CR-171635

E83-10310

AgRISTARS

IT-L3-04390
JSC-18592

Inventory Technology Development

A Joint Program for Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing

January 1983

A NONPARAMETRIC CLUSTERING TECHNIQUE WHICH ESTIMATES THE NUMBER OF CLUSTERS

(E83-10310) A NONPARAMETRIC CLUSTERING TECHNIQUE WHICH ESTIMATES THE NUMBER OF CLUSTERS (Lockheed Engineering and Management) 19 p HC A02/MF A01 CSCL 02C
N83-27300
Unclas
G3/43 00310

D. B. Ramey
Lockheed Engineering and Management Services Company, Inc.

Earth Resources Applications Division
Lyndon B. Johnson Space Center
Houston, Texas 77058

ORIGINAL PAGE IS
OF POOR QUALITY

1. Report No.
IT-L3-04390; JSC-18592

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle
A Nonparametric Clustering Technique Which Estimates the Number of Clusters

5. Report Date
January 1983

6. Performing Organization Code

7. Author(s)
D. B. Ramey

8. Performing Organization Report No.
LEMSCO-18387

9. Performing Organization Name and Address
Lockheed Engineering and Management Services Company, Inc.
1830 NASA Road 1
Houston, Texas 77258

10. Work Unit No.

11. Contract or Grant No
NAS 9-15800

12. Sponsoring Agency Name and Address
National Aeronautics and Space Administration
Lyndon B. Johnson Space Center
Houston, Texas 77058 Technical Monitor: J. L. Dragg

13. Type of Report and Period Covered
Technical Report

14. Sponsoring Agency Code

15. Supplementary Notes
The Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing is a joint program of the U.S. Department of Agriculture, the National Aeronautics and Space Administration, the National Oceanic and Atmospheric Administration (U.S. Department of Commerce), the Agency for International Development (U.S. Department of State), and the U.S. Department of the Interior.

16. Abstract

In applications of cluster analysis, one usually needs to determine the number of clusters, K, and the assignment of observations to each cluster. This paper presents a clustering technique based on recursive application of a multivariate test of bimodality which automatically estimates both K and the cluster assignments.

17. Key Words (Suggested by Author(s))

bimodality
cluster analysis
data reduction
density estimate
minimum spanning tree
modes
normal mixtures
split criterion
test statistic

18. Distribution Statement

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
19

22. Price*

*For sale by the National Technical Information Service, Springfield, Virginia 22161

ORIGINAL PAGE IS
OF POOR QUALITY

IT-L3-04390
JSC-18592

A NONPARAMETRIC CLUSTERING TECHNIQUE WHICH ESTIMATES THE NUMBER OF CLUSTERS

Job Order 72-422

This report describes Technological Development activities of the Inventory Technology Development project of the AgRISTARS program.

PREPARED BY

D. B. Ramey

APPROVED BY

Thomas C Baker Jr
T. C. Baker, Supervisor
Statistical Analysis Section

Mike Pore
M. D. Pore, Manager
Inventory Technology Development Department

LOCKHEED ENGINEERING AND MANAGEMENT SERVICES COMPANY, INC.

Under Contract NAS 9-15800

For

Earth Resources Applications Division

Space and Life Sciences Directorate

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION
LYNDON B. JOHNSON SPACE CENTER
HOUSTON, TEXAS

January 1983

LEMSCO-18387

PREFACE

The Agriculture and Resources Inventory Surveys Through Aerospace Remote Sensing is a multiyear program of research, development, evaluation, and application of aerospace remote sensing for agricultural resources, which began in fiscal year 1980. This program is a cooperative effort of the U.S. Department of Agriculture, the National Aeronautics and Space Administration, the National Oceanic and Atmospheric Administration (U.S. Department of Commerce), the Agency for International Development (U.S. Department of State), and the U.S. Department of the Interior.

The research which is the subject of this document was in part performed at Yale University with the support of the Yale Statistics Department. Further work was performed for the Earth Resources Applications Division, Space and Life Sciences Directorate, Lyndon B. Johnson Space Center, National Aeronautics and Space Administration. The tasks performed by Lockheed Engineering and Management Services Company, Inc., were accomplished under Contract NAS 9-15800, while the tasks performed at Yale University were supported by the National Science Foundation, through grant number DC-75-08374.

The following scientists and other personnel contributed to this work:

M. D. Pore and J. H. Smith of Lockheed Engineering and Management Services Company, Inc.

The author gratefully acknowledges the work of the following in support of this task:

J. A. Hartigan of the Department of Statistics, Yale University, New Haven, Connecticut, for his contributions.

ii iii iv

CONTENTS

Section Page

1. INTRODUCTION.................... 1

2. CURRENT LITERATURE.................... 1

3. A NONPARAMETRIC TEST OF BIOMODALITY.................... 2

4. A NONPARAMETRIC CLUSTERING TECHNIQUE.................... 5

5. AN EXAMPLE.................... 7

6. CONCLUDING REMARKS.................... 7

7. REFERENCES.................... 13

TABLES

Table Page

I PERCENTILES OF THE B STATISTIC FOR $X_1, \cdots, X_n$ GAUSSIAN RANDOM VECTORS.................................. 3

II NUMBER OF CLUSTERS FOUND BY THE NONPARAMETRIC CLUSTERING TECHNIQUE FOR 50-PERCENT MIXTURES OF BIVARIATE NORMALS WITH IDENTITY COVARIANCE MATRIX AND SEPARATION D BETWEEN THE MEANS OF THE COMPONENT NORMAL DISTRIBUTIONS.................................. 6

III MEASUREMENTS ON 150 IRISES.................................. 8

FIGURES

Figure Page

1 Plots of the percentiles of the B statistic versus $\log_e(n)$

(a) $X_1, \cdots, X_n$ bivariate normal with identity covariance matrix..... 4

(b) $X_1, \cdots, X_n$ from a 50-percent mixture of bivariate normals with identity covariance matrices and separation 6 between their means.................................. 4

2 Histograms of four measurements on each of 150 irises

(a) Sepal length.................................. 9

(b) Sepal width.................................. 9

(c) Petal length.................................. 10

(d) Petal width.................................. 10

3 Bivariate plots of the iris data

(a) Sepal length versus sepal width.................................. 11

(b) Sepal length versus petal length.................................. 11

(c) Sepal length versus petal width.................................. 11

(d) First two principal components.................................. 12

(e) First and third principal components.................................. 12

1. INTRODUCTION

The problem of assigning multivariate observations $X_1, \cdots, X_n$ to two or more clusters based on some measure of similarity arises in many fields. This assignment often results in a data reduction by allowing the researcher to study clusters rather than individuals. For example, in agricultural applications of remotely sensed data, cluster analysis is frequently performed to group agricultural fields into clusters. This grouping allows the image analyst to label clusters rather than each field, thus reducing the amount of data to be processed. Cluster analysis is also performed to discover the number of crops in an agricultural scene.

Typically, cluster analysis is performed to determine both the number of clusters, K, and the best partition of $X_1, \cdots, X_n$ into K clusters. However, most current clustering techniques either do not directly address the problem of estimating K, or they make unrealistic assumptions on the data.

In this paper we propose a multivariate clustering technique related to the density methods which automatically estimates both K and the assignment of observations to clusters. The technique is based on the use of a nonparametric multivariate test of bimodality as a splitting criterion. Because the method makes no parametric assumptions regarding the underlying density, it is less sensitive to nonnormality than the maximum likelihood techniques. The computational burden is roughly the same.

2. CURRENT LITERATURE

The current literature on the problem of estimating K is too diverse to be covered in this paper. We present only a brief summary of the current approaches and refer the interested reader to Dubes and Jain (ref. 1) or Ramey (ref. 2) for a more comprehensive review.

The earliest and probably most widely used method of estimating K consists of using some criterion, such as the criterion for cluster formation, as the basis for a test of significance. A typical example may be found in Engelman and Hartigan (ref. 3), where a test is described which is based on the ratio of the between clusters to the within cluster sums of squares.

A second approach assumes $X_1, \cdots, X_n$ are a realization from a mixture of multivariate normals. The estimation problem is then formalized as the problem of determining the number of normal components in the mixture distribution. This approach was suggested by Day (ref. 4) and Wolfe (ref. 5). More recently, Binder (ref. 6) and Postaire and Vasseur (ref. 7) have developed a Bayesian version of this approach.

The third approach is based on the assumption that $X_1, \cdots, X_n$ come from some continuous density. Observations are said to come from the same cluster if they lie in the same connected high-density region. The number of clusters is taken to be the number of modes in the density function. Hartigan (ref. 8) suggests a univariate implementation of this approach based on isotonic regression techniques. Good and Gaskins (ref. 9) and Silverman (ref. 10) suggest other approaches to the univariate problem, while Goldberg and Shlien (ref. 11) treat the multivariate case using histogram estimates of the density. The technique presented in this paper is also an example of this density-based approach.

3. A NONPARAMETRIC TEST OF BIMODALITY

Ramey (ref. 2) develops a multivariate likelihood ratio test of bimodality which will be referred to as the B statistic. The development is as follows.

First, suppose $X_1, \cdots, X_n$ are random vectors from some multivariate distribution with continous density function f. Calculate the maximum likelihood estimate f_u of f, subject to the constraint of unimodality, by assuming f is supported by the links of the minimum spanning tree. Recall the minimum spanning tree on $X_1, \cdots, X_n$ is the graph of minimum total length which spans $X_1, \cdots, X_n$. If $X_1, \cdots, X_n$ come from a distribution with continuous density, then with probability 1, the minimum spanning tree is unique. This assumption is made to narrow the class of density estimates sufficiently so that the constrained maximum likelihood estimate of f is easily calculated. Next, calculate a second maximum likelihood estimate f_b of f, subject to the constraint of bimodality, by again assuming f is supported by the minimum spanning tree.

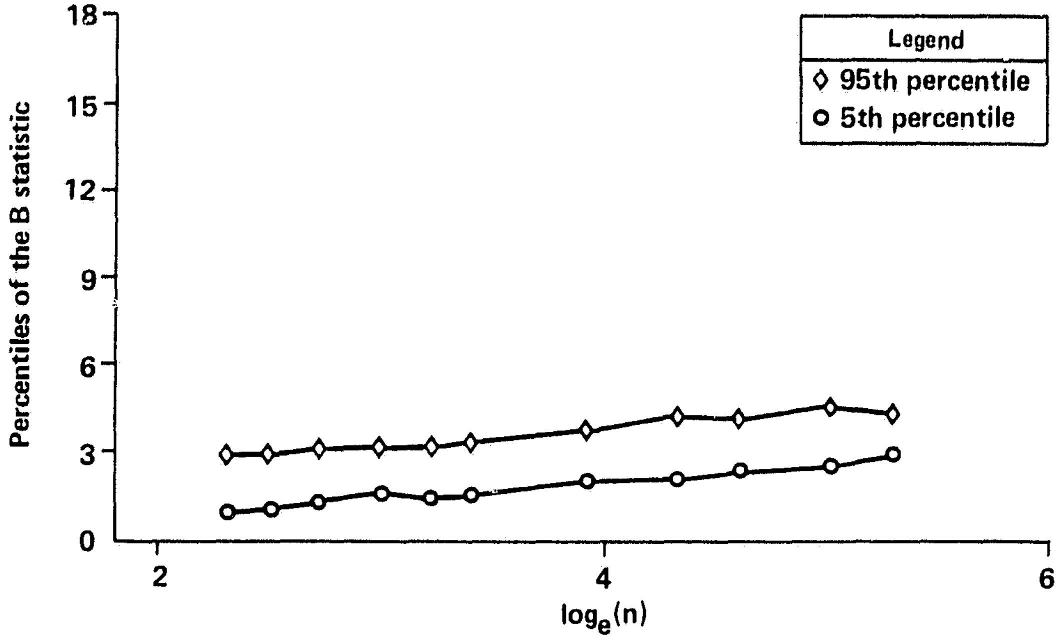
Given the two estimates f_u and f_b , then the B statistic is the difference in the two log likelihoods. The theoretical distribution of B is not known at this time; Ramey gives tables of the empirical distribution for the bivariate case for various values of n and for both Gaussian and uniform [on $(0,1)^2$] random vectors. Table I reproduces the results for samples from the bivariate normal. The results for uniforms on $(0,1)^2$ are similar.

Figure 1 displays plots of the distribution of the B statistic versus $\log_e(n)$ for both the unimodal case of $X_1, \cdots, X_n$ from a bivariate normal, and the bimodal case of $X_1, \cdots, X_n$ from a 50-percent mixture of bivariate normals. Only the 5th and 95th percentiles are shown. The plots indicate that for moderate values of n (for instance, n = 150), the test distinguishes between the two parent distributions at least 99 percent of the time.

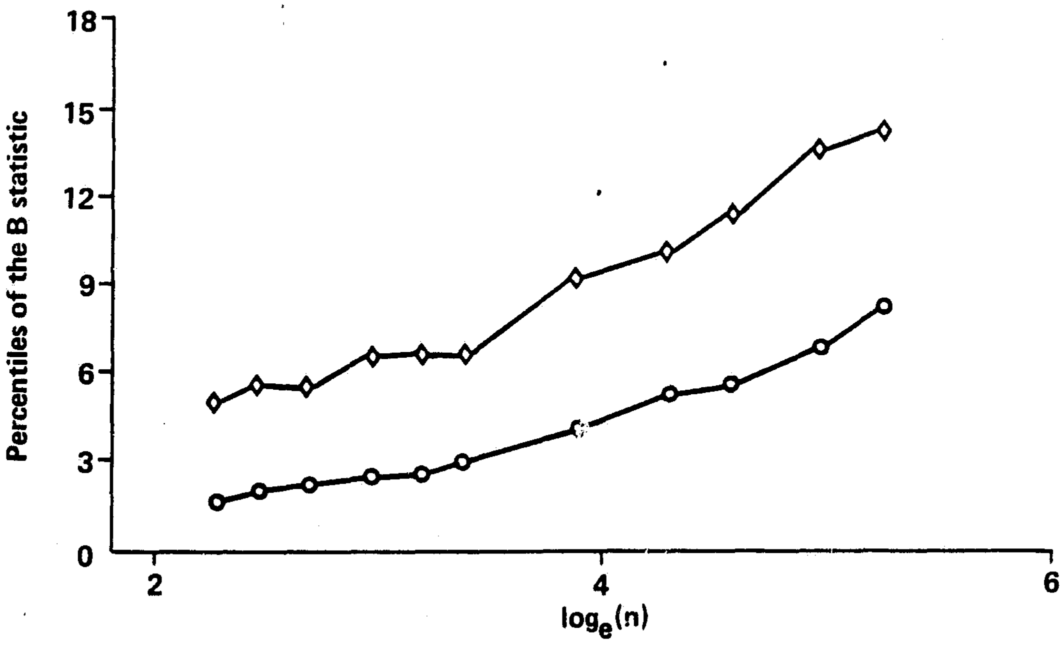
TABLE I.- PERCENTILES OF THE B STATISTIC FOR
$X_1, \cdots, X_n$ GAUSSIAN RANDOM VECTORS

[Based on 100 samples of size n from the bivariate normal]

n	Percentiles			
	50%	75%	90%	95%
10	1.63	1.99	2.46	3.06
12	1.80	2.05	2.43	2.86
15	1.99	2.33	2.83	3.07
20	2.24	2.49	2.93	3.10
25	2.07	2.46	2.88	3.12
30	2.26	2.65	3.08	3.33
50	2.64	3.01	3.36	3.67
100	2.85	3.21	3.59	4.23
150	3.01	3.39	3.83	4.10
200	3.23	3.74	4.12	4.56



(a) X_1, •••, X_n bivariate normal with identity convariance matrix.



(b) X_1, •••, X_n from a 50-percent mixture of bivariate normals with identity covariance matrices and separation 6 between their means.

Figure 1.- Plots of the percentiles of the B statistic versus $\log_e(n)$.

As can be seen in Ramey (ref. 2),

$$B < .562 \log_e(n) + 1.65$$

with high probability for bivariate random vectors from either a uniform or a normal distribution. While a formal proof does not yet exist, it seems clear that this relationship will hold for higher dimensions, since increasing the dimensionality of the space should decrease B. In addition, the relationship should hold for other unimodal distributions since the uniform is a limiting case of both bimodal and unimodal distributions.

4. A NONPARAMETRIC CLUSTERING TECHNIQUE

The B statistic of the previous section may be used recursively as a splitting criterion. That is, we take an initial partition of the data consisting of the entire set of observations and test for bimodality. If the test is significant, then we obtain a new partition by splitting the data set into two subsets and repeat the process until no further significant B statistics may be calculated. The entire algorithm may be stated:

a. Fit the minimum spanning tree to $X_1, \cdots, X_n$. Thus, initially, there is one tree defined on the nodes.

b. Calculate the B statistic for one of the trees on the data. A side result of this calculation is the maximum likelihood estimate f_b of f, subject to bimodality and to the constraint that it is supported by the minimum spanning tree. The function f_b will assign 0 mass to some link of the spanning tree. This link is a candidate for removal from the tree.

c. If B is not significant, remove the tree on which B was calculated and its nodes from further consideration, and then go to step e.

d. If B is significant, then remove the link with 0 mass from the tree under consideration, giving two trees for further testing.

e. If no trees remain, then terminate; otherwise, select one of the remaining trees and repeat steps b through d.

Table II illustrates the performance of this clustering technique for various 50-percent mixtures of bivariate normals with identity covariance matrices and separation D between the distribution means. For no separation, that is, for a bivariate normal, and for 25 or more nodes, the procedure correctly indicates one cluster at least 99 percent of the time, while for n = 75 or more and separation 6 between the means of the two components, the procedure correctly finds exactly two clusters at least 90 percent of the time.

In interpreting the results for separations of 2 and 4 between the means, it should be remembered that normal mixtures are not bimodal unless the separation between the normal component means is sufficiently large (slightly

TABLE II.- NUMBER OF CLUSTERS FOUND BY THE NONPARAMETRIC CLUSTERING TECHNIQUE FOR 50-PERCENT MIXTURES OF BIVARIATE NORMALS WITH IDENTITY COVARIANCE MATRIX AND SEPARATION D BETWEEN THE MEANS OF THE COMPONENT NORMAL DISTRIBUTIONS

[Based on 100 samples]

Separation between the means	Number of clusters found	Number of observations (n)				
		25	50	75	100	150
0.	1	100	99	100	100	100
	2	0	0	0	0	0
	3	0	1	0	0	0
	4	0	0	0	0	0
2.	1	99	99	100	100	99
	2	1	1	0	0	1
	3	0	0	0	0	0
	4	0	0	0	0	0
4.	1	89	82	80	78	80
	2	9	17	20	22	20
	3	2	1	0	0	0
	4	0	0	0	0	0
6.	1	40	15	7	3	0
	2	54	85	90	94	99
	3	6	0	2	3	1
	4	0	0	1	0	0

greater than 2.7 for bivariate normals with identity covariance matrices). Thus the procedure should find only one cluster for separation 2.

5. AN EXAMPLE

A typical use of cluster analysis is the analysis of biological measurements to assist in defining taxonomic relationships between species. The use of the nonparametric technique is illustrated on a data set consisting of four measurements on each of 150 irises. These data, from Fisher (ref. 12), are displayed in table III.

Before performing the cluster analysis, we calculate histograms on each of the measurements and plot all pairs of variables and the first three principal components. The histograms are displayed in figures 2(a) through 2(d), and selected plots in figures 3(a) through 3(e). From the graphical analysis, we see a separation between *Iris setosa* and the group consisting of *Iris versicolor* and *Iris virginica*. However, while *versicolor* and *virginica* are distinguishable on the basis of the measurements, they do not form separate clusters.

The proposed clustering algorithm was used on this data set. The cluster consisting of the 50 *Iris setosa* measurements was split off first. The B statistic was calculated to be 5.68, which is greater than the minimum split criterion of 4.47 for 150 observations. Then each of the resulting two clusters was considered for an additional split. However, in each case the B statistic was less than the minimum required for a split. For the *setosa* cluster, B = 1.65, while for the other cluster, B = 2.31. Thus, the cluster analysis confirms our visual impression of two groups in the data. We do not attempt a biological interpretation of this result.

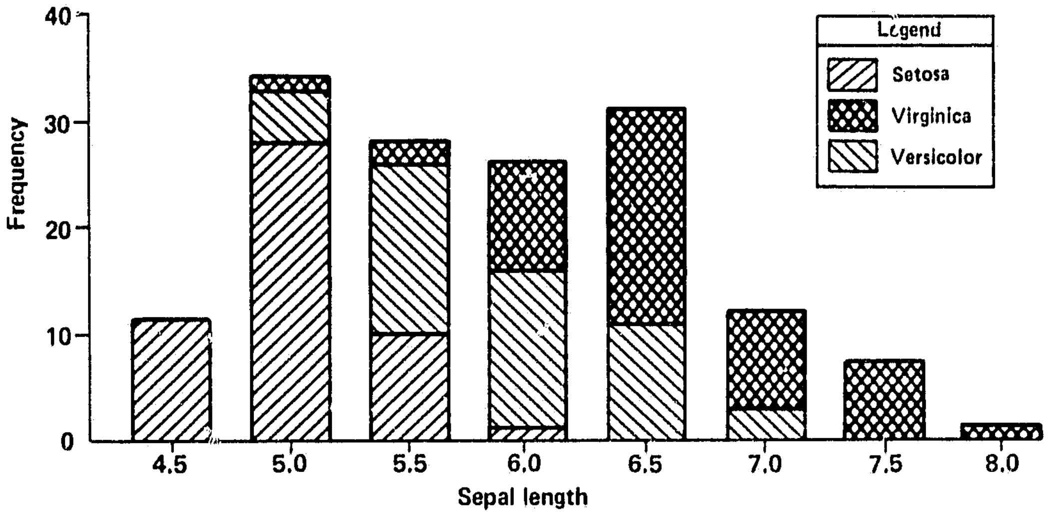
6. CONCLUDING REMARKS

The above algorithm has been programmed in Fortran IV using the IBM extended compiler. All testing and sample runs have been made on a National Advanced

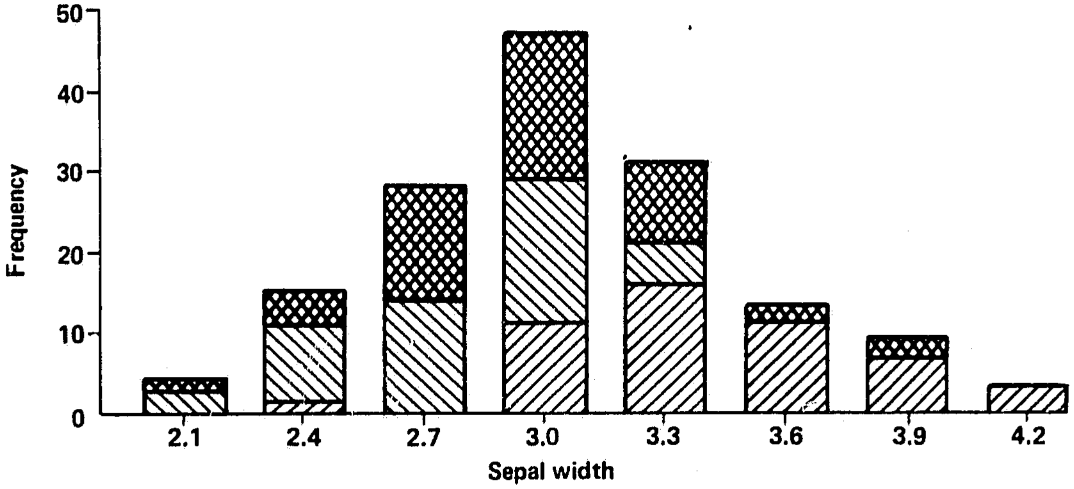
TABLE III.- MEASUREMENTS ON 150 IRISES

[Given in centimeters (from ref. 12)]

Iris setosa				*Iris versicolor*				*Iris virginica*			
Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width	Sepal length	Sepal width	Petal length	Petal width
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.4	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.8	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.0	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

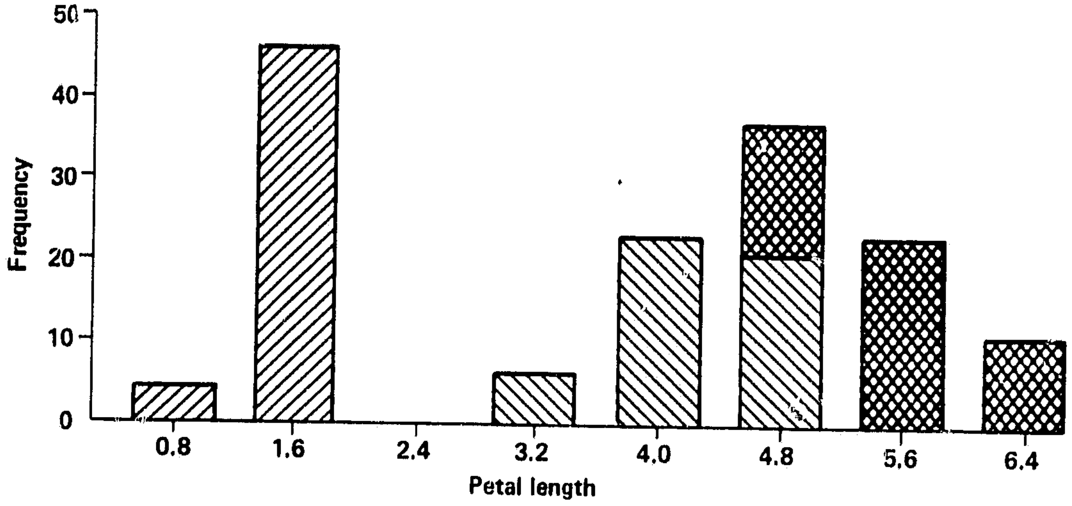


(a) Sepal length

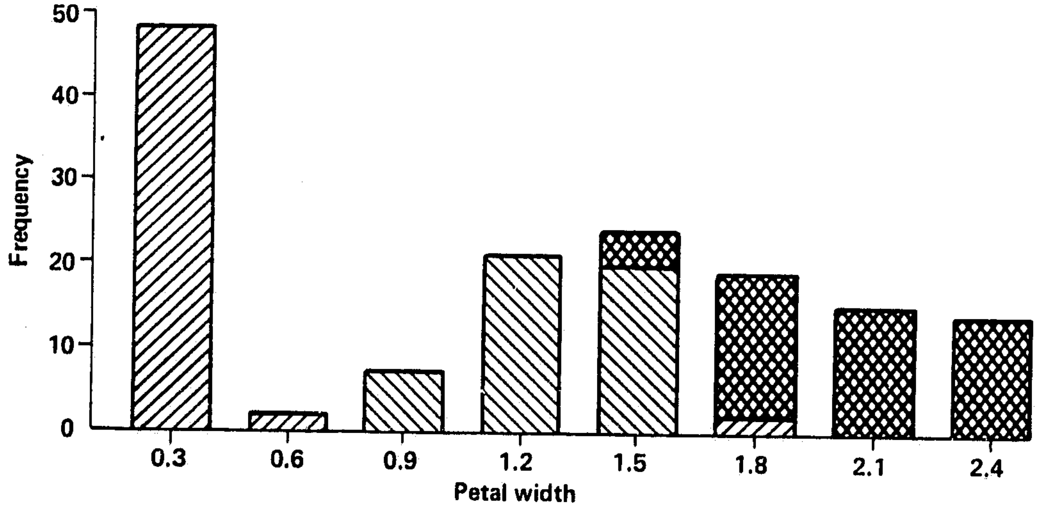


(b) Sepal width

Figure 2.- Histograms of four measurements on each of 150 irises.

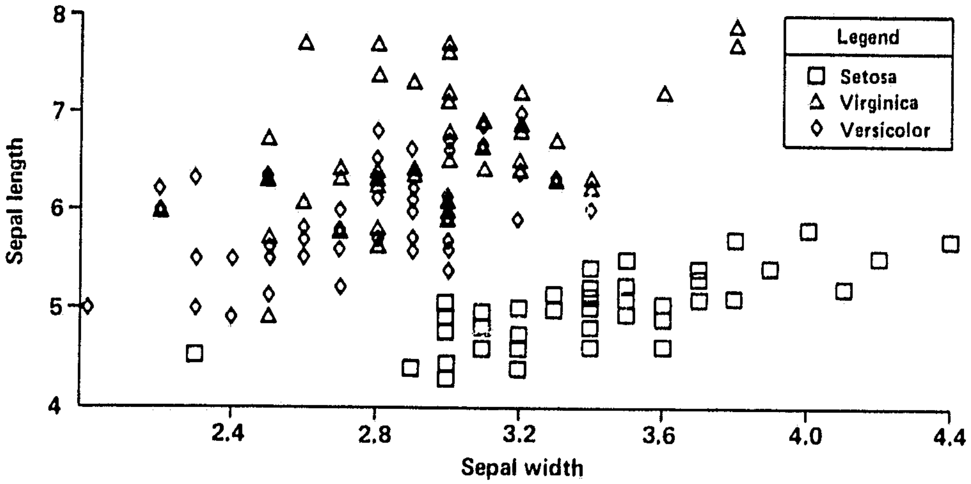


(c) Petal length

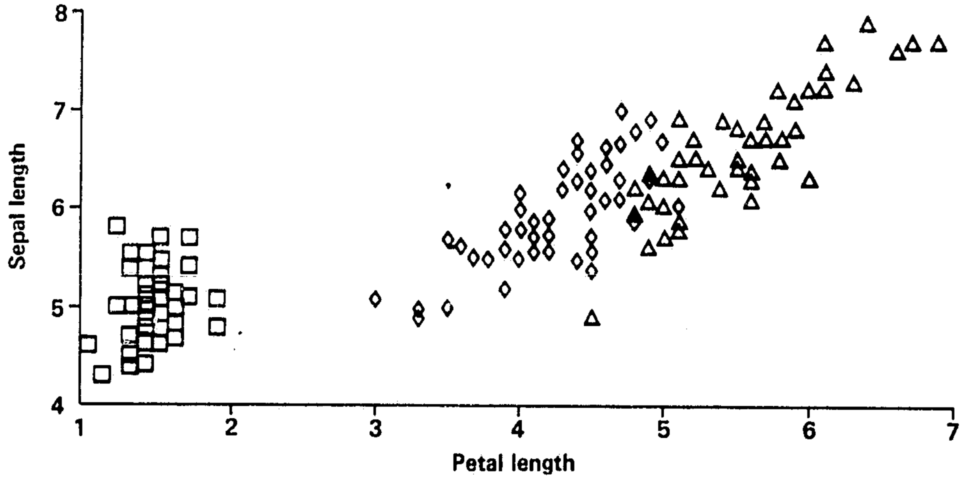


(d) Petal width

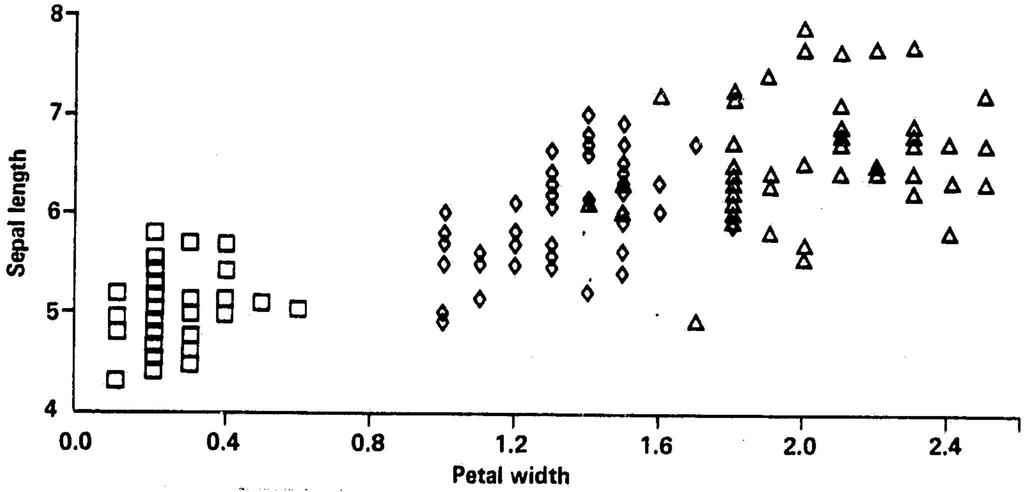
Figure 2.- Concluded.



(a) Sepal length versus sepal width

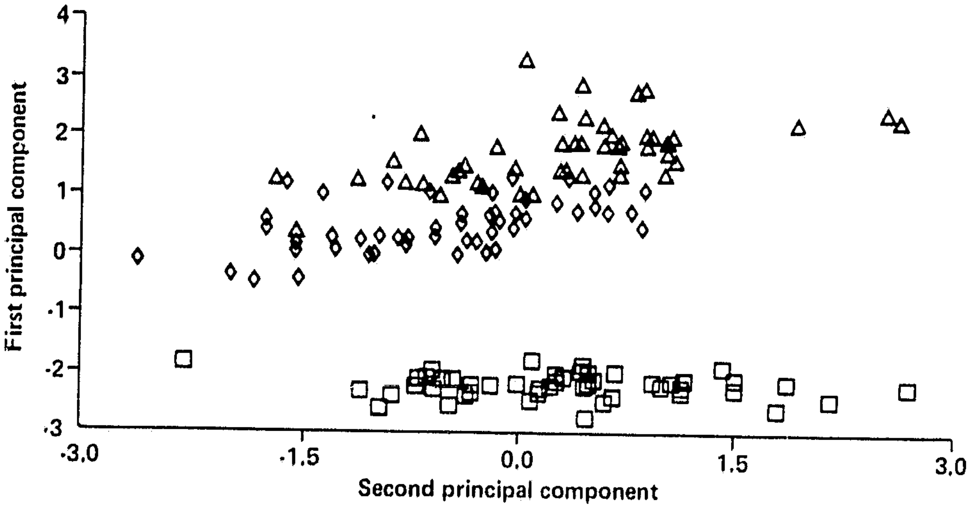


(b) Sepal length versus petal length

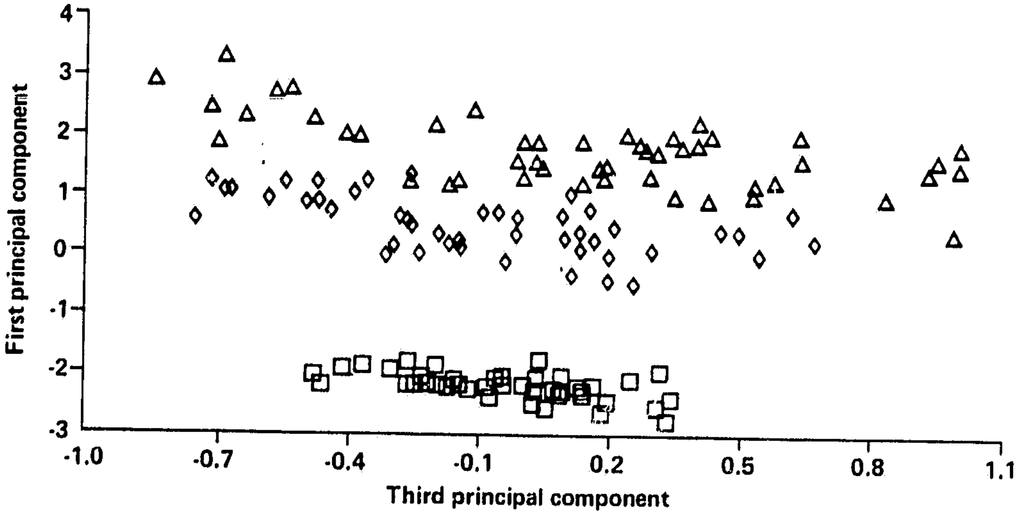


(c) Sepal length versus petal width

Figure 3.- Bivariate plots of the iris data.



(d) First two principal components



(e) First and third principal components

Figure 3.- Concluded.

Systems AS-3000 computer. Arrangements can be made with the author to obtain copies of the program.

7. REFERENCES

1. Dubes, R.; and Jain, A. K.: Validity studies in clustering methodologies. Pattern Recognition, vol. 11, 1976, pp. 235-254.

2. Ramey, D. B.: A nonparametric test of bimodality with applications to cluster analysis. Unpublished Ph.D. dissertation, Yale University, New Haven, 1982.

3. Engelman, L.; and Hartigan, J. A.: Percentage points of a test for clusters. Journal of the American Statistical Association, vol. 64, 1969, pp. 1647-1648.

4. Day, N. E.: Estimating the components of a mixture of normal distributions. Biometrika, vol. 56, 1969, pp. 463-474.

5. Wolfe, J. H.: Pattern clustering by multivariate mixture analysis. Multivariate Behavioral Research, vol. 5, 1970, pp. 329-350.

6. Binder, D. A.: Bayesian cluster analysis. Biometrika, vol. 65, 1978, pp. 31-38.

7. Postaire, J. A.; and Vasseur, C. P. A.: An approximate solution to normal mixture identification with application to pattern classification. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 3, 1981, pp. 163-179.

8. Hartigan, J. A.: Distribution problems in clustering. Classification and Clustering, Van Ryzin, ed., Academic Press (New York), 1976.

9. Good, I. J.; and Gaskins, R. A.: Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. Journal of the American Statistical Association, vol. 75, 1980, pp. 42-56.

10. Silverman, B. W.: Using kernel density estimates to investigate multimodality. Journal of the Royal Statistical Society, vol. B-43, 1981, pp. 97-99.

11. Goldberg, M; and Shlien, S.: A clustering scheme for multispectral images. IEEE Transactions on Systems, Man, and Cybernetics, vol. 8, 1978, pp. 86-92.

12. Fisher, R. A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics, vol. 7, 1936, pp. 179-188.