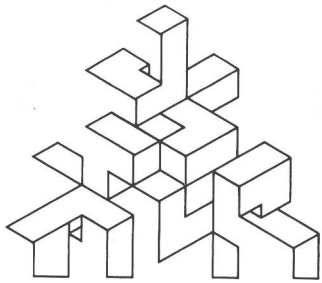


Cluster Analysis for Researchers

H. Charles Romesburg

Lifetime Learning Publications
Belmont, California
a division of Wadsworth, Inc.
London, Singapore, Sydney, Tokyo, Toronto, Mexico City

To Lafayette College, University of Pittsburgh, and Utah State University

Designer: Rick Chafian
Copyeditor: Kirk Sargent
Illustrator: John Foster
Composition: Science Typographers, Inc.

Printed in the United States of America

1 2 3 4 5 6 7 8 9 10—88 87 86 85 84

Library of Congress Cataloging in Publication Data

Romesburg, H. Charles
Cluster analysis for researchers.

Bibliography: p.
Includes index.
1. Cluster analysis. I. Title.
QA278.R66 1984 519.5'3 83-24835
ISBN 0-534-03248-6

Cluster Analysis for Researchers

Preface

PURPOSE OF THIS BOOK

This book explains and illustrates the most frequently used methods of *hierarchical cluster analysis* so that they can be understood and practiced by researchers with limited backgrounds in mathematics and statistics.

Widely applicable in research, these methods are used to determine clusters of similar objects. For example, ecologists use cluster analysis to determine which plots (i.e. objects) in a forest are similar with respect to the vegetation growing on them; medical researchers use cluster analysis to determine which diseases have similar patterns of incidence; and market researchers use cluster analysis to determine which brands of products the public perceives in a similar way. In all fields of research, there exists this basic and recurring need to determine clusters of objects.

This book will ground you in the basic methods of cluster analysis and guide you in all phases of their use. You will learn how to recognize when you have a research problem that requires cluster analysis, how to decide upon the most appropriate kind of data to collect, how to choose the best method of cluster analysis for your problem, how to obtain a computer program to perform the necessary calculations, and how to interpret the results.

INTENDED AUDIENCES

This is primarily a self-study text for researchers and graduate students in all fields of the physical sciences, the social sciences, the life sciences, as well as those in planning, management, and engineering. In addition, this is a reference book for applied statisticians and mathematicians, especially those who want to learn how researchers are using cluster analysis to solve a variety of research problems. Finally, this book will serve as a text, or a supplemental text, in applied statistics and mathematics programs for students enrolled in courses that aim to help them improve their abilities in using mathematical methods in *real* applications.

SPECIAL FEATURES

As well as examining the mathematics of cluster analysis, this book analyzes the *qualitative* aspects of cluster analysis, which help researchers to make the best choices of both data and methods for use in their research problems.

Nearly one hundred applications of cluster analysis are described in ways that nonspecialists will understand. Drawn from diverse research, these applications illustrate qualitative aspects of choosing and using methods in which researchers must be proficient if they are to make valid applications. By studying this wealth of examples, your ability to creatively use cluster analysis and other mathematical methods will improve dramatically.

ORGANIZATION OF MATERIAL

The material is divided into four parts and two appendices as follows:

Part I presents the basic calculations of cluster analysis and illustrates a variety of its uses. You will see how it can be used to a) develop interesting scientific questions; b) create research hypotheses; c) test research hypotheses; d) make general-purpose and specific-purpose classifications; and e) facilitate planning, management, and engineering.

Part II presents alternative methods for tailoring cluster analysis to specific applications. These include methods for standardizing the data, for estimating similarities among the objects to be clustered, and for clustering the objects. Also included is a chapter that shows how to report the results found using cluster analysis.

Part III shows how to use cluster analysis to make classifications. While Part I introduced this function of cluster analysis, Part III proceeds to supply considerable detail.

Part IV discusses and illustrates how to make the subjective decisions required to frame and validate applications. It shows how researchers are guided by norms within their research communities, both when they choose methods and when they validate applications.

Appendix 1 describes books and articles about cluster analysis and multivariate methods so that you can continue to learn beyond the basic methods presented in this book.

Appendix 2 describes computer programs for cluster analysis, including those in the widely available SAS and BMDP statistical packages, and including the CLUSTAR and CLUSTID programs available from Lifetime Learning Publications as a companion to this book.

There are exercises at the end of key chapters with answers given at the back of the book. For self-study, you should work the exercises and also independently rework the examples in the text.

ACKNOWLEDGMENTS

Development of the examples of applications was partly supported by a National Science Foundation grant (NSF/SER-8160606) for "Development of a Course in the Principles of Mathematical Applications, with Emphasis on Cluster Analysis." I thank the reviewers of earlier versions of the manuscript for their comments, and I also thank Lana Barr for typing it.

H. Charles Romesburg

Cluster Analysis for Researchers

To complement ***Cluster Analysis for Researchers***, Lifetime Learning is pleased to offer this statistical software for applied researchers . . .

CLUSTAR/CLUSTID:

Computer Programs for Hierarchical Cluster Analysis

by Kim Marshall and H. Charles Romesburg, Ph.D.,
both of the College of Natural Resources at Utah State University

CLUSTAR and CLUSTID are separate Fortran IV programs developed for researchers in the physical, natural, and social sciences, as well as for those in planning, management, and engineering, who use cluster analysis in their work.

CLUSTAR performs hierarchical cluster analysis. It finds clusters of objects that are defined by quantitative, qualitative, or mixed-scale attributes. CLUSTAR's features include:

- fast computing—problems with 300 objects and 300 attributes typically run in less than two minutes
- ten similarity coefficients for quantitative attributes
- fourteen similarity coefficients for qualitative attributes
- the ability to handle mixtures of quantitative and qualitative attributes
- a variety of options for standardizing the data
- the four most-used clustering methods: average linkage, single linkage, complete linkage, and Ward's minimum variance methods
- format control to read SPSS, SAS and BMPD data files. Can be used to complement these statistical packages
- produces publication-quality dendograms on both line printer and CalComp plotter
- rearrangement of data and similarity matrices according to the order in which objects cluster in the dendogram
- a variety of matrix correlation methods
- clearly labeled output
- clearly written User's Manual containing extensive examples
- designed for interactive as well as batch operating modes

CLUSTID is used to identify objects into a classification that has been created using CLUSTAR. In addition to the features of CLUSTAR, it provides:

- means, standard deviations, and ranges of data for each attribute by cluster
- identification of new objects into clusters (classes) and assesses how well they fit

(continued on overleaf)

System Requirements. The Fortran IV source code for CLUSTAR/CLUSTID is available on 9-track (EBCDIC) magnetic tape. It is designed to compile and run on a mainframe (or 32-bit scientific desktop) computer that supports a Fortran IV or Fortran 77 compiler, with no changes, or at most minimal changes, to the source code. The programs, which are well documented, have specifically been tested on the following systems: VAX 11 series, IBM 4300 series, UNIVAC 1100 series.

Pricing and Documentation. The price of CLUSTAR/CLUSTID on tape is $175.00 and includes a User's Manual, documentation and listing of source code, and example test problems with output. The User's Manual is very detailed and contains specific instructions for implementation and use, including worked-out examples. Additional copies of the User's Manual can be purchased at $8.95 per copy.

Note: These programs contain all the methods described in *Cluster Analysis for Researchers*. The tapes, while complementary to the book, are not essential for learning and understanding the methods described in the book.

ORDER FORM

Please send me

_____ *CLUSTAR/CLUSTID: Computer Programs for Hierarchical Cluster Analysis* (0-534-03420-9). I enclose $175.00, which includes the price of the *User's Manual*, and cost of shipping/handling.

_____ additional copy/ies of *User's Manual for Clustar/Clustid Computer Programs* (0-534-3251-6) @ $8.95 each. Payment is enclosed (includes shipping/handling costs)

_____ copy/ies of *Cluster Analysis for Researchers* (0-534-03248-6) @ $36.00 each.

☐ Bill me. I'll return any books I don't want within the 15-day trial period without further obligation and the invoice will be cancelled. For books I keep, I'll pay the amount above plus my local sales tax and a small amount for postage and handling.

☐ Check enclosed. Publisher pays postage and handling. 15-day return privilege still applies.

☐ Charge my ☐ VISA ☐ MasterCard ☐ American Express

Card # _____ Exp. Date _____

Prices subject to change without notice. Offer valid in U.S. and Canada only. Residents of CA, KY, MA, MI, NC, NJ, NY, and WA, please add sales tax.

Please sign here: _____

We cannot process your order without your signature.

Please print to ensure proper delivery.

Name _____

Company _____

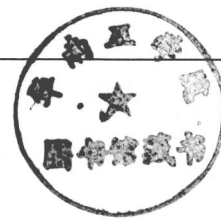
Address _____

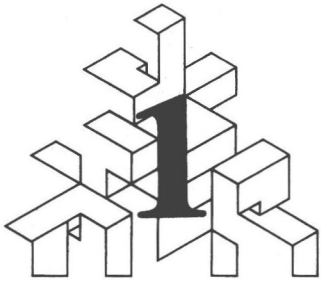
City/State/Zip _____

Detach and mail to:

LIFETIME LEARNING PUBLICATIONS
10 DAVIS DRIVE
BELMONT, CA 94002

Contents





A Road Map to This Book

OBJECTIVES

This book is an introduction to cluster analysis for researchers. It

- Presents the most often used methods of hierarchical cluster analysis so that they can be understood and applied by researchers with limited backgrounds in mathematics and statistics.
- Shows a variety of uses of cluster analysis that span the biological sciences, the social sciences, and the natural sciences, as well as planning and management.
- Explains how to use cluster analysis validly in research.

This chapter discusses

- What cluster analysis is about.
- What level of understanding you can expect to gain from this book.
- How you should use this book to meet your needs.