o, θ_i must be estimated from the data. Selection of the model type is accomplished by choosing the value of p and q to be used. The notation (p, q) is widely used to specify the model type. Basic candidate model types that may be used are (2, 2), (1, 1), (1. 0), and (0, 1). (In some cases other models and seasonal forms may be required.) It should be noted that the form (0, 1) is best when the series itself is purely random (as is often the case for such crimes as homicide or rape). This model type includes a form of the so-called exponential smoothing model.

Estimation of the parameters to be used in the ARIMA model is a major part of the actual calculation required. One way to estimate the parameters is based on the method of moments widely used to develop estimates in statistics. These are called Yule-Walker estimates. They insure that the model will yield some number of moment values equal to the corresponding sample moments obtained from the data. The required calculations are reasonable but the estimates are not considered as good as those obtained by minimizing the mean square error between model values and actual values. However this minimization is a major problem of numerical analysis. Sometimes the Yule-Walker values are used as initial points from which to start the optimization process. If this is considered too extensive a calculation the optimization is started from selected initial values. Since the difference function being minimized depends on all N values in a series of length N a complicated function of the estimators is involved. This will be discussed a little more in Section 10.

If one takes expected values in the model form shown above, the following result is obtained:

$$E(z_t) = \sum_{i=1}^{p} \Phi_i E[z_{t-i}] + \theta_o$$

This occurs because the a_t random variables are assumed to have zero expected value (mean). Denoting $E(z_t)$ by the series mean μ one obtains:

$$\theta_o = (1 - \sum_{i=1}^{p} \Phi_i)\mu.$$

The quantity θ_o is what has been designated above as the residual crime level. When the amount of crime at time t does not depend directly on previous values, the parameters Φ_i are all zero. In this case the time series of reported crime is a pure moving average form of model, and θ_o is the average value μ, estimated as the sample mean. When a good model can be produced in which no Φ_i terms are present, the residual crime level θ_o is the same as the mean crime level μ. However, for those cases in which the better model employs some previous time series values (Φ_i not all zero), the residual crime level differs from the mean crime level. This occurs because in such cases more of the incidence of crime is accounted for by the model terms corresponding to previous series values. There is a residual crime level unaccounted for by either the influence of previous levels of crime or any purely random (shock) terms. The level may be less or greater than the sample mean, depending on the numerical values of the parameters Φ_i (which may be positive or negative).

One may denote the three parts of the model as the

- correlated part $C = \sum_{i=1}^{p} \Phi_i z_{t-i}$

- residual crime level part θ_o

- purely random (white noise) part $R = a_t - \sum_{i=1}^{q} \theta_i a_{t-i}$

A model is considered to be good if R has the properties of white noise, as illustrated in the sample cases discussed below. The effect of each of the three parts does not relate to model correctness (i.e., goodness or suitability). However, they may well relate to model utility. In some cases a good model may be a useful operational tool while in other cases an equally good model (judged on the basis of the white noise characteristic of R) may be less useful. Model utility depends on the relative strength of each of the three parts. Since the application of time series model analysis to offense crime data is only beginning, there presently does not exist any body of information as to how to utilize such models. For applying such models in evaluation work there is interest in considering what the three parts of the model may contribute regarding measurement of change in crime level or pattern. For this purpose one can conjecture the following aspects of the three parts, with a caution that there is no evidence at this time that these conjectures are correct:

- There is relatively little that one can do to alter the purely random part R; contributions to crime governed by this process show the least response to changing police activity.

- The residual crime level θ_o depends most strongly on the crime environment rather than on changes in police activity; if it is to be changed, rather general measures must be taken to alter the total environment.

- The correlated part C is most likely to change as a result of altered police activity, changes being indicated in parameter values and/or p, the order of the process.

These are preliminary considerations, stated here to indicate some of the ways in which time series models of reported crime might be expected to contribute to evaluation of change due to altered police activity. Several difficulties can be identified with these concepts. The residual crime level θ_o depends on the same parameters as C does so that both should reflect the same kind of changes; however, it is reasonable to suppose that θ_o may be less sensitive than C to changes in the parameter values Φ_i. Another difficulty is the considerable variation in θ_o values over candidate models (and on techniques used for parameter estimation). This is a weakness of the residual crime level, in that it can have rather different values for models that are almost equally good so far as their white noise characteristics are concerned.

Time series may have some cyclic or seasonal variation that contributes to the magnitude of the series values. Retail sales have peaks before Christman and electricity use is high in August due to air conditioning demand. Reported crime does not show strong seasonal effects such as those that stand out for retail sales. However it is possible that some type of cyclic variation is present for certain crime types. This is discussed more fully in Section 9.

The ARIMA methodology can produce a stochastic model that includes cyclic variation (seasonal trends) by using higher order difference terms. These may be either autoregressive or moving average type terms. In the examples of autocorrelations the presence of cyclic effects was described. When study indicates the presence of effects and identifies the appropriate lag value the ARIMA methodology includes the difference factors in constructing the model. Of course the introduction of such cyclic factors complicates the model, requiring additional parameter estimation. Thus longer series are required to provide the necessary data for estimation.

It is important to balance the potential improvement in a model that might be provided by cyclic factors against the added complexity and possible downgrading of parameter estimation. When there are strong indications of cyclic trend they must be accounted for but with most crime data such indications are slight and it is very much a matter of judgment how they should be dealt with.

The following two model formulations illustrate how models are selected and tested for goodness. Both a good and a bad model will be described in practice the bad model would not be used. Additional models selected from candidate models of various crime types are given in Section 7. Details such as those given below are not presented for the other illustration cases.

An example of a good model is the (1, 1) model for burglary in District 1, for the 87-week pre-COMSEC period. The (1, 1) model has the following mathematical form:

$$z_t = \Phi_1 z_{t-1} + \theta_o + a_t - \theta_1 a_{t-1}.$$

Data for this case give estimates for the three parameters $\Phi_1 = .4$, $\theta_o = 19.57$, $\theta_1 = .1$ and the sample mean and standard deviation are $\mu = 32.76$ and $\sigma = 7.97$. Note that the residual burglary level θ_o is considerably less than the sample mean. Thus the model form is accounting for a number of instances of burglary that do not have to be lumped into the residual crime level.

The major purpose of this, and the next, example is to clarify how a model is judged to be good or not good, thereby illustrating the diagnostic phase of ARIMA. The most important features of the model are the white noise properties of the random residual. This is the part of the time series that is not attributed to correlated or fixed levels, it is the estimated a_t series. White noise should have zero mean value and zero autocorrelation values. Of course, one only obtains estimates of all these quantities, based on the data and the model structure. These estimates must be used to test the white noise property of the residual. An autocorrelation estimate is assumed to be not significantly different from zero if it lies within one standard deviation, of its distribution, from zero (positive or negative). The sample distributions are rather complicated, but for far-size samples the standard deviation is taken to be about $1/\sqrt{n}$ where n is the length of the time series (in the present example n=87 and $1/\sqrt{n} = .11$).

For the (1, 1) model of burglary in District 1, the residual mean is -.05 which is close enough to zero to be acceptable (though smaller values are commonly obtained). The properties of the residual autocorrelation are shown in Figure 6.1 where the plus and minus $1/\sqrt{n}$ values indicate the significant region. Only autocorrelation estimates falling outside these levels are of any significance at all. In this case only five out of a total of 20 values fall into the significance region and none fall very far into that region. This is a very acceptable sample white noise autocorrelation spectrum. In Figure 6.1, "lag" refers to how much separation is taken between time series values in computing the corresponding autocorrelation estimate. No autocorrelation value falls beyond two standard deviations ($2\sqrt{n}$) from zero which would be a highly significant occurrence as illustrated in the next example.

Though the white noise property is essential for a good model, there are some other, more subjective criteria that one may also consider. One of these is seeing how well the model forecasts for a few weeks ahead. A bad model is often indicated by strange (e.g., negative) forecast values. On the other hand a good model gives values near the sample mean and usually becomes constant in a very few weeks (two or three). Though possibly not a useful forecast (because of the relatively simple form of models being used here), such reasonable behavior indicates a sound model representation.

Another criterion is the contrast between the original time series and that series with the random residual (white noise in a good model) subtracted out. In a good model the result is a much smoother series than the original. It represents the part of the time series that is not purely random, combining a constant level with a correlated (one value affecting future values) random series. As has been remarked above, this remaining part may well represent the levels of crime for which various activities can produce change. The purely random part cannot be affected in the same way (e.g., one cannot forecast that part at all on the basis of other data). Thus smoothing seems like a good model feature; however, too smooth a result would possibly indicate an inability to produce changes, at least by certain methods (of course, a "police state" can strongly affect even the purely random part).

Figure 6.2 shows the series for burglary in District 1 and the smoothed series after removal of the random residual, which has been seen to be close to white noise for the (1, 1) model. Considerable smoothing occurs indicating a good model. However, variation still exists indicating that there is more to burglary than purely random effects and that actions can be taken to use information contained in the burglary time series in efforts to reduce burglary. Thus, not only is the model a good one in theoretical terms, but it is potentially useful as an operational aid.

A final consideration is the standard deviation of the random residual 7.55 in the present example. The standard deviation has nothing to do with goodness of the model, but relates to its potential usefulness. Since the random residual, once identified by a good model form, is an instrinsic part of

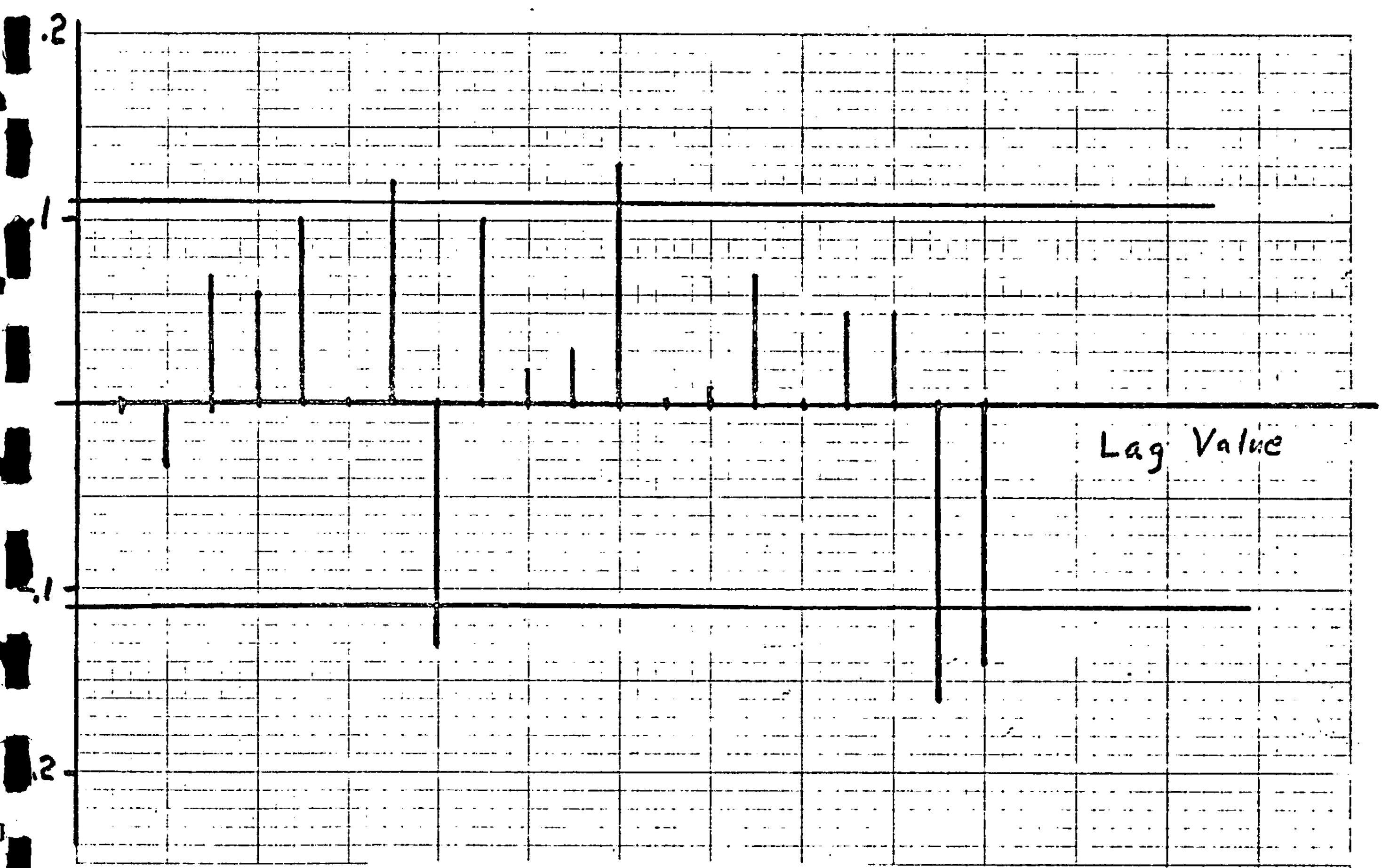


FIGURE 6.1 BURGLARY DISTRICT 1 MODEL (1, 1)
RESIDUAL AUTOCORRELATION

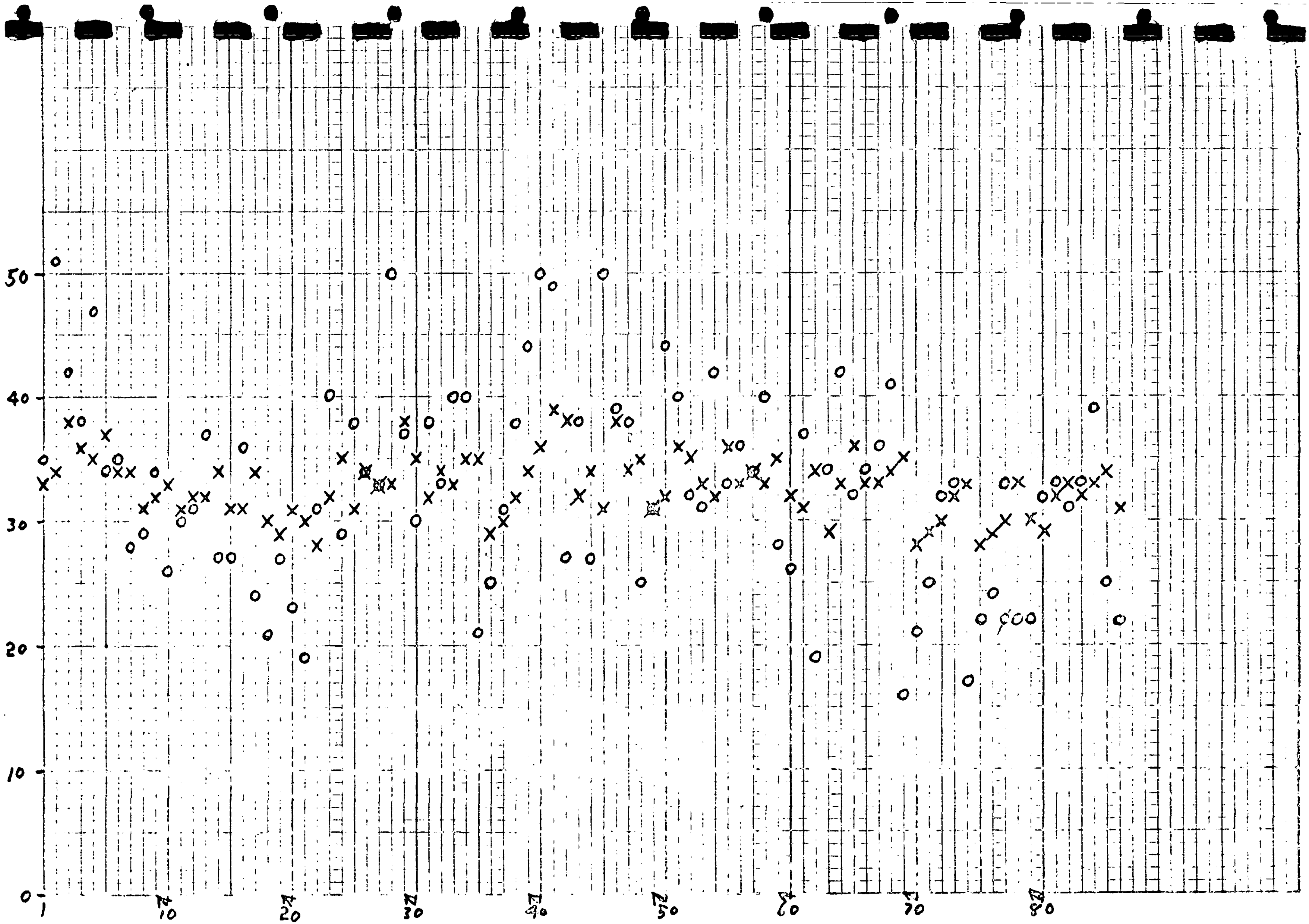


FIGURE 6.2 BURGLARY DISTRICT 1

the complete time series, one has to deal with it. If the purely random residual has large variability, then there is a strong uncorrelated effect that may be difficult to reduce. At least it seems likely that different procedures may be required in attacking the purely random part from those that are shown to be useful in attacking the correlated and constant parts. Certainly our interest here in identifying the different parts is to have an evaluation tool that can address those aspects of change that may be expected to be affected by a program without having to demand that a program also produce changes in processes over which it could not be expected to have the same effect.

In contrast to the good and potentially useful model (1, 1) for burglary discussed above, it may be noted that rape is essentially a purely random series. A good model for rape is of the form (0, 1) with $\theta_1 = 0$ so that $z_t = \theta_o + a_t$. The residual is essentially the series itself shifted from a mean of θ_o to a mean of zero. Removal of the a_t series gives perfect smoothing (constant θ_o value) and the white noise spectrum is ideal (all autocorrelations are zero). However, this very good model seems to have little potential utility because it only bounds the occurrences of rape by sample standard deviation values.

Another comparison with the burglary model is provided by a bad model for petty larceny in D1 for the 87-week pre-COMSEC period (in this period the definition of petty larceny was based on $50 value). It is a (0, 1) model with $\theta_o = 38.98$ and $\theta_1 = -.59$. In this model the residual crime level θ_o equals the sample mean μ. Sample standard deviation for the larceny series is 10.24.

The random residual has mean .01 which is satisfactory but has a very poor autocorrelation spectrum as shown in Figure 6.3. Two values are more than two standard deviation limits. This is completely unsatisfactory autocorrelation and the residual cannot be considered as representing a white noise process. Thus the model is not good (unacceptable). For this case one can consider the series remaining after removal of the residual as shown in Figure 6.4. There is nothing like the smoothing effect seen before in Figure 6.2; in fact, there is very little smoothing.

When the random residual has a poor autocorrelation spectrum, the significant peaks may be due to poor model selection (or parameter estimation). However, they may also be due to periodic trends in the data. When this is the case, one can often identify similar (i.e., same lag values) peaks in the original series autocorrelation. This is one method for identification of periodic (seasonal) effects.

Other difficulties in the model may result from attempting to model a nonstationary series or a series that should be transformed in order to yield better model representation.

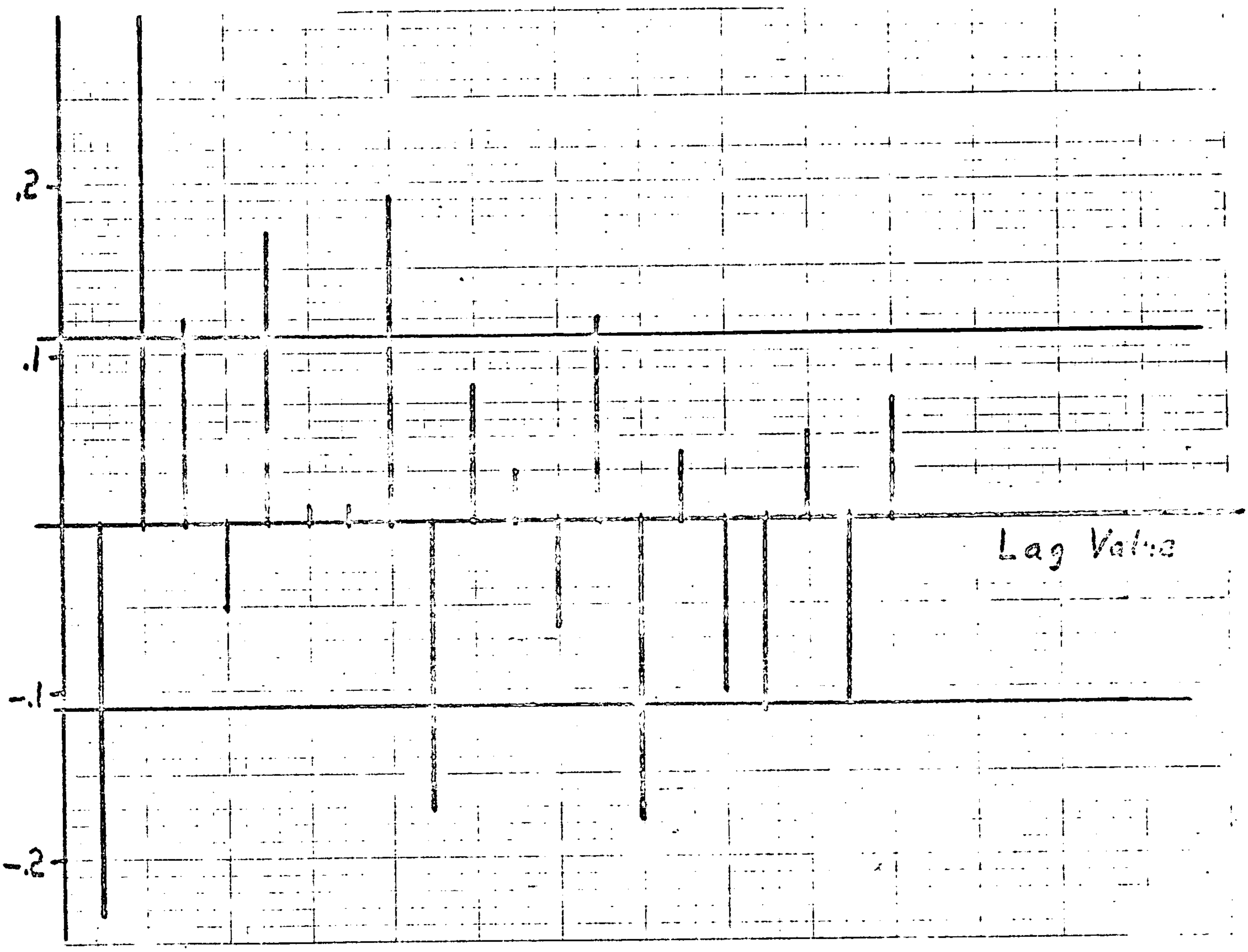


FIGURE 6.3 LARCENY LESS THAN $50 DISTRICT 1
MODEL (0, 1) RESIDUAL AUTOCORRELATION

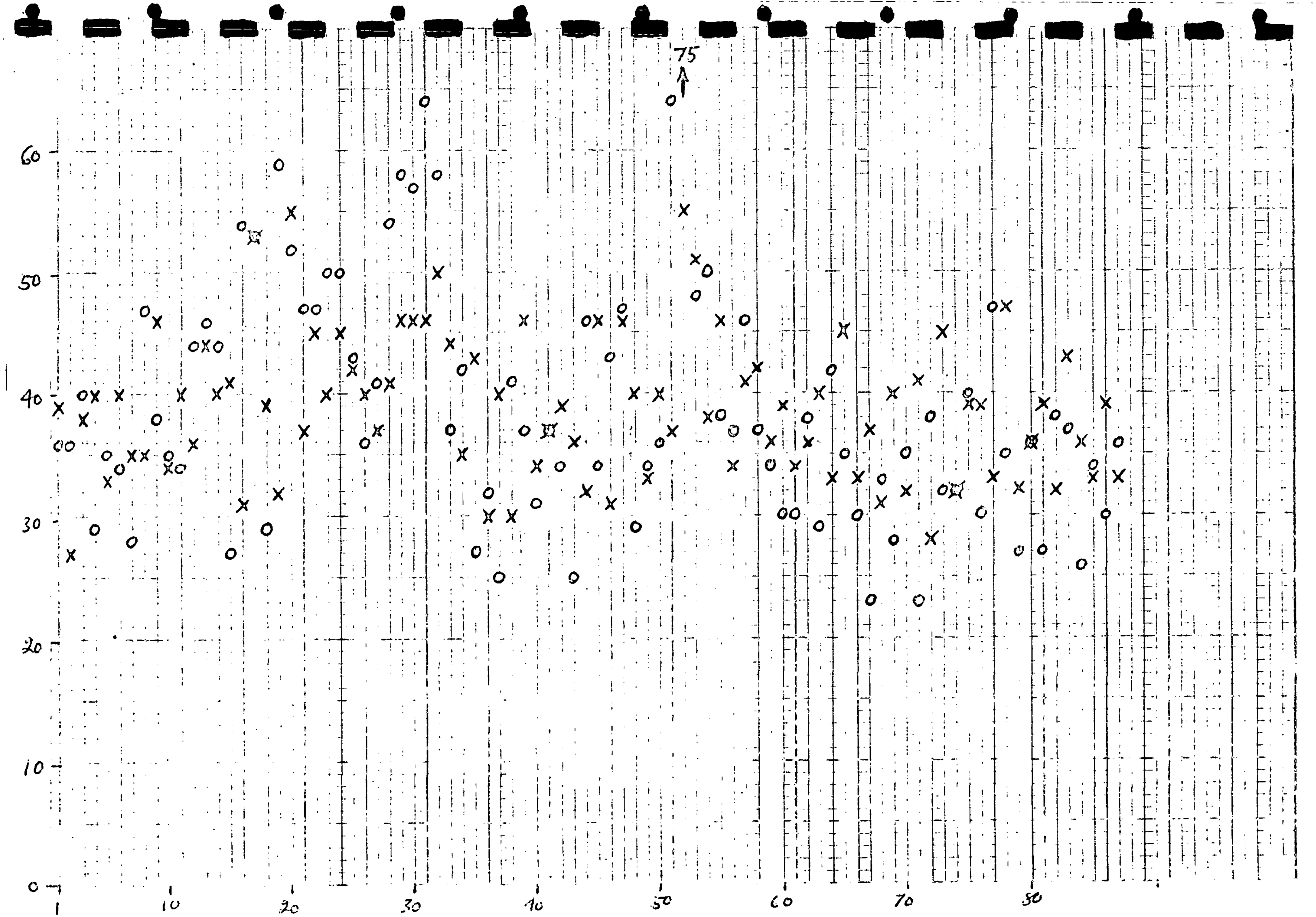


FIGURE 6.4 LARCENY LESS THAN $50 DISTRICT 1

Section 7. Illustrations of ARIMA Models

This section illustrates the mathematical model approach to statistical analysis of time series comparisons, for example as employed in the study of reported crime for evaluation purposes. An ARIMA model may be developed for each crime type and region of interest both before and after a program initiation for which evaluation is desired. Due to the relative effort in producing good stochastic models, examples may be restricted. The illustrations are for representative crime types in District 1 (D1) and the rest of the city (CD1). The models given were produced without extensive preliminary model formulation studies. Thus, the examples do not include such things as data transformation, differencing for improving stationarity, or the detailed study of autocorrelation and partial autocorrelation statistics. These are illustrated elsewhere in the report. Instead of the detailed preliminary procedures, a number of candidate models having simple form were made and their residual properties noted. If a good model could not be found, a more careful study was made and some additional candidates were formed for that case. From this collection of models, it was possible to obtain a good model in most cases under consideration for this illustration. The selected model for each case is shown in Tables 7.1 through 7.5 -- one table for each of the five crime types considered. The notation employed in these tables is explained below.

Model type describes the level of autoregressive terms p and moving average terms q by an ordered pair (p, q). Seasonal variation is indicated as autoregressive lag or moving average lag values, written after the ordered pair. Parameter values are given below the type description. The order is indicated by a subscript, autoregressive values are denoted by Φ and moving average values are denoted by θ. The mean value estimated by the ARIMA software is denoted by μ. Each such value is followed by the sample mean $\bar{x}$ in parentheses which is used as an indication of "proper convergence" as discussed below. The quantity called "trend" in the tables is a constant parameter value which is included as part of the operational software. It is not the θ_o value included in the model forms given previously. The relation between "trend" and θ_o is discussed below.

Description of the residual gives indications of how good the model is. In an ideal model the residual is uncorrelated and has a zero mean value. The standard deviation does not relate to goodness of the model, but rather to its utility as a descriptor of the random process. Large standard deviations in a good residual (uncorrelated, zero mean) indicate considerable purely random effects difficult to "control" or predict. The tables show residual mean and standard deviation values. They also show autocorrelation values of the residual which are greater than $1/\sqrt{n}$, taken to be a measure similar to one standard deviation of purely random variation (in autocorrelation values). In practice it is common to use $2/\sqrt{n}$ as the measure of significant residual autocorrelation, values exceeding this measure are indicated by an asterisk (*) on the tables. There are a number of factors that contribute to the residual autocorrelation values, keeping them from zero, other than actual autoregressive effects. It is for this reason that only rather strong (large) values should be considered significant in evaluating the "goodness" of a particular model.

In many cases the model selected as best from among the candidate

TABLE 7.1 BEST MODELS FOR REPORTED RAPE BY REGION

REGION	D1		CD1	
PERIOD	Pre-COMSEC	Post-COMSEC	Pre-COMSEC	Post-COMSEC
BEST MODEL TYPE PARAMETERS	(0, 1) μ = .9 (1.0) trend = .11 θ_1 = .08	(0, 1) μ = 1.0 (1.1) trend = .11 θ_1 = -.07	(1, 1) Φ_1 = .82 μ = 3.0 (3.5) trend = .11 θ_1 = .72	(0, 1) μ = 3.5 (3.6) trend - .13 θ_1 = -.11
RESIDUAL MEAN	.002	.001	-.002	.001
STANDARD DEVIATION	.97	1.1	2.0	2.1
AUTOCORRELATION: SIGNIFICANT VALUES WITH LAGS	-.15 (11) +.15 (12) .18 (14) .16 (18)	.36 (3)* .23 (12)* -.16 (18)	-.14 (1) .14 (3) -.17 (4) .14 (5) -.16 (10) .18 (14)	-.23 (4)* -.15 (6) .22 (7)* .16 (15) -.16 (17)
COMMENTS	Good Model	Marginal Model the .36 Value at Lag 3 is high.	Good Model	Satisfactory Model some autocorrelations are high.

* Indicates a significant residual autocorrelation $\geq 2/\sqrt{n}$

TABLE 7.2 BEST MODELS FOR REPORTED ROBBERY BY REGION

REGION	D1		CD1	
PERIOD	Pre-COMSEC	Post-COMSEC	Pre-COMSEC	Post-COMSEC
	(1, 0) autoregressive lag 4 ϕ_1 = .012 $\phi(4)$ = .10 μ = 9.4 (9.4) trend = .10	(1, 0) Φ_1 = .21 μ = 8.4 (8.6) trend = .11	(1, 1) Φ_1 = .80 μ = 20.4 (24.2) trend = .11 θ_1 = .75	(1, 0) Φ_1 = .18 μ = 19.2 (18.9) trend = -.03
RESIDUAL MEAN	.0009	.0000	2.6	.0002
STANDARD DEVIATION	3.2	3.6	6.8	5.1
AUTOCORRELATION: SIGNIFICANT VALUES WITH LAGS	-.16 (10) .18 (11) -.18 (13) -.22 (16)*	.18 (3) -.19 (14)	.17 (3) -.18 (7) -.19 (16)	-.15 (19) -.29 (20)*
COMMENTS	Satisfactory Model	Good Model	Good Model	Satisfactory Model however the lag 20 value is high.

* Indicates a significant residual autocorrelation $\geq 2/\sqrt{n}$

TABLE 7.3 BEST MODELS FOR REPORTED AGGRAVATED ASSAULT BY REGION

REGION	D1		CD1	
PERIOD	Pre-COMSEC	Post-COMSEC	Pre-COMSEC	Post-COMSEC
	(1, 1) $\Phi_1 = .71$ $\mu = 4.4$ (4.9) trend = .13 $\theta_1 = .65$	(1, 1) $\Phi_1 = .24$ $\mu = 5.1$ (5.3) trend = .10 $\theta_1 = -.06$	(1, 1) $\Phi_1 = .22$ $\mu = 9.6$ (9.8) trend = .12 $\theta_1 = .25$	(1, 1) $\Phi_1 = .82$ $\mu = 13.3$ (13.4) trend = .01 $\theta_1 = .54$
RESIDUAL MEAN	.05	.002	-.01	-.14
STANDARD DEVIATION	2.7	2.8	3.5	5.3
AUTOCORRELATION: SIGNIFICANT VALUES WITH LAGS	.17 (6) .17 (8) .17 (9) -.21 (18)	.18 (4) .23 (8)* .16 (10) -.14 (11) .16 (12) .14 (17) -.14 (20)	-.15 (12)	.20 (3) -.17 (4) .27 (5)* .20 (6) -.15 (7) .14 (17) .16 (18)
	(0, 1) is acceptable $\mu = 4.7$ trend = .15 $\theta_1 = -.003$ Largest residual autocorrelation -.23 (18)	(0, 1) is acceptable $\mu = 5.1$ trend = .15 $\theta_1 = -.05$ Largest residual autocorrelation .24 (8)	(0, 1) is acceptable $\mu = 9.6$ trend = .19 $\theta_1 = .037$ Largest residual autocorrelation -.17 (12)	(0, 1) unsatisfactory has autocorrelation values as high as .37
	(1, 1) Good Model	(1, 1) Good Model	(1, 1) Good Model	(1, 1) Satisfactory

* Indicates a significant residual autocorrelation $\geq 2/\sqrt{n}$

TABLE 7.4 BEST MODELS FOR REPORTED BURGLARY BY REGION

REGION	D1		CD1	
PERIOD	Pre-COMSEC	Post-COMSEC	Pre-COMSEC	Post-COMSEC
BEST MODEL TYPE PARAMETERS	(1, 0) Φ_1 = .30 μ = 32.8 (33.0) trend = .12	(1, 0) Φ_1 = .048 μ = 22.3 (22.5) trend = .12	(0, 1) μ = 173.3 (163.9) trend = -9.4 θ_1 = -.23	(0, 1) μ = 226.6 (175.9) trend = -50.6 θ_1 = .61
RESIDUAL MEAN	.0001	.0001	-.01	.42
STANDARD DEVIATION	7.8	5.7	20.4	18.3
	-.15 (8) -.14 (19) -.19 (20)	-.14 (2) .22 (3)* .21 (6) -.18 (7) .37 (9)* .30 (12)* .14 (15) -.19 (16) -.18 (19)	.27 (2)* .22 (5)* -.20 (15) -.18 (17)	-.17 (2) -.15 (7) -.14 (13) -.22 (19)*
COMMENTS	Good Model	Unsatisfactory Model, but the best obtained from candidates. Several lag values are high.	Marginal Model Mean value is not too good.	Satisfactory Model so far as autocorrelation goes. The mean and trend values indicate problems with the model.

* Indicates significant residual autocorrelation $\geq 2/\sqrt{n}$

TABLE 7.5 BEST MODELS FOR REPORTED AUTO THEFT BY REGION

REGION	D1		CD1	
PERIOD	Pre-COMSEC	Post-COMSEC	Pre-COMSEC	Post-COMSEC
	(1, 1) $\Phi_1 = .89$ $\mu = 10.5\ (13.6)$ trend = .29 $\theta_1 = .71$	(1, 1) $\Phi_1 = .19$ $\mu = 9.7\ (9.8)$ trend = .12 $\theta_1 = .06$	(1, 1) $\Phi_1 = .97$ $\mu = 109.9\ (52.0)$ trend = -2.0 $\theta_1 = .87$	(1, 1) $\Phi_1 = .77$ $\mu = 51.6\ (43.9)$ trend = -1.6 $\theta_1 = .60$
RESIDUAL MEAN	.065	.001	-.04	.16
STANDARD DEVIATION	4.2	3.7	10.6	7.8
	-.14 (15)	-.17 (5) -.14 (7) -.15 (10) -.23 (14)* -.18 (16) .24 (18)* -.14 (20)	.15 (9) .16 (16)	.19 (2) .17 (11) -.17 (12) .14 (16) -.14 (20)
	Good Model. Mean value is not too good	Satisfactory Model	Good Model for autocorrelations, however the mean value is <u>not</u> satisfactory.	Good Model. Mean value is marginal.

* Indicates a significant residual autocorrelation $\geq 2/\sqrt{n}$

models was either simpler in form or only slightly better than other candidates. There are no absolute rules for such selection and the procedure is to attempt to get a good model. The ARIMA model approach is entirely satisfactory for making forecasting models. However, it can decrease the value of such models for evaluation purposes because a "changed" model need not reflect significant changes in the time series. Strong changes can be discovered very well. Other cases lead to indications and measures of possible change. The work reported here illustrates an initial investigation of the use of ARIMA models for use in evaluation of police operations. A number of questions arise relative to such use.

The quantity θ_o specified in Section 6 may be considered as a possible indicator of crime level as discussed there. The "trend" term present in the operational ARIMA software indicates actual change or nonstationarity in the process. This term should be small in the models because stationarity is a basic requirement for ARIMA model methodology. In most of the "selected" models, the trend term is in fact small. When this is not so, a major change (nonstationarity) is indicated. The crime level expression also includes the trend term and must be considered to be changing when the trend is not small. These level measures are related as follows:

$$\theta_o = (1 - \sum_{i=1}^{p} \Phi_i)\mu + T$$

where T is the "trend" term. This relation only applies to nonseasonal models. A full study of θ_o for seasonal models and application to interpretation of data falls beyond the scope of this study.

Some conclusions and discussions following from the examples are:

1. REPORTED CRIME CAN PROVIDE GOOD ARIMA MODELS

Reported crime does lend itself to the formulation of ARIMA models.

Reference to the previous tables shows that many satisfactory models may be obtained. One model is considered unsatisfactory. This is the model for burglary in D1 for the post-COMSEC period (for which data are probably nonstationary).

Considering the complexity of the ARIMA methodology and the limited analyses employed (no transformations or differencing), this seems to be a positive result. There are two factors that are likely to be contributing to the occurrence of the less than good cases: The 78-week post-COMSEC period is definitely marginal in sample size for computations of this complexity; better results are very likely to be obtained from large time series runs, and the parameter estimation procedure itself is a complex numerical analysis problem.

Convergence of the parameter estimation procedure is determined by imposing control values on variables; in some cases these produced models in which the mean value was not properly estimated (as seen by its differing from the sample mean). In such cases one cannot be confident that good model parameters (in the sense of minimum mean square error) have been obtained.

2. ARIMA MODELS INDICATE PROBLEMS OR CHANGES IN REPORTED CRIME

ARIMA models of reported crime indicate problems or changes in reported crime. In many cases this reflects effects identified by detailed statistical studies of series autocorrelations as discussed in Section 4. However, in some cases the model complements or lends additional insights to the autocorrelation studies. Examples:

The failure to achieve good models for rape in the post-COMSEC periods in both D1 and CD1 indicates some causative effect. It may be the 78-week sample size. However, there is also a difference in this effect between D1 and CD1 with D1 yielding a somewhat worse model. The poor residual autocorrelation for rape cannot be overcome by simple changes in the model that incorporate seasonal effects because such models were among the candidate models and gave even poorer results. It seems that rape may not be a simple (0, 1) model (random) though such a model is very intuitively acceptable. It is close to being such a model but further considerations are in order to determine why the (0, 1) model is not better.

Robbery is well represented by simple autoregressive models. Therefore, it has relatively little purely random effects and may be expected to yield to various actions. It seems to be a statistically well-behaved (under control) type of random process.

Burglary gives some problems in model formulation indicating a lack of stationarity in this crime type. More detailed (transformation or differencing) studies would be required to develop truly satisfactory models for burglary. It is likely (but not assured) that such methods would be successful, but they were beyond the scope of the COMSEC study.

Auto theft shows some problems with model formulation, particularly with estimation of the mean values. It is known that the auto data were in rather poor shape compared to data for other crime types (e.g. missing or improperly reported data). This would present difficulty to the parameter estimation program resulting in termination before convergence to proper minimum mean square error values for parameters.

3. ARIMA MODELS MAY CONTRIBUTE TO EVALUATION

ARIMA models contribute to evaluation when compared with each other within a particular crime type.

It is felt that the ARIMA models are most useful in static evaluation as complementary indicators used together with simple and detailed statistical studies. The models are relatively sophisticated tools and can only be properly used when one has some idea about the case under study. Such additional knowledge is provided by the other statistical studies. Part of the reason for the complementary character of ARIMA models is their newness as evaluation tools, characterizing reported crime.

Reference (2) establishes the potential utility of stochastic models illustrated by the examples of this report. It demonstrates their use as complements to other studies (e.g., detailed statistics) and provides a background of methodology and examples upon which to develop further, more detailed analyses of reported crime as stochastic processes.

The forecasting ability of ARIMA models was not utilized in (2) or the evaluation studies illustrated above, only their representational character was used. However, it is likely that once a good ARIMA (or other stochastic) model has been developed, for a particular crime type and region, it could be used to forecast future crime values with reasonable accuracy. Such forecasts would probably be meaningful for only a few time units into the future. This capability suggests the utilization of stochastic models in various operational applications, in particular as a quick evaluation technique that could indicate changes from "expected" values (as forecast) which might be due to altered police operations. This is discussed further in Section 8.

Section 8. Utilizing the Forecasting Ability of Stochastic Models

Three distinct levels of statistical analysis have been identified and described in this report for the study of time series data. The most advanced of these levels is the formulation of stochastic models which represent the series data. A stochastic model ideally contains all the information present in the data used for its formulation. Actually any model will be less than ideal but should well represent the data if it is a good model. Thus a model tells as much or more about the data than simple or even detailed statistics do by themselves. On the other hand models require considerable additional effort for their formulation and for many purposes are not as useful in communicating results as the simpler methodologies.

Forecasting is the one aspect of time series analysis for which some kind of mathematical model is required. In order to forecast, the series must be represented by a model which can produce values corresponding to future time periods. If there exists no theory or deterministic model components the best forecasting is based on the data themselves as in an ARIMA type model. The concept of forecasting data such as reported crime must be taken in a proper perspective. It is very different than forecasting the exact position of a space satellite for example. For the satellite a close approximation to its position can be obtained by solving detailed equations of motion derived from physical laws. The exact position then requires a statistical forecast in the form of a correction time series. There is no similar underlying theory for reported crime and the full value must be forecast from the data alone. Moreover crime data are likely to have a much greater degree of purely random components than is normally found in engineering data. In such a situation even a theoretically good model will yield values of considerable variation often missing the true values by a substantial amount. This is in the nature of the data and not a fault of the model (provided it is, indeed, a good model).

Under these conditions one may consider whether or not it is useful to apply the forecasting ability of stochastic models to reported crime or similar data. This will depend on the quality of data available, the methodology available, and the extent and goals of the proposed study. Forecasting is often felt to be desirable but one should clearly understand the effort involved and the amount of uncertainty that is likely to be present. When a good forecasting model can be constructed, the forecast values can be useful in a number of operational activities including planning and evaluation.

In particular the possibility of what may be called dynamic evaluation becomes possible with forecasting. By contrast the usual, static, evaluation makes comparisons between time data collected over rather extended periods. Two major shortcomings with this normal procedure are the necessity for waiting a considerable period before making any meaningful comparison, and the possibility of all sorts of additional effects entering in over such periods. If a good model of the time series can be constructed than future values can be forecast. As actual values occur they can be compared with the forecast to see what changes may be taking place for evaluation indicators.

Any particular situation must be studied on its own to determine how feasible the construction of forecasting models might be. On the basis of the time series analyses used as examples in this report and described more fully in Reference (2) it seems likely that reported crime data can sometimes be developed into useful forecasting models. The values of such models will depend on the quality of the data, the crime type, available methodology, and often the experience of the analyst as well. It is important to stress that even good models may not provide useful forecasts if it is in the nature of the data to be highly random as in the case of homicide or aggravated assault. However even in such cases one can forecast nominal levels and variations which are extremely unlikely to be exceeded. This kind of "control chart" indication of expected random activity, though potentially useful, is not the kind of forecasting under consideration in this section.

The kind of results that may be encountered in applying ARIMA methodology to various types of data is illustrated in the remainder of this section. First three cases of the artificial Sump example data will be discussed then several examples of actual reported crime will be open.

Examples Sump 1 and Sump 2 have been previously used to illustrate concepts in other sections. They differ in that Sump 2 has a linear trend with a slope of .2 and a deterministic increase in level of 2 units. In each of the illustrations two time origins are used, one at 80 and one at 96. The value 80 is within the series allowing forecast values to be compared with actual. The time origin at 96 does not allow any such comparison because each series is 96 values long. Since all the ARIMA models used for illustrations in this report are simple in structure the forecasts quickly obtain a steady state value near the series mean. In fact this happens after three or four time periods. Such behavior is common for forecasting models of this type and in practice one tries to update forecasts every two or three periods. The models and forecasts discussed here are for purpose of illustration so that extensive forecasting methodology has not been employed.

Sump 1 is a (1,1) ARIMA series generated using parameters .25 and 1.5 for the autoregressive and moving average coefficients respectively with a series mean of 32.3. The ARIMA model estimated (and used) parameter values of .31 and 1.7 respectively with an estimated mean of 35.9. It seems rather close to the actual series. Table 8.1 shows the forecast values for the two time origins and the actual values when known.

Periods ahead	1	2	3	4	5
T = 80 time origin					
Forecast	39.24	37.95	37.56	37.44	37.40
Actual	44.02	36.88	42.30	36.68	37.91
T = 96 time origin					
Forecast	40.29	38.28	37.66	37.47	37.41

Table 8.1. Sump 1 Forecasts

Sump 2 was generated in the same way as Sump 1 using the same autoregressive and moving average parameters. However, it had a definite trend line build in as well as the deterministic increase in level. The mean of Sump 2 data is 52.6. The (1,1) ARIMA model estimated the parameters as .95 and

1.8 with an estimated mean of 39.72. Because of the non-stationarity introduced by the trend the model is not expected to be too good. Table 8.2 shows the forecasts that it produced.

Periods ahead	1	2	3	4	5
T = 80 time origin					
Forecast	59.31	59.30	59.30	59.30	59.30
Actual	68.19	61.32	67.01	61.66	63.15
T = 96 time origin					
Forecast	61.82	61.78	61.56	61.44	61.32

Table 8.2. Sump 2 Forecasts

The data of Sump 2 can be subjected to a difference operation to reduce the effect of the non-stationary trend. When this is done the resulting series, Sump 2D has a somewhat improved forecasts as shown in Table 8.3.

Periods ahead	1	2	3	4	5
T = 80 time origin					
Forecast	62.01	61.80	62.13	62.41	62.68
Actual	68.19	61.32	67.01	61.66	63.15
T = 96 time origin					
Forecast	65.51	65.69	65.98	66.26	66.54

Table 8.3. Sump 2D Forecasts

Comparison of Tables 8.2 and 8.3 shows that the differencing does help produce a model with improved forecasting capability. In both cases the series mean is not well defined because of the non-stationary nature of the data (due to the presence of the linear trend). Thus the forecast does not go to the series mean as it did in the case of the stationary data of Sump 1. The models are not very satisfactory and should have the trend removed before the ARIMA process is applied to produce a good forecasting model.

Some forecast model results for reported crime data are given in Table 8.4. All of the cases shown made use of the models illustrated in Section 7 and judged to be good representational models in that they estimated to mean values closely (indicating satisfactory parameter estimation) and had good white noise properties. However in each case the variance present in the random residual was fairly high as one must expect in such data. The residual variances are: 2.7, .97, 3.3, 7.8, and 4.2 for Aggravated Assault, Rape, Robbery, Burglary, and Auto Theft respectively.

Table 8.4 is an illustration of how simple ARIMA models forecast when applied to reported crime data. The forecasts are not particularly encouraging since they tend to simply give values close to the series mean. However there are several things to observe about the forecasting of reported crime data:

- The effect of purely random variation is great in all crime types and this will limit the ability of an ARIMA model to track actual values.

- Though the ARIMA forecasts do about the same as a simple mean extrapolation bounded by standard deviation ranges, the forecasts are not bad. Within statistical variation, induced by the purely random component, they are satisfactory.

- Forecasting of reported crime data needs an enlargement of methodology beyond the strictly ARIMA approach. The ARIMA forecasts are most successful when the series data have correlated structure. To forecast reported crime additional structure must be developed by using other (leading indicator) time series or some other theory of crime generation.

- The forecasting of reported crime is only now starting to be investigated. The examples show that some reasonable level of forecast is possible. It remains to improve on the ARIMA approach.

CRIME TYPE	TIME PERIOD	TIME ORIGIN 50 Forecast	50 Actual	60 Forecast	60 Actual	70 Forecast	70 Actual	80 Forecast	80 Actual
AGGRAVATED ASSAULT	1	4.99	5.0	4.88	7.0	5.09	2.0	4.50	3.0
	2	4.96	2.0	4.88	6.0	5.03	4.0	4.61	4.0
	3	4.94	11.0	4.88	6.0	4.99	7.0	4.69	5.0
	4	4.92	7.0	4.88	8.0	4.96	4.0	4.74	1.0
	5	4.91	2.0	4.88	3.0	4.94	4.0	4.78	1.0
RAPE	1	.78	0.0	1.10	1.0	1.03	2.0	1.02	2.0
	2	1.02	1.0	1.02	1.0	1.02	0.0	1.02	1.0
	3	1.02	2.0	1.02	3.0	1.02	1.0	1.02	1.0
	4	1.02	1.0	1.02	3.0	1.02	0.0	1.02	0.0
	5	1.02	1.0	1.02	2.0	1.02	0.0	1.02	0.0
ROBBERY	1	8.94	7.0	9.95	11.0	9.86	6.0	9.31	11.0
	2	9.35	15.0	9.25	11.0	9.15	11.0	9.85	9.0
	3	9.25	12.0	9.86	9.0	8.84	8.0	9.45	8.0
	4	9.45	7.0	9.45	14.0	9.55	15.0	9.15	9.0
	5	9.44	11.0	9.55	5.0	9.54	12.0	9.48	11.0
BURGLARY	1	32.3	44.0	31.5	26.0	27.9	21.0	29.7	39.0
	2	32.8	40.0	32.5	37.0	31.4	25.0	31.9	34.0
	3	32.9	32.0	32.8	19.0	32.5	32.0	32.6	45.0
	4	32.9	31.0	32.9	34.0	32.8	33.0	32.8	33.0
	5	32.9	42.0	32.9	42.0	32.9	17.0	32.9	39.0
AUTO THEFT	1	11.7	11.0	9.8	18.0	12.9	16.0	13.2	15.0
	2	11.8	11.0	10.2	7.0	12.9	13.0	13.2	6.0
	3	12.0	6.0	10.5	10.0	13.0	14.0	13.2	12.0
	4	12.1	12.0	10.8	11.0	13.0	18.0	13.2	14.0
	5	12.2	13.0	11.1	7.0	13.0	15.0	13.2	8.0

Table 8.4. Reported Crime Forecasts

Section 9. Additional Considerations for Cyclic Trends

In any stochastic model of time series one aspect that has not been discussed in detail is the so-called seasonal variation. Use of this term may include cyclic, periodic or repeating variation in the time series. Such variation can cause misleading interpretations of time series since one may think a change in values is due to a deterministic or random effect, when in actuality, it only reflects a change due to some cyclic effect. (Clothing and toy sales are typical examples.) Detection of such variations can be extremely difficult and establishing them on significant statistical grounds can be even more difficult.

Subjective observers are prone to "see" cyclic variations in time series plots and often the presence of cycles is reported on the basis of such observation. This is unsatisfactory because of the lack of quantitative evidence in many cases. As a substitute for such evidence, one often cites "reasons" for the variation, but such justification has no real statistical validity and cannot stand up against quantitatively based criticism. On the other hand, as previously indicated in this report it is possible to accommodate seasonal (cyclic) variations in analytic models.

To do so in the ARIMA technique increases the complexity of the model, requires computation of more parameter estimates and, therefore, requires longer time series to supply the amount of data necessary for parameter estimation. Large difference operations, acting over many time units, are the mathematical expression of seasonal variation. They can be useful in theoretical models, but introduce some problems in practice. In particular, the reported crime series of length 87 used for illustrations are rather short for seasonal variation models. However, in some cases, improved models have been obtained by including seasonal variation (indicated by a study of the autocorrelation lag values).

Seasonal variation is further complicated by the presence of spurious data (which should be identified and removed), special events (like major holidays), trading day effects (that give a different number of workdays to special weeks or months), combination of cyclic variations, etc. Various methods have been developed for the analysis of seasonal (cyclic) variation. One of the most widely used is the Census Bureau method known as Census Two. This method and some other work on seasonal variation is discussed in this Section.

In discussing periodic variation, the possibility of so-called spectral analysis comes to mind. Its applicability is mostly to fairly well-behaved stochastic processes such as are encountered in the communications and engineering control fields. Indeed the major discussions of such methods are found in books and articles dealing with such time series. It is a widely held view among time series analysts that spectral analysis will not yield satisfactory models of time series representing economic, social or similar data. It seems that the stochastic features of such series cannot be represented by any reasonable form of spectral model.

It is interesting to note that the spectral analysis method has been employed in another way by Rosenblatt. (Reference 7.) Rather than being used as a basis for model formulation, it is used as an evaluation method to test the effectiveness of Census Two and an alternative Bureau of Labor Statistics procedure for the determination of seasonal variation.

One should note that much of the extensive work on seasonal variation is directed toward developing good representational models. In that work there is essentially no concern with forecasting. This is compatible with static evaluation needs. However, if one desired a forecasting capability, the basic Census Two methodology would require modification. (This has been done by several groups.)

Weekly crime data do not show strong cyclic characteristics. For this reason monthly data may be investigated in order to look for seasonal effects. Though monthly crime data are not very satisfactory for detailed analysis of reported crime they do provide a data base for cyclic studies. The remainder of this section describes the Census (Method) Two, version X-eleven (X-11) program and illustrates its use on reported crime data.

THE CENSUS BUREAU X-11 PROGRAM

Over the past twenty years the Bureau of the Census has developed a number of approaches for detecting and adjusting seasonal variation in time series data. Their major approach is widely known as Census II (Method Two). It has been implemented by a number of computer program versions since 1955. The current version, developed by the Census Bureau is called X-11 (X-eleven). This program is designed to detect and adjust for seasonal variation and trend. it is most widely used in analysis of economic time series. Such use is presently under the auspices of the Bureau of Economic Analysis (Department of Commerce) and is carried out by the Bureau (often for outside "customers") or by others who have obtained the software from the Bureau,* at nominal cost. The Bureau of Economic Analysis supports use of the X-11 by making available expert advice on the use and interpretation of the program.**

The X-11 program is designed for use with monthly data (it can deal with quarterly data by means of three month aggregates) and such series are the only type of time series that can be employed (e.g., weekly series cannot be dealt with by X-11). This is not a trivial limitation that one can easily circumvent, the details of the program assume monthly data and built-in operations act accordingly.

The program considers time series to have three basic components: seasonal (S), trend cycle (C), and irregular (I). It allows, as an option, consideration of a fourth component: trading day (D).

One can use a multiplicative or additive type model to represent the original series data (0) yielding the following two possible forms: (1) Multiplicative 0 = CSID; Additive 0 = C+S+I+D.

This section gives a brief description of what X-11 does and illustrates how its output can be used on reported crime data. A great deal of output can be obtained from the program, but, in fact, only a few of the available tables and charts are used by most time series analysts. Input and operation of the

---

* The applications described in this report employed X-11 software obtained from the Bureau, operating at the Polytechnic Institute of New York

** The present section owes a great deal to Morton Somer of the Bureau who gave the author advice and encouragement in the use and interpretation of X-11.

program is also described here so that the Section updates, summarizes, and complements the basic document describing X-11, Reference (8).

Use of X-11 is illustrated by study of four crime types: Robbery, Burglary, Aggravated Assault, and Rape for District 1 in Cincinnati over a seven-year period (1968-1974) during which that District was employed in testing an innovative change in police operations known as the Community Sector Team Policing (COMSEC) program.

There are two distinct choices one must make in using X-11: multiplicative or additive model and trading day adjustment or not. How to do this is described subsequently. Output is similar in all cases and is described below. Only those parts of the Standard Output commonly used are discussed. Table designations do not follow the normal numbering of this report. Instead they reflect standard X-11 usage as described in Reference (8).

1. OUTPUT WHERE THERE IS NO TRADING DAY ADJUSTMENT

--Table B1. Lists the original data by month and year, it gives yearly totals, monthly averages, table total, overall average and overall standard deviation (basic sample statistics).

--Table C17. Unusual (extreme) values in the irregular (random) part of the data are identified and assigned weights by which they are modified in subsequent calculations (rather than disregarding them as an alternative). This employs a 12 month moving average technique as described in Reference A (under B17 page 13). This table is the same for both multiplicative and additive models.

--Table D8. Gives comparisons of season (s) to irregular (I) components. In the multiplicative model S-I ratio's are given with 100 percent the nominal value. In the additive model S-I differences are given with 0 (value) the nominal value. These measure the relative strength of seasonal effect to purely random (irregular) values. Unmodified values are used with regard to the extreme weights shown in Table C17.

This table is one place to look for seasonal effects. One considers the column for a particular month. If essentially every year entry is above or below nominal values there is a strong indication that that month has a seasonal effect. Hypotheses can be made and tested on these statistics in an analysis of variance (F-test) type table which is included. A statement about seasonality is also given by the program (one can question the statistical validity of such statements, relative to given data, however they do supply some level of quantification in the indication of seasonal effect).

--Table D9. Shows replacement values used for extreme SI ratios or differences. Table D9A is not used. This table indicates the effect on calculations of the extreme identification and weighting process. It has no direct role in analysis of the data.

--Table D10. This is an improved version of Table D8 which uses replaced values for extremes, smooths over total (years), and forces ratio to average 100 (over months of the year) in multiplicative models. It forces differences to average 0 over months in additive models. This forcing renders the above and below nominal values more meaningful and allows one to interpret their significance. It would be difficult to do so if the average (nominal) was different from the nominal (recalling the very basic assumption of seasonality which states that above normal values will balance below normal values over a complete year).

--Table D10A. Produces seasonal factors for each month, one year ahead. This is done by taking one half the difference between the last two values and adding this to the last value (if negative of course it is subtracted).

--Table D11. Gives data with the seasonal component removed. It is divided out in the multiplicative model and subtracted out in the additive model. These values still include the trend cycle and irregular components.

--Table D12. Gives data with both the seasonal and irregular components removed. These values represent the trend cycle part of the time series data.

--Table D13. Gives the irregular component without seasonal or trend cycle effects. For multiplicative models it expresses the irregular as a weight factor with nominal value of 100. For additive models it gives random values with nominal value of 0.

--Table E4. Is useful as a check on the validity of work done by the X-11 program. Values of 100 for the multiplicative and 0 for the additive indicate meaning results. The second column of printout is redundant and should be omitted.

--Table F1. Gives moving average based on the number of months for cyclical dominance (MCD) value, discussed under Table F2. Some analysts find these values useful. They require a rather detailed appreciation of the MCD concept.

--Table F2. Is a rather complicated table which may be used by experts in X-11 type analyses but need not concern others. The MCD value is computed within this table. It represents the number of months (span in months) required to produce a change over in influence between I and C. For some span of months I has a greater effect than C. Then C becomes more significant. It is the transition span that determines MCD (used in Table F1). The second part of Table F2 is similar but uses variance comparison rather than percent change. A changeover effect is also to be found in this table. The I/C ratio also shows a transition from greater than to less than unity. When MCD is indicated as more than 6 the value 6 is used because a greater span is assumed to lose too much information. The MCD can be used to estimate the trend cycle. A Henderson curve may also be used, but it loses some value at end points. This material is further developed in Reference 10.

--Chart G1. Gives a useful visual display of the seasonally adjusted series, and the (Henderson) trend cycle.

--Chart G2. Gives several things, particularly the seasonal factors. There is one chart for each month. By study of these charts one can get an impression of the presence of seasonal factors (by month) and of their effect (increase or decrease).

2. OUTPUT WHEN TRADING DAY ADJUSTMENT IS PRESENT

When the trading day adjustment is active the number of days in each month is considered in relation to the position they occupy in the week (Monday, Tuesday, etc.). When this option is used many of the tables reflect appropriate consideration of the trading day effect D. It is a factor in the multiplicative model and a term of the additive model. In addition the standard output includes three additional tables:

--Table C14. Indicates any extreme value (associated with component I) to be excluded from the trading day considerations.

--Table C15. Summarizes statistical analysis of trading day effect.

--Table C16. Gives the adjustment factors for trading day effects.

These tables are rather technical and need not be considered when using the trading day option. The other tables are properly modified by trading day effects when this option is used and provide the analyst with "corrected" values.

Illustrations Using Reported Crime Data

As part of the evaluation of COMSEC, an innovative police operation in Cincinnati, some examples of X-11 operation on reported crime data were developed. These are discussed in this section to illustrate use of X-11 output and to record the results of those examples.

Data used were for District 1 in Cincinnati. Monthly data for the period January 1968 through December 1974 were used (the COMSEC program started in March 1972). Four crime types were selected on the basis of interest and available data. Types selected were: Robbery (RB), Burglary (BU), an aggregate of four reported classes of burglary, Aggravated Assault (AA), and Rape (RA). Rape was selected because it is generally felt to be purely random (irregular) but some people feel it may contain some seasonal variation. Thus it acted as a test case with some possibility of side interest as well.

In order to find out as much as possible about the four crime types in an X-11 study methodology the X-11 was used in four different analysis modes in each case. This also provides a full range of output examples and guidelines for use of the X-11 program.

Cases are distinguished by letter designation added to the basic letter designations for crime types. Cases considered are shown in the following table in which their letter designations are specified for easy reference.*

---

* Inclusion of trading day effect studies was suggested by Morton Somer of the Bureau of Economic Analysis. This effect is often important in economic or social scientific time series analyses.

TABLE 1: CASES CONSIDERED

	MULTIPLICATIVE MODEL		ADDITIVE MODEL	
	Without Trading Day	With Trading Day	Without Trading	With Trading Day
Robbery	D1RBP	D1RBPT	D1RBS	D1RBST
Burglary	D1BUP	D1BUPT	D1BUS	D1BUST
Aggravated Assault	D1AAP	D1AAPT	D1AAS	D1AAST
Rape	D1RAP	D1RAPT	D1RAS	D1RAST

Letter designation D1 in each case refers to the use of District 1 data. The designators are used as titles in X-11 runs; they are not used for series identification. Series identification must agree between control (option) card (instructions) and data cards specification as described in the next section. Thus use of the above titles as identifiers would require four sets of the same data to be used. In fact each crime type had one set of data, identified by crime type designators only. Each was used in turn for the four distinct X-11 type runs (P, PT, S, ST).

Output illustrations will be discussed by crime type. However several observations can be made which apply equally well to every case. In addition the more important tables and charts can be illustrated for a particular case so as to serve as guides to the reader. The case of Aggravated Assault (AA) will be used for this purpose. Output for that case is shown in the following set of tables and charts with accompanying comments of a general nature. Chart G1 is discussed briefly for each crime type case and one example is shown, for Aggravated Assault.

Tables D8 and D10 are major places to look for seasonal variation. In the AA example they are shown for a multiplicative model with nominal value 100 percent. Additive models have nominal values of zero in the tables. Table D10 is the more reliable since it reflects appropriate modifications for extreme values carried out by the X-11 program. To interpret these tables one looks for months which consistently, over the span of years, have higher or lower values than nominal. For example AA Table D10 shows low values for January and high values for September, while March shows no indication of any seasonal effect (some values are high while others are low). Table D8 gives a statistical test for indication of a seasonal effect at a confidence level of one percent. Table D10 gives one year ahead seasonal factors. This is not a forecast (the Bureau is not allowed under the law to do forecasting) but does indicate reasonable factors, based upon the available historical data.

Table D13 shows the irregular component. Examples for both the multiplicative and additive models are shown for AA to illustrate how they differ. One is based on a nominal value of 100, the other on zero. A reasonable irregular component should roughly divide the number of cases evenly between those below and those above the nominal value. One observes this to be true for the tables shown.

Table E4 indicates how well the program corrections and analyses are operating. Nominal values should be obtained for each year. Typical examples are shown for both the multiplicative and additive models.

Tables F1 and F2 are not illustrated here or used in the present set of analyses. They do not contribute to these studies but are often useful in analysis of economic time series.

Charts G2 are illustrated for AA by two examples. There is one chart for each month and the plots show indications of high, low, and transition effects. In the examples shown February indicates a low effect, consistent over seven years. By contrast June shows a high effect increasing over the years. Charts G2 together with Tables D8 and D10 are the major indicators of seasonal effect. Chart G1, shown with the crime case, is a major representation of trend effects.

Before considering the results for specific crime types, some general results which apply to each of the four types are given:

--In each crime type case the seasonal effects and trend values were essentially the same for multiplicative and additive models. One must interpret the two models from a different point of view and use different nominal values. However the basic conclusions do not depend on the model form.

--Trading day component had no effect on any case for either model. There was certainly no effect for the case of Rape. In the other cases there were slight effects to be found (most noticeable for Burglary). But these were mere hints of effect only, they never had any real effect on the seasonal factors or the trend component (or the irregular component).

Illustrative results for each crime type will now be considered. In each case, chart G1 will be summarized showing the trend effect and the seasonally adjusted series, it will be shown for Aggravated Assault only.

TABLE D-8

D1AA — SERIES D1AA

D8. Final Unmodified SI Ratios

YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	AVGE
1968	81.0	67.1	112.7	86.1	83.6	123.7	82.9	152.5	134.4	110.8	55.2	73.7	97.0
1969	71.5	149.2	112.7	61.0	161.8	95.0	96.2	218.7	128.0	107.0	103.8	76.3	115.1
1970	56.2	86.0	123.6	128.4	76.0	87.4	143.6	101.6	105.3	114.1	115.0	121.0	104.9
1971	91.3	60.7	71.6	111.7	156.7	92.8	106.0	85.3	139.3	128.1	95.2	117.6	104.7
1972	86.3	91.1	59.6	83.0	69.8	122.4	138.4	106.6	142.8	110.5	97.4	68.5	98.0
1973	86.6	44.8	115.4	130.1	84.1	132.6	83.9	175.4	95.9	112.4	55.5	108.2	102.1
1974	91.9	81.3	99.2	86.8	128.5	100.9	92.0	119.5	136.9	116.6	94.4	65.8	101.1
AVGE	80.7	82.9	99.2	98.1	108.6	107.8	106.1	137.1	126.1	114.2	88.1	90.2	

TABLE TOTAL- 8673.9

STABLE SEASONALITY TEST

	SUM OF SQUARES	DGRS. OF FREEDOM	MEAN SQUARE	F
BETWEEN MONTHS	22508.813	11	2046.256	2.784**
RESIDUAL	52922.000	72	735.028	
TOTAL	75430.813	83		

**STABLE SEASONALITY PRESENT AT THE 1 PER CENT LEVEL

TABLE D-10

D1AA — SERIES D1AA

D10. Final Seasonal Factors

YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	AVEG
1968	72.4	74.8	108.9	104.6	88.5	100.0	100.3	119.6	124.4	112.8	99.8	94.2	100.0
1969	74.4	75.3	104.5	103.7	87.7	101.3	103.1	117.2	126.5	113.5	99.3	93.5	100.0
1970	76.5	75.2	100.1	105.9	87.1	103.2	105.9	111.5	125.9	114.3	98.4	95.4	100.0
1971	79.0	75.9	96.7	106.3	86.6	105.8	108.1	107.4	125.2	114.9	97.1	95.0	99.8
1972	83.0	74.9	93.8	107.0	88.6	108.7	107.1	104.2	125.0	116.2	94.7	94.9	99.8
1973	86.6	75.0	91.6	103.8	90.2	112.1	105.9	105.6	126.9	116.5	92.4	90.6	99.8
1974	88.9	75.2	90.6	101.9	91.4	115.3	104.8	107.3	128.0	116.0	90.4	86.8	99.7

TABLE TOTAL- 8389.3 MEAN- 99.9 STD. DEVIATION- 14.4

D10A. SEASONAL FACTORS, ONE YEAR AHEAD

YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	AVGE
1975	90.1	75.3	90.1	100.9	91.9	117.0	104.3	108.1	128.5	115.8	89.4	84.9	99.7

TABLES D-13

D13. FINAL IRREGULAR SERIES D1AA — SERIES D1AA

YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	S. D.
1968	110.8	88.7	102.5	81.6	93.7	123.1	82.6	128.1	109.1	99.4	56.1	79.3	19.8
1969	97.1	199.1	108.0	58.7	183.9	93.6	93.0	186.2	101.2	94.4	104.6	81.6	47.0
1970	73.3	114.1	123.1	121.0	87.1	84.5	134.9	90.6	82.9	98.8	115.5	125.6	20.0
1971	114.8	79.7	74.2	105.8	182.8	88.5	98.4	79.3	110.4	110.1	96.6	121.9	28.1
1972	102.6	120.5	63.2	77.4	78.9	113.0	129.7	102.6	114.3	95.0	102.5	71.7	20.1
1973	99.4	59.4	125.4	125.0	93.2	118.7	79.8	167.9	76.3	97.0	60.0	118.4	30.0
1974	101.9	106.2	107.6	83.8	138.8	86.8	87.7	111.8	107.9	101.7	106.1	77.1	15.7
S. D.	12.3	42.6	21.8	23.6	47.9	15.3	20.8	43.9	13.7	5.0	23.4	22.6	

TABLE TOTAL- 8679.1 MEAN- 103.3 STD. DEVIATION- 27.4

D13. FINAL IRREGULAR SERIES D1AAS — SERIES D1AA

YEAR	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	S. C.
1968	0.	-3.	1.	-3.	-1.	0.	-4.	7.	2.	-0.	-9.	-4.	4.
1969	0.	15.	2.	-5.	14.	-0.	-0.	15.	-1.	-2.	1.	-4.	8.
1970	-4.	3.	4.	6.	-2.	-4.	9.	-2.	-4.	0.	4.	5.	4.
1971	2.	-3.	-5.	2.	15.	-3.	-0.	-5.	3.	3.	-0.	4.	5.
1972	-0.	4.	-7.	-5.	-3.	3.	7.	-0.	3.	-1.	2.	-6.	4.
1973	-0.	-5.	4.	4.	1.	3.	-3.	12.	-6.	-0.	-5.	4.	5.
1974	1.	2.	3.	-3.	12.	-2.	-3.	3.	4.	0.	-1.	-13.	6.
S. D.	1.7	6.4	4.3	4.0	9.2	3.1	4.8	8.0	3.6	1.3	4.0	6.4	

TABLE TOTAL- 60. MEAN- 1. STD. DEVIATION- 5.

TABLES E-4

E 4. RATIOS OF ANNUAL TOTALS, UNMODIFIED — D1AA

YEAR	UNMODIFIED
1968	100.0
1969	98.0
1970	100.7
1971	99.3
1972	99.9
1973	100.2
1974	100.3

E 4. DIFFERENCES OF ANNUAL TOTALS UNMODIFIED — D1AAS

YEAR	UNMODIFIED
1968	0.1
1969	0.0
1970	-0.2
1971	-0.4
1972	-0.5
1973	-0.6
1974	-0.7

```
G 2.    CHART                    D1AA
                (x) - D 8.  FINAL UNMODIFIED SI  RATIOS
____________________________________________________________________________
                (0) - D 9.  FINAL SI RATIOS         MODIFIED FOR EXTREMES
                (+) - D10.  FINAL SEASONAL FACTORS
                (*) - COINCIDENCE OF POINTS
   SCALE-ARITHMETIC

      44.       60.       76.       92.      108.      124.      140.      156.
      *         *         *         *         *         *         *         *

FEBRUARY

 1968                    *  *
 1969                           *
 1970                              *
 1971             *
 1972                     *           *
 1973  x          o       *
 1974                          *
 1975
      *         *         *         *         *         *         *         *
      44.       60.       76.       92.      108.      124.      140.      156.
```

```
                                 D1AA
G 2.    CHART
                (x) - D 8.  FINAL UNMODIFIED SI RATIOS
____________________________________________________________________________
                (0) - D 9.  FINAL SI RATIOS         MODIFIED FOR EXTREMES
                (+) - D10.  FINAL SEASONAL FACTORS
                (*) - COINCIDENCE OF POINTS
   SCALE-ARITHMETIC

      44.       60.       76.       92.      108.      124.      140.      156.
      *         *         *         *         *         *         *         *

JUNE
 1968                                                    *
 1969                                     *
 1970                                *
 1971                                   *
 1972                                                   *
 1973                                                         *
 1974                                        *
 1975
      *         *         *         *         *         *         *         *
      44.       60.       76.       92.      108.      124.      140.      156.
```

1. ROBBERY

- Table D8. Indicated seasonal effect present at one percent level. March had low values, July and December high values. December was particularly high (consistently over the seven year period).

- Table D10. Agrees with D8 and hence adds to the strength of the effect indicated there.

- Table D13. Shows a reasonable irregular part. The multiplicative model has 41 values below nominal (100 percent) and 43 above. The additive model has 41 negative and 43 positive values (about a nominal zero value). This indicates a good analysis by X-11 into the various model components.

- Table E4. Shows values close to nominal, hence indicates reasonable analysis procedure.

- Chart G1. The graph of the trend cycle is of major interest on this chart. The trend is steady for January 1968 through December 1969 then rises sharply till January 1970 where it remains steady until March 1973, dropping to a minimum in February 1974. The trend then undergoes a rise to the previous high level and remains through December 1974. From this chart there does not seem to be any significant change in the trend cycle component of Robbery after March 1972 (COMSEC initiation period). The dip in February 1974 is not likely to be an indication of true changes in level since it is followed by a rather steep rise to previous high levels.

- Chart G2. Shows February as irregular going from low to no effect over the seven year period. March shows some low effect, July some high effect. December shows a strong high effect.

Tables D8 and D10 together with Chart G2 indicate a high seasonal effect for December. A much smaller seasonal effect is indicated for July (somewhat high) and March (somewhat low).

Chart G1 indicates two major levels of the trend component, a steady, rather low level prior to December 1969 and a higher level since August 1970 (with a rapid steady rise between these periods). The second period is rather steady with a few low dips which seem to be a rather insignificant fluctuation with the exception of February 1974 which is rather low.

2. BURGLARY

- Table D8. Gives no evidence of seasonal effect at the one percent level.

- Table D10. Agrees with D8.

- Table D13. Shows a reasonable irregular component. The multiplcative model shows 51 values below and 33 values above normal. The additive model shows 46 negative, 5 zero, and 33 positive values.

- Table E4. Indicates a reasonable analysis with all values close to nominal.

- Chart G1. The trend cycle shows a change in level starting in October-November 1969 which reaches a high by May 1971. It remains high until August 1972, drops to a low in April 1974, then increases until the end of the study period in December 1974.

- Chart G2. Shows some variation effects. There are no consistent high or low cases so there is no indication of seasonal effects. The most interesting variation over the seven year period shows change over in 1972, October goes from low to high, November goes from high to low, and December goes from high to no effect.

No seasonal effects are indicated. The trend cycle underwent a definite downward movement after March 1972 which started to reverse in April 1974. This may be associated with real changes in the nature of Burglary (displacement effects, shifts to other crime types, etc.) which in turn may be due to COMSEC activity. The upward trend may indicate an accommodation (or learning process) by criminals to the new police operations. It may also be a reflection that displaced burglary is not as desirable (to the criminal) outside of District 1. In this sense the rise may be a reverse displacement effect.

Chart G2 indicates some changes in monthly effects about the 1972 value. No real seasonal pattern is indicated in any case.

Though no strong trading day effects were found in any of the studies reported on here, the case of Burglary came near to having some trading day variation. Values in some cases did differ between models with or without the trading day component. However, such variation was not large enough to produce significant effects.

3. AGGRAVATED ASSAULT

- Tables D8 and D10. Show evidence of seasonal effect at the one percent level. January is low, February is also low but not as consistently as January. August and September are high with the exception of one year each. October is consistently high.

- Table D13. Indicates a reasonable irregular component. The multiplicative model has 41 values below nominal and 43 above. The additive model has 33 negative values, 13 zero, and 38 positive values.

- Table E4. Shows all values near nominal indicating a reasonable analysis.

- Chart G1. Shows the trend cycle to be an irregular cycle moving above and below a value of about 24 incidents (recall this is only the trend component and does not relate directly to the reported crime values). About January 1974 the trend cycle started to increase, reaching what appears to be a new, high level by September 1974, where it remained to the end of the study period in December 1974.

- Chart G2. Shows a number of seasonal effects. Low months are: January, February, March, and to some degree December. High months are: June, August, September, and to some degree October.

There is a seasonal effect indicated for Aggravated Assault, low in January, rather higher in August and September, remaining high in October.

The trend cycle shows a significant increase in the 1974 period.

4. RAPE

This crime type was included in the present study as a check comparison on the other cases. It is widely felt to be essentially random in nature. However some studies have indicated a measure of seasonal effect and a number of people have expressed the view that there may be such an effect.

Actual data for Rape includes zero values so that a multiplicative model cannot be used with such data. In this study a multiplicative model was run using data with each value increased by unity. The additive model used actual data. Both models agreed so far as seasonal effect study and general nature of the trend cycle were concerned.

- Table D8. Gives no evidence of seasonal effect at the one percent level.

- Table D10. Shows a rather different situation than Table D8. Though there are no very strong effects one can distinguish some months: very high (May), high (January, June, October), low (February, July, September, and November).

- Table D3. Shows a reasonable irregular component. The multiplicative model has 42 values below and 42 values above nominal. The additive model has 33 negative, 18 zero, and 33 positive values. This is a strong indication of the rather random nature of Rape which results in a particularly good irregular part (the major part for this crime type).

- Table E4. Gives all cases near nominal indicating a reasonable analysis.

- Chart G1. The trend cycle is irregular showing a very low rise of about one or two incidents per year added to the level value. Variation about that value also shows increase. There is some indication that as a random statistical quantity Rape is going "out of control" so that extreme values may occur more often than otherwise expected. This is seen by contrasting the early period up to January 1971 with the period following that date.

Of course one should consider the complete composition of Rape rather than the trend cycle only to gain a full appreciation of possible changes.

- Chart G2. Indicates a number of high and low values with a great deal of variation. May is very high consistently over the seven-year period June is rather high, July is low. October and December change from high to low over the years of study.

The trend cycle may indicate some changes in the levels of rape that can be expected to occur.

Though no strong seasonal effect is indicated there is evidence of a high tendency in May carrying over to some degree into June. This lends evidence

7/75 D1AA

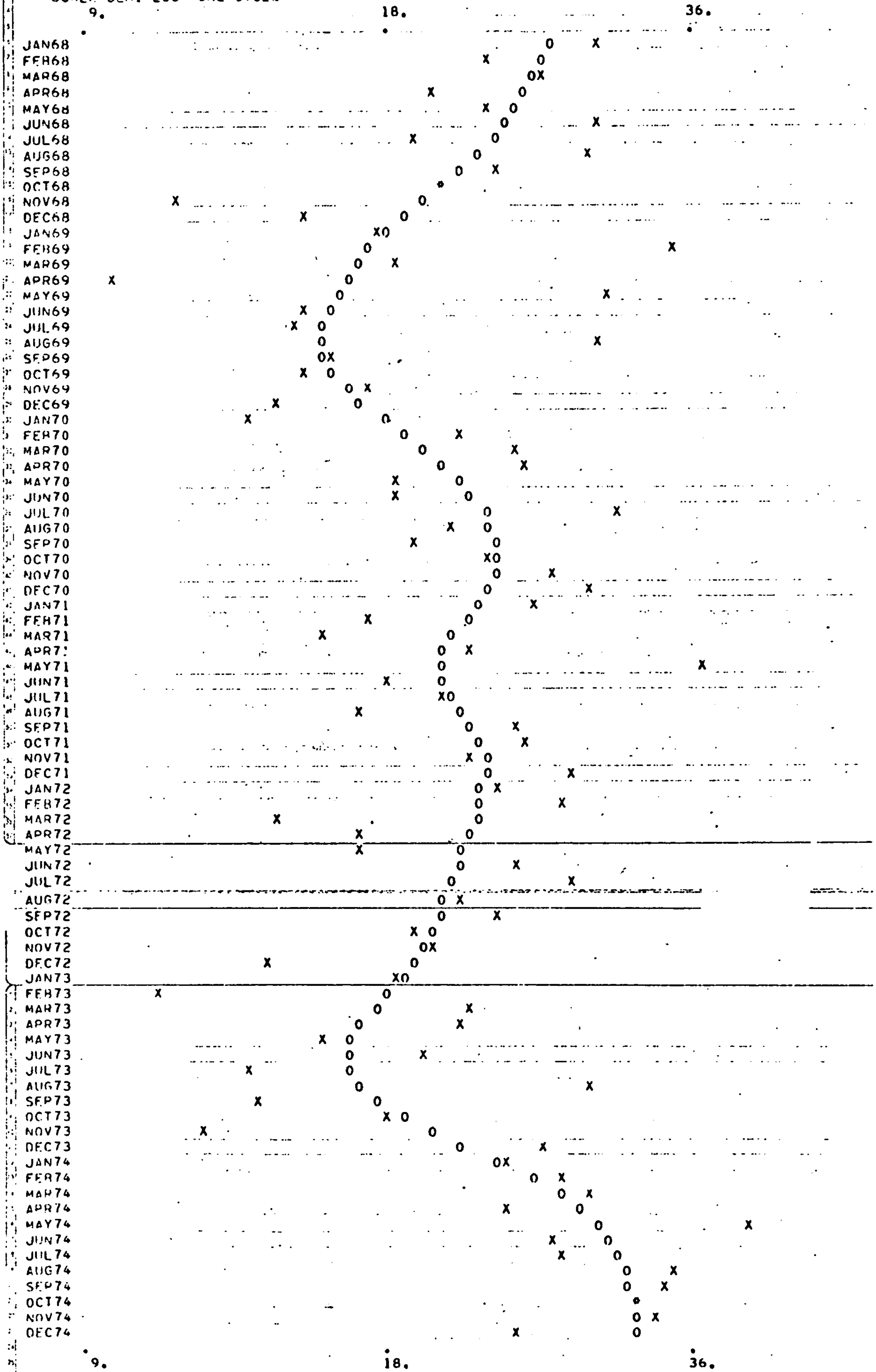
G 1. CHART

(X) - D11. FINAL SEASONALLY ADJUSTED SERIES

(O) - D12. FINAL TREND CYCLE

(*) - COINCIDENCE OF POINTS

SCALE-SEMI-LOG ONE CYCLE



from the data to the view that spring is a time of year in which a crime such as Rape might be expected to be unusually high.

For Rape the trading day effect was particularly absent, having exxentially no effect at all in analyses (as one would expect).

Section 10. Software for Statistical Analysis of Time Series

One reason for the relatively restricted use of statistical analysis until recent times was the extensive calculations required in order to apply theory that has long been available. The advent of computers has greatly changed this situation. By using computers extensive calculations become feasible. However some aspects of time series analysis are so involved that the necessary computer programs, called software, require considerable effort in their development. This has limited the use of ARIMA type model analysis until the last three or four years during which several versions of appropriate software have become available and various commercial suppliers of computer service are making proprietary software available at a fee.

All aspects of time series analysis either require or are greatly simplified by using computers and appropriate software, provided the available data are established in a data base useful to computer operation. Data plots form an important introduction to any time series study and they are very tedious to do by hand, particularly if many cases are under study. Computers can make time series plots as illustrated in Section 3. They can also calculate means, standard deviations, and simple comparison statistics. The calculations of autocorrelations almost requires a computer. Though the calculations can be done "by hand" (using a calculator), the work is time consuming and chances of error are high.

All of these applications of computers to time series analysis require only standard software, available at any computer facility in various forms. Therefore such software need not be given any further consideration here. Implementation of the concepts of ARIMA modeling are much more involved. The major effort is in calculation of the estimates of the model parameter using minimization of the mean square error function.

This section will describe how some available software may be tested for use with one's computer facility and data. It will aslo indicate some of the commercial services available.*

THE AVAILABILITY AND RELIABILITY OF ARIMA SOFTWARE

The ARIMA methodology requires rather complicated computer programs, called software, for its implementation. A nonseasonal version of ARIMA software was written and used by the author in reported crime time series analysis. This version used the method of moments (Yule/Walker) for computing estimates of model parameters. A number of satisfactory baseline models were obtained by using this software and it provided some insights into ARIMA methodology. Those models are described in Reference 9; the software will be called the Marshall version in subsequent discussion.

Subsequent considerations indicated that it would be useful to have the ability to include seasonal differences in the ARIMA model. Furthermore, parameter estimation based on a minimum mean square error procedure is

---

*The services included are only some of the available services. Their inclusion does not imply any judgments about the services and is meant only to guide prospective uses regarding the kind of commercial service available at the present time.

considered to be better than the Yule/Walker method. This is due to the very pragmatic nature of ARIMA model formulation. The goal is to get a good model according to stipulated criteria, particularly with white noise residual.

Minimum mean square techniques use the model structure itself, together with the data, to generate estimates of model parameters. The price one must pay for the more detailed estimates, over the relatively simple Yule/Walker estimates, is increase of computational complexity. The mean square error expression is generated by what is called back forecasting and is minimized by numerical procedures known as gradient search methods. This is a rather extensive computational task which will be discussed further below.

To employ a seasonal version of ARIMA methodology that uses minimum mean square estimation requires one of the following approaches:

--develop a complete version of ARIMA software
--obtain appropriate software from some other source, or
--make use of a commercial system having the desired version of ARIMA software.

A reason for developing one's own version of ARIMA software is the shortage of working versions of ARIMA software available at acceptable cost. But considerable effort is required and it may be most effective to use existing software. A version is available from the University of Wisconsin Computing Center as Number 517 of its Supplementary Program Series. At nominal cost, the center supplies a report, "Computer Programs for the Analysis of Univariate Time Series Using the Methods of Box and Jenkins." This report included listings of all associated programs. The resulting software will be called the Wisconsin version in the following discussion.

Though it is often not a simple matter to utilize "other people's programs," the magnitude of the task required in modifying the Marshall version resulted in a decision to make use of the Wisconsin version for the reported crime examples. This was done for all the time series analyses and models used as examples in this report (given in Sections 7 and 8). It may become costly to use a commercial system having ARIMA software when many models are to be developed, as in the COMSEC study. This is because commercail systems charge for each model. Moreover, there are problems with input of data to commercial systems that are not part of the major system in which one has data base storage. One use of the third option is as a check on "in-house" software by using several software forms for a few cases. The availability of two commercial systems, both of which were considered for the COMSEC studies, will be described below.

1. NATIONAL COMPUTER SOFTWARE SYSTEMS

National Computer Software Systems, Inc.* (NCSS) is a commercial supplier of computer time and software systems. Their services are widely used in business and industry in a variety of ways. In the spring of 1973, NCSS added a software implementation of the ARIMA technique to their line of services. NCSS introduced this service in a series of three seminars conducted by Box and Jenkins in May 1973. The initial presentation was in New

*300 Westport Avenue, Norwalk, Conn. 06851

York City on May 14-17, 1973. At that presentation, two days were devoted to lectures in which Box and Jenkins developed a number of the concepts discussed in References 6 and 11. The final two days were used to allow potential users of the NCSS system, called SPX/Time, to learn about that system using their own data on computer terminals.

The SPX/Time system is based on proprietary products of ISCOL, Ltd. (UK) and is copyrighted by NCSS (in 1973). There are indications that the software will operate reasonably well to carry out the three major phases of the ARIMA technique as described in Section 6. It also includes transforming and differencing of data for preliminary analysis. Aspects of seasonal modeling are also available as part of the system. The subjective aspects of the technique must be supplied by the user and this may limit the general commercial use of the system.

Many people with the training required for making proper subjective judgments will also be able to produce their own software implementation and are likely to prefer to do so rather than work with a commercial system. This is the point of view that was taken for the COMSEC evaluation studies and has been expressed by several other groups, both users and nonusers of commercial services. There are a number of reasons for this viewpoint: One has greater understanding of self-developed systems and can introduce features that are particularly appropriate to problems under study, it is convenient and, in most situations, will be less costly to use a self-developed system when many time series models are required. (These observations also apply to software developed by others, such as the Wisconsin program.) Though the cost of developing such a system might seem to indicate that one should employ a service such as NCSS, the situation under which the system is developed can greatly affect such a consideration.

Some details on the background of the ARIMA technique and on SPX/Time are given in Reference (10). The software system will be referred to as the NCSS version of the ARIMA methodology in the following discussion.

It should be observed that the NCSS system includes software implementation of the so-called transfer function methods dealing with the study of several time series as they relate to each other. These methods (discussed in References 4 and 10) show great promise for use in the study of societal systems.

2. COMNET IMPLEMENTATION

COMNET is another computer services company which can provide ARIMA software. This is a version developed as part of the "Econometric Software Package" (ESP) and will be called the ESP version of ARIMA software. This version does not seem to be as general as the NCSS version nor the "in-house" version represented by the Wisconsin version. However, it was used as a check and for special problems.

The major disadvantages of both the NCSS and ESP versions are the transfer of input data (the time series) and cost for large numbers of time series. They are useful systems when studying a few series.

Another commercial version is provided by Dialogue Inc.* as part of their PLATO system. The ARIMA aspects of this system are described in Reference (11). In this system the subjective step of model identification is done by the software resulting in what is called automatic ARIMA model building. This is an important feature, to the extent that it works, because the identification step requires skill and knowledge that are not often found in groups otherwise able to utilize ARIMA technology. The success of PLATO has been observed in a few simple cases but should be investigated by prospective users on their own.

The Wisconsin version was implemented for a PDP-10 computer, requiring considerable modification of the version supplied. It was used in various sites. Due to a number of technical problems,** changes had to be made in the version as supplied by The University of Wisconsin. Since these were done at the Urban Institute the operational system is referred to as the UI (Wisconsin) version.

George Washington University (GWU) also has implemented the Wisconsin version. Material supplied by GWU implies that the Wisconsin version is used directly as given. However, this is unlikely to be the case due to technical problems similar to those encountered in the UI implementation. The actual operational version is referred to as the GW (Wisconsin) version.

No other versions of the ARIMA software were considered for this study. Of course, other versions exist, and more are being developed as the ARIMA procedure becomes widely known and applied. A summary of ARIMA software and its role in reported crime studies is given below to illustrate how several versions of complex software may be used to check each other and gain an impression of the reliability present in the main operational version (UI Wisconsin in the illustration cases).

--<u>Marshall Version:</u> No seasonal capability, crude estimation method. Used for baseline studies in early work. Not used in final analyses given as examples in this report.

--<u>NCSS Version:</u> Very general but relatively expensive to use. Methodology helpful as background material. System not used in COMSEC studies.

--<u>ESP Version:</u> Fairly general, but expensive for large-scale use. Used as a check on several test cases in COMSEC studies.

--GW (Wisconsin) Version: Used as a check on several test cases. May have some problems related to software implementation.

--UI (Wisconcin) Version: Used for all COMSEC time series model work. Has full (seasonal) capability and uses minimum mean square error estimates. May have a few (technical) problems still, but was taken to be operational in its present form for development of ARIMA models.

---

* 233 Broadway, Room 809, New York, N. Y. 10007

** Due primarily to the fact that the Wisconsin version was written for a computer having very different characteristics (e.g., word length) than the PDP-1

The ARIMA software is sufficiently complicated to require some kind of testing before one can have a reasonable degree of confidence in the results. A full-scale testing effort of the UI (Wisconsin) version was beyond the scope of work that produced the examples of Section 7. However, a number of tests were made as described below. In these tests, it is necessary to understand two major aspects of the software reliability issue:

--Detailed involvement with software implementation; including making changes and debug test runs, leads to an informed appreciation of what the programs are doing and how they are operating.

--Numerical aspects of matrix inversion and gradient search, particularly the latter, are not trivial and can have significant effects on the model results.

The minimum mean square error is sought by means of a gradient method of a particularly effective kind (Marquardt Algorithm). In practice, the minimum is never found. One specifies when the computation should stop by employing one of three rules:

--stipulate the number of iterations,
--stipulate minimal cutoff change in the mean square, or
--stipulate minimal cutoff change in parameter values.

Thus, one does not know that a "best" solution (minimum mean square error) has been obtained, but only that a solution satisfied the stipulated criteria. Residual mean square values are obtained but these do not give a very good measure of how well the parameter estimates have been in minimizing the error expression.

In terms of the background material above, the following will indicate how the present UI (Wisconsin) version was tested and why it was felt to be operational.

Considerable work was done in implementing the version. In the process, a number of changes were made in the Wisconsin version. One major change was made in the matrix inversion subroutine which was felt to be a true error in the original. (Thus, anyone using the original without detailed study might have errors.) Such detailed association with the software lends a measure of reliability to its operation.

Four tests cases were run using UI (Wisconsin) version, GW (Wisconsin) version, and ESP version. In each case rather different results were obtained for model parameters. General characteristics were sometimes similar and sometimes not. The UI (Wisconsin) version was judged to be the most operationally correct for the following reasons.

ESP version does not compute mean values as part of the estimation process. Such computation provides an internal check on the estimation procedure and in particular on its convergence to the minimum mean square value as shown by numerical values in the tables of Section 7 and discussed there.

Both GW and UI versions compute the mean, but the GW values do not always give correct values (mean values are known from other, elementary, calculations). In most cases the UI version does give a good estimate of the mean. Poor values indicate that convergence has not occurred and indeed the GW version stopped its cáluclations well before the UI version in most cases.

As another operational check, some values were compared with the Marshall version which was a completely different set of software. Values were in reasonable agreement.

In addition to the checks discussed above, a computer program was developed to produce simulated time series with known properties. Such series data were subjected to the ARIMA software to see how well the resulting model represented the known input series. This approach is discussed in Section 11.

It was judged likely, on the basis of the above tests, that the UI (Wisconsin) version was reasonably good for developing ARIMA models. This illustrates an approach to establishing the reliability of complex software. Such considerations should be made whenever software of comparable complexity is employed in analysis.

Section 11. Synthetic Time Series

There are several reasons for generating synthetic time series with specified characteristics. They can be used to form examples such as the water level in the boiler sump used throughout this report. When the basic form of a reported crime series is established from data, or assumed, simulated series can be produced for training purposes. Time series of known form can illustrate various statistical characteristics, particularly in autocorrelations and partial autocorrelations. Special features of seasonal variations and non-stationary trend can be directly related to specific aspects of the synthetic data. An additional use of simulated time series of known form is as test cases for complex software to see how well it carries out such functions as parameter estimation since the parameter values are known for the simulated series.

A synthetic time series is a set of values that can be thought of as representing values occurring at distinct times and that have been generated by some completely described deterministic process. Such a series is distinct from real time series that are obtained by recording actual values occurring in prescribed time intervals such as the number of robberies in a police district by week. The major distinction is that the synthetic series is artificial rather than "natural" and is reproducible, being based on deterministic generation procedures. One could generate a series that was not reproducible by employing some form of natural random generation (such as radioactive emission or second-hand position on a clock at arbitrary times). Such series may be felt to be more realistic as simulations of naturally occurring series. However, it is necessary for the scientific use of simulated series that they be reproducible while at the same time having features that act like true randomness.

Such series employ so-called random number generators which are, in fact, pseudorandom. They have the statistical form of purely random number but are completely reproducible.

As indicated above there are various applications for synthetic time series including:

- as input to broader simulation studies;
- to assist in the general study of time series types and how various effects in series generation relate to values obtained in the series; and
- in operational testing of computer software designed for analysis of time series.

The simulated series reported on here were developed for the second and third of these applications and used in this report to generate the boiler sump data. Major interest in having such series was to test software for autoregressive calculations and parameter estimation in Autoregressive Integrated Moving Average (ARIMA) models of time series and provide examples to illustrate the various parts of this report.

This section describes the way in which time series may be simulated and reports on the use of such series in testing ARIMA software. It gives conclusions and recommendations for the use of ARIMA software based on a limited number of simulation studies.

It is beyond the scope of present work to use simulated series in extensive testing of ARIMA software or in theoretical features of stochastic model forms. However, such studies would contribute significant knowledge to time series model formulation. The work reported here forms an introduction to the basic methodology for such extended study.

Time series of interest for this study are assumed to be of the ARIMA form. Simulated series were not produced for seasonal or other special variation.* The nonseasonal mathematical series employed here express the value z_t in terms of previous series values such as z_{t-1}, purely random amounts a_t (called random shocks) that cannot be accounted for by the previous series values, and a non-random constant term. If the model uses p previous series values and q previous shocks, then the value of z_t is expressed in the form:

$$z_t = \sum_{i=1}^{p} \varphi_i z_{t-i} + \theta_o + a_t - \sum_{i=1}^{q} \theta_i a_{t-i} \qquad (1)$$

To define such a series, then p + q + 1 parameters φ_i, θ_o, θ_i must be specified. Selection of the series type is accomplished by choosing the value of p and q to be used. The notation (p, q) is used here for series type.

Terms arising from previous z values are autoregressive and the φ_i are autoregressive parameters. Terms depending on the random shocks are moving average with θ_i the moving average parameters. The parameter θ_o represents the mean level of the series corrected for autoregressive effects. It is referred to as the mean in this section. In fact, it may include a trend term and be modified by a factor $1 - \sum_{i=1}^{p} \varphi_i$.

From the above formulation, it is seen that the desired time series may be generated using a formula in the form of Equation 1. Input to be supplied are the parameters Φ_i, θ_i, and the series of random shocks a_t. Both the z_t values and the a_t series must be generated sequentially. A simple computer program will generate (z_t) following Equation 1 if one provides for generation of the random imput series (a_t).

The a_t values are produced by a standard (built-in) random number generator function. These give values in the range zero to unity. Since one usually wants larger random values, a scale factor B is included as input to

*For the boiler sump examples the synthetic series generation programs were modified so as to include the fixed subtraction and addition of water at cyclic intervals as used in the previously discussed illustrations. Trend effects were also included.

effectively enlarge the random interval to (O, B). It is also desirable to have both positive and negative random values. The algebraic sign of a_t can be assigned in various ways. In the simulation studies reported on here, the sign was determined by giving positive and negative equal likelihood of occurrence. A second random number is used in each case to determine the sign of a_t.

PROGRAM DESCRIPTIONS

The computer programs used to carry out the formulation of (z_t) following Equation 1 are designated SER. F4 and a subroutine version denoted by GSR.F4 (subroutine GNSER) to be used with various calling programs. These are used as follows (listings are supplied at the end).

1. PROGRAM SER. F4

This program is designed for terminal use. It generates a simulated time series of the form shown by Equation 1. Output is printed out term by term. Input is requested as needed. There is an option to supply integer data output (in floating point form).

- Input: NC--number of cases (series to be generated (if negative, only get one)); NP--number of autoregressive terms: NQ-number of moving average terms; PH(I)--autoregressive parameters; TH(I)--moving average parameters; B--scale factor for random numbers; A--starting value (z_1); XM--mean; N--length of series (maximum of 200); and SEITCH--Y or N (integer data or not).

2. PROGRAM GSR. F4

This program is a subroutine version of SER. F4, for terminal use with various calling programs. It generates a simulated time series of the form shown by Equation 1. Output is the series returned to the calling program. Call is to: SUBROUTINE GNSER (X, N, INF, OUTF) where X is a one dimensional array of dimension N (dimension limit 200), N is the length of series to be generated (maximum of 200), INF and OUTF are specified in the calling program as the input and output devices.

- Input: NP--number of autoregressive terms; NQ--number of moving average terms: PH(I)--autoregressive parameters; TH(I)--moving average parameters; B--scale factor for random numbers; A--starting values (z_1); XM--mean, and SWITCH--Y or N (integer data or not).

EXAMPLE OF USE OF THE ABOVE PROGRAMS

- Input: NP = 1, NQ = 1, PH = .05, TH = .2, B = 3, A = 30, XM = 25; SWITCH = Y to give integer data, N = 80.

- The resulting series has the following initial 30 terms; 30, 27, 24, 26, 26, 24, 24, 27, 23, 24, 26, 24, 25, 25, 24, 25, 23, 25, 26, 25, 28, 25, 27, 23, 24, 25, 25, 23, 27, and 24.

Greater fluctuation can be introduced by increasing the size of the scale factor B. This is illustrated by another example with all input the same as the above except that the scale factor B = 8. The first 30 terms are: 30, 28, 21, 26, 26, 21, 21, 30, 18, 20, 27. 20, 24, 23, 21, 23, 19, 23, 26, 23, 31, 25, 30, 18, 21, 24, 25, 18, 20, and 22 which show greater fluctuation than the prveious series with a B value of 3.

The time series of water levels for the boiler sump examples employed the synthetic time series program with NP = 1, NQ = 1, PH = .25, TH = 1.5, B = 5, A = 30, and XM = 28. The series were all 96 values long. Modifications were made to the series which allowed addition or subtraction of fixed amounts at intervals six units apart by amounts AA and BB respectively. A trend with fixed intercept at AT and slope of CC was also provided. The cases used as illustrations employed the values:

	SUMP 1	SUMP 2	SUMP 3	SUMP 4
AA	-5	-5	-10	-10
BB	2	2	2	2
AT	0	2	0	0
CC	0	.2	0	.1

OPERATIONAL TESTING OF ARIMA SOFTWARE

ARIMA software consists of programs to calculate autocorrelation and partial autocorrelation values for time series data. Such values are used to select appropriate model forms and to test the resulting residual series for white noise characteristics. The ARIMA software also employes a minimization of the mean square error to estimate parameters. This is a relatively complicated and potentially sensitive type of calculation. Though there are a number of operational tests, one may carry out on ARIMA software (including careful analysis of the computer codes, comparison with other software system results, and convergence to known parameters such as the mean), it is desirable to subject it to the common procedure of testing with a known input. Due to the recursive nature of the input (as shown by Equation 1), one must use a simulated time series for this purpose. The parameters and type of series are known for such input. One can then test the software by noting how well it functions in specifying series characteristics and most particularly in estimation of model parameter values (which are known for simulated series input).

Extensive testing of this kind is beyond the scope of the present study. However, some tests have been made which illustrate the use and value of such techniques. These are described below.

Because of its complexity (size), the ARIMA software is contained in a collection of subroutines known collectively on our system as BOX.F4, designed for terminal use. To test such software by using simulated time series (generated from SUBROUTINE GNSER), some form of call program is required. These are described below.

Two call programs have been employed for terminal operation of the UI-ARIMA (Modified Wisconsin Version discussed in Section 10) software package BOC. F4. One, designated ACT. F4, provides autocorrelation and partial autocorrelation values only. The other, designated ARMT. F4, gives model parameters and information regarding the residual series and other aspects of the model.

1. PROGRAM ACT. F4

This is a call program that uses GNSER for series input. It is used for simulated series input. It calls parts of BOX. F4 to produce autocorrelation and partial autocorrelation values. Input is GNSER input together with BOX. F4 input (briefly specified at the end of this sub-section). ACT. F4 is for terminal operation. It can also derive input series from a binary file created by appropriate software systems.* Thus it also requires the following input:

NOB is the number of observations
"Generate" or "Read" 0 or 1 for GNRD.
Series name for identification--BOX. F4 input as required.

2. PROGRAM ARMT. F4

This program has the option of calling GNSER to generate a simulated input series for which parameter estimation is made. It is for terminal operation and uses BOX. F4 subroutines.

- Input: NOB--number of observations; "generate" or "read" series input option; list input or not option; series name; BOX. F4 input as required.

EXAMPLE

To test the ARIMA software five simulated series were used with input situations and results as reported below. Even this small number of cases provided useful information about the ARIMA software that one would never have without employing simulated series input.

The convergence process to estimate parameters seems to need a number of iterations. One must take care not to limit the number of iterations when specifying MIT or by giving a condition on EPSI which terminates due to small changes in the mean square error while parameter values are still changing. This was the situation for Case 5 above and it resulted in a poor estimate for PH.

Starting values for the parameters do not seem critical. In fact, there was no difference between Case 1 and Case 2 even though the actual parameter values were used as initial values in Case 1. However, due to the high dimension of the minimization problem local minima and similar difficulties may give trouble.

The length of the series (sample size) is felt to be important and the

*Such systems should be part of any computer center that provides terminal operation. The work illustrated in this report was done on the Brookings Institution (Washington, D. C.) computer where the software is called PLANET.

TABLE 11.1 ARIMA METHODOLOGY APPLIED TO SIMULATED INPUT

Input	Case 1	Case 2	Case 3	Case 4	Case 5
B	80	80	60	100	120
NP	1	X*	X	X	X
	1	X	X	X	X
PH	.05	X	X	X	X
TH	.2	X	X	X	X
	3	X	X	X	X
A	30	X	X	X	X
M	25	X	X	X	X
ITCH	Y	X	X	X	X
ries Name	Simulate 7	Simulate 8	Simulate 9	Simulate 10	Simulate 11
LOG, NRD NDS, NSEA)	0, 0, 0, 0	X	X	X	X
C	1, 0, 1, 1, 1, 0	X	X	X	X
PA	1, 0, 0, 1	X	X	X	X
PA	.05, .1, .1, .2	.1, .1, .1, .1	X	X	X
PS1, EPS2, IT, IPDEST, IPRES)	0.0, .001, 60, 0, 0	X	X	X	.00001, .001, 60, 0, 0
AC, NPAC, DAC, MCSE, NAPL, IWTPA, NCHI)	20, 15, 0, 1. 12, 0, -1	X	X	X	X
F, NTO, NU, ICI, DEST, IWTPF)	3, 1, 0, 4, 0, 0	X	X	X	X
NT	80	X	60	100	120
omments on runs	Floating underflow, one case.	Floating underflow, one case. Square root of negative number, two cases.	Square root of negative number, two cases.	Floating underflow, one case.	Floating, underflow, one case. Square root of negative number, two cases.
Estimated rameters	.09 25.16 .11 .23	.09 25.16 .12 .23	.21 25.17 .11 .24	.07 25.36 .12 .18	.17 25.35 .12 .24
pproximate rations**	51	50	38	58	47

*An X means that the same value was used as for the case to the left.

*Some iterations have many more substeps than others so these figures do not indicate exact differences in computational effort. All cases except Case 5 terminated because parameter value changes became too small. n Case 5, termination was due to too small a change in the sum of square value (the test value was .00001 in his case and 0.0 in other cases).

above cases illustrate this fact. For the simple cases considered, to only one level in each parameter type, the series of length 60 gave very poor results for PH and the series of length 100 gave improved results over the 80 points in Cases 1 and 2. The extra length (120) of Case 5 should have given further improvement in PH; however, the change in termination criteria resulted in a worse result.

Thus, even the restricted study of only five cases provides the following conclusions:

- The operational ARIMA software used for COMSEC evaluation studies and as illustrations in this report seems to be working as desired.
- Termination of estimation too early is a major source of poor model parameter values.
- Series should be as long as possible and be increased for more complicated model types requiring estimation of several parameters.

PROGRAM LISTINGS AND ARIMA INPUT

The programs, SER.F4, ACT.F4, and ARMT.F4, are listed here to illustrate the nature of such programs. It can be seen that they are relatively simple in form. This adds to the desirability of using similar programs for helping to test complicated ARIMA software.

The listings are followed by a statement of the input requirements for the operational ARIMA software. This will depend on the version used but the UI-Wisconsin version, given here, is typical of the nature of input requirements to do ARIMA model building.

For completeness of illustration, the following gives the input requirements involved in the ARIMA SOFTWARE (UI version):

SUMMARY OF BOX. F4 INPUT NEEDS

Input values are requested by call programs and used by BOX. F4 subroutines.

{ NOB--Number of data points in the series.

{ Z(I) I = 1, NOB--Series data.

{ SERIES(J) J = 1, 14--Name of the series data.

NLOG--0 original data used ≠ 0 log of data used.
NRD--number of regular differences (d in $(1-B)^d$).
NSD--number of seasonal differences (d_1 in $(1-B^s)^{d_1}$.
NSEA--order of seasonal differences if any (S in $(1-B^s)$).

{ INC(J) J = 1, 6--number of each of the six possible parameters
regular autoregressive (p in number) INC(1) = p
seasonal autoregressive (p_1 in number) INC(2) = p_1
mean of the series, INC(3) = 0 or 1
deterministic trend constant, INC(4) = 0 or 1
regular moving average (q in number) INC(5) = q
seasonal moving average (q_1 in number) INC(6) = q_1.

{ IOPA(J) J = 1, NP--(computed from INC) powers of B for each nonzero parameter (0 value for "constants") in order left to right in the general model; i.e., as listed above for INC input (nonzero values only).

{ PA(J) J = 1, NP--initial values of parameters for estimation (values must be nonzero), one for each specified as an entry for IOPA.

EPSI--maximum relative change in residual sum of squares.
EPS2--maximum relative change in each parameter.
MIT--maximum number of iterations allowed.
IPDEST--0 to suppress plot of data <--plot.
IPRES--0 to suppress plot of residual <--plot.

NAC--number of autocorrelations of residuals.
NPAC--number of partial autocorrelations of residuals (maximum of NAC).
NRDAC--number of differences of residuals for which autocorrelations calculated.
MCSE--0 if no standard error of residual autocorrelation.
NAPL--number of residual autocorrelations per line.
IWTPA--0 suppress plots of residual autocorrelation <--plot.
NCHI--number of autocorrelation to compute χ^2 statistic ≤ 0 if not wanted.

NF--number of forecasts desired.
NTO--number of time origins.
NU--number of new observations for updating.
ICI--width of confidence limits:

1	2	3	4	5	values for ICI
50	75	90	95	99	percent confidence internal.

IPDFST--0 to suppress plotting of series data <--plot.
IWTPF--0 to suppress plotting of forecasts <--plot.

NT(I) I = 1, NTO--forecast time origins.

ZN(J) J = 1, NU--new data for update of forecasts, omit if NU = 0.

A. PROGRAM SER. F4

```
      DIMENSION X(200),TH(6),PH(6),RA(200)
      INTEGER INF,OUTF
      INTEGER SWITCH,YES
      DATA YES/'Y'/
      INF=5
      OUTF=5
50    WRITE(OUTF,147)
        READ(INF,135) NC
        WRITE(OUTF,148)
        READ(INF,135) NP
        WRITE(OUTF,149)
        READ(INF,135) NQ
        IF(NP.EQ.O) GO TO 55
        WRITE(OUTF,150)
        READ(INF,130) (PH(I),I=1,NP)
55      IF(NQ.EQ.O) GO TO 56
        WRITE(OUTF,151)
        READ(INF,130) (TH(I),I=1,NQ)
56      WRITE(OUTF,136)
        READ(INF,130) B
        WRITE(OUTF,140)
        READ(INF,130) A
      WRITE(OUTF,152)
      READ(INF,130) XM
      WRITE(OUTF,145)
        READ(INF,135) N
      WRITE(OUTF,153)
      READ(INF,154) SWITCH
        X(1) = A
        RA(1)=0.0
        DO 100 I = 1,N
      IF(SWITCH.EQ.YES) X(I)=FLOAT(INT(X(I)))
      WRITE(OUTF,155) I,X(I)
        R=R*RAN(D)
        R1=RAN(D)
      IF(R1.LE.0.5) R=-R
      RA(I+1)=R
```

```
      ZT=0.0
      ZS=0.0
      IF(NP.EQ.O) GO TO 85
      IF(I.GE.NP)  GO TO 75
      DO 72  JJ=1,I
      ZT=ZT+PH(JJ)*X(I-JJ+1)
72    CONTINUE
      GO TO 85
75    DO 80  J=I,NP
      ZT=ZT+PH(J)*X(I-J+1)
80    CONTINUE
85    IF(NQ.EQ.O) GO TO 95
      IF(I.GE.NQ)  GO TO 88
      DO 86 KK=1,I
      ZS=ZS+TH(KK)*RA(I-KK+1)
86    CONTINUE
      GO TO 95
88    DO 90  K=1,NQ
      ZS=ZS+TH(K)*RA(I-K+1)
90    CONTINUE
95   X(I+1)=ZT-ZS+R+XM
      Y=X(I+1)
      J=I+1
      WRITE(OUTF,120) R1,R
      WRITE(OUTF,110) J,Y
100  CONTINUE
      IF(NC.GT.O) GO TO 50
      CONTINUE
110  FORMAT(I,10X,F)
120   FORMAT(2X,'R1',2X,F,2X,'R',2X,F)
130  FORMAT(6F)
135  FORMAT(6I)
136   FORMAT(2X, 'B:',$)
140   FORMAT(2X, 'A:',$)
145   FORMAT(2X, 'N:',$)
147   FORMAT(1H-,2X,'CASE :',$)
148   FORMAT(2X, 'NP:',$)
149   FORMAT(2X, 'NQ:',$)
150   FORMAT(2X, 'PH:',$)
151   FORMAT(2X, 'TH:',$)
152  FORMAT(2X, 'XM:',$)
153  FORMAT(2X, INTEGER SERIES (Y/N)? ',$)
154  FORMAT(A1)
155  FORMAT(2X, 'I = ',I,10X,'X(I) = ',F)
      END
```