Preface

In the last ten years, there has been an explosive development of clustering techniques and an increasing range of applications of these techniques. Yet there are at present only three books in English on clustering: the pioneering *Numerical Taxonomy* by Sokal and Sneath, oriented toward biology; *Cluster Analysis* by Tryon and Bailey, oriented toward psychology; and *Mathematical Taxonomy* by Jardine and Sibson. The principal contribution of statisticians has been in the area of discriminant analysis, the problem of assigning new observations to known groups. The more difficult, more important, and more interesting problem is discovery of the groups in the first place. Although modern clustering techniques began development in biological taxonomy, they are generally applicable to all types of data. They should be used routinely in early descriptions of data, playing the same role for multivaiate data that histograms play for univariate data.

The origin of this book was the general methodology lecture on clustering, given at the invitation of Dr. S. Greenhouse to the December 1970 meeting of the American Statistical Association. The notes for the lecture were used in a seminar series at Yale in early 1971 and later for a number of seminars given on behalf of the Institute of Advanced Technology, Control Data Corporation, at various places in the United States and overseas. These intensive two-day seminars required the preparation of more detailed notes describing algorithms, discussing their computational properties, and listing small data sets for hand application of the algorithms. After considerable evolution, the notes became the basis for the first draft of the book, which was used in a course at Yale University in 1973.

One difficulty of the two-day seminars was describing the various algorithms explicitly enough for students to apply them to actual data. A comprehensible but unambiguous description is needed before the algorithm can sensibly be discussed or applied. The technique eventually developed was the step-by-step description used in this book, an amalgamation of verbal description and Fortran notation. These descriptions form the skeleton of the book, fleshed out by applications, evaluations, and alternative techniques.

The book could be used as a textbook in a data analysis course that included some work on clustering or as a resource book for persons actually planning to do some clustering. The chapters are pretty well independent of each other, and therefore the one or two chapters containing algorithms of special interest may be read alone. On the other hand, the algorithms become increasingly complex as the book proceeds, and it is easier to work up to the later chapters via the early chapters.

Fortran programs implementing the algorithms described in the chapter are listed at the end of each chapter. An attempt has been made to keep these programs machine independent, and each program has been run on several different data sets, but my deficiencies as a programmer and comment card writer could make the programs tricky to use. The ideal user is an experienced Fortran programmer who is willing to adapt the programs to his own needs.

I am indebted to G. E. Dallal, W. Maurer, S. Schwager, and especially D. A. Meeter, who discovered many errors in facts and style in the first draft of the book. I am also indebted to Mrs. Barbara Amato, who cheerfully typed numerous revisions.

April 1974 JOHN A. HARTIGAN