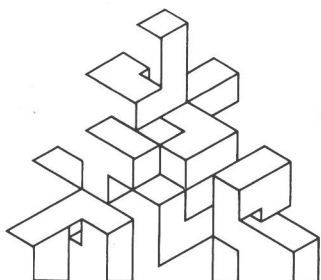




O212.4  
R763

8565525



# Cluster Analysis for Researchers

**H. Charles Romesburg**



E8565525



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.

© 1984 by Wadsworth, Inc. All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transcribed, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, Lifetime Learning Publications, Belmont, California 94002, a division of Wadsworth, 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  
ISBN 0-534-03248-6

519.5'3

83-24835

---

---

# **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 \_\_\_\_\_



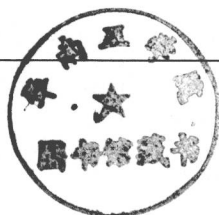
Detach and mail to:

**LIFETIME LEARNING PUBLICATIONS**  
10 DAVIS DRIVE  
BELMONT, CA 94002

---

# Contents

---



<b>Preface</b>	<b>xi</b>
<b>1. A Road Map to This Book</b>	<b>1</b>
1.1 What Cluster Analysis Is About	2
1.2 The Methods of Cluster Analysis in This Book	2
1.3 Special Features of This Book	3
1.4 How to Use This Book	5
<b>PART I. OVERVIEW OF CLUSTER ANALYSIS</b>	<b>7</b>
<b>2. Basics—The Six Steps of Cluster Analysis</b>	<b>9</b>
2.1 Step 1—Obtain the Data Matrix	10
2.2 Step 2—Standardize the Data Matrix	11
2.3 Step 3—Compute the Resemblance Matrix	11
2.4 Step 4—Execute the Clustering Method	14
2.5 Step 5—Rearrange the Data and Resemblance Matrices	23
2.6 Step 6—Compute the Cophenetic Correlation Coefficient	24
Summary	27
Exercises	28
<b>3. General Features of Cluster Analysis</b>	<b>29</b>
3.1 Relation of Cluster Analysis to Numerical Taxonomy	30
3.2 Using Cluster Analysis to Make Classifications	31
3.3 Scales of Measurement for Attributes	34
3.4 Cluster Analyzing Attributes	35
3.5 Computers and Cluster Analysis	36
Summary	37

<b>4. Applications of Cluster Analysis in Retroductive and Hypothetico-Deductive Science</b>	<b>38</b>
4.1 Research Goal 1: Create a Question	40
4.2 Research Goal 2: Create a Hypothesis	44
4.3 Research Goal 3: Test a Hypothesis	46
4.4 The Importance of Background Knowledge	51
Summary	52
<b>5. Applications of Cluster Analysis in Making Classifications</b>	<b>53</b>
5.1 Using Cluster Analysis to Make General-Purpose Classifications	55
5.2 Using Cluster Analysis to Make Specific-Purpose Classifications	59
Summary	65
<b>6. Applications of Cluster Analysis in Planning and Engineering</b>	<b>66</b>
6.1 Differences between Science, Planning, and Engineering Applications	67
6.2 Uses of Cluster Analysis in Planning	68
6.3 Uses of Cluster Analysis in Engineering	71
Summary	72
Exercise	73
<b>PART II. HOW TO DO CLUSTER ANALYSIS IN DEPTH</b>	<b>75</b>
<b>7. Standardizing the Data Matrix</b>	<b>77</b>
7.1 Reasons for Standardization	78
7.2 Standardizing Functions	78
7.3 Standardizing for <i>Q</i> -analysis and <i>R</i> -analysis	88
7.4 Data Transformations and Outliers	89
Summary	91
Exercises	92
<b>8. Resemblance Coefficients for Quantitative Attributes</b>	<b>93</b>
8.1 Data Profiles	94
8.2 Resemblance Coefficients for Quantitative Attributes	96
8.3 How Resemblance Coefficients are Sensitive to Size Displacements Between Data Profiles	104
8.4 Examples of Applications Using Quantitative Resemblance Coefficients	108

8.5 How Standardizing Can Compensate for Unwanted Size Displacements in Data Profiles	114
8.6 How to Handle Missing Data Matrix Values	115
Summary	117
Exercises	117
<b>9. Clustering Methods</b>	<b>119</b>
9.1 SLINK Clustering Method	120
9.2 CLINK Clustering Method	123
9.3 Clustering with Similarity and Dissimilarity Coefficients	128
9.4 Ward's Minimum Variance Clustering Method	129
9.5 Other Clustering Methods	135
9.6 Chaining	137
Summary	139
Exercise	139
<b>10. Resemblance Coefficients for Qualitative Attributes</b>	<b>141</b>
10.1 Qualitative Resemblance Coefficients	142
10.2 The Role of 0-0 Matches	154
10.3 Multistate Attributes	158
10.4 The Information in Qualitative Attributes	159
Summary	162
Exercises	162
<b>11. Special Resemblance Coefficients for Ordinal-Scaled Attributes</b>	<b>164</b>
11.1 Kendall's Tau Coefficient, $\tau_{jk}$	165
11.2 An Example of Kendall's Tau Coefficient	167
Summary	169
Exercise	169
<b>12. Resemblance Coefficients for Mixtures of Quantitative and Qualitative Attributes</b>	<b>170</b>
12.1 Strategies	171
12.2 An Example of the Combined Resemblance Matrix Approach	173
Summary	176
<b>13. Bypassing the Data Matrix</b>	<b>178</b>
13.1 Reasons for Bypassing the Data Matrix	179
13.2 When Using the Data Matrix Is Inconvenient	180

13.3 When Using the Data Matrix Is Impossible	181
Summary	184

#### **14. Matrix Correlation** **185**

14.1 Research Goals Involving the Correlation of Resemblance Matrices	188
14.2 Cluster-Analyzing Resemblance and Cophenetic Matrices	190
14.3 Alternatives to the Cophenetic Correlation Coefficient	190
Summary	192

#### **15. How to Present the Results of a Cluster Analysis** **193**

15.1 Key Information	194
15.2 Optional Information	195
15.3 Ways of Presenting the Tree	198
Summary	199

### **PART III. HOW TO USE CLUSTER ANALYSIS TO MAKE CLASSIFICATIONS** **201**

#### **16. How to Make Classifications** **203**

16.1 Steps in Making a Classification	204
16.2 Kinds of Classifications	206
16.3 Weighting Attributes	211
16.4 How to Determine Where to Cut the Tree	213
Summary	215

#### **17. How to Identify Objects into a Classification** **217**

17.1 How to Identify Objects	218
17.2 Strategies for Clustering "Too Many" Objects	220
Summary	221

#### **18. Philosophy of Classification and Identification** **223**

18.1 Classifications as Islands in Attribute Spaces	224
18.2 The Uses of Classifications in Research	225
18.3 How Identification Helps Improve Classifications	227
18.4 How Attribute Spaces Affect Classifications	228
18.5 How to Choose and Sample an Attribute Space	229
Summary	233

---

<b>PART IV. CLUSTER ANALYSIS—PHILOSOPHY</b>	<b>235</b>
<b>19. The Six Steps of Research</b>	<b>237</b>
19.1 A Model of Research	238
19.2 Objectivity and Subjectivity in Research	241
Summary	242
<b>20. The Roles of Information, Tolerances, and Norms in Research</b>	<b>244</b>
20.1 Information	245
20.2 Tolerances and Norms	246
Summary	252
<b>21. Framing and Validating in Cluster Analysis</b>	<b>253</b>
21.1 The Nature of Framing and Validating	254
21.2 Primary and Secondary Validity	256
Summary	259
<b>22. Examples Illustrating How to Frame and Validate Applications of Cluster Analysis</b>	<b>261</b>
22.1 Research Goal 1: Create a Question	262
22.2 Research Goal 2: Create a Hypothesis	263
22.3 Research Goal 3: Test a Hypothesis	266
22.4 Research Goal 4: Make a General-Purpose Classification	267
22.5 Research Goal 5: Make a Specific-Purpose Classification	269
22.6 Research Goal 6: Facilitate Planning and Management	274
Summary	276
<b>23. The Orders of Patterns of Similarity</b>	<b>277</b>
23.1 What a Pattern of Data Is	278
23.2 Three Orders of Patterns of Similarities	278
23.3 How Patterns Become Information	282
23.4 Some Examples of the Orders of Patterns in Cluster Analysis	283
Summary	286

---

<b>APPENDIX 1. BOOKS AND ARTICLES ON CLUSTER ANALYSIS AND OTHER MULTIVARIATE METHODS</b>	<b>288</b>
A1.1 Books	288
A1.2 Articles	291
<b>APPENDIX 2. COMPUTER PROGRAMS FOR CLUSTER ANALYSIS</b>	<b>292</b>
A2.1 SAS Programs	295
A2.2 BMDP Programs	296
A2.3 CLUSTAN Programs	298
A2.4 NTSYS Programs	299
A2.5 CLUSTAR and CLUSTID Programs	300
A2.6 Other Cluster Analysis Programs	304
<b>ANSWERS TO EXERCISES</b>	<b>305</b>
<b>GLOSSARY OF TERMS</b>	<b>314</b>
<b>REFERENCES</b>	<b>318</b>
<b>INDEX</b>	<b>330</b>



---

## 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.