COMPUTER-AIDED DATA ANALYSIS

A PRACTICAL GUIDE

WILLIAM R. GREEN

COMPUTER-AIDED DATA ANALYSIS

A PRACTICAL GUIDE

WILLIAM R. GREEN

Placer Development Limited
Vancouver, British Columbia

A Wiley-Interscience Publication

JOHN WILEY & SONS

New York • Chichester • Brisbane • Toronto • Singapore

Library of Congress Cataloging in Publication Data:

Green, William R.
Computer-aided data analysis.

"A Wiley-Interscience publication."
Bibliography: p.
Includes index.
1. Multivariate analysis—Data processing. I. Title.

QA278.G75 1985 519.5'35 84-21931
ISBN 0-471-80928-4

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PREFACE

This book was written to provide an introduction to the data analysis techniques in common use in the applied sciences. The algorithms can be easily implemented on a computer, and most people who have access to a computer may find that many programs are already available to perform these tasks. The emphasis here will be on the effective use of computer methods—merely having a set of working programs does not guarantee that the nonspecialist computer user will find efficient procedures for analyzing a particular set of data. There are a great many books in print giving the mathematical operations—and programming techniques for implementing them—but very little has been published to help people learn how to use such methods effectively.

For the past 10 years, I have been involved in attempts to train geologists, engineers, market analysts, and others in computer methods. While these people may be very familiar with statistical concepts, and be used to working with computer output, they often have not been directly involved in running computer programs. The general experience is that everyone has to learn the "tricks of the trade" the hard way, by trial and error. If they are lucky, experienced users in the same organization may be able to help speed up this process. The purpose of this book is to make such experience more widely available.

The advent of "personal" computers and other inexpensive hardware means that many more people are now running their own computer studies, while in the past they would have relied on expert staff in a computer department to do this work. Professionals in geology, forestry, agriculture, commodities and stock trading, and many other fields are becoming com-

puter users. They cannot of course devote full time to learning computer methods, as their primary functions remain in other areas.

There are two main groups who may find this book useful. The first consists of those who wish to use a computer to aid in data analysis, but who are not themselves computer experts (students and people in professional occupations, for example). For these people, my intention is to show the capabilities they should look for in the programs available to them, and to provide practical suggestions on how to use standard programs. The second group consists of analysts and programmers who develop data analysis systems. I hope they will gain insight into how their programs might be applied, to aid in future enhancements.

After the introduction in Chapter 1, the book is divided into three main sections. The first deals with the basics of analyzing data on a computer. Chapter 2 discusses the preparatory stages, which include developing an understanding of the computer system and how to use it. In addition, the data must be put into a computer-compatible form, which is frequently taken for granted, although of fundamental importance in obtaining useable results. In Chapter 3, frequency distributions are introduced, with simple data displays such as histograms providing the key. Chapter 4 moves on to the problem of investigating data with more than one variable. Here, correlation matrices and two-variable scatter plots are of great value. Chapter 5 contains suggestions for designing programs to allow these basic tools to be used effectively. In addition to flexible analysis software, a good set of utility routines (for data sorting, selection, listing, etc.) is essential.

Many problems in analysis involve data which have a spatial component, that is, which can be located on a map. Generally this involves plotting data on a two-dimensional coordinate system, which is the general topic of Part II. Chapter 6 outlines the principles of computer graphics. In Chapter 7, the requirements for effective application of graphics to data analysis are studied. Chapter 8 deals with specific techniques for combining the basic statistical methods with plotting to determine spatial structure and locate anomalous data. In Chapter 9 this is extended by considering additional methods for enhancing the visual impact of computer plots.

Part III is a brief review of more advanced analysis techniques. Chapter 10 covers advanced statistical methods. These include geostatistics, which allows rigourous statistical procedures to be applied in spatial data analysis. In addition, Chapter 10 contains an overview of powerful multivariate methods like factor analysis, and considers the application of computer graphics to aid in interpreting the results. Chapter 11 considers special analysis methods designed for array-oriented data, including time-series analysis and image analysis for remote sensing. Finally, in Chapter 12 the subject turns to interpretation of the data, which frequently involves an excercise in model-building.

Throughout the book are example applications of these methods, drawn largely from the earth sciences, but also including other types of data. The nature of the data in the examples is secondary to the techniques, which are quite general. A review of principles for effective use of computer methods is given in appendix A, along with an outline of a working data analysis system, given in Appendix B. Finally, there is a list of references and sources of more detailed information on many of the specialized topics introduced here, and a glossary of basic terms used in data processing and computer graphics.

It is a great help to any computer user to be able to draw on the experience of others. I have been fortunate in this regard, since I have received a great deal of useful advice from associates at Chevron and Placer. As this book stems from my work at these companies, I am indebted to all who have helped me. While they are too numerous to list here, I would like to extend my thanks to each one.

A few people deserve special thanks. Bart Cramer (of Chevron in Houston) showed me many useful techniques for plotting data, taught me the great power of simple tools such as histograms and scatter plots, and demonstrated the advantage of a flexible approach. Bill Robertson (of Chevron in Calgary) was also a valuable source of practical ideas. At Placer, Peter Kowalczyk, Jerry Thornton, and Bruno Barde provided a thorough test of many of the basic concepts in this book, and helped in developing an easy-to-use set of programs. Peter Bradshaw instigated the general concept of providing an interactive analysis system for occasional computer users at Placer.

The encouragement I received from my superiors for learning and developing new ideas was also a valuable asset. At Chevron, Paul Skakun, Dave Peacock, Al Singleton, and Garth Greenwood, among others stand out in this regard. At Placer, Ed Rychkun and John Brunette have given me the freedom to develop new ideas. If there are any names I have omitted, I will apologize in advance.

In preparing the manuscript, the Univac 1100 was indispensable, and Al Kemp helped me to prepare some of the figures: their assistance is greatly appreciated. The plotting program CPS1 (from Radian Corporation, of Austin, Texas) has been invaluable to Placer as a company, and to me in preparing the computer graphics displays in this book.

Finally I must thank my family (Nunzia, Cristina, and Daniel) for their support.

WILLIAM R. GREEN

British Columbia, Canada
January 1985

CONTENTS

CHAPTER 1

APPLICATIONS OF DATA ANALYSIS

Throughout history, man has sought to increase his knowledge and understanding of his surroundings. Such efforts usually begin with observations of some natural event, followed by attempts to define the underlying causes of that event. A natural extension of these acts is an attempt to predict when (or where) particular events will occur. Although pure curiosity is a strong incentive, more practical motives are often involved as well. For example, early astronomical studies produced economic benefits by predicting the seasons, but must also have infused feelings of intellectual achievement in primitive astronomers.

In the broadest sense, data analysis can be considered to include all such studies. The observations we wish to study constitute the "data" (following the Latin, an individual observation is a "datum"). The normal definition of "analysis" is the separation of a whole into parts. In this context there may be many aspects to analysis, including identification of recurring patterns, definition of previously unknown features, and determination of the causes of the phenomenon under investigation.

The ultimate goal is a complete understanding of an event and all of its physical processes. Since the universe may be considered infinitely variable, if viewed on a sufficiently fine scale, it is necessary to define these processes in terms of simplified physical models. The final result of data analysis, then, may be a set of parameters that is consistent with the data.

SCIENTIFIC DATA ANALYSIS

The scientific approach to data analysis is largely a matter of defining rules to make the processes of acquiring and studying data repeatable. This implies that anyone trained in a particular field should be able to obtain the same data by following the same observational procedures, and should get the same results from the same set of data. Many branches of mathematics were developed to formalize analysis, and indeed we might consider mathematics itself to be an embodiment of the rules necessary for studying data. The other sciences can be viewed as the principles needed to model phenomena.

Organized scientific studies have been going on for centuries, and much of our present knowledge of the universe has been derived from the great volumes of data collected in the past. Many current projects aim to refine the historical work by increasing the accuracy of the observations and improving the analysis techniques. As a result, many sciences are very dependent on sophisticated electronic devices to acquire data, and powerful computers to analyze the data and reduce them to an understandable form. Often the methods of analysis are not in themselves new, although the volumes of data treated are orders of magnitude greater than would have been considered manageable even fifteen years ago.

Modern capabilities for collecting enormous volumes of data have resulted in a great reliance on statistical analysis methods. The statistical behavior of a set of data (e.g., the range and distribution of values) is first determined, while particular data values are studied only in relationship to this "average." In some fields, there are so many data that individual values may never be looked at on their own. In seismic exploration, the raw data are the amplitudes of seismic waves taken at short increments of time, and at many locations on the ground. A typical survey might have several hundred "records," each containing 200,000 or more "samples."

Progressive improvement in data acquisition and analysis eventually results in a need to revise the physical explanation of the phenomenon. This often results from observation of previously undetected subtle behavior not accounted for in current theory. There are many examples of this effect in scientific history. The Ptolemaic model of planetary motions proved adequate for centuries, but was replaced by Kepler's model of elliptical orbits which provided a better explanation for deviations from circular motion. Newton's laws of motion were a marvel when first developed, but had to be extended by Einstein to include the relativistic effects first observed in the late 19th Century.

AIMS OF DATA ANALYSIS

Once a scientific problem has been defined, the first step is to record observations of the events involved. After a number of data have been collected (perhaps over a period of time or at different locations), it may be possible to determine if there are recognizable patterns. If so, data can be extrapolated to predict future events, or to infer useful properties related to the observations. There were a number of notable successes quite early in man's history, including the prediction of the motions of the sun and moon, the discovery of areas suitable for agriculture, and the location of mineral resources.

The first aim of analysis, then, could be described as finding whether a set of data is in any sense predictable. This can be defined as determining the "structure" of the data. There may be many components to the structure; the key problem is to find those that are most significant. The characteristics of a set of data that has a structural component might include a tendency to take on preferred values, a similarity to adjacent values (in time or space), and a dependence on the value of other variables observed at the same location.

Finding a structural component in a set of data often implies that the data values are functionally related to another variable which defines position. This might be time (day to day variations in temperature or stock prices, for example), linear position (air pollution measurements along a highway), spatial position (population per unit area throughout a country), or indeed any coordinate system that is relevant to the problem. The structure is that part of the measured value which can be predicted knowing the coordinates of the observation in the chosen coordinate system.

If a structure can be identified, it can be removed from the data to produce a "residual" field. In many cases, the first level of attack on the residuals is to search for unusual values (those much higher or lower than the average, for example). These abnormal events are called "anomalies." The next step is to determine if the anomalies themselves contain a structural component that might be used to refine the physical model.

Strong structural features in the data may obscure the presence of secondary effects which are equally interesting. A classic example can be drawn from astronomy. A planetary orbit is dominated by the gravitational attraction of the sun, although small perturbations due to other bodies in the Solar System can also be detected. A path that takes account only of the sun must be worked out before these variations can be studied as a separate occurrence. In fact, this type of analysis led to the discovery of the outer planets, by indicating the approximate part of the sky in which to look. It

is widely believed that there are still other planets to be discovered beyond Pluto, since our current knowledge of the orbits of the outer planets cannot be totally explained in terms of the identified bodies.

In some cases, identification of anomalies is an important step in itself, since they can point to areas of intrinsic interest. In exploring for minerals, geochemical measurements of trace element concentrations in rock or soil samples are frequently used in the first stages. After accounting for regional trends, the geologist might locate all of the high values for one element on a map to delineate areas for more detailed study. There are many possible variations, of course, such as examining ratios, or requiring that several elements have coincident anomalous readings. The key principle is that recognition of the unusual situation can be an incentive for further investigation.

Another type of structure may be present if the data can be grouped into distinct classes based on the data values. In this case, it may not be possible to predict values without first accounting for the characteristics of the groups. In the geochemical example, the "background" levels used to search for anomalies may be very different when samples from different types of rock are considered. When this is the case, the groups may be determined by independent observation, as when a geologist notes the predominant type of rock. Failing this, grouping may be attempted by investigating the statistical behavior of data. This approach is the aim of multivariate statistical methods such as factor analysis. These techniques are widely applied in analysis of the spectral data recorded by the LANDSAT satellites. Data clustering and pattern recognition can often be used to indicate areas suitable for agriculture or mineral development, by comparing poorly explored territory to locations whose potential is known.

Another problem in isolating the significant features of the data is the presence of observation errors. While these may be systematic functions of the data collecting procedure, more often they can be considered to be random "noise" added to a real "signal." The first objective of analysis is to remove the noise so that later studies can assume the data to be error-free. A major goal of any data collection procedure is to reduce the magnitude of the errors as much as possible. In general, however, one cannot expect to record noise-free data (due to the physical limitations of the instruments and the presence of natural background "noise" levels).

In any study, there is always the possibility that an unknown phenomenon will be present. In prehistoric times, the causes of everyday events were unknown, so the first steps in studying the universe naturally dealt with events such as weather, seasons, and the motions of celestial bodies. In modern science, new phenomena may not be apparent until considerable effort has been expended in analysis. In some fields, a major aim of analysis is in