

**STATISTICAL INFERENCES ON INSECT PEST RESISTANCE INDICES:
IMPLICATIONS FOR SELECTION IN CROP PLANTS**

By

MARGARET NABASI RYE

BSc. Hons. (MUK), MSc. (READING).

A thesis submitted in fulfillment of the requirement for the Degree of
Doctor of Philosophy of Makerere University.

NADIO
29 AUG 1997
PO BOX 11000
KAMPALA
ACC. NO. 433

Department of Mathematics
Faculty of Science
Makerere University.

August, 1995

DECLARATION

I, the undersigned, declare that this thesis is my own original work which has not been submitted for any degree in any university and all sources of material and help have been duly acknowledged.

Signed:

Date

Margaret Nabasiye

RECEIVED
2001 JUN 25
10:00 AM
LIBRARY

DEDICATION

To the memory of my late father

Francis Drake Gita

whose departure from this world

was not a 'statistically random occurrence'.

ACKNOWLEDGEMENT

I wish to thank my employers, the National Agricultural Organisation (NARO) for granting me the study leave. For his various roles and assistance, I extend my gratitude to Dr. I. Kibirige-Ssebunya, Director, Kawanda Agricultural Research Institute (KARI).

I would also like to show my sincere gratitude to the Director of the ICIPE and Chairman of the African Regional Postgraduate Programme in Insect Science (ARPPIS) for granting me the opportunity to undertake this study at ICIPE, the successive ARPPIS coordinators, Drs Z.T. Dabrowski and V.O. Musewe, for their encouragement and material support as well as the lecturers of ARPPIS 1992 for equipping me with the entomological foundation that I required in this field of study. I deeply acknowledge the DSO of The Netherlands for financial support.

I wish to thank most sincerely Dr A. Odulaja, my principal ICIPE supervisor and Head of Biomathematics Unit, for the deep interest he took in this work, the immense guidance, critical evaluation and valuable suggestions he rendered in all aspects of the research work and its write up; and Dr. S.O. Ajala, my second ICIPE supervisor for valuable insights, constructive criticisms and encouragement as well as offering his data for the study. My University supervisor, Dr. L.S. Luboobi, is deeply appreciated for his assistance and following my

progress through valuable discussions, suggestions and encouragement throughout the study and preparation of the thesis.

To Dr. S.K. Nokoe, Biometrician, IITA, I extend my deep appreciation for his helpful contributions to the development of ideas, enlightening discussions and encouraging remarks at various stages of the research work. Thanks are also due to Prof. J.A. Odebiyi, Department of Agricultural Biology, University of Ibadan, for free-handedly providing some of the data used in this study.

The assistance of Messrs D. Munyinyi, H. Meena, O. Okello, M. Gathoga and M. Akello of the Biomathematics Unit of the ICIPE, each in his specific area, is greatly appreciated. In spite of their busy duty schedules, they were ever willing to, and indeed did, solve my computing problems which ranged from software, programming to hardware. I sincerely wish to thank them for their patience, friendliness, and tireless efforts in providing technical solutions to my problems as if they were their own. Let me take this opportunity to thank Mrs G. Gathura whose moral support until her departure from the unit was very helpful. Thanks are also due to Ms V. Manene, Training unit, for her patience and material support. My fellow colleagues of the ARPPIS 1992 class and other contemporary ARPPIS members, with whom I interacted, deserve my gratitude.

My special gratitude goes to Drs Z.M. Nyiira and E. Karamura for encouraging me to join this training program. Mr. A.P. Lugolobi, KARI, who prepared me for the coursework deserves special mention.

Lastly, but by no means least, I must extend my sincere gratitude for the invaluable contribution from my family, relatives and friends. I am particularly indebted to my dear husband, Yekoyasi, whose cooperation, patience, understanding and constant encouragement were indispensable. My special thanks go to my children namely, Geoffrey, Eric and Kate for their patience during my long absence from home. To these, I now leave the future task of writing higher-degree theses in their hands, God willing. I also take this opportunity to thank my mother, Mrs M. Gita, and my aunt, Mrs K. Kalyango, without whose assistance in various ways at virtually all stages of my academic life I would not have attained this goal.

TABLE OF CONTENTS

DECLARATION	ii
DEDICATION	iii
ACKNOWLEDGEMENT	iv
TABLE OF CONTENTS	vii
LIST OF TABLES	xiii
LIST OF FIGURES	xvii
ACKNOWLEDGEMENT	xxvi
Synonymous terms	xxviii
ABSTRACT	xxix
CHAPTER 1: INTRODUCTION	
1.1 GENERAL BACKGROUND	1
1.2 HOST PLANT RESISTANCE	2
1.3 RESISTANCE MEASURES	4
1.4 CLASSIFICATION INTO RESISTANCE GROUPS	5
1.5 STATEMENT OF THE PROBLEM AND IMPLICATIONS	6
1.6 OBJECTIVES AND SCOPE OF THE STUDY	8
CHAPTER 2: LITERATURE REVIEW	
2.1 CLASSIFICATION PROCEDURES	11
2.1.1 Single-trait classification	11
2.1.2 Multi-trait classification	14

2.1.2.1	<i>Selection indices</i>	14
2.1.2.2	<i>Other approaches</i>	21
2.2	DISTRIBUTIONAL PROPERTIES OF RESISTANCE MEASURES	24
2.3	STUDY APPROACH	28

CHAPTER 3: METHODOLOGY

3.1	UNIVARIATE PROCEDURES	30
3.1.1	Graphical procedure	30
3.1.1.1	<i>Description</i>	30
3.1.1.2	<i>Optimum number of groups</i>	32
3.1.2	Cluster analysis	33
3.1.2.1	<i>Description</i>	33
3.1.2.2	<i>Optimum number of groups</i>	34
3.1.3	Hypothesis testing and test statistics	35
3.1.3.1	<i>Description</i>	35
3.1.3.2	<i>Formulation of hypotheses</i>	36
3.1.3.3	<i>Testing the hypotheses</i>	38
3.1.3.3.1	<i>Quasi Bayesian Test (QBT) statistic</i>	38
3.1.3.3.2	<i>Likelihood ratio test (LRT) statistic</i>	40
3.1.3.4	<i>Optimum number of groups</i>	47
3.2	MULTIVARIATE PROCEDURES	48
3.2.0	The multivariate selection problem	48
3.2.1	Classification using the indices	49

3.2.2	Alternative approach	50
3.2.2.1	<i>Non-significantly correlated variables</i>	50
3.2.2.2	<i>Significantly correlated variables</i>	51
3.2.3	Classification procedures	52
3.2.3.1	<i>Graphical technique</i>	53
3.2.3.2	<i>Cluster analysis</i>	53
3.2.3.3	<i>Likelihood ratio test statistic</i>	54
3.2.4	Optimum number of groups	61
3.3	COMPARISON OF PROCEDURES	62

CHAPTER 4: DATA SETS AND THEIR PROPERTIES

4.1	REAL DATA	63
4.1.1	Description of real data	63
4.1.2	Empirical distribution of data	65
4.1.3	Analysis of variance	70
4.1.4	Correlations among variables	74
4.1.5	Comment on real data	74
4.2	SIMULATED DATA	76

CHAPTER 5: APPLICATIONS AND RESULTS

5.1	EVALUATION OF UNIVARIATE PROCEDURES	80
5.1.1	The graphical procedure	81
5.1.2	Cluster analysis	81
5.1.3	Hypothesis testing procedure	82
5.1.3.1	<i>Generation of groupings</i>	82

5.1.3.2	<i>Computation of test statistics</i>	84
5.2	RESULTS AND DISCUSSION OF UNIVARIATE PROCEDURES	85
5.2.1	DATA SET A	86
5.2.1.1	Leaf feeding (A - LF)	86
5.2.1.2	Dead heart (A - DH)	87
5.2.1.3	Stem tunnelling (A - ST)	88
5.2.2	DATA SET B	100
5.2.2.1	Leaf feeding (B - LF)	100
5.2.2.2	Dead heart (B - DH)	101
5.2.2.3	Stem tunnelling (B - ST)	102
5.2.3	DATA SET C	112
5.2.3.1	Leaf feeding (C - LF)	112
5.2.3.2	Dead heart (C - DH)	113
5.2.3.3	Stem tunnelling (C - ST)	114
5.2.4	DATA SET D	125
5.2.4.1	Time to 50% emergence (D - T)	125
5.2.4.2	Number of adult insects (D - N)	127
5.2.4.3	Weight loss (D - WL)	127
5.2.5	Summary of results for univariate analyses	138
5.3	MULTIVARIATE ANALYSES	141
5.3.1	RESULTS AND DISCUSSION	143
5.3.1.1	DATA SET A	143
5.3.1.2	DATA SET B	161

5.3.1.3	DATA SET C	174
5.3.1.4	DATA SET D	184
5.3.2	Summary of results for multivariate analyses	195
5.4	COMPARISON OF PROCEDURES USING SIMULATED DATA	197
5.4.2	Number of groupings	197
5.4.2	Results	199
5.4.2.1	<i>Univariate classifications</i>	199
5.4.2.2	<i>Multivariate classifications</i>	203
5.4.3	Discussion of results from the simulations	214
CHAPTER 6: GENERAL DISCUSSION AND CONCLUSIONS		
6.1	DISTRIBUTIONAL ASSUMPTIONS	218
6.2	ANALYSES OF VARIANCE	218
6.3	CLASSIFICATION PROCEDURES	219
6.3.1	Single-trait	219
6.3.1.1	<i>Normal score plots</i>	219
6.3.1.2	<i>Tests statistic</i>	220
6.3.1.3	<i>Cluster analysis</i>	222
6.3.2	Multi-trait selections	222
6.3.2.1	<i>Procedures</i>	222
6.3.2.2	<i>Indices</i>	223
6.4	CONCLUSIONS	223
6.5	RELEVANCE OF RESULTS FROM THE STUDY	225

6.6	SUGGESTIONS FOR FURTHER WORK	226
	REFERENCES	228
	APPENDICES	246

LIST OF TABLES

Table 2.1	ANOVA and expected mean squares (EMS)	17
Table 3.1	Summary of the hypotheses testing for the equality of g means ($g = 2, 3, \dots$) with sizes $n_1, n_2, \dots, n_g$, ($j = 1, \dots, g$, $k = 1, \dots, n_j$, $n = \sum n_j$).	44
Table 3.2	One way analysis variance table.	46
Table 3.3	Multivariate analysis of variance table.	59
Table 4.1a	Analysis of variance summary for data set A.	71
Table 4.1b	Analysis of variance summary for data set B.	72
Table 4.1c	Analysis of variance summary for data set C.	73
Table 4.2	Correlations among damage variables in the different data sets.	75
Table 4.3	Simulation factor combinations.	79
Table 5.1	Number of possible groupings for different combinations of number of entries (n) and number of classes (g).	83
Table 5.2a	Computed values of g^2w for the damage variables (data set A).	86

Table 5.2b	Computed values of g^2w for the damage variables (data set B).	100
Table 5.2c	Computed values of g^2w for the damage variables (data set C).	112
Table 5.2d	Computed values of g^2w for the variables (data set D).	125
Table 5.3	Summary of optimal groupings by different procedures for the different data sets.	140
Table 5.4a	A summary of results of PCA of data set A damage variables.	143
Table 5.4b	Composition of groups based on the indices (data set A).	148
Table 5.5a	A summary of results of PCA of data set B damage variables.	161
Table 5.5b	Composition of the groups based on indices (data set B).	165
Table 5.6a	A summary of results of PCA of data set C damage variables.	174
Table 5.6b	Composition of the groups based on indices (set C).	177

Table 5.7a	A summary of results of PCA of data set D variables.	184
Table 5.7b	Composition of the groups based on indices (set D).	187
Table 5.8	Summary of optimal groupings by different indices.	196
Table 5.9	Number of groupings examined using LRT and QBT procedures.	198
Table 5.10	Frequency of indication of correct and incorrect classifications of univariate data sets (C and N represent correct and incorrect classifications, respectively).	200
Table 5.10a	Total frequencies of correct (C) and incorrect (N) classifications for different levels of s.	201
Table 5.10b	Total frequencies of correct (C) and incorrect (N) classifications by different procedures for the 2-group model.	202
Table 5.11	Frequency of correct and incorrect classifications when $r \geq 0.8$	204
Table 5.11a	Total frequencies of correct (C) and incorrect (N) classifications for different levels of s.	205
Table 5.11b	Two-way table of method by level of s.	205

Table 5.12	Frequency of correct and incorrect classifications when $0.5 < r < 0.7$	207
Table 5.12a	Total frequencies of correct (C) and incorrect (N) classifications for different levels of s.	208
Table 5.12b	A two-way table of method by levels of s.	208
Table 5.13	Frequency of correct and incorrect classifications when $r \leq 0.3$ and $p = 2$	210
Table 5.14	Frequency of correct and incorrect classifications when $r \leq 0.3$ and $p = 3$	211
Table 5.14a	Total frequencies of correct (C) and incorrect (N) classifications for different levels of s.	212
Table 5.14b	Frequencies of correct and incorrect determinations for different method-index combinations.	213
Table 5.15	Effect of intercorrelations among variables on performance of procedures.	214

LIST OF FIGURES

Figure 4.1a	Frequency distribution for A-LF showing a skewed distribution.	68
Figure 4.1b	Frequency distribution for C-LF showing a normal approximation for the variable.	68
Figure 4.1c	Frequency distribution for C-DH, highly skewed with a concentration at zero.	69
Figure 4.1d	Frequency distribution for C-ST, a skewed distribution normalised by arcsine transformation.	69
Figure 5.1a	Plot of A-LF for the 51 entries against their respective normal scores.	90
Figure 5.1b	Plot of LRT statistic for the 2-group model against hypothetical cutpoints for A-LF.	91
Figure 5.1c	Plot of PF and CCC for optimum classifications into g groups with respect to A-LF.	92
Figure 5.1d	Plot of QBT statistic for the 2-group model against hypothetical cutpoints for A-LF.	93

Figure 5.2a	Plot of A-DH for the 51 entries against their respective normal scores.	94
Figure 5.2b	Plot of LRT statistic for the 2-group model against hypothetical cutpoints for A-DH.	95
Figure 5.2c	Plot of PF and CCC for optimum classifications into g groups with respect to A-DH.	96
Figure 5.3a	Plot of A-ST for the 51 entries against their respective normal scores.	97
Figure 5.3b	Plot of LRT statistic for the 2-group model against hypothetical cutpoints for A-ST.	98
Figure 5.3c	Plot of PF and CCC for optimum classifications into g groups with respect to A-ST.	99
Figure 5.4a	Plot of B-LF for the 55 entries against their respective normal scores.	103
Figure 5.4b	Plot of LRT statistic for the 2-group model against hypothetical cutpoints for B-LF.	104

Figure 5.4c	Plot of PF and CCC for optimum classifications into g groups with respect to B-LF.	105
Figure 5.5a	Plot of B-DH for the 55 entries against their respective normal scores.	106
Figure 5.5b	Plot of LRT statistic for the 2-group model against hypothetical cutpoints for B-DH.	107
Figure 5.5c	Plot of PF and CCC for optimum classifications into g groups with respect to B-DH.	108
Figure 5.6a	Plot of B-ST for the 55 entries against their respective normal scores.	109
Figure 5.6b	Plot of LRT statistic for the 2-group model against hypothetical cutpoints for B-ST.	110
Figure 5.6c	Plot of PF and CCC for optimum classifications into g groups with respect to B-ST.	111
Figure 5.7a	Plot of C-LF for the 39 entries against their respective normal scores.	116

Figure 5.7b	Plot of LRT statistic for the 2-group model against hypothetical cutpoints for C-LF.	117
Figure 5.7c	Plot of PF and CCC for optimum classifications into g groups with respect to C-LF.	118
Figure 5.8a	Plot of C-DH for the 39 entries against their respective normal scores.	119
Figure 5.8b	Plot of LRT statistic for the 2-group model against hypothetical cutpoints for C-DH.	120
Figure 5.8c	Plot of PF and CCC for optimum classifications into g groups with respect to C-DH.	121
Figure 5.9a	Plot of C-ST for the 39 entries against their respective normal scores.	122
Figure 5.9b	Plot of LRT statistic for the 2-group model against hypothetical cutpoints for C-ST.	123
Figure 5.9c	Plot of PF and CCC for optimum classifications into g groups with respect to C-ST.	124

Figure 5.10a	Plot of D-T for the 77 entries against their respective normal scores.	129
Figure 5.10b	Plot of the LRT statistic for the 2-group model against hypothetical cutpoints for D-T.	130
Figure 5.10c	Plot of PF and CCC for optimum classifications into g groups with respect to D-T.	131
Figure 5.11a	Plot of D-N for the 77 entries against their respective normal scores.	132
Figure 5.11b	Plot of the LRT statistic for the 2-group model against hypothetical cutpoints for D-N.	133
Figure 5.11c	Plot of PF and CCC for optimum classifications into g groups with respect to D-N.	134
Figure 5.12a	Plot of D-WL for the 77 entries against their respective normal scores.	135
Figure 5.12b	Plot of the LRT statistic for the 2-group model against hypothetical cutpoints for D-WL.	136

Figure 5.12c	Plot of PF and CCC for optimum classifications into g groups with respect to D-WL.	137
Figure 5.13a	Plot of the 51 entries against their values for PC1 and PC2 (letters represent the entry codes).	150
Figure 5.13b	Plot of PF and CCC for optimum classifications into g groups using the first two PCs for data set A.	151
Figure 5.13c	Distribution of entries among the 7 classes along the PC1 and PC2 axes.	152
Figure 5.14a	Plot of A-RMI for the 51 entries against their respective normal scores.	153
Figure 5.14b	Plot of the LRT statistic for the 2-group model against hypothetical cutpoints for A-RMI.	154
Figure 5.14c	Plot of PF and CCC for optimum classifications into g groups with respect to A-RMI.	155
Figure 5.14d	Distribution of entries in various groups (based on RMI) along the PC axes.	156
Figure 5.15a	Plot of A-SHI for the 51 entries against their respective normal scores.	157

Figure 5.15b	Plot of LRT statistic for the 2-group model against hypothetical cutpoints for A-SHI.	158
Figure 5.15c	Plot of PF and CCC for optimum classifications into g groups with respect to A-SHI.	159
Figure 5.15d	Distribution of entries in various groups (based on SHI) along the PC axes.	160
Figure 5.16	Plot of PF and CCC for optimum classifications into g groups based on damage variables (standardised).	167
Figure 5.17a	Plot of B-RMI for the 55 entries against their respective normal scores.	168
Figure 5.17b	Plot of LRT statistic for the 2-group model against hypothetical cutpoints for B-RMI.	169
Figure 5.17c	Plot of PF and CCC for optimum classifications into g groups with respect to B-RMI.	170
Figure 5.18a	Plot of B-SHI for the 55 entries against their respective normal scores.	171

Figure 5.18b	Plot of LRT statistic for the 2-group model against hypothetical cutpoints for B-SHI.	172
Figure 5.18c	Plot of PF and CCC for optimum classifications into g groups with respect to B-SHI.	173
Figure 5.19a	Plot of the 39 entries against their values for PC1 and PC2 (letters represent the entry codes).	178
Figure 5.19b	Plot of PF and CCC for optimum classifications into g groups using the first two PCs for data set C.	179
Figure 5.19c	Plot of the 39 entries against their values for PC1 and PC2 (figures indicate their respective classes	180
Figure 5.20a	Plot of C-RMI for the 39 entries against their respective normal scores.	181
Figure 5.20b	Plot of LRT statistic for the 2-group model against hypothetical cutpoints for C-RMI.	182
Figure 5.20c	Plot of PF and CCC for optimum classifications into g groups with respect to C-RMI.	183

Figure 5.21a	Plot of the 77 entries against their values for PC1 and PC2.	189
Figure 5.21b	Plot of PF and CCC for optimum classifications into g groups using the first two PCs for data set D.	190
Figure 5.21c	Plot of the 77 entries against their values for PC1 and PC2 (figures indicate their respective classes in the 6-group model).	191
Figure 5.22a	Plot of D-I for the 77 entries against their respective normal scores.	192
Figure 5.22b	Plot of the LRT statistic for the 2-group model against hypothetical cutpoints for D-I.	193
Figure 5.22c	PF and CCC for optimum classifications into g groups with respect to D-I.	194

LIST OF ABBREVIATIONS AND SYMBOLS

μ	population mean
σ^2	variance of population
ANOVA	analysis of variance
CA	cluster analysis
CCC	cubic clustering criterion
DH	dead heart
d.f.	degrees of freedom
EMS	expected value of mean square
$f(x)$	probability (density) function
g	number of groups (clusters)
I	Dobie's index of susceptibility
IGP	grouping based on Dobie's index of susceptibility
LF	Leaf foliar damage
LRT	likelihood ratio test
MS	mean square
n	number of observations (entries)
n_j	number of observations (entries) in group j
N	Number of adult insects
p	number of variables (measures)
PC1	first principal component
PC2	second principal component
PC(s)	principal component(s)

PCA	principal component analysis
PCGP	grouping based on PCs
PF	pseudo F statistic
r	correlation coefficient
R^2	coefficient of determination
RMI	rank mean index
RMIGP	grouping based on RMI
RSI	rank summation index
s^2	sample variance
s.e.	standard error
SHI	Smith-Hazel index
SHIGP	grouping based on SHI
SS	sum of squares
SSCP	sum of squares and cross products
ST	stem tunnelling
STDGP	grouping based on standardised variables
VCA	variance component analysis
B	between groups variance covariance matrix
C	variance-covariance matrix
T	Time to 50% emergence
T	total sum of squares and cross products matrix
w	within groups sum of squares
WL	weight loss
W	pooled within groups sum of squares and cross products

$\text{tr}(W)$	trace of matrix W
$ W $	determinant of W
Π	product (of) sign
Σ	summation sign

Synonymous terms

Any gene combination still undergoing tests will be referred to as a "genotype", while a "variety" will refer to one that has already been tested and released for use. However, the term "entry" will be used for an individual, be it a variety, genotype, etc., when such distinction is unnecessary.

The term "variable", will be used interchangeably with "measure", "trait" and "character". Likewise, "group" with "class" and "cluster"; "classification" with "grouping", "partition" and "division".

ABSTRACT

Utilisation of plant resistance for insect pest management requires identification of resistant genotypes and development of cultivars combining high levels of resistance with other desirable characteristics. Consequently, a large number of genotypes are usually screened under suitable conditions to identify those with resistance to the insect pest in question using appropriate measures.

A basic limitation to conventional selection methods is the choice of cut-off point(s) for efficient partitioning of genotypes into resistance/susceptible groups and the assumption that resistance traits contribute equally and independently to yield loss. This makes it difficult to effectively distinguish and/or delineate genotypes into resistance/susceptible groups and may thus hinder selection and consequently delay or derail the development of resistant genotypes.

This study considered objective statistical approaches to the classification and selection problems. The aim was to optimise the selection, by maximising the differences between the selected group and the other group(s) while minimising the differences within the selected group. Procedures for partitioning crop genotypes into resistance classes, based on observed resistance levels, were proposed and the potential for their use in selecting genotypes was evaluated.

For the single-trait selection, the three categories of procedures investigated include a graphical method where plots of ordered means against the corresponding normal scores were used to determine the optimum number of groups and their limits. The second is based on cluster analysis techniques. In the third procedure, classification models were formulated and hypothesis testing using the likelihood ratio and the quasi bayesian test statistics facilitated the identification of the most plausible points to subdivide the entries under study into distinct groups.

Cluster analysis (CA) and likelihood ratio test (LRT) procedures were investigated for multi-trait selection, focussing on interrelationships among the resistance variables. Use of principal component analysis (PCA) generated indices (principal components (PCs)) was compared with some 3 commonly used indices, viz, rank summation index (RSI), Smith-Hazel index (SHI) and Dobie's index of susceptibility (I).

The practical application of the procedures was investigated using two groups of real field data and the comparison of their relative performance was made through simulation studies.

The results show that the hypothesis testing procedure, overall, performed best. The LRT statistic plots for the 2-group model, in addition to locating the optimum cutpoint for the model, gave a clear picture of the entire structure of the data, especially when distinct groups existed. Its power to detect the 1-group model, as shown

from simulation studies, is an added advantage over the others. Use of the quasi bayesian test statistic gave poor results.

Cluster Analysis, in general, gave fairly similar results to the LRT procedure. The method was fairly robust to distributional assumptions and easy to use.

Normal score plots were fairly precise at determining the optimum number of groups in the data and, used along with other procedures, the plots gave satisfactory and useful results.

The results further showed that, for multivariate classifications, the appropriate procedure and index to adopt depended heavily on intercorrelations among the variables. Simulation studies indicated that classification was greatly facilitated by the existence of high correlations among variables. In this case, promising results were obtained using PCs as the basic variables for classification, in comparison to RSI, SHI and I.

Overall, the optimal conditions for all the classification procedures included equal groups sizes and high intercorrelations among variables.

Application of the proposed procedures to real data demonstrated that the conventional methods of pre-setting selection conditions, in form of group limits, selection intensity, etc., could lead to enormous loss of useful materials.

It was concluded that, other than being alternatives for solving the classification and subsequent selection problems, the procedures

proposed in this study would better serve as complementary solutions.

It is therefore recommended that, where possible, more than one technique be applied before deciding on the optimal solution.

CHAPTER 1

INTRODUCTION

1.1 GENERAL BACKGROUND

Crop production level is generally low under subsistence farming conditions in the developing world where farmers have limited access to various inputs. Insect pests constitute one of the major constraints responsible for this low production level with yield losses of up to 80% having been reported (Singh & Allen, 1980; Agrawal & Taneja, 1989; Singh & Rana, 1989). From an agricultural point of view, the definition of an insect pest is essentially economic, and refers to any insect which does economic damage to crops and animals, or is harmful to human health.

The use of chemical pesticides has been the usual control option for insect pests. However, pesticides are expensive and oftentimes require special equipment and critical timing to apply (Davis *et al.*, 1988). Moreover, undesirable side effects, including residual toxicity in the consumed products and adverse effects on other crop plants in mixed cropping systems and also on parasites, predators and other beneficial organisms, are known (Rana & Murty, 1971; Ashraf & Fatima, 1990). In addition, the rapid development of resistance by insect pests to pesticides has been reported (Bell & Stuart, 1990; Ellis *et al.*, 1994). Hence, emphasis is now being placed on minimal use of chemicals for pest control. Other control options, such as biological

control, are of limited effectiveness especially for pests that feed within stems, such as stem borers, whose larvae are protected to a large extent from natural enemies, unfavourable environmental conditions and pesticides (Agrawal & Taneja, 1989; Mansour & Bar-Zur, 1992).

The use of insect pest resistant crop varieties offers one method of reducing the farmer's dependence on insecticides. The method is economical, simple and convenient, ecologically acceptable, effective and has cumulative effects that renders it long lasting (Rana & Murty, 1971; Horber, 1972; Pathak & Saxena, 1979; Davis *et al.*, 1988; Horn, 1988; Agrawal *et al.*, 1990; Alexander & Bramel-Cox, 1991; Dent, 1992; Oghiake & Odulaja, 1993a; Ellis *et al.*, 1994; Thorne *et al.*, 1994). Being compatible with chemical, biological, and other control measures promotes its use in insect pest management systems. Therefore, host plant resistance probably offers the best insect pest management option.

1.2 HOST PLANT RESISTANCE

Insect pest-resistance in plants is that heritable property which reduces the infestation and damage by an insect species to a level relatively lower than those of other plants under similar environmental conditions (Painter, 1958; Panda & Heinrichs, 1983). Host plants differ in their levels of resistance/susceptibility to insect pests. Whereas resistant plants suffer relatively low degrees of infestation

and damage, susceptible plants are heavily attacked by insect pests and suffer severe damage and yield losses. On the other hand, tolerant plants suffer low yield losses despite the relatively heavy infestation and/or damage.

Painter (1951) subdivided plant resistance into nonpreference, antibiosis and tolerance. Nonpreference refers to situations whereby an insect simply neither feeds nor oviposits extensively on the resistant plant, but rather, departs to more favoured varieties or species. Such a resistant plant may have special features which affect the behaviour of the insect during orientation for food, shelter and oviposition thus making it less attractive or unacceptable to the pest. Antibiosis refers to the physical or biochemical characteristics of the plant that exert adverse effects on the pests' biology/life history such as survival, development and reproduction. Tolerance is the ability of a plant to support a population of insect pests, withstand insect damage and continue to yield at productive levels. While antibiosis and tolerance are both attributes of the resistant plant, nonpreference is viewed from the insect's perspective and connotes choice; the insect either prefers one variety or plant to another or does not. Kogan and Ortman (1978) have suggested an analogous term 'antixenosis' which is a characteristic of the plant and describes the inability of a plant to serve as a host to an insect herbivore, as a result of which the insect pest is forced to select an alternative host plant (Smith *et al.*, 1994).

1.3 RESISTANCE MEASURES

Utilisation of plant resistance for insect pest management requires identification of resistant lines/genotypes, elucidation of mechanisms of resistance and development of cultivars combining high levels of resistance with other desirable characteristics. Consequently, a large number of genotypes are usually screened under artificial and natural conditions to identify, using appropriate measures, those with resistance to the pest in question.

Quantification of relative levels of resistance requires development of appropriate rating scales and measurement systems. Criteria for measuring host plant resistance are generally based on either the effect of the pest on the plant, plant survival and production under infestation and/or the effect of the plant on the behaviour and biology of the pest (Davies, 1985; Saxena, 1989). For example in the case of resistance of maize to stem borers, measures of interest include pest abundance (number of eggs laid, larval/pupal density, etc.) and damage measures (foliar lesions, dead heart, stem tunnelling etc.). However, the latter are of greater interest because they directly affect yield.

Various experiments primarily aimed at identification of the most useful damage measure that can be effectively used in breeding programs to select for resistance have been conducted. It has however been argued (Barrow, 1985 & 1987; Ampofo *et al.*, 1986; Ajala, 1994) that resistance factors present in leaves may not

necessarily be operating in the stem and that resistance in the early growth stages of the plant may not be operating in a later stage (Sneath & Sokal, 1973). Moreover, it is known that a genotype may, for example, be highly resistant to foliar damage but less for stem tunnelling and vice versa for another genotype. In effect, a genotype by trait interaction exists (Ajala & Saxena, 1994). Therefore, to have an overall improvement in the level of resistance that will eventually protect the plant at all stages of feeding activities by the insect, resistance to more than one damage measure is desired. Such an approach will require the use of appropriate indices for simultaneous selection for resistance.

1.4 CLASSIFICATION INTO RESISTANCE GROUPS

Genotypes are usually classified as resistant or susceptible to pests based on infestation and/or damage measures. Since resistance is a relative term (in contrast to immunity, which is absolute and a quality of a non-host), its evaluation often hinges on comparison of the entries tested against checks whose resistance or susceptibility levels have been previously determined (Horber, 1980; Sanford *et al.*, 1993).

Classification can be based on either a single measure or a combination of measures. Both single or multi-trait classification approaches are useful depending on the research objectives. A

general review of the approaches for single-trait and multi-trait classification and selection is presented in Chapter 2 of this thesis.

1.5 STATEMENT OF THE PROBLEM AND IMPLICATIONS

There are various problems associated with the existing classification and selection methods. Inferences are often made using various resistance measures and/or indices, without consideration for the distributional properties and the statistical assumptions required for the techniques employed.

The most critical shortcoming of the existing classification methods and use of resistance indices probably lies in the choice of the number of classes and the class limits or cut-off points. A common practice is to preset the number of groups, the group sizes (eg. "10% of test entries") or limits (eg. "with measured values less than 3"), regardless of the overall performance of the test entries. Such a classification, based solely on ranking, does not take into consideration the magnitude of the differences and variability in the resistance measures of the genotypes being compared. Moreover, fixing class limits prior to experimentation may result into nonsensical selection since all or none of the genotypes may fall within the fixed limits (Nabasirye *et al.*, 1993a).

Two extreme examples may help to illustrate the problem: (a) suppose ten genotypes, half of which were to be selected, were scored for foliar damage (scale 1-9) and gave the following mean

scores: 1 2 2 3 7 | 8 8 8 9 9, ranked in order of magnitude. It would be very difficult to justify selection of the fifth entry, which scored 7, given such a distribution of the mean scores; (b) starting with three predetermined groups defined by 1-3: resistant, 4-7: moderate and 7-9: susceptible, and the mean scores obtained are: 4 4 4 4 6 6 6 6 7 7. Since it was pre-conditioned to select only entries defined as resistant (i.e., 1-3), no selection would be made.

Even where classification is based on some statistics (e.g., mean $\pm$ s.e. (Bhumannavar *et al.*, 1989)), there are often no tests to support these decisions. In some trials, standard varieties are included and test entries are judged with reference to these, in form of relative indices. Obviously, this is a reasonable approach considering the fact that the primary objective is to improve on the existing levels of pest resistance. However, it only applies when such standards are known. In the case of a new pest, screening begins with no known standard and hence, a different approach is required.

Another fundamental limitation to the use of some of the existing classification methods is with regards to the difficulty in assigning weights to various traits included in the indices. A less obvious limitation in some studies involving multi-trait selection approaches is the assumption that damage measures contribute equally in reducing yield; that is, they have equal weights (Ampofo, *et al.* 1986; Saxena, 1989). It is often the case that indices are constructed without due consideration of the interrelationships among

the various variables. Moreover, use of selection indices has not solved the problem of classification because selection cutpoints are still often arbitrarily chosen.

As a consequence of these problems, it becomes difficult to effectively distinguish and/or delineate genotypes into resistance/susceptible groups and may thus hinder effective selection and consequently delay or derail the development of resistant genotypes. It therefore becomes important to develop approaches and procedures that will improve the efficiency of selecting resistant genotypes.

1.6 OBJECTIVES AND SCOPE OF THE STUDY

This study considers statistical approaches to the aforementioned problems and is undertaken with the aim of developing some statistical basis for evaluating and classifying crops genotypes into resistance classes using damage levels. The objectives of the study are to:

1. investigate the probability distributions of the resistance measures that will facilitate appropriate statistical inferences on them,
2. develop procedures for partitioning genotypes into resistance classes to facilitate selection and

3. compare several alternative procedures and evaluate the potential for their use in selecting genotypes based on resistance levels.

Optimal classification, which comprises the determination of optimal number of classes and defining class limits (cut-points), will depend on the magnitude of variability within the variables used for classification. Proper use of variability will depend on the statistical distribution of the measure(s) of resistance which may be determined empirically or reasonably assumed.

It should be noted that the aim of any selection programme is to reduce N genotypes available for selection to n so as to maximise the average level of resistance of the n selected. This, in effect, suggests that maximum gain from selection may be achieved by selecting genotypes whose resistance levels are high, do not differ much among themselves, yet differ widely from the rest of the genotypes.

In Chapter 2, the literature on methods of classification of genotypes is reviewed. In Chapter 3, the univariate (single-trait) and multivariate (multi-trait) classification procedures are respectively developed. In Chapter 4, the data sets used in the study are described. In Chapter 5, empirical assessment of the procedures is carried out through their application to two categories of real data and their comparison through simulation studies. Finally in Chapter 6, a summary of the findings from the study is presented, and further research areas on the topic are suggested.

CHAPTER 2

LITERATURE REVIEW

Classification is one of the basic aims of all scientific research and is something that the human brain does intuitively in most aspects of life. Identification of the selected group and the rest of the genotypes essentially involves a prior classification of genotypes into at least two groups based on some criteria.

The simplest definition of classification is the formation of items into groups (Digby & Kempton, 1987). According to Gupta (1992), classification impresses upon the arrangement of the data into different classes according to their points of similarity thus introducing homogeneity or uniformity amidst diversity.

Many authors use the term 'classification' synonymously with 'cluster analysis' and 'taxonomy'. However, some authors use the term for the assignment of individuals to a pre-established set of categories. This type of problem, which is related to discriminant analysis, is usually called assignment, identification or diagnosis. We shall use 'classification' to denote grouping techniques where classes are determined from the data (that is, the boundaries of the classes are not prespecified but instead are derived according to patterns established from the resistance measures). We will not consider the assignment problem in this study.

2.1 CLASSIFICATION PROCEDURES

Classification in the univariate population (single-trait) involves determination of the best class boundaries/limits and the number of classes. In a multivariate population (multi-trait), it involves making choices on classification variates or a function of variates, the number of classes, and the method of constructing the classes.

Approaches to single-trait and/or multi-trait classification and selection have been reported by several workers (e.g. Marriott, 1970; Gordon, 1973; Dhooria, 1981; Smith *et al.*, 1981; Alghali, 1983; Mendes *et al.*, 1985; Ampofo *et al.*, 1986; Kamali *et al.*, 1989; Bhummanavar *et al.*, 1989; Saxena, 1989; Smith, 1990; Giga & Mazarura, 1991; Odulaja & Nokoe, 1993; Oghiakhe & Odulaja 1993a & 1993b; Ajala, 1994; Ajala, & Saxena, 1994), but with little consensus or consistency of methodology among various users.

2.1.1 Single-trait classification

Single-trait based selection is important at preliminary evaluation stages of varietal testing, when emphasis is on selecting genotypes with desired levels of resistance for use in developing varieties that will combine overall resistance with other desirable characters. It is also particularly very useful in tandem selection, through which individual traits are improved successively for a number of generations (Hallauer & Miranda, 1988).

The easiest and most widely used procedure utilises ranks. For instance, the interest may be in selecting the best ten or a given fraction (selection intensity) of the test population. In some cases, genotypes whose values of the variable of interest are less (or greater) than a given value are selected (Ajala *et al.*, 1995). Dhooria (1981) used a five-point scale with the upper and lower limits of each class fixed arbitrarily. Likewise, Kamali *et al.* (1989), using a 9-point rating scale, classified varieties with 1-3 as resistant, 4-6 as intermediate and 7-9 as susceptible. In a few cases, genotypes have been classified based on the frequency distribution of the data. For instance, Bhummanavar *et al.* (1989) based their classification on the mean and standard deviation of the number of mites per leaf to group a citrus germplasm into five resistance groups. A similar approach was used by Mendes *et al.* (1985) and Smith (1990) to establish class boundaries between tolerant, intermediate and intolerant maize cultivars tested for tolerance to soil acidity. Performance class boundaries between various classes were established by adding or subtracting the standard error of a cultivar from the overall population mean.

In other fields of study, the classification problem has involved hypothesis testing. Various test statistics have been proposed in the literature. For example, in clinical research, investigators often assume that some quantitative measure allows a classification of a population into two groups (e.g. a risk group and a normal group).

When interest is in survival time, the P-value of the log-rank statistic has been used (Lausen & Schumacher, 1992). The quasi bayesian statistic of Chernoff & Zacks (1964) has been widely employed for two groups (Gardner, 1969; Sen & Srivastava, 1975; James, *et al.*, 1987).

Similar methods have been applied to detect shifts in production processes, to study impacts of treatments (Sen & Srivastava, 1975), in prognosis, diagnostics (Gail & Green, 1976) and classification used in regression tree algorithms (Lausen & Schumacher, 1992).

The likelihood ratio test statistic is a popular measure of model adequacy. It has been used for univariate problems (Gardner, 1969; Sen & srivastava, 1975; Worsley, 1979; James, *et al.*, 1987; Mudholkar & McDermott, 1989), though not related to pest resistance. Cate *et al.*, 1971 used the statistic to establish class limits for dividing soil test correlation data into two classes (low, high). The procedure splits the data into two groups, using successive tentative critical levels to ascertain that particular critical level which would maximise the overall predictive ability (R^2) with the means of the two classes as the predictor values.

Each of these approaches help in situations where a known single variable is involved and on the assumption that this single variable is capable of effective classification. A common procedure is to classify on the basis of the most important variable. However, in most cases involving damage due to insects, several variables interact

to reduce yield and thus the need for simultaneous selection using several traits. Furthermore, selecting for one trait at a time assumes that the various infestation/damage measures are independent of each other (i.e. genetically and environmentally uncorrelated). The contrary has often been reported. If genetic associations exist, selection for one trait will cause changes in other traits. These indirect changes are known as correlated responses (Widstrom, 1974; Hallauer & Miranda, 1988; Ajala *et al.*, 1995). Zeinali-Khanghah *et al.* (1993) selected varieties using morphological and developmental traits to effect an indirect selection for yield based on correlated response, the main reason being that direct selection for yield required relatively expensive yield trials.

2.1.2 Multi-trait classification

To have an overall improvement in the level of resistance, selection for more than one damage trait is desired. Multi-trait classification has usually been achieved through the use of indices. Other approaches have included graphical and clustering methods.

2.1.2.1 *Selection indices*

The goal of many plant breeding programs is simultaneous improvement of a crop for several traits. Consequently, plant breeders must consider a number of traits during the selection process. Efforts

have been made over the years to develop indices for use in the selection process.

Index selection was originally developed by Smith (1936) to cope with the complex task of improving breeding material by selecting simultaneously for several quantitative traits. Hazel (1943) later modified and extended the index by outlining a method of estimating genetic variances and covariances and by defining the aggregate genotype (i.e. the genetic worth of an individual) as a linear function of genetic values, each weighted by its relative economic value.

The Smith-Hazel index is given by the equation

$$SHI_i = \sum_{j=1}^p b_j X_{ij} \quad \text{----- (2.1)}$$

X_{ij} represents the phenotypic (observed mean) value of the i th entry for the j th trait, expressed in standard units (z-score); $i = 1, 2, \dots, n$ entries; $j = 1, \dots, p$ traits; b_j , which represents the index coefficient for trait j , is given by:

$$[b_j] = B = P^{-1} G a \quad \text{----- (2.2)}$$

where $P = [p_{jk}]$ and $G = [g_{jk}]$, $j, k = 1, \dots, p$, are variance-covariance matrices of phenotypic and genotypic values, respectively; and $[a_j]$ is a vector of economic weights, given by the user. For example, if the

resistance measures are considered to be of equal economical importance, a will be a unit vector.

Variance component analysis is performed to obtain the genotypic and phenotypic variances and covariances. For this purpose, combined (across environments) analysis of variance (ANOVA) and analysis of covariance (ANOCOVA) tables are required. If y_{klm} denotes the m th observation for the k th genotype in the l th environment, the complete linear model for the data is represented as

$$y_{klm} = \mu + \alpha_k + \beta_l + (\alpha\beta)_{kl} + \gamma_{m(l)} + \epsilon_{klm} ,$$

----- (2.3)

where

- μ = mean effect
- α_k = effect of genotype k , $k = 1, 2, \dots, g$
- β_l = effect of environment l , $l = 1, 2, \dots, e$
- $(\alpha\beta)_{kl}$ = interaction effect
- $\gamma_{m(l)}$ = effect of replicate m , $m = 1, 2, \dots, r$ in environment l
- ϵ_{klm} = random error.

In this analysis, a mixed model is assumed where the effects of genotypes and environments are considered fixed and random, respectively. The mean-squares in the standard ANOVA for a balanced classification may be utilised for estimating variance components. The method, in general, is to equate the observed mean squares with their expected values (Crump, 1951; Miller-Garvin &

Viands, 1994) and solve the resulting equations for the appropriate components. Table 2.1 shows the skeletal ANOVA, where the notation for the variance components corresponds to that under the linear model in equation (2.3).

Table 2.1 ANOVA and expected mean squares (EMS)

Source	d.f.	MS	EMS
Env (E)	e-1		
Rep (R)	e(r-1)		
Genotype (G)	(g-1)	M_α	$\sigma^2_\epsilon + r\sigma^2_{\alpha\beta} + re\sigma^2_\alpha$
G x E	(g-1)(e-1)	$M_{\alpha\beta}$	$\sigma^2_\epsilon + r\sigma^2_{\alpha\beta}$
Error	e(g-1)(r-1)	M_ϵ	σ^2_ϵ
Total	erg-1		

The EMS equations are used to estimate the phenotypic, genotypic, genotype x environment interaction, and error variances (σ^2_{ph} , σ^2_α , $\sigma^2_{\alpha\beta}$ and σ^2_ϵ , respectively) as:

$$\hat{\sigma}^2_{ph} = [e(r-1)M_\epsilon + (e-1)M_{\alpha\beta} + M_\alpha]/re, \quad \text{----- (2.4a)}$$

$$\hat{\sigma}^2_\alpha = (M_\alpha - M_{\alpha\beta})/re, \quad \text{----- (2.4b)}$$

$$\hat{\sigma}^2_{\alpha\beta} = (M_{\alpha\beta} - M_\epsilon)/r, \quad \text{----- (2.4c)}$$

$$\hat{\sigma}^2_\epsilon = M_\epsilon. \quad \text{----- (2.4d)}$$

Similarly, expectations of mean cross products (EMCP) from the ANOCOVA for pairs of variables are used to estimate phenotypic, genotypic, $G \times E$, and error covariances.

The Smith-Hazel approach is considered the optimum index when accurate estimates of variances and covariances are available (Williams, 1962; Crosbie *et al.*, 1980). Williams (1962) suggested what is now called the Base index in which economic weights were introduced. Plant breeders, however, often do not have reliable estimates of variances and covariances or the information that is needed to assign relative economic weights (Crosbie, *et al.* 1980). Therefore several researchers have suggested using indices that are economic weight-free and/or parameter-free.

Elston (1963) proposed a multiplicative index, called the Elston weight-free (EWF) index. The index value for each genotype is obtained as the product of the phenotypic deviations for each trait in the index. It is written as:

$$EWF_i = \prod_{j=1}^p (x_{ij} - k_j) ,$$

----- (2.5)

where x_{ij} is the observed value of the j th trait on the i th genotype, k_j is the lower bound for trait j and p is the number of traits. Modifications to the index were made by Subandi *et al.* (1973) and Compton and Lonnquist (1981).

Pesek and Baker (1969) suggested that breeders usually would be able to specify a desired gain rather than an economic weight for a trait. The authors hence developed an index that uses desired gains to determine the relative weights by maximising the expected response in the population to the gain specified by the breeder.

Dobie (1974) proposed the following index:

$$I_i = 100 (\log_e F_i) / D_i \quad \text{----- (2.6)}$$

where I_i = index of susceptibility for the i th variety

F_i = number of adult insects on the i th variety

D_i = mean development time on the i th variety

The index takes into account both the number of F_1 progeny developing during the experimental period and the average developmental time of the progeny. Alghali (1983), Horber (1980) and Giga and Mazarura (1991) have used the index in storage pest studies.

Mulamba and Mock (1978) developed a parameter-free index, rank summation index (RSI), to select for improved plant density tolerance in maize (Zea mays L.). The index, for a particular genotype, is obtained as a sum of the ranks of the genotype in each of the traits. The index is given by

$$RSI_i = \sum_{j=1}^p R_{ij}$$

----- (2.7)

where R_{ij} , is the observed ranking for the i th genotype with respect to the j th trait; $i = 1, 2, \dots, n$ entries; $j = 1, \dots, p$ traits. The index value lies between p and pn . The use of RSI eliminates the need to assign relative economic weights. In this study, we prefer taking the mean of the ranks over the p variates for each entry, which gives what we refer to as the Rank Mean Index (RMI):

$$RMI_i = \frac{1}{p} \sum_{j=1}^p R_{ij}$$

----- (2.8)

RMI is easier to perceive as it lies between 1 and n .

Further studies have compared relative efficiencies of the different indices using various criteria and RSI is one of the few that has been singled out for its (i) simplicity of use, (ii) freedom from need to estimate genetic parameters, and (iii) providing good selection differential and predicted gains (Crosbie *et al.*, 1980).

Many other proposals and modifications of these indices now exist. Classifications have then been effected by applying the univariate procedures mentioned in Section 2.1.1 on the adopted indices.

Relative infestation/damage indices are sometimes derived, with reference to some known check genotypes included in the

experiments, which are often averaged to give an Overall Resistance/Susceptibility Index (ORSI) (Reddy 1985a & 1985b; Ampofo *et al.*, 1986; Saxena, 1989).

2.1.2.2 *Other approaches*

Graphical approach

The value of a graphical examination of a data set cannot be overemphasised. With two variables, one can make a bivariate chart of the data for visual examination of the configuration of the points representing the genotypes. According to Chatfield and Collins (1980) this is probably the best way of finding natural classes. Odulaja and Nokoe (1993) presented a good example of a graphical method for classifying varieties based on yield potential and yield loss due to pests.

Sometimes, the groupings are obvious visually but yet difficult to detect using a numeric approach like cluster analysis algorithms (Marriott, 1974; Mardia *et al.*, 1979; Chatfield & Collins, 1980). This is particularly true when the clusters are curved or elongated. Most clustering procedures are quite good at detecting spherically shaped clusters but not so good at finding those of other shapes.

With higher dimensional data, graphical representation is generally not convenient, except making all possible bivariate charts (called the grand tour) and examining them (Rao, 1994). However, one can raise the question whether the vectors of variables can be

replaced by a few functions of the variables, preferably two or three, without much loss of information. Attempts to do this have resulted in numerous techniques, the most popular of which is the principal component analysis (PCA) (Green, 1978; Krzanowski, 1979; Chatfield & Collins, 1980; Chheda & Fatokun, 1982; Flury, 1984; Pla, 1991; Souza & Sorrells, 1991; Oghiaké & Odulaja, 1993a; Amaro, 1994).

The most common approach is to plot the principal components (PCs), in pairs, or possibly in more than two dimensions, and examine the scatter diagrams. Perhaps even more important, an examination of the graphs can confirm or cast doubts on the existence of a classification produced by other methods. The approach is perhaps not emphasised in the literature and has not received much attention in pest resistance studies, possibly because of the subjective element involved.

Cluster analysis (CA)

Clustering methods classify a large number of observations into a few relatively homogeneous groups. There is extensive literature on cluster analysis (CA) and computer algorithms for implementing the different methods of analysis (e.g. Friedman & Rubin, 1967; Gower, 1967; Marriott, 1971 & 1974; Scott & Symons, 1971; Sneath & Sokal, 1973; Blashfield & Aldenderfer, 1978; Arnold, 1979; Milligan & Cooper, 1985; Manly, 1986; Digby & Kempton, 1987; Kaufman & Rousseeuw, 1990; Dagnelie & Merckx, 1991; Amaro, 1994; Rao, 1994, among others).

The different cluster analysis methods fall into two families: hierarchical or non-hierarchical. There are two methods of forming a hierarchical classification. Whereas agglomerative methods initially consider each of the n units as being a separate group and proceed by repeatedly combining the two closest groups until only a single group (of n units) remains, divisive methods operate by successive splitting of groups, starting with one group of n individuals and finishing with n groups of one individual each.

The non-hierarchical methods are iterative partitioning techniques for finding a classification which will optimise a measure of homogeneity within each cluster and heterogeneity between clusters. The techniques usually involve a method of initiating a predetermined number of clusters. Several methods of initiating clusters have been proposed.

One major problem evident from the literature is that cluster analysis partitions the entities whether or not there are natural groupings (Arnold, 1979). Thus precautions must be taken to ensure that the user has "sharp differing, relatively homogeneous groups" and is not reporting artifacts. The precautions are difficult to take, however, because clustering techniques were first developed in an applied field (biological taxonomy) and are rarely accompanied by the expected statistical clothing of significance tests, probability models, or optimal procedures (Arnold, 1979). This calls for alternative methods that have an element of hypothesis testing. However, some

clustering algorithms use test statistics as stopping rules (Friedman & Rubin, 1967; Marriott, 1971; Scott & Symons, 1971, Milligan & Cooper, 1985).

A few studies have employed some of these multivariate statistical techniques for the classification of individuals into groups, though not all directly involved with resistance to insect pests. For example, Cornelli *et al.* (1993) suggested clustering techniques in the process of selection, testing and identifying superior genotypes. Oghiake & Odulaja (1993b) used the technique to classify cowpea cultivars for resistance to Maruca testulalis. Binelli *et al.* (1992) based their method on canonical variate analysis to discriminate between maize inbreds with regard to different aspects of resistance to stalk rot. Principal component analysis has been employed to generate uncorrelated variables which are then used in cluster analysis to classify varieties and other test material (Chheda & Fatokun, 1982; Smith, 1990; Pla, 1991; Souza & Sorrells, 1991; Amaro, 1994; Rao, 1994).

2.2 DISTRIBUTIONAL PROPERTIES OF RESISTANCE MEASURES

The probability distribution of resistance measures should be determined before any parametric statistical inference can be made. Many authors have attempted to determine the probability distribution of various resistance measures (e.g. Anscombe, 1949 & 1950; Bliss & Fisher, 1953; McGuire *et al.*, 1957; Harcourt, 1963; Taylor, 1961 &

1984; Widstrom *et al.*, 1971; Lynch *et al.*, 1980). For example, the distribution of corn borer egg masses and larvae has been investigated by Beall (1940), McGuire *et al.* (1957) and Chiang & Hodson (1959), among others. This distribution was reported to be best fitted by the Poisson distribution, but shows a tendency towards contagion, or aggregation, suggesting a negative binomial distribution.

While the Poisson distribution may be appropriate for insect counts on varieties with similar resistance levels, the negative binomial may be more realistic for counts obtained from varieties with varying resistance levels. Other applications of the negative binomial to biological data have been enumerated by Quenouille (1949) and Elliott (1977).

The binomial distribution serves as an appropriate mathematical model in a number of experimental situations. In the context of pest damage, presence of an attribute, for example, dead heart, may take this form.

When one has some indications of the distribution of a population by probabilistic (theoretical) reasoning, practical experience, or otherwise, it is often possible to fit such theoretical distributions to frequency distributions obtained from a sample of the population. One way of identifying the theoretical distribution to consider is by plotting frequency distribution (histogram) of the data set which may provide

information on the shape of the distribution involved. For bell-shaped curves, the skewness and kurtosis may be computed.

Goodness-of-fit tests may be used to determine whether a set of observed data conforms to a specified distribution. The testing problem is essentially the same as deciding whether there are significant differences between the assumed distribution and the sample values.

An important significance test for goodness of fit of the theoretical distribution, the chi-square test, determines whether the observed and expected frequencies agree (at least approximately) or not. Some general (nonparametric) goodness of fit tests formally defined in the statistical literature may also be used. The Kolmogorov-Smirnov test (Neave, 1981; Zar, 1984; Mendell et al., 1993) and the Wilk-Shapiro's test (W) (Shapiro & Wilk, 1965; Shapiro et al., 1968; Zar, 1984; Mendell et al., 1993) as well as the skewness and kurtosis tests (Snedecor & Cochran, 1980; Mendell et al., 1993) are often available in most statistical computing programs (SAS Institute Inc., 1985).

When the data fails to fit a normal distribution, an appropriate transformation may be used to convert it to data that is approximately normal. For positively skewed data, the logarithmic transformation is normally used while the square root transformation is used for negatively skewed data (Southwood, 1978; Snedecor & Cochran, 1980; Sokal & Rolf, 1981; Gomez & Gomez, 1984; Zar, 1984). For

proportions, which are usually highly skewed, the arcsine transformation has been recommended (Snedecor & Cochran, 1980; Gomez & Gomez, 1984; Zar, 1984). However, care must be taken when multimodality is indicated as this may suggest a mixture of distributions, which may individually be normal.

Very often, theoretical assumptions about the distribution of the data are made and transformations based on these assumptions applied. For instance, data on leaf damage scores has been assumed to follow a Poisson distribution and a square root transformation applied to normalise the scores (Mansur *et al.*, 1993). Likewise, the arcsine transformation has been applied to data on stem tunnelling and dead hearts, since these measures are often presented as percentages (Widstrom, 1967; Widstrom *et al.*, 1971; Chheda & Fatokun, 1982, among others).

Validity of distributional assumptions of resistance measures may require testing. This has been done by a few researchers before effecting a transformation (e.g. Widstrom, 1967; Dagnelie & Merckx, 1991).

When fitting theoretical distributions to observed data, it has often been assumed that a single distribution is involved. More than one distribution may, in reality, characterise the population under study. Situations may arise where the data may exhibit a composite of distributions giving rise to a multimodal distribution. For example, the normal mixture has been suggested as a possible distribution for

responses determined by a major gene in genetics research (Souza & Sorrels, 1991; Mendell *et al.*, 1991 & 1993). On the other hand, if resistance was determined by the additive effects of several genes (i.e. polygenic) or of many random environmental factors of equal importance, then the expected distribution would be normal (Falconer, 1989). It has been severally demonstrated that the normal distribution gives poor fits to genetically determined responses, and that the normal mixture fits such data better. Other applications were reviewed by Everitt (1994). Such distributions are often considered appropriate models for populations known to consist of a number of relatively distinct subpopulations. Where the sub-population structure is only suspected, finite mixture densities can be used as the basis for cluster analysis and other investigations (Scott & Symons 1971; Mendell, *et al.*, 1991 & 1993). The approach will be adopted in this study for the problem of classifying genotypes into resistance groups.

2.3 STUDY APPROACH

The lack of adequate and sound statistical basis for varietal selection using most of the conventional approaches has been clearly revealed in the above review. Appropriate statistical methodologies must therefore be employed to provide suitable evidence upon which to base the classification and the subsequent decision to select a particular group of genotypes.

This study addresses this issue by proposing appropriate statistical procedures that will identify the optimum number of groups and the position of cutpoints for classifying crop varieties.

CHAPTER 3

METHODOLOGY

Introduction

Most existing classification procedures use a predetermined number of classes. Our interest in this study lies in deriving the classifications from the observed variation in the data. In this chapter, approaches and procedures for solving the classification problem that take into account the inherent variation in the data are proposed. The general approach is based on goodness-of-fit tests for determination of optimum classification, in form of number of classes and the position of the cutpoints. Existing, modified and newly developed approaches will be illustrated using real problem data and some simulation studies.

3.1 UNIVARIATE PROCEDURES

3.1.1 Graphical procedure

3.1.1.1 *Description*

Let Y be some measure of resistance, and y_i be the mean value for entry i ($i = 1, \dots, n$). When the data consists of an unstructured single set of means with a standard error s , there is no prior pattern except the null one, $E(Y_i) = E(Y_j) \forall i$ and j , that the entries are not significantly different. Possible posterior patterns would include:

- a) the means are a sample from a normal distribution but with a variance greater than s^2 ;

b) the means divide into two or more groups which may be samples from normal distributions, i.e. $E(Y_i) \neq E(Y_j)$ for some $i \neq j$.

One way of investigating the possibility of a mixture of distributions or any grouping of the entries, as implied by (b), is to plot the ordered means $(y_1, y_2, \dots, y_n)$ against the expected values of the order statistics in a sample of the same size from a standard normal distribution $(z_1, z_2, \dots, z_n)$. This can be achieved by plotting the ranked mean values on a normal probability paper (Plackett, 1971; Perry, 1986). As an alternative, the means may be arranged in rank order and plotted against the corresponding normal scores tabulated by Fisher and Yates (1963). The normal scores, which are approximations to the exact expected order statistics for the normal distribution, are computed from the ranks using the formula

$$x_i = \psi(r_i - 3/8)/(n + 1/4), \quad \text{----- (3.1)}$$

where x_i is the normal score for entry i ,

ψ is the inverse cumulative normal (probit) function,

r_i is the rank for entry i , and

n is the number of entries ranked.

On the same graph, the line

$$\hat{y} = \bar{y} + sx, \text{ where } \bar{y} = \frac{1}{n} \sum y_i$$

----- (3.2)

is drawn, where s is the pooled standard error of the estimated means. The plot is then inspected for any possible groupings using the line (equation (3.2)) as a reference.

For patterns of type (a) the points should lie approximately on a straight line, and for the null hypothesis this line is given by equation (3.2). A sharp deviation from the straight line, which suggests patterns of type (b), is enough evidence for grouping the means into different classes. After the grouping, each class may be replotted using the corresponding normal scores determined by the number of entries in the class. No attempt, however, should be made to further subdivide the class when the slope approximates that of the base line in equation (3.2).

3.1.1.2 *Optimum number of groups*

The graphical procedure described above has hitherto been used as an alternative to multiple range tests, in the interpretation of differences indicated by F-ratios in ANOVA, but not to determine the optimum number of groups. We here propose that visual inspection of the normal score plots may give a clue as to the likely number of groups in the data set. This can be achieved by examining the points of sharp increments in the observed values (y-axis) in relation to the

expected values (x-axis) facilitated by the line given by equation (3.2). The number of such points is the likely optimum number of groups in the data set. However, such an inspection could be subjective, and may require verification by other methods using test statistics.

3.1.2 Cluster analysis

3.1.2.1 *Description*

Another approach is to consider the classification as a clustering problem in one dimension. We investigate the possibility of using univariate clustering methods by adopting the analogues of the multivariate criteria. In this procedure, entries are assigned to clusters so as to minimise the differences within clusters. We define the distance, d_{ij} , between entries i and j by

$$d_{ij} = (\bar{y}_i - \bar{y}_j) , \quad \text{----- (3.3)}$$

and use some clustering algorithms to group together entries whose distances do not differ from some critical value defined by an appropriate criterion. Some standard algorithms for minimising the sum of squared deviations from the cluster means in the univariate application may be used. The focus in the current study is on iterative partition (non-hierarchical) methods which divide the observations into clusters such that every observation belongs to one and only one cluster.

Inferences based on clustering techniques may require further investigation using significance tests.

3.1.2.2 *Optimum number of groups*

Several suggestions have been made in the literature for determining 'natural' number of clusters in the data set (Marriott, 1971; Scott & Symons, 1971; Milligan & Cooper, 1985; Dagnelie & Merckx, 1991; among others). Various statistics have also been suggested. In this study, two of them which have been shown to be among the best (SAS Institute Inc., 1987) will be used. These are the Pseudo-F (PF) (Calinski & Harabasz, 1974) and the cubic clustering criterion (CCC) (Sarle, 1983) which are respectively computed as:

$$PF = \frac{R^2/(g-1)}{(1-R^2)/(n-g)} , \quad \text{----- (3.4)}$$

$$CCC = \ln \left(\frac{1-E(R^2)}{1-R^2} \right) \frac{\sqrt{nP/2}}{[0.001+E(R^2)]^{1.2}} , \quad \text{----- (3.5)}$$

where R^2 is the proportion of variance accounted for by the clusters and its expected value, $E(R^2)$, is determined under the assumption that the data have been sampled from a uniform distribution. The parameter P is an estimate of the dimensionality of the between cluster variation. Milligan and Cooper (1985) have chosen the constant terms on the basis of extensive simulation studies.

Plotting values of the statistics against g values and considering the common peak(s) to the statistics will facilitate the identification.

3.1.3 Hypothesis testing and test statistics

3.1.3.1 *Description*

One way of approaching the classification problem is to rephrase it as a hypothesis testing problem (Nabasiye, *et al.*, 1993b). The procedure involves formulating hypotheses in form of classification models and using appropriate test statistics which measure the "goodness of fit" of each model to identify the most plausible model based on the observed data. We hereby develop such classification models and propose methods of testing the corresponding hypotheses.

A classification procedure can be defined in terms of a function, $G(y)$, that allocates entries to various classes based on the classification variate. If, for example, insect count is the variable of interest, a perfect classification procedure, when two classes are considered, is one that assigns entries with counts less than or equal to C , say, to the relatively resistant class and assigns the rest to the susceptible class. That is

$$G(y) = 0 \text{ for } y \leq C \quad \text{----- (3.6a)}$$

$$G(y) = 1 \text{ for } y > C \quad \text{----- (3.6b)}$$

where 0 represents a resistant class and 1 a susceptible class of entries. Such a classification into two classes has a wide application in crop varietal selection (Curnow, 1961).

This can further be extended to a classification into three classes, which may represent highly resistant, moderately resistant and susceptible entries, with

$$G(y) = 1 \text{ for } y \leq C_1 \quad \text{----- (3.7a)}$$

$$G(y) = 2 \text{ for } C_1 < y \leq C_2 \quad \text{----- (3.7b)}$$

$$G(y) = 3 \text{ for } y > C_2 \quad \text{----- (3.7c)}$$

Further extension could be made to more than three groups. The problem is how to fix C or C_1 and C_2 , etc. which define the class limits. We propose one approach based on hypothesis testing.

3.1.3.2 Formulation of hypotheses

Classification via hypothesis testing can be achieved by examining the differences between the hypothesised groups. Consider a set of n entries. Let Y be some variable which can be used to classify the entries into at least two groups and y_i be the mean value for entry i ($i = 1, 2, \dots, n$). The aim is to test whether the class means can be considered to be equal.

For the purpose of classification into various groups, the problem, after ranking the mean values in ascending order, reduces to testing the following hypotheses corresponding to different appropriate models as in Section 3.1.3.1.

Notation: It should be noted that H_1 , as defined below, is the null hypothesis conventionally denoted by H_0 . The subscript 1 is meant to

depict the 1-group model. Likewise H_g will refer to the g -group model.

(a) 1-group model,

$$H_1 : \mu_1 = \mu_2 = \dots = \mu_n = \mu$$

$$: E(Y_i) = \mu \quad (i=1,2,\dots,n)$$

(b) 2-group model,

$$H_2 : \mu_1 = \dots = \mu_r < \mu_{r+1} = \dots = \mu_n$$

$$: E(Y_i) = \mu_a, \quad i \leq r$$

$$= \mu_b, \quad i > r$$

(c) 3-group model,

$$H_3 : \mu_1 = \dots = \mu_s < \mu_{s+1} = \dots = \mu_t < \mu_{t+1} = \dots = \mu_n$$

$$: E(Y_i) = \mu_c, \quad i \leq s$$

$$= \mu_d, \quad s < i \leq t$$

$$= \mu_e, \quad i > t$$

.

.

.

(d) n -group (maximal) model,

$$H_n : \mu_1 < \mu_2 < \dots < \mu_n$$

$$: E(Y_i) = \mu_i, \quad i=1,2,\dots,n$$

3.1.3.3 *Testing the hypotheses*

For Classification into two groups, we test H_1 against H_2 . We test H_1 against H_g for classification into g groups. The parameters of the models (e.g. μ_a and μ_b , in the 2-group model), which are in turn determined by the cutpoint(s) (e.g. r , for the 2-group model), are unknown. The aim then is to estimate and test some hypotheses about these cutpoints. In general, for classification into g groups (g -group model), we need to estimate and test hypotheses about $g-1$ cutpoints.

Two test statistics will provide a measure of how well a given model describes the given data set. The quasi bayesian statistic of Chernoff and Zacks (1964) will be used to classify entries into 2 groups, while the likelihood ratio test statistic will be used for two or more groups.

Assumptions: In addition to normality, the hypothesis testing procedures will assume equal groups variances, which may be unknown.

3.1.3.3.1 *Quasi Bayesian Test (QBT) statistic*

We begin with the mean values for the entries, y_i , ranked in ascending order and we consider the standardised residual of y_{i+1} from the mean of $y_1, \dots, y_n$, obtained as

$$z_i = [i/(i+1)]^{1/2} (y_{i+1} - \bar{y}), \quad (i=1, \dots, n-1) .$$

----- (3.8)

Under H_1 , the recursive residuals $z_1, \dots, z_{n-1}$ are independent standard normal random variables (Sen & Srivastava, 1975). With the assumption that the cutpoint, r , has a uniform prior distribution over $\{1, \dots, n\}$ (that is, there is an equal chance for the cutpoint to be placed between any two ranked entries), Chernoff and Zacks (1964) gave

$$QBT_r = \sum_{i=1}^r [i(i+1)]^{1/2} z_i, \quad r = 1, \dots, n-1$$

----- (3.9)

where z_i is the recursive residual defined in equation (3.8). An equivalent but more convenient test statistic is given by:

$$\sum_{i=1}^r i(y_{i+1} - \bar{y}), \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad r = 1, \dots, n-1.$$

----- (3.10)

The distribution of the statistic is trivial (normal distribution).

It is required to compute values of the QBT statistic for all possible groupings into two ($g=2$). The grouping with the minimum value is considered optimal. That is, the grouping that gives the minimum value of the statistic provides the best classification of the entries into two resistance groups representing relatively resistant and susceptible groups.

3.1.3.3.2 *Likelihood ratio test (LRT) statistic*

Likelihood function

Suppose a population has a probability density function, denoted by $f(y, \theta)$, with a population parameter θ , to be estimated by sample statistic. Assuming there are n independent observations $y_1, \dots, y_n$, the joint density function for these observations, which is called the likelihood (function), is

$$L = f(y_1, \theta) f(y_2, \theta) \dots f(y_n, \theta) \quad \text{----- (3.1.1)}$$

For example if observations $y_1, y_2, \dots, y_n$ are drawn from the normal distribution with mean μ and variance σ^2 , the joint likelihood is given by

$$L = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left[-\frac{1}{2\sigma^2} \sum (y_i - \mu)^2 \right]. \quad \text{----- (3.1.2)}$$

Estimators of the parameters which maximise the likelihood or, equivalently, the log-likelihood function can be obtained.

Model adequacy

The adequacy of a model is defined relative to a maximal model H_n , which has the same number of parameters $(\mu_1, \mu_2, \dots, \mu_n)$ as the entries and so provides a complete description of the data. That is, $\theta_n = [\mu_1, \mu_2, \dots, \mu_n]^T$. We compare this with another model specified by a parameter vector of length g with $g < n$, ($g = 1, 2, \dots, n-1$). That is $\theta_g = [\mu_1, \dots, \mu_g]^T$.

Notation

Denote the likelihood function for model g , $L(\theta_g; \mathbf{y})$, by L_g and the corresponding log likelihood function evaluated at the maximum likelihood estimator for model g by l_g .

In terms of the likelihood function, the model adequately describes the data if $L_g = L_n$, or otherwise when $L_g \ll L_n$. This suggests using the generalised likelihood ratio statistic, $\lambda = L_n/L_g$, or equivalently $\log \lambda = l_n - l_g$, as a measure of goodness of fit. Large values of $\log \lambda$ provide evidence for inadequacy of the g -parameter model for the data.

Log-likelihood ratio statistic

We define the statistic

$$D_g = 2 \log \lambda = 2 (l_n - l_g) \quad \text{----- (3.13)}$$

This statistic is often referred to as the (scaled) deviance (Dobson, 1983). It can be shown that

$$D_g \sim \chi_{n-g}^2,$$

if the model is adequate, where χ_m^2 is the central chi-square distribution with m degrees of freedom. If the model is inadequate, D_g will be larger than predicted by χ_{n-g}^2 . In this case, D_g has a non-central chi-squared distribution.

We test H_1 against H_2 , using the difference of the log-likelihood ratio statistics

$$D = D_1 - D_2 = 2(l_2 - l_1) .$$

----- (3.14)

If both models $H_1: \theta' = \theta_1$; and $H_2: \theta'' = (\theta_1, \theta_2)$ each describe the data as adequately as $H_n: \theta^{(n)} = (\theta_1, \dots, \theta_n)$, then

$$D = D_1 - D_2 = 2(l_2 - l_1) \sim \chi^2_1 .$$

If the observed value of D is in the critical region (i.e. greater than the upper tail $100(1-\alpha)\%$ point of the χ^2_1 distribution) then we would reject H_1 in favour of H_2 , indicating that the 2-group model provides a significantly better description of the data (even though it may not completely describe the data, as does the maximal model).

To evaluate the statistic, we assume that resistance variates $y_1, \dots, y_n$ are independent and normally distributed with common variance σ^2 . Then the log likelihood function is

$$l = - \frac{1}{2\sigma^2} \sum_{i=1}^n [y_i - E(y_i)]^2 - \frac{n}{2} \log(2\pi\sigma^2) .$$

----- (3.15)

For the maximal model, $E(Y_i) = \mu_i$, $i = 1, \dots, n$; and the estimate of μ_i is

$$\hat{\mu}_i = y_i .$$

----- (3.16)

Therefore

$$\begin{aligned} \log(L) &= l = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 - \frac{n}{2} \log(2\pi\sigma^2) \\ &= -\frac{n}{2} \log(2\pi\sigma^2), \text{ since } \hat{\mu}_i = y_i. \end{aligned}$$

----- (3.17)

Considering model $H_1: E(Y_i) = \mu$ for all i , the estimate of μ is

$$\hat{\mu} = \bar{y}.$$

----- (3.18)

Therefore

$$l_1 = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \bar{y})^2 - \frac{n}{2} \log(2\pi\sigma^2).$$

----- (3.19)

Thus, from equation (3.13),

$$D_1 = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \bar{y})^2, \quad D_1 \sim \chi_{n-1}^2.$$

----- (3.20)

Otherwise, for example, if the means of the y_i 's are heterogeneous, D_1 will be larger than the expected from the χ_{n-1}^2 distribution. The statistics corresponding to H_1 , H_2 and H_0 are summarised in Table 3.1. It can easily be shown that

$$D_1 - D_2 = (n/2\sigma^2)(\bar{y}_1 - \bar{y}_2)^2.$$

However, this cannot be used to test H_2 directly if σ^2 is unknown.

$$j=1, \dots, g, k=1, \dots, n_j, n = \sum n_j.$$

44

	H_1	H_2	H_g
Model	$\theta_1 = [\mu]$	$\theta_2 = [\mu_1, \mu_2]$	$\theta_g = [\mu_1, \dots, \mu_g]$
Estimates	$\hat{\mu} = \bar{y} = \frac{1}{n} \sum_j \sum_k y_{jk}$	$\hat{\mu}_j = \bar{y}_j = \frac{1}{n_j} \sum_k y_{jk}$	$\hat{\mu}_j = \bar{y}_j = \frac{1}{n_j} \sum_k y_{jk}$
Log likelihood statistic	$D_1 = \frac{1}{\sigma^2} \sum_j \sum_k (y_{jk} - \bar{y})^2$	$D_2 = \frac{1}{\sigma^2} \sum_j \sum_k (y_{jk} - \bar{y}_j)^2$	$D_g = \frac{1}{\sigma^2} \sum_j \sum_k (y_{jk} - \bar{y}_j)^2$
Sampling distribution	$D_1 \sim \chi_{n-1}^2$	$D_2 \sim \chi_{n-2}^2$	$D_g \sim \chi_{n-g}^2$
Test statistic	-	$f = \frac{(D_1 - D_2)/1}{D_2/(n-2)} \sim F_{1, n-2}$	$f = \frac{(D_1 - D_g)/(g-1)}{D_g/(n-g)} \sim F_{g-1, n-g}$

This variance may be estimated by the pooled sample variance,

$$s^2 = \frac{\sum \sum (y_{jk} - \bar{y}_j)^2}{n-2} = \frac{\sigma^2 D_2}{n-2} .$$

----- (3.21)

Hence from equation (3.14) and the subsequent expressions, given that both models are adequate,

$$f = \frac{(D_1 - D_2)/1}{D_2/(n-2)} \sim F_{1,n-2} ,$$

----- (3.22)

where $F_{j,k}$ is the F distribution with j and k degrees of freedom.

If H_1 is not true, D_1 has a non-central chi-square distribution so that f has a non-central F distribution. Therefore, values of the statistic f which are large compared to $F_{1,n-2}$ distribution provide evidence leading to the rejection of H_1 . Similarly, for the 3-group model we have

$$D_3 = \frac{1}{\sigma^2} \sum_j \sum_k (y_{jk} - \bar{y}_j)^2 \sim \chi^2_{n-3} ,$$

----- (3.23)

and the corresponding F-statistic is given by

$$f = \frac{(D_1 - D_3)/2}{D_3/(n-3)} \sim F_{2,n-3} .$$

----- (3.24)

In general for the g -group model, the aim is to determine whether the groups represent different resistance levels, estimate the groups mean

resistance levels and test whether they differ significantly. The F-statistic is

$$f = \frac{(D_1 - D_g)/(g-1)}{D_g/(n-g)} \sim F_{g-1, n-g}.$$

----- (3.25)

Maximisation of the F-statistic provides the optimum classification with the assumed number of groups.

The LRT statistics may be obtained by running a one-way ANOVA to generate values of the total, between and within groups, sum of squares. Using the notation above, we construct the ANOVA table as shown in Table 3.2.

Table 3.2 One way analysis variance table.

Source	d.f.	SS	MS	F
Groups	$g-1$	$D_1 - D_g$	$(D_1 - D_g)/(g-1)$	f^*
Error	$n-g$	D_g	$D_g/(n-g)$	
Total	$n-1$	D_1		

f^* is given by equation (3.25).

To use the procedure, all possible ways of arranging the n entries into a given value of g classes ($g = 2, \dots, n-1$) must be

considered. The grouping that minimises the within groups variance, and thus maximises the observed between groups variance or their functions, (eg. the corresponding F-value and R^2) provides the best classification of the entries into g groups. In other words, the grouping with the largest F-value, which essentially maximises differences among the various groups, provides the best evidence for rejecting the 1-group model in favour of g groups.

3.1.3.4 *Optimum number of groups*

The reduction in variance due to the partitioning into g groups may be used to determine the optimal number of groups. One statistic, g^2W (where W is the within groups sum of squares and cross products (SSCP) matrix), suggested by various authors (eg. Friedman & Rubin, 1967) for multivariate situations will be adopted. The univariate analogue of the statistic, using the notation above, is g^2D_g . However, it is preferred to use $g^2D_g/(n-g)$ which takes into consideration the varying degrees of freedom with changing values of g (it may make little difference when n is very large). The minimum value of the statistic ($g^2 \times$ within groups mean squares = g^2w) indicates the optimum partition. Thus we take the value of g which minimises g^2w to be the optimum number of groups.

Furthermore, the choice between $g-1$ or g groups ($g = 2, 3, \dots$) may be made using the F-ratio comparing the reduction in within groups variation when the $g-1$ groups model is adopted against the g

groups model. That is,

$$f = \frac{D_{g-1} - D_g}{D_g / (n - g)} \sim F_{1, n-g} ,$$

----- (3.2)

where D_j is the within groups SS for the 'best' partition into j groups ($j = 2, \dots$) as determined above. Non-significance of the statistic implies the $g-1$ groups model is adequate, or rather the g groups model is not significantly better than the $g-1$ groups model.

A plot of the LRT statistic for the two-group model may indicate the number of groups. However, procedures for determining the optimum number of groups using the quasi bayesian statistic are unknown.

3.2 MULTIVARIATE PROCEDURES

There may be, and usually is, need to simultaneously consider several measures collected on each entry in grouping them. Grouping using individual variables may result in conflicting signals as to the relative status of the entries. Moreover, the variables may not be exerting independent influence on the entries. It is therefore essential and is the purpose of this section, to consider multivariate procedures for classifying entries into resistance groups.

3.2.0 The multivariate selection problem

Given p different resistance traits measured on each entry, the

problem is how to select the best entries with respect to all the p traits combined. This problem becomes more acute when there are negative correlations among different traits. We need to select n of the N entries so as to jointly optimise, the measures of the p traits in the selected entries. The desired selection is one in which excellence with respect to one or more of the traits can, to an extent, compensate for, but not overrule, weaknesses with respect to the other traits.

If the variates are highly positively linearly correlated, then classification for a specific variate will yield reasonable classification for the other variates and the univariate approach will suffice.

The main interest centres on the mean vector and its covariance matrix. The number and position of boundaries clearly will depend on the covariance matrix. Whereas too many classes would provide too fine a classification, too few would leave much structure unrevealed. Therefore, there has to be a criterion for determining the optimum classification.

3.2.1 Classification using the indices

This study considers some three indices that have been proposed by different authors for selection for resistance to insect pests. These are the Rank Summation Index (RSI), Smith-Hazel index (SHI) and (Dobie's) Index of susceptibility (I). Expressions for the indices were given in Section 2.1.2. The three indices will serve as

standards, where appropriate, for comparative purposes in this study.

The univariate procedures earlier discussed will be applied to each of the indices.

3.2.2 Alternative approach

A better approach to index construction should ensure that variables that are highly correlated are not given excessive weight. In this section, we propose and adopt multivariate methods that utilise the inherent relationships between resistance measures for multivariate-based classification.

3.2.2.1 *Non-significantly correlated variables*

Suppose there are two resistance measures, X_1 and X_2 . The correlation coefficient $r \approx 0$ between X_1 and X_2 indicates no linear relationship between the variables. We propose a linear index that gives equal weights to both measures.

Since measures are made on different scales, it is necessary to standardise them so that any particular measure does not dominate the index and thus the subsequent classification. If the measurements on X_2 are w times as variable as those on X_1 , we might consider ranking entries on the basis of the index $wX_1 + X_2$ or equivalently, $X_1 + X_2/w$. Often, each measure is weighted by its standard deviation to achieve the same objective while z-scores may also be used.

3.2.2.2 *Significantly correlated variables*

On the other hand, X_1 and X_2 may be significantly correlated. For example, $r \approx \pm 1$ occurs when the two variables are measuring essentially the same (or opposite) quality. In these situations, if the correlation is positive, any of the variables, X_1 or X_2 , can be used to classify the entries. That is, classification based on X_1 suffices for X_2 as well. In terms of indices, this in effect is a linear index in which one variable is assigned weight 1 while the other gets 0. That is,

$$I = w_1X_1 + w_2X_2, \text{ with } w_i = 0 \text{ or } 1. \quad \text{----- (3.27)}$$

For $r \approx -1$, w_i has to be based on some other criterion, possibly representing the relative economic importance of the variables.

In real situations, a wide range of values are possible and the choice of weights for the various traits included in the indices becomes crucial. We consider Principal component analysis as an objective alternative for construction of indices that utilises the inherent intercorrelations among resistance measures.

Principal component analysis (PCA)

The PCA may be used to construct indices based on combinations of the variables so that a substantial amount of the variation in the data is accounted for as concisely as possible.

The objective is to take p variables $X_1, X_2, \dots, X_p$ and obtain combinations of these to produce uncorrelated indices, $Z_1, Z_2, \dots, Z_p$. This lack of correlation is a useful property because it means that the

indices are measuring different 'dimensions' in the data. However, they are also ordered so that Z_1 displays the largest amount of variation, Z_2 displays the second largest, and so on, i.e. $\text{var}(Z_1) > \text{var}(Z_2) > \dots > \text{var}(Z_p)$.

PCA is however not scale-invariant. Hence, the correlation matrix rather than the covariance matrix, is usually analysed.

A problem arises as to how many principal components (PCs) retain in the subsequent analyses. When analysing a correlation matrix where the sum of the eigenvalues is p , many researchers use the rule that eigenvalues < 1 may be disregarded. In effect, this retains components which individually account for at least the variability in each of the original variables. This arbitrary policy is a useful rule of thumb but has no theoretical justification. The present study uses an arbitrary course to retain only PCs that cumulatively account for at least 80% of the total variability. If the first PC is a to meet this requirement, which is only possible with very high correlations between the variables, the univariate procedures developed earlier may be used for the classification of the entries. Otherwise, with more than one component, multivariate methods using the retained components are required.

3.2.3 Classification procedures

Multivariate analogues of the univariate procedures developed in Section 3.1 may be adopted, with modifications where required.

3.2.3.1 *Graphical technique*

The reduction in dimensions afforded by PCA can facilitate graphical procedures. Thus if the first two PCs explain a large proportion of the total variance, we can plot the scores on these two components for each entry and visually observe any possible groupings. Such plots can augment the other classification procedures.

If more than two components are needed, other classification approaches may be required.

3.2.3.2 *Cluster analysis*

A similar procedure to the one discussed in Section 3.1 may be used on a p -dimensional set of data. This may be performed on standardised variables if the original measures have different scales or on the few PCs identified using PCA. Disjoint (discrete) cluster analysis may be performed based on Euclidean distances (d_{ij}) computed from the p variables as

$$\begin{aligned} d_{ij}^2 &= (y_i - y_j)^T (y_i - y_j) \\ &= \sum_{k=1}^p [(y_{ik} - y_{jk})^2] \quad , \end{aligned} \quad \text{----- (3.28)}$$

where d_{ij}^2 is the squared Euclidean distance between entries i and j , y_i is a p -vector of means for entry i .

3.2.3.3 *Likelihood ratio test statistic*

We assume that the n entries are drawn independently from a mixture of multivariate normal distributions. Let ν be an unknown parameter vector with n components, where the i th component indicates the distribution from which the i th entry was drawn. In the classification problem, it is required to group together the observations from the same distribution.

Classification models

Let the sample consist of n observations $Y = (y_1, \dots, y_n)$ where y_i represents measurements on p characteristics. Suppose that the observations are independent and that each may arise from any one of g possible p -variate normal distributions with means $\mu_1, \dots, \mu_g$ and covariance matrices $C_1, \dots, C_g$.

The joint distribution of Y is completely determined by μ_j, C_j ($j = 1, \dots, g$) and the grouping or classification parameter $\nu = (\nu_1, \dots, \nu_n)$, where ν_i takes on values $1, \dots, g$; g being the number of groups. Denoting the collection of all the parameters by

$$\theta = (\nu, \mu_1, \dots, \mu_g, C_1, \dots, C_g),$$

----- (3)

the fundamental question of whether there are more than one group can be considered as a test of the null hypothesis, H_1 , against H_0 which suggests g distinct groups. That is,

(a) the 1-group model,

$$H_1 : \theta_1 = \mu, C$$

$$: v_1 = \dots = v_n$$

(b) the g-group model,

$$H_g : \theta_g = (\mu_1, \dots, \mu_g, C_1, \dots, C_g)$$

$$: v_i = j, j = 1, \dots, g$$

The log likelihood function, $l(\theta)$ is given by

$$l(\theta) = -\frac{1}{2} \sum_{j=1}^g \left[\sum_{M_j} (y_{ij} - \mu_j)^T C_j^{-1} (y_{ij} - \mu_j) + n_j \log |C_j| \right],$$

----- (3.30)

where M_j is the set $\{y_{ij}\}$ assigned to j th group by v ,

n_j is number of observations in M_j , and

$|C_j|$ denotes the determinant of C_j .

Estimation of model parameters

The classification problem is to estimate and test hypotheses about v and hence the groups $M_1, \dots, M_g$. For a given partition of Y into g groups, $M_1, \dots, M_g$, the likelihood is maximised by substituting the ordinary ML estimates of μ_j and C_j . The estimate of μ_j , regardless of the assumption about C_j , is

$$\hat{\mu}_j = \bar{y}_j = \frac{1}{n_j} \sum_{M_j} y_{ij} ,$$

----- (3)

that is, the mean of the n_j observations in group M_j . When this is substituted for μ_j in the log likelihood expression in (3.30) it follows that the ML estimate of $\mathbf{v}$ can be obtained by minimising

$$\sum_{j=1}^g \left[\text{tr}(\mathbf{W}_j \mathbf{C}_j^{-1}) + n_j \log |\mathbf{C}_j| \right] ,$$

----- (3)

with respect to $\mathbf{C}_j$, where

$$\mathbf{W}_j = \sum_{M_j} (y_{ij} - \bar{y}_j)^T (y_{ij} - \bar{y}_j) .$$

If it is assumed that the dispersion matrix $\mathbf{C}$ is equal for all the g groups, i.e. $\mathbf{C}_j = \mathbf{C}$ ($j=1, \dots, g$), then the expression (3.32) reduces to

$$\text{tr}(\mathbf{W} \mathbf{C}^{-1}) + n \log |\mathbf{C}|, \quad \mathbf{W} = \sum_{j=1}^g \mathbf{W}_j .$$

----- (3)

$\mathbf{W}$ is the pooled within groups sum of squares and cross products (SSCP) matrix. The ML estimate of $\mathbf{C}$ for fixed $\mathbf{v}$ is equal to $\mathbf{W}/n$. On substitution in expression (3.30) it follows that the optimal subdivision is given by the grouping that minimises $|\mathbf{W}|$. This is a widely used criterion, originally suggested by Friedman and Rubin (1967). It has been found to work well in several empirical studies (Marriott, 1971; 1974; Scott & Symons, 1971; Arnold, 1979). Techniques that

maximise the ratio of determinants ($|T|/|W|$), have served the same purpose, where T is the total scatter matrix given by

$$T = \sum (y_i - \bar{y})^T (y_i - \bar{y}) . \quad \text{----- (3.34)}$$

Manly (1986) suggested a function of this statistic given by

$$\phi = [n - 1 - \frac{1}{2}(p - g)] \log(|T|/|W|) , \quad \text{----- (3.35)}$$

where n = number of entries,

p = number of variables (traits),

g = number of groups,

$|T|$ = determinant of the total SSCP matrix, and

$|W|$ = determinant of the within groups SSCP matrix.

This statistic can be tested for significance by comparison with $\chi^2_{p(g-1)}$.

Matrices T and W

Let x_{ijk} denote the value of variable X_k for the i th individual in the j th group,

$\bar{x}_{jk}$ denote the mean of X_k in the same group, and

$\bar{x}_k$ denote the overall mean of X_k for all data combined.

Then the element in row r and column c of T is

$$t_{rc} = \sum_{j=1}^g \sum_{i=1}^{n_j} (x_{ijr} - \bar{x}_r)(x_{ijr} - \bar{x}_c) .$$

----- (3.)

The element in the r th row and c th column of W is

$$w_{rc} = \sum_{j=1}^g \sum_{i=1}^{n_j} (x_{ijr} - \bar{x}_{jr})(x_{ijr} - \bar{x}_{jc}) .$$

----- (3.)

The test based on ϕ involves the assumption that the distribution of the p variables is multivariate normal, with a constant within group covariance matrix. It is probably a fairly robust test in the sense that moderate deviations from this assumption do not unduly affect the characteristics of the test (Manly, 1986).

If the assumption that the dispersion matrix C is equal for all g groups is invalid, i.e. the covariance matrices are not equal, then their ML estimates are given by W_j/n_j for $j=1, \dots, g$, which can be substituted in expression (3.33) above. In this case the optimal grouping minimises

$$\prod_{j=1}^g |W_j|^{n_j} .$$

----- (3.)

Computational procedures

Given a set of ungrouped data, we could consider all possible arrangements into g groups. Whereas it is easy to order (rank)

univariate observations, there is usually no natural basis for ordering multivariate data. It is impractical to search for all possible groupings as the rankings within variables may not be the same. Groupings of the entries may be based on the first PC as it carries maximal variance, after which the necessary statistics are computed for each of the groups.

The approach requires estimation, by maximum likelihood, of the p means, p variances and $p(p-1)/2$ covariances of the hypothesised groups and obtaining the value of the test statistic. A one-way multivariate ANOVA can be carried out in the usual way to give matrices of estimated variances and covariances within and between groups (Table 3.3).

Table 3.3 Multivariate analysis of variance table.

Source	d.f	dispersion matrix
Between groups	$g-1$	B
Within groups	$n-g$	W
Total	$n-1$	T

If the dispersion matrix corresponding to **W** is the same for all groups in the population, a suitable test statistic for differences

between group means is the ratio $|W|/|T|$, sometimes called the Wilks' criterion (λ). One of the groupings will make $|W|$ a minimum and correspondingly give a minimum value to the test statistic $|W|/|T|$ (since $|T|$ is constant for a given set of observations, regardless of classification). This gives the optimal subdivision of into g groups. Wilks' λ , a function of the likelihood ratio statistic, be preferred because its distribution is known (at least approximately) for samples from multivariate normal distribution, and forms the basis for standard multivariate analogue of variance-ratio tests of significance.

Apart from the heavy computations required for this procedure there are other problems that may arise from the data being analysed. In practice, data are often recorded with limited accuracy. This may have a very substantial effect on the subsequent grouping. For example, if a continuous variate is coded on a 5-point scale (eg. 1,2,3,4,5), a subdivision into 5 groups on basis of that variate alone makes $|W|$ zero, and is therefore optimal. In less extreme cases, it remains true that the less accurately a measurement is recorded, the more influence it is likely to have on the grouping.

When many variates are measured, high correlations often exist between them and the dispersion matrix may be singular or ill-conditioned in the sense that departures from strict linear dependence are smaller than errors of measurement. In this case the grouping procedure breaks down, and it is necessary to reduce the dimension

the data. An easy way of doing so is by using the first k PCs in the procedure. The value of k should be large enough to include virtually all the information in the original data.

3.2.4 Optimum number of groups

It is intuitively obvious that if the data are drawn from a population which is strongly clustered around a small number of modes, the optimal subdivision into the same number of groups will produce a large reduction in $|W|$. If the underlying population is unimodal, or uniform, the reduction is much smaller.

An optimal subdivision of a uniform distribution into g groups reduces $|W|$ by a factor g^2 (Marriot, 1971). A strongly multimodal distribution will reduce $|W|$ much more than this when it is appropriately divided, while most ordinary unimodal distributions will give less reduction. We propose that the best approach to the problem is probably to plot $g^2 |W|$ against g . If the graph decreases steadily, there is no indication of any grouping, but if for one value of g the value of $g^2 |W|$ falls well below the neighbouring values (i.e. when there is a marked drop in the statistic), it suggests that this corresponds to an optimal grouping. Alternatively we may obtain similar information by tabulating the values of the statistic for different values of g and pick the last to decrease, if there is any. If $g^2 |W| / |T|$ is > 1 for all possible subdivisions, the distribution will be regarded as a single group, showing no evidence of being

composite. Milligan & Cooper (1985) examined various functions the $|W|$ statistic in determining the optimal number of groups in hierarchical cluster analysis and found that they gave fairly good results in empirical applications.

3.3 COMPARISON OF PROCEDURES

It is necessary to develop a criterion for comparing the proposed procedures as well as the performance of the three indices included in the study. The usually employed Smith's (1936) criterion is based on the average gain expected from use of a selection index (Williams, 1962; Harris, 1964). It involves at least 2 selection cycles to estimate the observed gain and compare it with the expected gain. However, this criterion does not apply if the program under study does not involve repeated selections from the same population.

The classification procedures described in this Chapter may be compared by simulating known groups in data sets and determining the power of the procedures at recognising the groups. Real problem data may also serve this purpose to a lesser degree. In Chapter 5, we illustrate the procedures and some existing approach using both real and simulated data.

CHAPTER 4

DATA SETS AND THEIR PROPERTIES

4.1 REAL DATA

4.1.1 Description of real data

Two groups of real problem data were examined. The first group comprises field data from a series of experiments on maize breeding for resistance to Chilo partellus in a study conducted at the Mbita Point Field Station (MPFS) and Ungoye Field Site (UFS) of the ICIPE on the shores of Lake Victoria in Western Kenya. There are three sets in this group as described below:

- (a) Set A consists of 51 entries comprising 6 inbred lines and all possible single crosses derived from them, namely 15 main crosses (F_1 s), 15 reciprocal crosses (F_1 s) and 15 F_2 s, evaluated in two environments.
- (b) Set B consists of 7 open-pollinated maize populations and all possible crosses among them (21 main crosses, 21 reciprocal crosses) and 6 local varieties, to give a total of 55 entries that were evaluated in two environments.
- (c) Set C consists of 39 local unselected Kenya open-pollinated maize populations evaluated in three environments.

Trials in this group were carried out in randomised complete blocks with three replicates. The three sets of data were earlier described by Ajala and Saxena (1994). One of the major objectives of these

experiments was to determine the potential of the open-pollinated sources of resistance as candidates for recurrent selection aimed at improving levels of resistance to the stem borer and also for good agronomic qualities including yield (Ajala, 1993).

In these experiments, single- and multiple-row plots were planted. A row was divided into two equal parts. At 3 weeks after emergence, plants in one half of the rows were artificially infested with egg batches of C. partellus, reared on artificial diet in the laboratory, containing about 30-40 eggs at the black-head stage. The second half of the row (uninfested) served as the control.

Measures obtained on each entry include foliar damage rating, deadheart and stem tunnelling as damage variables, and grain yield both the infested and control. Leaf foliar (LF) damage, observed at four weeks after infestation, was scored on a 1-9 scale rating (1 = visible damage, 9 = severe damage). Deadheart (DH) is the proportion of plants showing deadhearts in a plot. Stem tunnelling (ST) was estimated at harvest by splitting each plant longitudinally to measure the stem length tunnelled and this was expressed as a percentage of the plant height.

The second group comprises data from laboratory assessed susceptibilities of 77 maize varieties to Sitophilus zeamais (set D). Data on number of adults (N), time to 50% emergence (T) (estimated from the middle of the oviposition period to the emergence of 50% of the F_1 generation) and weight loss (WL), in form of means was

available for analysis. In addition to these variables, Dobie's index of susceptibility (I) was obtained as: $I = \ln(N) * 100/T$. Dobie's index of susceptibility is based on the assumption that the more F_1 progeny and the shorter the duration of development, the more suitable the grain is as a host. Hence, the more susceptible varieties have a higher index.

4.1.2 Empirical distribution of data

Given a set of data, the first step was determination of its underlying distribution and application of appropriate transformations, where necessary. PROC CHART and UNIVARIATE, with NORMAL option, in SAS (SAS Institute Inc., 1985) were used to construct histograms and test for normality (using the Wilk-Shapiro's test (W)), respectively.

Leaf foliar (LF) damage

Departure from normality was observed for LF in data set A in which the variable exhibited positive skewness (Fig 4.1a). Normality was achieved by the logarithmic transformation which implied a lognormal distribution for the variable. The tabulated data on this variable in data sets B and C showed a good approximation to normality (e.g. Fig 4.1b).

Deadheart (DH)

No realistic normal approximation to the data was indicated for this variable in any of the data sets (e.g. Fig 4.1c). DH showed highly skewed distributions, had a wide range of values, with a concentration at zero. Application of the square root transformation to data set E resulted into a distribution that approached normality. Virtually no improvement resulted from power and arcsine transformations on the variable in the other data sets. Exclusion of cases where all replicates had zero values, gave fairly good approximations to the normal distribution, suggesting a possibility of distribution mixtures. In effect the remaining (non-zero) values would be measuring severity (intensity) of DH. To avoid omitting varying levels of this variable from the test entries, the variable was regarded as discrete (which is the case before the number of plants with the attribute is expressed as percentage of the total number of plants in the plot) and scored as follows were a few groups, defined by, for example,

score = 0 for $x = 0$ (no DHs)

1 for $0 < x \leq 15$

2 for $15 < x \leq 25$

etc.,

where x is the observed percentage of plants with deadhearts. Means of the scores obtained were then used in subsequent analyses.

Stem tunnelling (ST)

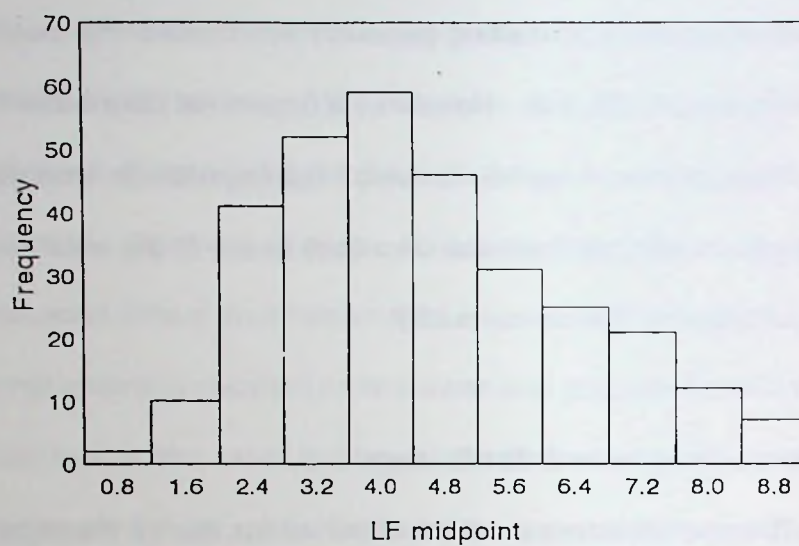
Data on stem tunnelling generally exhibited some positive skewness (e.g. Fig 4.1d). However, a lognormal distribution fitted well to data in sets A and B, to which the logarithmic transformation was applied with good results. For data in set C the arcsine transformation achieved normality.

Discussion on observed distributions

The normal approximation obtained for the LF damage data is due to the fact that each plant was scored individually and a mean score for the plot was then computed. Using the Central Limit Theorem (CLT), when n is large, the distribution of the mean tends to normality, regardless of the distribution of the individual scores.

ST, on the other hand involves a ratio of two varying measures of the stem (tunnel length and stem length (plant height)). If the tunnel length is the same for a set of plants, the length of the stem may vary with different plants resulting in ST being inversely proportional to the length of the stem. Multiplicative variables are usually normalised by taking the logarithms of the variates. DH, also measured as a percentage, has a constant denominator (total number of plants in a plot) for all entries. Therefore the differences in number of plants with the attribute among entries will be relatively maintained when converted to percentages. Perhaps this explains the differences in distributions of ST and DH.

(a)



(b)

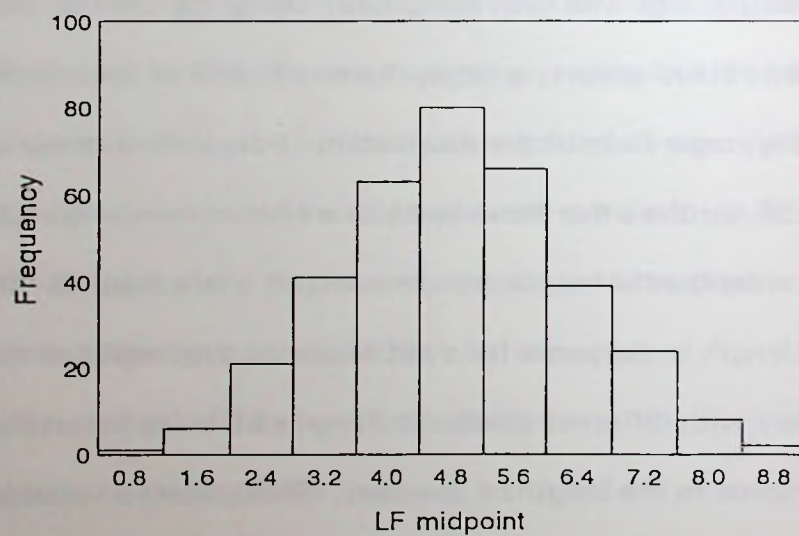
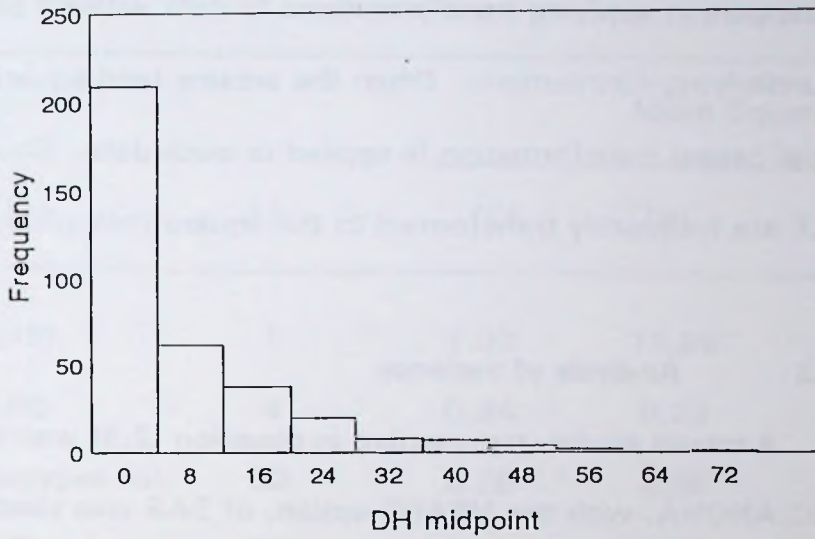


Figure 4.1 Frequency distribution for: (a) A-LF showing a skewed distribution; and (b) C-LF showing a normal approximation for the variable.

(c)



(d)

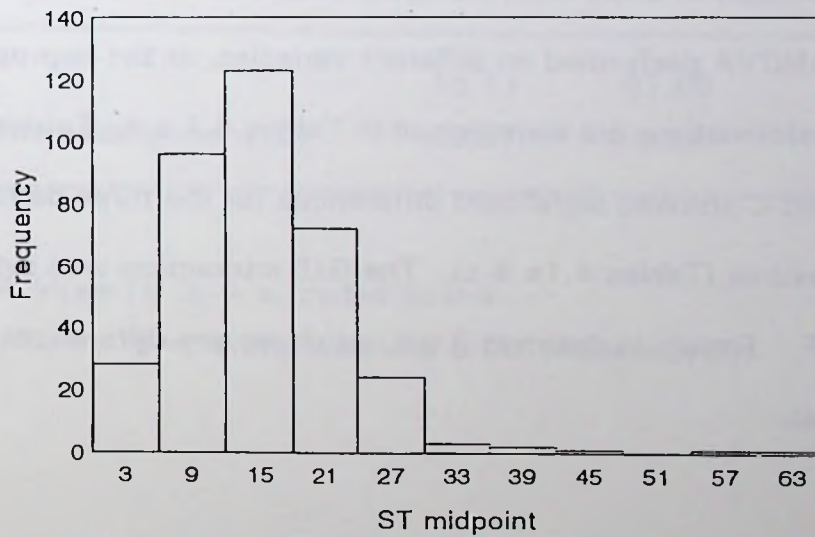


Figure 4.1 Frequency distribution for: (c) C-DH, highly skewed with a concentration at zero and (d) C-ST, a skewed distribution normalised by arcsine transformation.

Results of these empirical distributions of the variables revealed the dangers in applying transformations to data without prior study of the underlying distributions. Often the arcsine (and square root, in special cases) transformation is applied to such data. Scores as they are for LF are frequently transformed to the square root scale.

4.1.3 Analysis of variance

A mixed model, represented in equation (2.3) was assumed. PROC ANOVA, with the MEANS option, of SAS was used for the analysis of variance (combined across environments) and the estimation of entry means and the respective standard errors. Results of ANOVA performed on different variables, or the appropriate transformation, are summarised in Tables 4.1 a-c. Entries in data set A and C showed significant differences for the three damage measures (Tables 4.1a & c). The GxE interaction was significant for C-LF. Entries in data set B did not show any differences in LF (Table 4.1b).

Table 4.1a Analysis of variance summary for data set A.

Source	d.f	Mean Squares		
		⁺ LF	[*] DH	⁺ ST
Env (E)	1	1.93	18.88 ^{***}	125.93 ^{***}
Rep(R)	4	0.34	0.23	0.07
Genotypes (G)	50	0.28 ^{***}	2.36 [*]	0.70 ^{**}
G x E	50	0.08	1.32	0.28
Error	200	0.06	0.93	0.22
CV		15.11	87.99	17.60
Variance component (G)		28.37	4.47	5.53

⁺x = $\ln(x+1)$, ^{*}x = x, coded as 0-4

^{*}, ^{**}, ^{***}: p < 0.05, 0.01 and 0.001, respectively.

Table 4.1b Analysis of variance summary for data set B.

Source	d.f	Mean Squares		
		*LF	**DH	*ST
Env (E)	1	40.77	0.01	24.45
Rep(R)	4	16.87	0.14	7.33
Genotypes (G)	54	1.26	0.12***	0.25*
G x E	54	0.94	0.06	0.19
Error	216	1.00	0.04	0.14
CV		23.64	19.19	13.63
Variance component (G)		3.63	2.64	2.62

*x = $\ln(x+1)$, **x = $\sqrt{(x+1)}$, .

*, **, *** = p < 0.05, 0.01 and 0.001, respectively.

Table 4.1c Analysis of variance summary for data set C.

Source	d.f	Mean Squares		
		LF	[*] DH	^{**} ST
Env (E)	2	25.14	28.77	835.87
Rep(R)	6	14.47	6.01	202.86
Genotypes (G)	38	2.41 [*]	1.87 ^{**}	38.54 [*]
G x E	76	2.15 [*]	0.79	27.53
Error	228	1.41	0.78	25.20
CV		22.12	67.19	22.91
Variance component (G)		1.39	9.77	3.29

^{*}x = x, coded as 0-4, ^{**}x = arcsin($\sqrt{x}$).

^{*}, ^{**}, ^{***} = p < 0.05, 0.01 and 0.001, respectively.

4.1.4 Correlations among variables

A study of the linear relationships between traits was carried out. PROC CORR of SAS was used to compute and test for significance of the correlations between pairs of damage variables. While data sets A and D have some highly correlated variables, data set B has noncorrelated variables, and set C has fairly low correlated variables (see Table 4.2).

4.1.5 Comment on real data

The data sets included in this study represent a range of sources of variation that occur in practical situations. These include distribution pattern, mean, variance and covariance structure, size of data and/or number of entries (n), number of resistance groups (g), size of each group (s) and the linear relationships among traits (r). These are inherent characteristics of the data which the investigator has no control of. Application of the classification procedures to the data will identify characteristics in the data responsible for differences in their performance.

Table 4.2 Correlations among damage variables in the different data sets.

Group I		LF	DH	ST
Data set A	DH	0.563***	-	-
	ST	0.771***	0.716***	-
Data set B	DH	0.001 _{ns}	-	-
	ST	0.005 _{ns}	-0.101 _{ns}	-
Data set C	DH	0.305 _{ns}	-	-
	ST	0.456*	-0.301 _{ns}	-

Group II		N	T	WL
Data set D	T	-0.412***	-	-
	WL	0.943***	0.441***	-

*, **, *** = $p < 0.05, 0.01, 0.001$, respectively.

ns = not significant.

4.2 SIMULATED DATA

Random normally distributed variables were simulated with a range of specified parameter values that included means and variances as well as correlations/covariances in multivariate cases. The following functions were used for this purpose:

$$X = \mu_X + \sigma_X Z_1,$$

$$Y = \mu_Y + \rho \sigma_Y Z_1 + \sigma_Y \sqrt{1 - \rho^2} Z_2,$$

where X and Y are the generated variables, Z_1 and Z_2 are independent standardised normal random variables, μ and σ the mean and standard deviation and ρ is the correlation coefficient between X and Y . This in effect, simulates sets of data by adding random variation from a normal distribution with zero mean and standard deviation 1 (Brown and Rothery, 1993).

Simulation of a mixture of distributions was made possible by specifying all the parameters pertaining to the component distributions. This entailed inclusion of the number of component distributions, the size of each component (group size) and the earlier mentioned parameters, within each group. A model SAS program for the simulations is presented in Appendix A1.

The simulated data sets used in the study contained either 2, 3, or 4 distinct groups. Data sets consisted of a total of 30, 60, 120, 240, 600 or 1200 observations (representing entries).

The design factors include

Data size n_i , $i = 1, 2, 3, 4, 5, 6$;

Number of groups, g_j , $j = 1, 2, 3, 4$;

Relative size of groups s_k ,

$k = 1$ for $j = 1$,

$k = 1, 2$ for $j = 2$, and

$k = 1, 2, 3$ for $j = 3, 4$:

s_1 had an equal number of entries in each group, s_2 required that one group must contain 10% of the data observations, whereas s_3 required that one group must contain 50% of the entries. The remaining entries were distributed equally (or as equally as possible) across the other groups represented in the data. Choice for the last two levels was made considering the common selection intensities investigators have used. Ten percent is among the most common, while 50% is frequently used at early breeding stages to reduce the number of genotypes to be tested during subsequent stages of the evaluation. The first level was made mainly on intuitive grounds.

For a full factorial design, the number of cells expected, at this level, would be $n \times g \times s = 6 \times 4 \times 3 = 72$. At g_1 , that is $g = 1$ (no grouping), the s levels do not apply. With $g = 2$, $s_1 = s_3$ as equal group size implies 50% entries in each group. Overall, the design had $(1 + 2 + 3 + 3) \times 6 = 54$ cells (Table 4.3). Two replicates (t) were generated for each cell of the design, producing 108 data sets for testing purposes.

Additional factors, considered in the multivariate case, are number of variables ($p = 2$ or 3) and the intercorrelations among variables. Correlation coefficients (r) of about 0.6 and 0.9 were used on some selected data sets. Included in these were all the combinations of $p = 2$ & 3 as well as $n = 30$ & 60 which gave another 36 $((1 + 2 + 3 + 3) \times 2 \times 2)$ data sets. The 54 combinations mentioned earlier (analogues of the univariate) had uncorrelated variables (i.e. $r \approx 0$). Thus the multivariate design had a total of 180 different data sets and two replicates ($= 180$). For $p = 1$ (i.e. univariate), two additional replicates were obtained by including some of the variables generated for multivariate classification purposes.

Table 4.3 Simulation factor combinations.

n	Number of groups (g) and group size							
	1	2	3	4	5	6	7	8
30	30	15, 15 3, 27	10, 10, 10 3, 13, 14 15, 7, 8	7, 7, 8, 8 3, 9, 9, 9 15, 5, 5, 5				
60	60	30, 30 6, 54	20, 20, 20 6, 27, 27 30, 15, 15	15, 15, 15, 15 6, 18, 18, 18 30, 10, 10, 10				
120	120	60, 60 12, 108	40, 40, 40 12, 54, 54 60, 30, 30	30, 30, 30, 30 12, 36, 36, 36 60, 20, 20, 20				
240	240	120, 120 24, 216	80, 80, 80 24, 108, 108 120, 60, 60	60, 60, 60, 60 24, 72, 72, 72 120, 40, 40, 40				
600	600	300, 300 60, 540	200, 200, 200 60, 270, 270 300, 150, 150	150, 150, 150, 150 60, 180, 180, 180 300, 100, 100, 100				
1200	1200	600, 600 120, 1080	400, 400, 400 120, 540, 540 600, 300, 300	300, 300, 300, 300 120, 360, 360, 360 600, 200, 200, 200				

Note: Rows within each set of combinations of n represent levels of sizes of groups, s (s_1, s_2, s_3). While s_1 denotes a set with equal groups, s_2 and s_3 denote data sets in which one of the groups contains one-tenth and half of the entries, respectively.

CHAPTER 5

APPLICATIONS AND RESULTS

Introduction

In this chapter we investigate the practical use of the procedures described in Chapter 3 using sets of real and simulated data described in Chapter 4. To make direct comparisons of the performance of the procedures, each data set was subjected to all procedures.

SAS, and in particular the SAS/STAT, facilities were used at various stages of the analysis, supplemented by LOTUS 1-2-3. Some programs were written in the BASIC language to carry out specific tasks. HARVARD GRAPHICS was employed for printing various graphs.

5.1 EVALUATION OF UNIVARIATE PROCEDURES

Data arrays

Prior to application of the classification procedures, entries were ranked in ascending order based on the mean values for each classification variable. This order would be maintained throughout each analysis. For example, after ranking n entries by stem tunnel (ST), the entry in position 1 was the most resistant with respect to that particular measure, and the entry in position n was the most susceptible. This logical start meant that entries close together or

ranking (and thus have small differences in their resistance levels) were more likely to be grouped together than those that are far apart in their rankings (and possibly have large differences between their resistance levels).

5.1.1 The graphical procedure

The normal scores were obtained using PROC RANK with NORMAL = BLOM option in the SAS program. The observed means were then plotted against the corresponding normal scores. The line given by equation (3.2) was computed and plotted on the same graph. Visual inspection of the plots, discussed in Section 5.2, indicated possibilities of groupings.

5.1.2 Cluster analysis

Cluster analyses were carried out by specifying a range of g values in the MAXCLUS option of PROC FASTCLUS of the SAS program. The procedure performs a disjoint grouping of entries based on Euclidean distances computed from the damage variable(s).

For this particular procedure, values of g ranged from 2 to $n/5$ (approximately) to allow investigations of any possible trend in the statistics, PF and CCC, employed to determine the optimum number of classes.

5.1.3 Hypothesis testing procedure

Classification models formulated in Section 3.1.3.2 were tested for their adequacy using the proposed test statistics. The likelihood ratio test (LRT) procedure required computation of test statistics considering all possible arrangements of n entries into g classes. The values of g investigated using this test statistic ranged from 1 to 6 in some cases). Similarly, quasi bayesian test statistics (QBT) had to be computed for all possible partitions of the n entries into 2 classes ($g = 2$).

5.1.3.1 *Generation of groupings*

This study employed an iterative technique. For example in the case of two classes, the entries were first partitioned into two such that the top $n-1$ are in class 1 and the last one in class 2. For the next grouping, the top $n-2$ entries are placed in class 1 and the last two in class 2. The process is repeated until we have only the top entry in class 1 and the other $n-1$ entries are in class 2. For example, if we arranged 51 entries in ascending order with respect to variable means, we could partition the entries in 50, i.e. $(n-1)$ different ways: 1-50, 1; 1-49, 50-51; ... ; 1, 2-50. In general, for any number of classes, g , a similar procedure was made to generate a series of all possible groupings. To partition n entries into g classes, the total number of partitions considered is given by

$$\binom{n-1}{g-1} = \frac{(n-1)!}{(g-1)! (n-g)!}.$$

----- (5.1)

For the different real data sets described in Chapter 4, the number of possible groupings to consider is shown in Table 5.1.

Table 5.1 Number of possible groupings for different combinations of number of entries (n) and number of classes (g).

Data set	n	Number of classes (g)			
		2	3	4	5
A	51	50	1225	19600	230300
B	55	54	1431	24804	316251
C	39	38	703	8436	73815
D	77	76	2850	70300	1282975 ¹

¹ Not all groupings were investigated due to limitations of computer memory.

A restriction was made to have at least two entries assigned to each class to avoid the degenerate case of infinite test statistics.

5.1.3.2 *Computation of test statistics*

Test statistics were computed for each assumed cutpoint case of two classes, for each pair of cutpoints for three classes, and for each set of $g-1$ hypothesised cutpoints for g classes.

QBT statistics were computed for all possible groupings into two and the grouping with the minimum value considered optimal. LOTUS 1-2-3 was fast at the computations which only required shifting the single cutpoint by one place to produce all the possible groupings and copying one equation into the appropriate cells to obtain the statistics.

The LRT statistics were obtained by running one-way ANOVA to generate values of between and within classes sum of squares, F-ratio and R^2 . For this purpose, some programs were specifically written in BASIC language to generate various groupings. An example of the programs is given in Appendix A2. The routine programs generated the partitions iteratively and computed the required statistics for each iteration. Maximum values of the test statistics indicated the best classification of entries into the given number of classes. Such a grouping, which essentially maximised differences among the various classes, provided the best evidence for rejecting the 1-group model in favour of the g -group model.

The minimum (or sharp fall in) value of the g^2w statistic (g^2w within classes mean squares) indicated the optimum partition number of classes, where this was achieved. The F-ratio comparing the

reduction in within classes variation when the g-1 classes model is adopted against the g groups model was used to choose between g-1 or g groups models.

5.2 RESULTS AND DISCUSSION OF UNIVARIATE PROCEDURES

Note:

- (i) Comparisons involving QBT will only be made with respect to the 2-group model, since use of the statistic assumes a 2-group model is adequate and the question is to locate the single cutpoint.
- (ii) In plots of various test statistics (LRT, CA (PF & CC) and QBT), the points are joined by lines, not meant to depict continuity of cutpoints, but to emphasize the trend which facilitates the determination of these points.
- (iii) Only one example of plots of QBT statistic is included for variable A-LF, since the plots provide no additional information from that given by the tabulated values of the statistic for different locations of the cutpoint.

The optimum classifications achieved by different procedures are summarised in Table 5.3.

5.2.1 DATA SET A

5.2.1.1 Leaf feeding (A - LF)

The normal score (NS) plot (Fig 5.1a) suggests 3 groups. The LRT plot (Fig 5.1b) seems to suggest 2 cutpoints - a definite one at the first 36 and a less definite one after the first 14 entries. The test statistics plot (Fig 5.1c) suggests 2 optimal classifications at 3 & 6. Computed values of g^2w , for $g = 2-5$, confirm the optimum number of groups being 3 (see Table 5.2a).

Table 5.2a Computed values of g^2w for the damage variables (data set A).

g	Variable		
	LF	DH	ST
1	0.047	5.64	32.99
2	0.059	5.56	37.22
3	0.049	3.72	42.76
4	0.061	4.02	33.99
5	0.089	3.97	32.87
6	-	-	22.50

While QBT gives groups of 28 & 23 (Fig 5.1d), LRT and CA give groups of 36 & 15 as the best partition into 2 groups.

Disagreement in location of cutpoints, on the assumption of a 2-groups model, also seems to suggest a different model. Both LRT and CA agree on composition (number of entries, n_j , in group j) of the 3 groups, whose estimates are given by:

j	n_j	mean (s.e.)
1	14	3.23 (0.09)
2	22	4.18 (0.08)
3	15	6.14 (0.18)

Using the conventional selection intensity of 10% based on LF, the ranking of the entries would have resulted into 5 entries being selected. Vital resources in the 9 entries left out would be lost.

5.2.1.2 Dead heart (A - DH)

There are a few points in the NS plot (Fig 5.2a), due to so many ties in the data. One point may be representing up to ten entries, while another one may be representing a single entry. Nevertheless, the plot suggests 3 classes. The same number is suggested by the LRT plot (Fig 5.2b) and CA (Fig 5.2c). The minimum value of g^2w is given by $g = 3$ (Table 5.2a). The group estimates are given by

j	n_j	mean (s.e.)
1	38	3.76 (0.36)
2	11	11.15 (0.91)
3	2	34.62 (5.12)

While QBT gives classes of 30 & 21, CA and LRT procedures give classes of 40 & 11 entries. The wide difference in results seems to suggest inadequacy of the 2-group model or, provide additional evidence to inadequacy of the model.

Conventional selection intensity of 10% based on this variable would have left out over 30 entries of potential benefit.

5.2.1.3 Stem tunnelling (A - ST)

The NS plot (Fig 5.3a) suggests at least 5 groups, with 2 definite groups on the extreme right, above the line. At least 3 groups are suggested by the LRT plot (Fig 5.3b), while CA (Fig 5.3c) gives 4 groups. The value of g^2w (Table 5.3a) shows a marked decrease in variance for $g = 6$. The QBT grouping into 30 & 21 differs from the LRT and CA grouping into 40 & 11 entries, dismissing the likelihood of the 2-group model. The group estimates are given by:

j	n_j	mean (s.e.)
1	4	11.04 (0.45)
2	19	14.06 (0.15)
3	9	16.77 (0.23)
4	9	19.82 (0.29)
5	7	25.46 (0.34)
6	3	33.27 (0.75)

It may be of particular interest to note how the two groups 5 & 6 (with large differences from others and among themselves), were clearly indicated by the NS plot (Fig 5.3a).

Although the best group contains 4 entries which is very close to 5, the number selected by the conventional ranking method, the inclusion of the fifth entry from group 2 while leaving out the other 18 in the same group may not be justified.

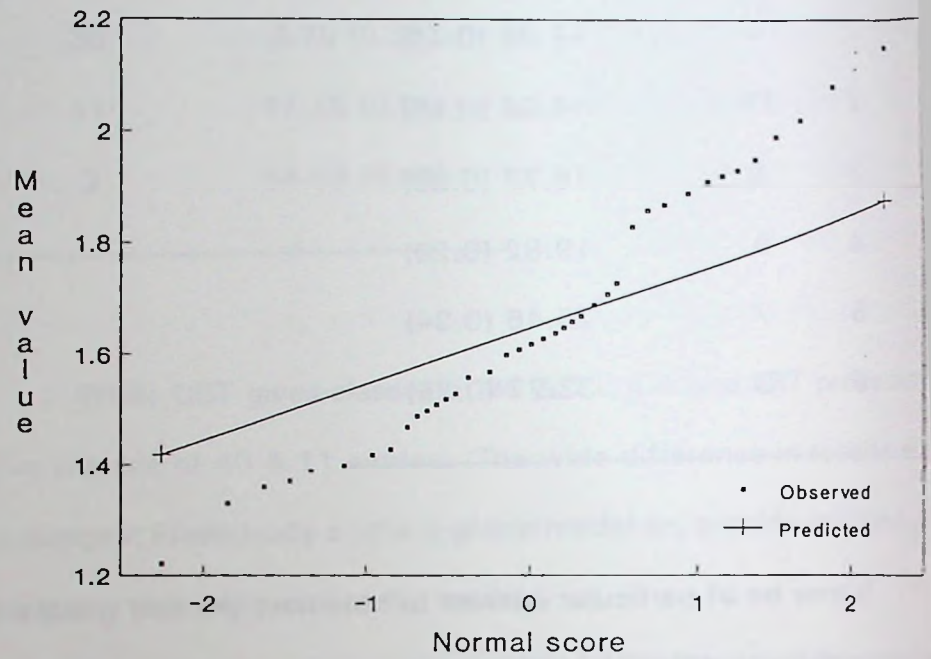


Figure 5.1a Plot of A-LF for the 51 entries against their normal scores.

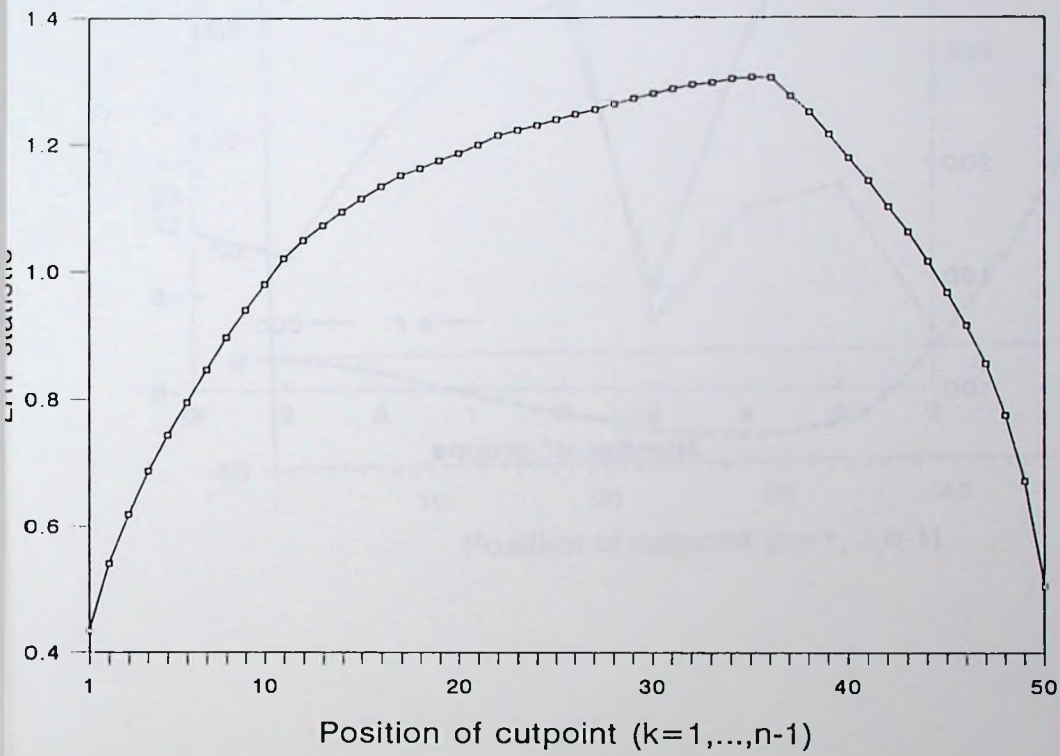


Figure 5.1b Plot of LRT statistic for the 2-group model against hypothetical cutpoints for A-LF.

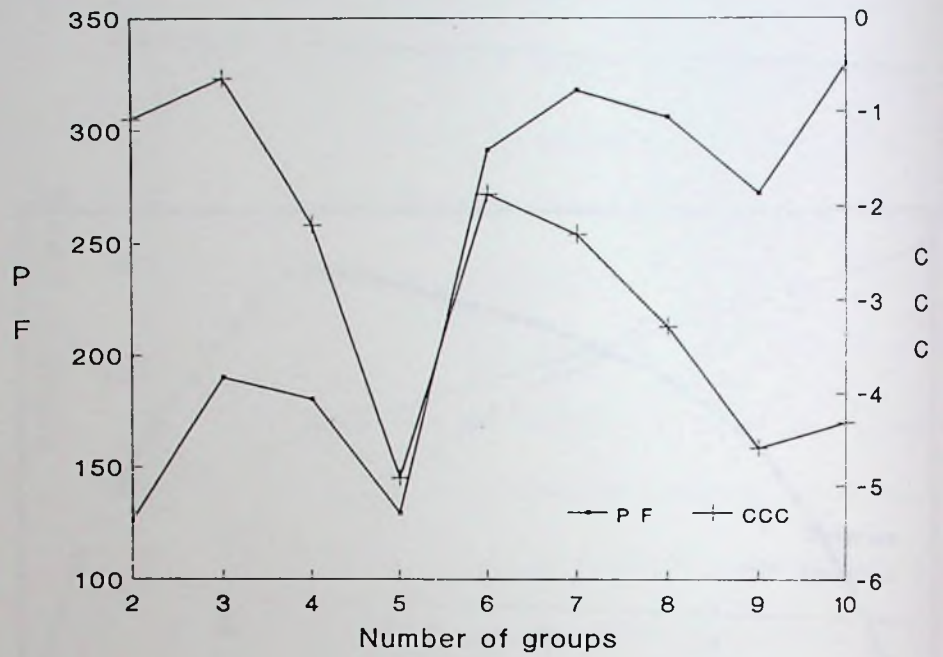


Figure 5.1c Plot of PF and CCC for optimum classifications groups with respect to A-LF.

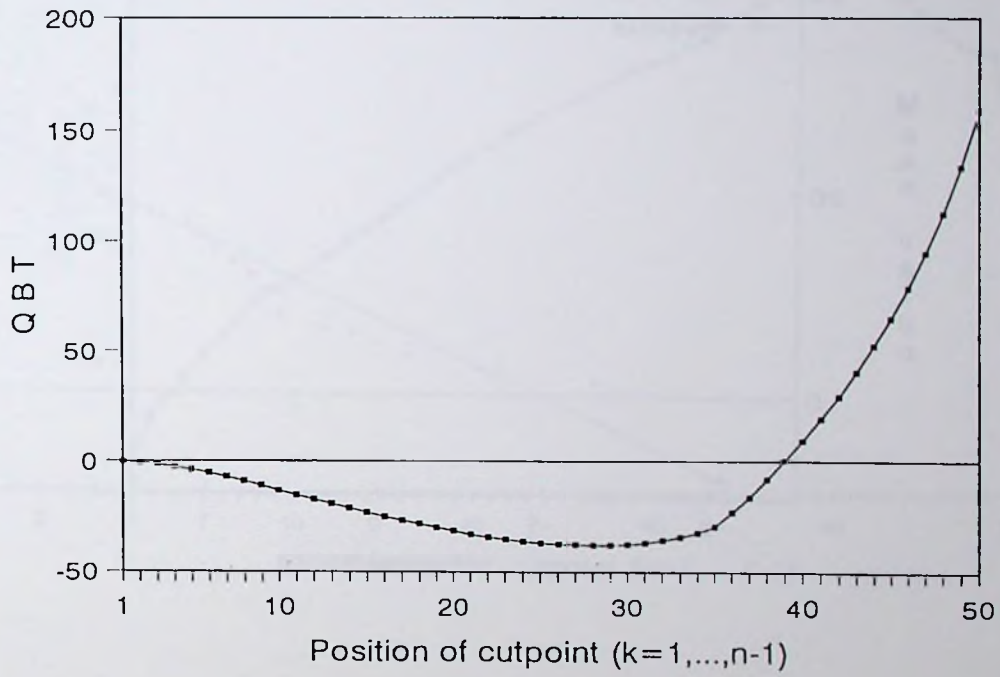


Figure 5.1d Plot of QBT statistic for the 2-group model against hypothetical cutpoints for A-LF.

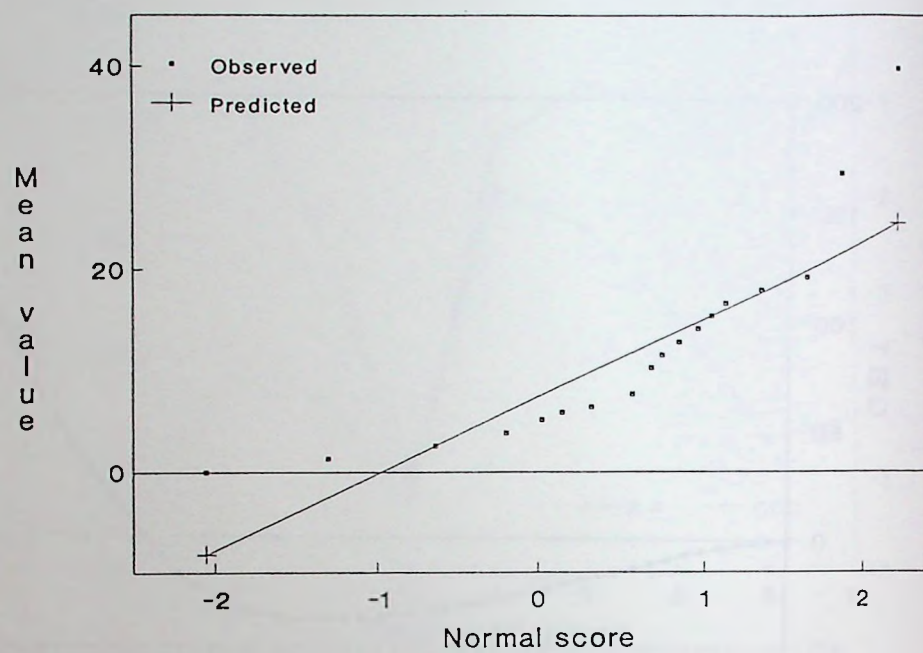


Figure 5.2a Plot of A-DH for the 51 entries against their re normal scores.

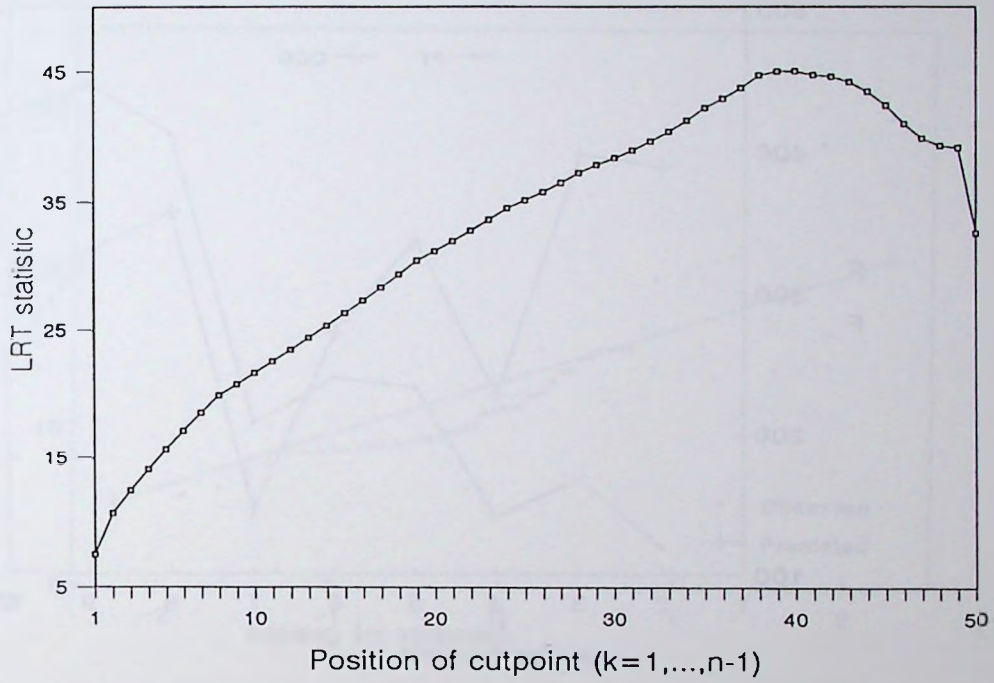


Figure 5.2b Plot of LRT statistic for the 2-group model against hypothetical cutpoints for A-DH.

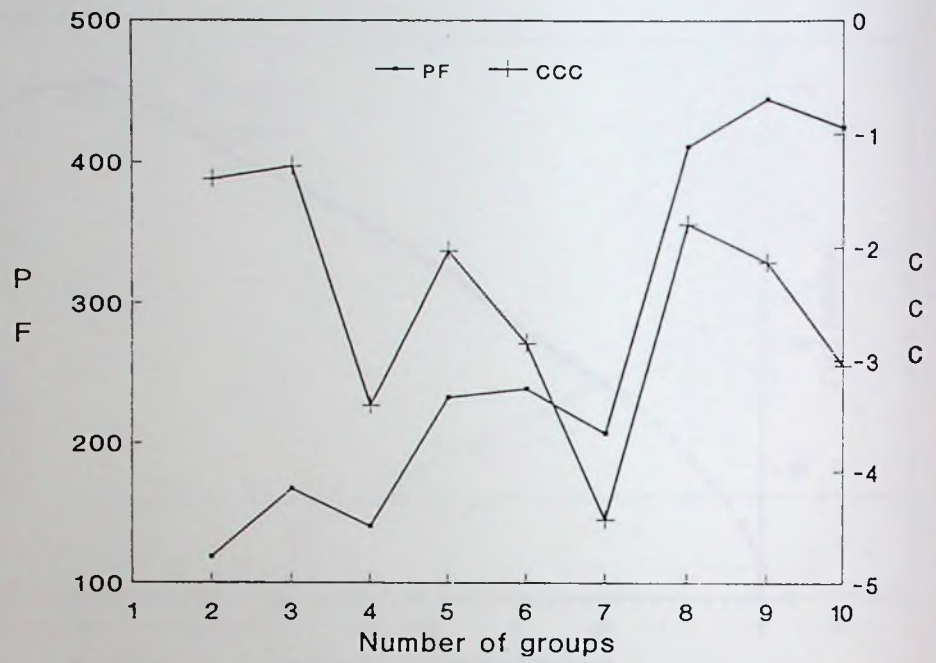


Figure 5.2c Plot of PF and CCC for optimum classifications groups with respect to A-DH.

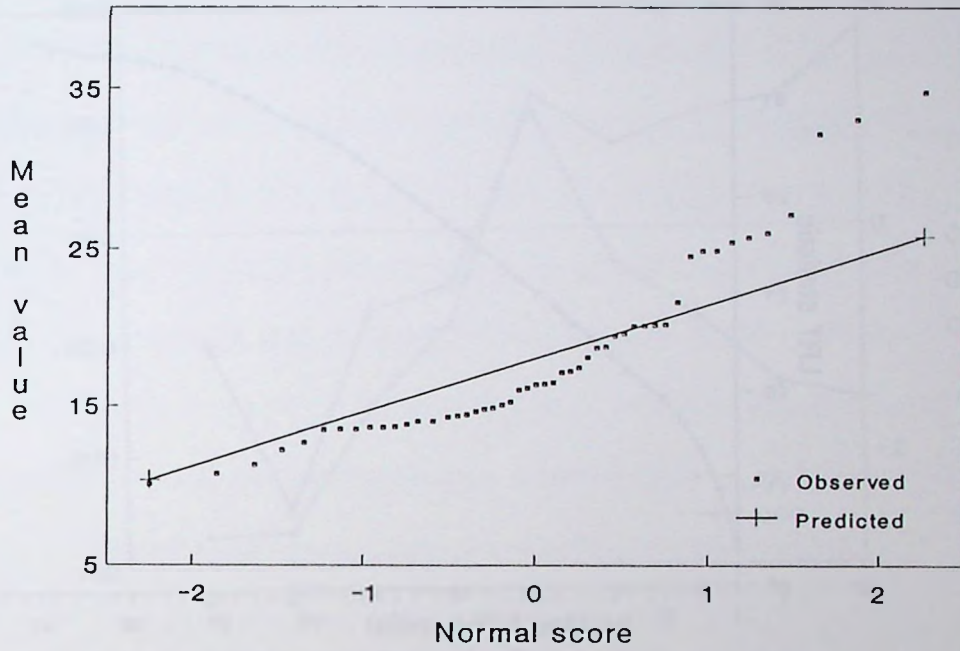


Figure 5.3a Plot of A-ST for the 51 entries against their respective normal scores.

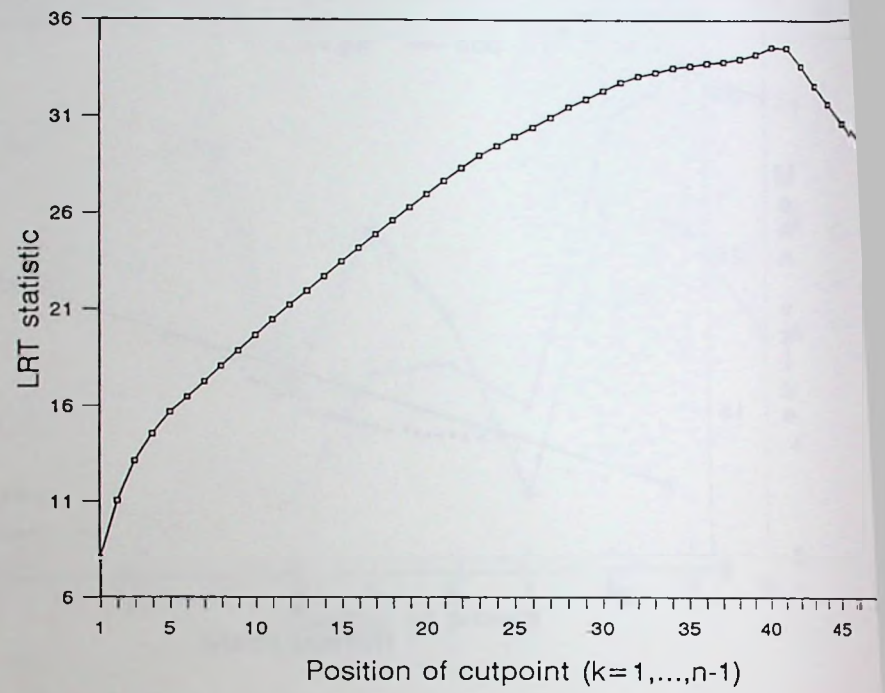


Figure 5.3b Plot of LRT statistic for the 2-group model against hypothetical cutpoints for A-ST.

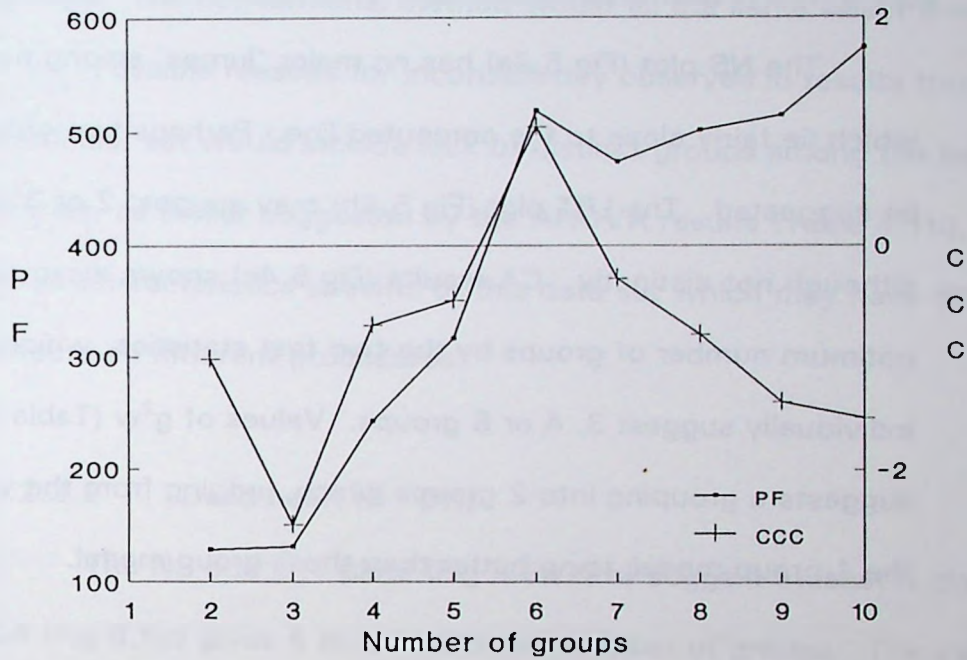


Figure 5.3c Plot of PF and CCC for optimum classifications into g groups with respect to A-ST.

5.2.2 DATA SET B

5.2.2.1 Leaf feeding (B - LF)

The NS plot (Fig 5.4a) has no major 'jumps' among points which lie fairly close to the computed line. Perhaps two groups may be suggested. The LRT plot (Fig 5.4b) may suggest 2 or 3 groups although not distinctly. CA results (Fig 5.4c) shows disagreement in optimum number of groups by the two test statistics, which individually suggest 3, 4 or 8 groups. Values of g^2w (Table 5.2b) suggests a grouping into 2 groups which, judging from the value of the 1-group model, is no better than the 1-group model.

Table 5.2b Computed values of g^2w for the damage variables (data set B).

g	Variable		
	LF	DH	ST
1	0.210	0.18	13.17
2	0.277	0.20	24.14
3	0.278	0.22	23.94
4	0.330	0.16	24.11
5	0.280	0.13	19.78

QBT divides the entries into groups of 28 & 27, while LRT suggests groups of 36, 19. The inconclusive results would lead to unreliable

selections as choice would have to be made between 2, 3 and 4 groups. The conventional method would all the same select 6 entries.

Possible reasons for inconsistency observed in results from this particular set would include lack of distinct groups among the tested entries, as earlier suggested by the ANOVA results (Table 4.1b), and other characteristics specific to this data set which may have different effects on different procedures.

5.2.2.2 Dead heart (B - DH)

The NS and LRT plots (Fig 5.5a & b) suggest at least 4 classes. CA (Fig 5.5c) gives 4 as the optimum number of groups. The values of g^2w suggest at least 5 groups (being minimum), though not significantly different from 4 groups value as confirmed by the F statistic value (1.86, $p > 0.05$), $F = (D_4 - D_5)/(D_5/50) \sim F_{1,50}$. The estimates for the 4-group model are given by

j	n_j	mean (s.e.)
1	36	0.076 (0.014)
2	13	0.368 (0.021)
3	4	0.918 (0.108)
4	2	2.000 (0.170)

The conventional 10% selection intensity would result into 5 selections or 11 if increased to 20%. It may be noted that 20 entries in this data set had mean = 0 and 16 had mean = 0.17, etc. This further indicates how the conventional method would give misleading results without consideration for the structure in the data. Separating the 6 (10%) or 11 (20%) entries from the set of 20 that scored exactly the same would be difficult.

5.2.2.3 Stem tunnelling (B - ST)

The NS and LRT plots (Figs 5.6a & b) suggest at least 5 groups. CA (Fig 5.6c) indicates optimal number of groups at $g = 6$ & 8. Values of g^2w (Table 5.2b) suggest that either the 1-group model or probably a model with more than 5 groups is optimal. While QBT gives groups of 27 & 28, LRT and CA give groups of 37 & 18 entries. Group estimates for the 6-group model are:

j	n_j	mean (s.e.)
1	6	12.56 (0.62)
2	18	15.35 (0.23)
3	20	17.48 (0.25)
4	7	20.18 (0.44)
5	2	22.78 (0.74)
6	2	30.10 (0.76)

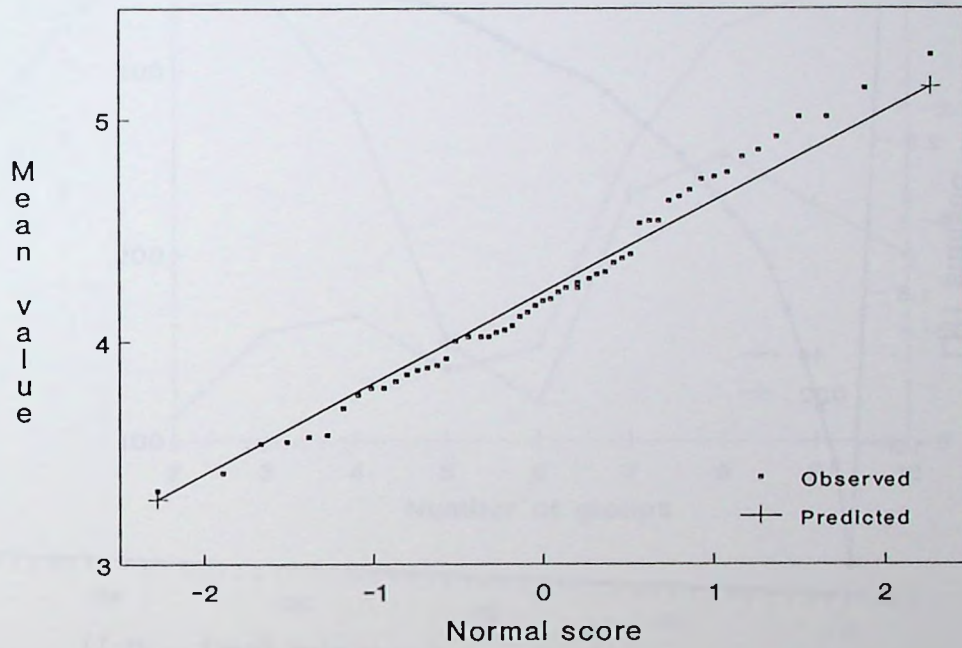


Figure 5.4a Plot of B-LF for the 55 entries against their respective normal scores.

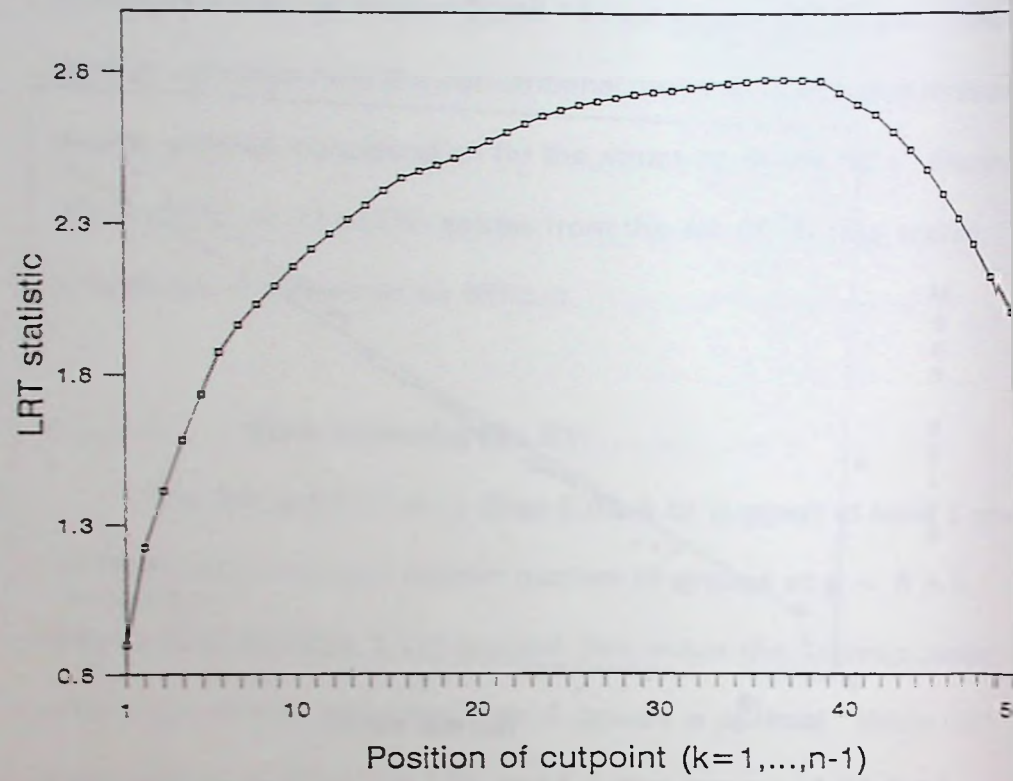


Figure 5.4b Plot of LRT statistic for the 2-group model against hypothetical cutpoints for B-LF.

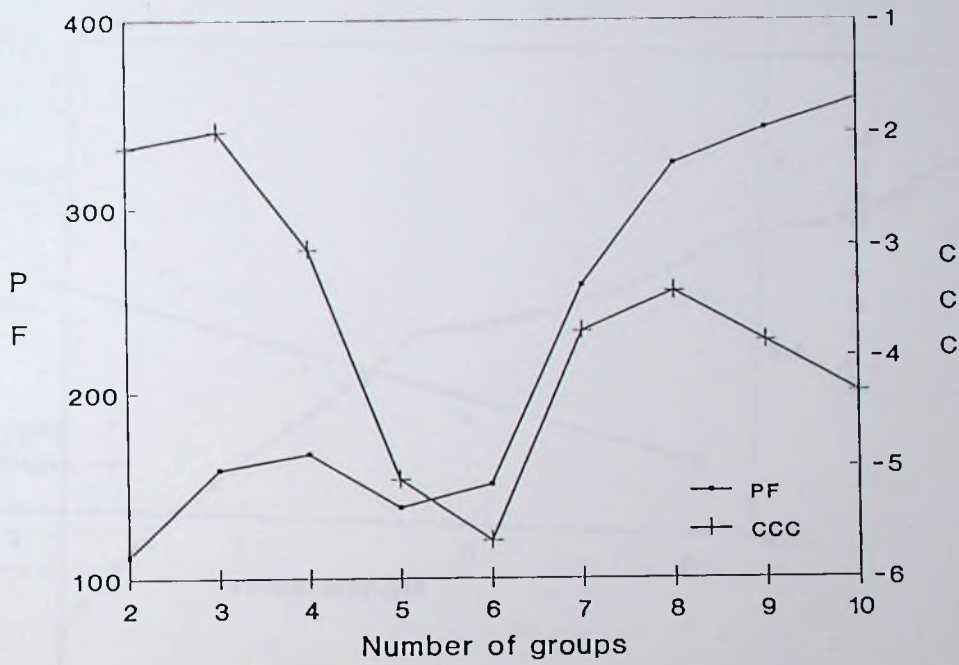


Figure 5.4c Plot of PF and CCC for optimum classifications into g groups with respect to B-LF.

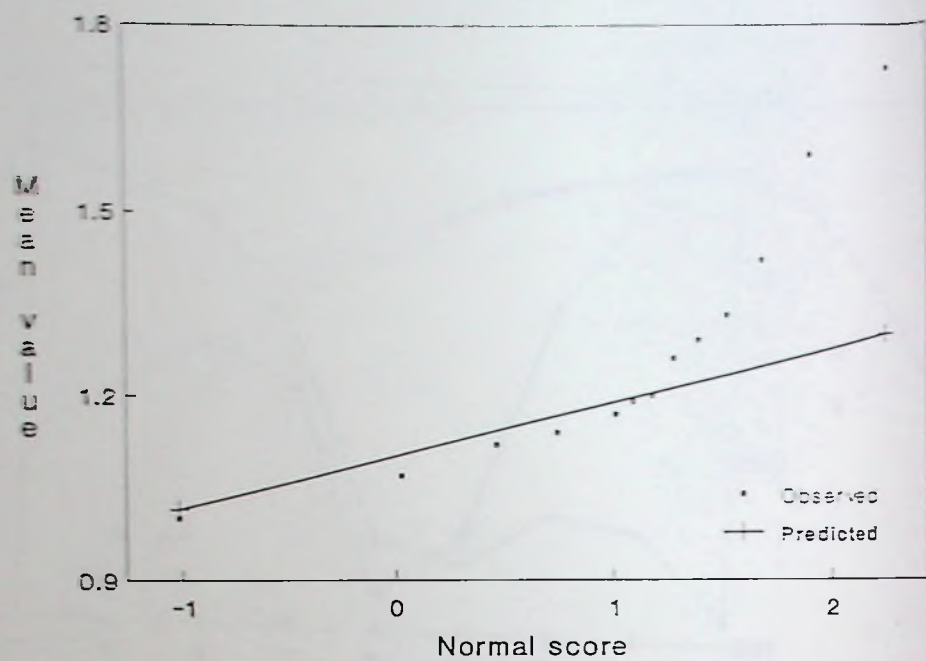


Figure 5.5a Plot of B-DH for the 55 entries against their res
normal scores.

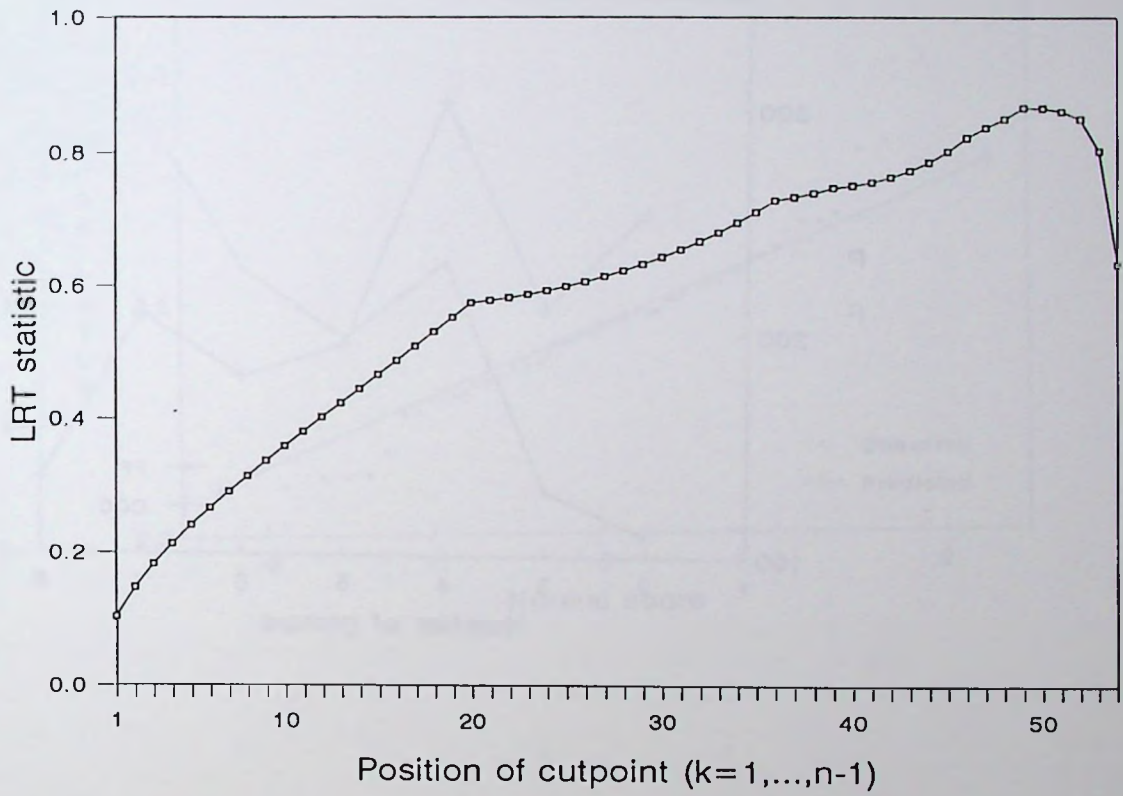


Figure 5.5b Plot of LRT statistic for the 2-group model against hypothetical cutpoints for B-DH.

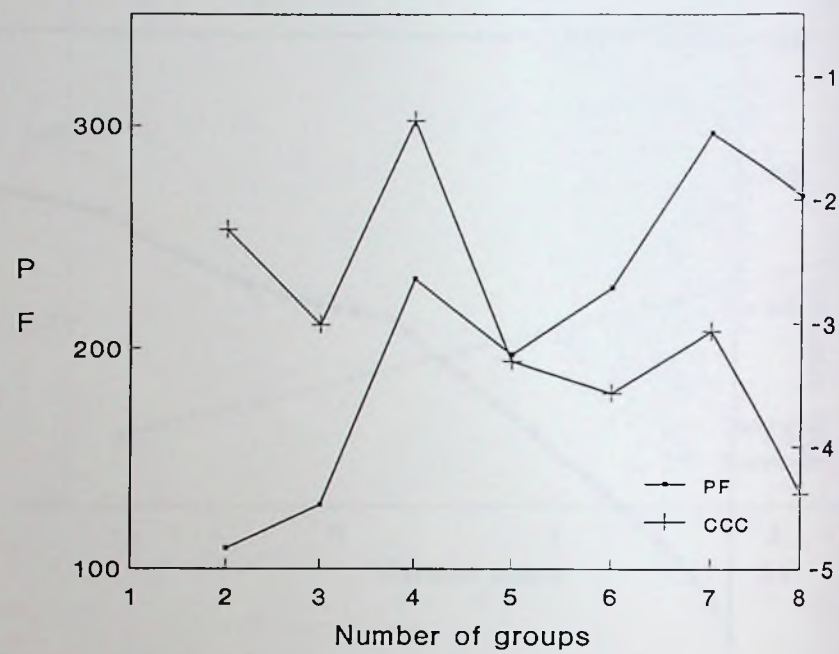


Figure 5.5c Plot of PF and CCC for optimum classifications groups with respect to B-DH.

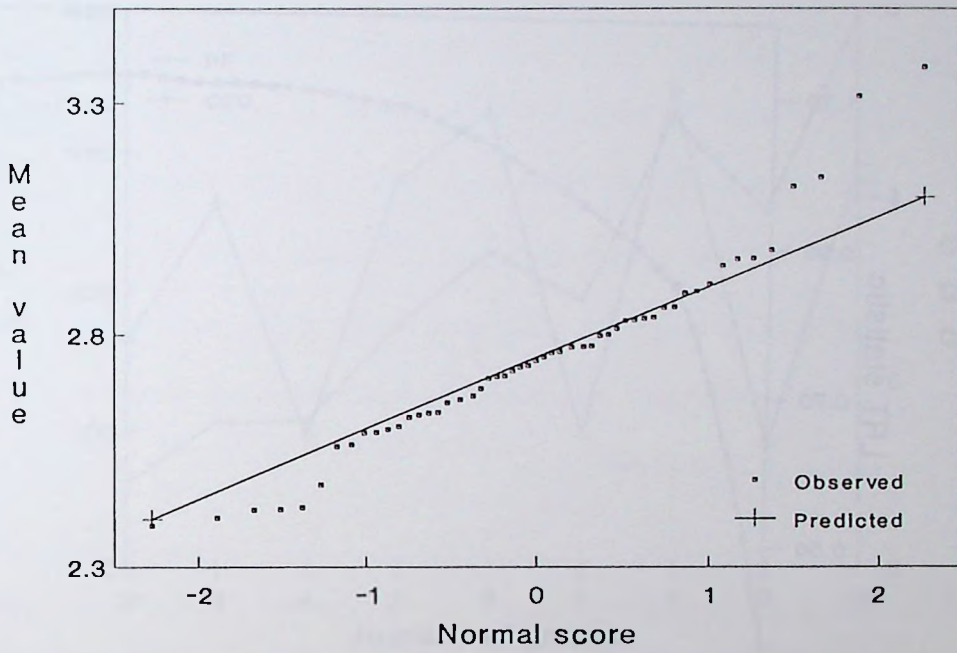


Figure 5.6a Plot of B-ST for the 55 entries against their respective normal scores.

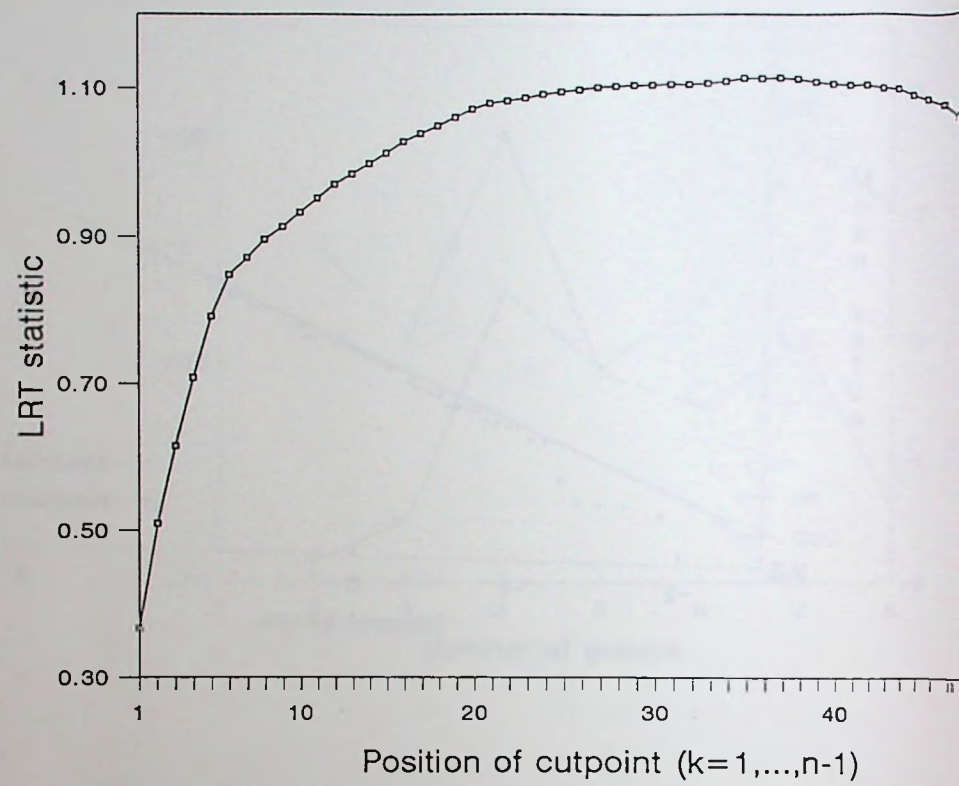


Figure 5.6b Plot of LRT statistic for the 2-group model against hypothetical cutpoints for B-ST.

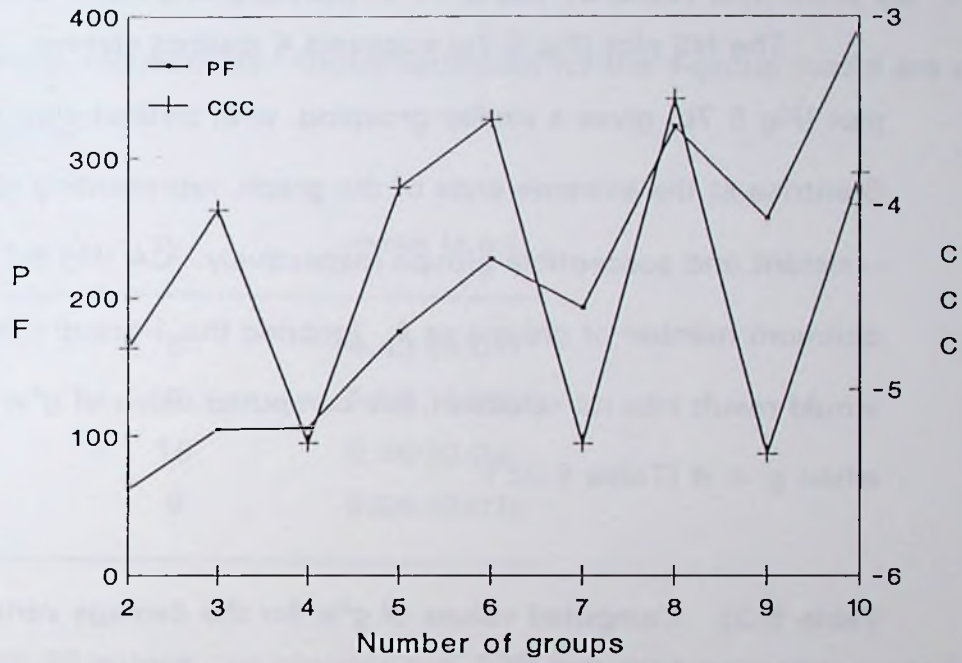


Figure 5.6c Plot of PF and CCC for optimum classifications into g groups with respect to B-ST.

5.2.3 DATA SET C

5.2.3.1 Leaf feeding (C - LF)

The NS plot (Fig 5.7a) suggests 4 distinct classes. The LR plot (Fig 5.7b) gives a similar grouping, with distinct classes of 5 and 8 entries at the extreme ends of the graph, representing highly resistant and susceptible groups respectively. CA (Fig 5.7c) gives optimum number of groups as 4. Ignoring the 1-group model, which would result into no selection, the computed value of g^2w is minimum when $g = 4$ (Table 5.2c).

Table 5.2c Computed values of g^2w for the damage variables (data set C).

g	Variable		
	LF	DH	ST
1	0.268	24.186	5.912
2	0.423	31.501	6.604
3	0.368	28.592	6.919
4	0.320	28.890	5.090
5	0.359	29.562	5.942

All the procedures converge to the same grouping. Inadequacy of the 2-groups model is further indicated by varying sizes/or

composition of the groups provided by various procedures, with QBT, LRT and CA giving groups of 17 & 22, 12 & 27, and, 13 & 26 entries, respectively. Group estimates for the 4-group model are as follows:

j	n_j	mean (s.e.)
1	5	4.43 (0.07)
2	8	5.03 (0.05)
3	18	5.46 (0.03)
4	8	6.08 (0.07)

With 39 entries, the conventional 10% selection intensity would have selected 4 entries leaving out, without justification, one important resource. Increasing this selection intensity to 20% would mean taking additional 3 entries from group 2 and leaving out the 5 others in this group.

5.2.3.2 Dead heart (C - DH)

At least 3 groups are suggested by the NS plot (Fig 5.8a). A minimum of 3 groups are also suggested by the LRT plot (Fig 5.8b). The CA statistics (Fig 5.8c) suggests 2 classes. Ignoring the 1-group model, as in Section 5.2.3.1, values of g^2w suggest $g = 3$ (Table 5.2c). While QBT gives groups of 24 & 15 entries, LRT and CA give

groups of 35 & 4 entries as the optimum classification into two classes. The 3-group model estimates are given by

j	n_j	mean (s.e.)
1	25	4.37 (0.39)
2	10	9.35 (0.41)
3	4	19.02 (1.59)

These results clearly indicate that, whatever model is adopted (2-group or 3-group), the conventional selection would have resulted into marginal loss of resources by leaving out a good number of promising genotypes.

5.2.3.3 Stem tunnelling (C - ST)

The NS and LRT plots (Figs. 5.9a & b) and CA (Fig 5.9c) suggest 4 classes. Furthermore, the value of g^2w at $g = 4$ (Table 5.2c) confirms the optimum number.

The grouping into two classes by LRT and CA (23 & 16 entries) is quite close to the one given by QBT (22, 17 entries). This may be due to the fact that a distinct cutpoint appears at or about point 2.3 while the other two are not very distinct on the plots. The CA test statistics (Fig 5.9c) also indicate a second (local) optimum division.

$g = 2$. The composition of the groups indicated below shows that the four groups are obtained by an initial (major) division into two groups of 23 & 16 entries which are further subdivided into two.

Group estimates are:

j	n_j	mean (s.e.)
1	13	12.15 (0.15)
2	10	14.01 (0.14)
3	10	16.06 (0.17)
4	6	18.09 (0.33)

It has been further demonstrated by these results that the conventional selection could lead to major loss of vital information and resources in these genotypes.

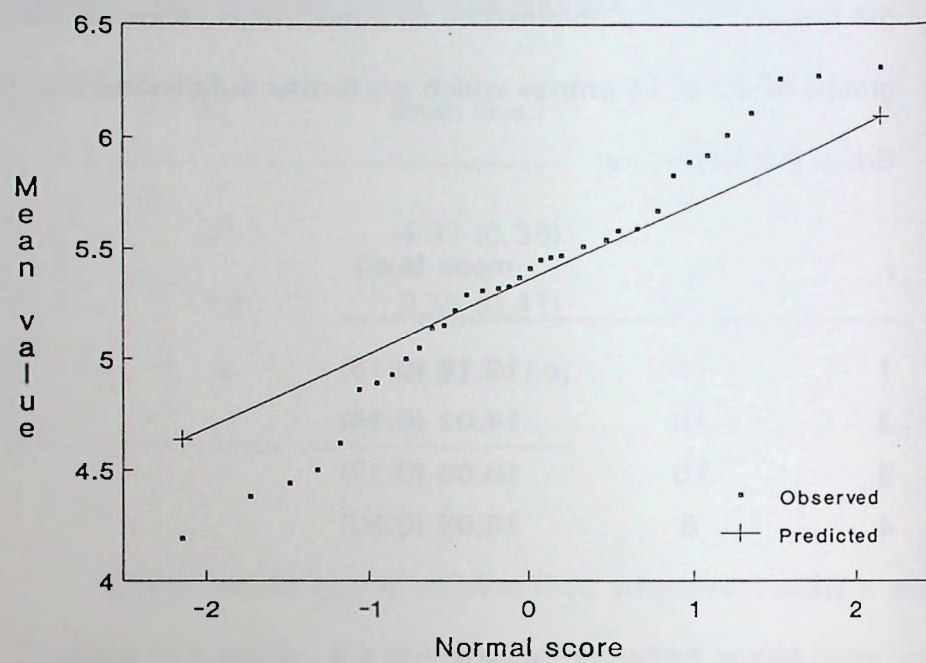


Figure 5.7a Plot of C-LF for the 39 entries against their respective normal scores.

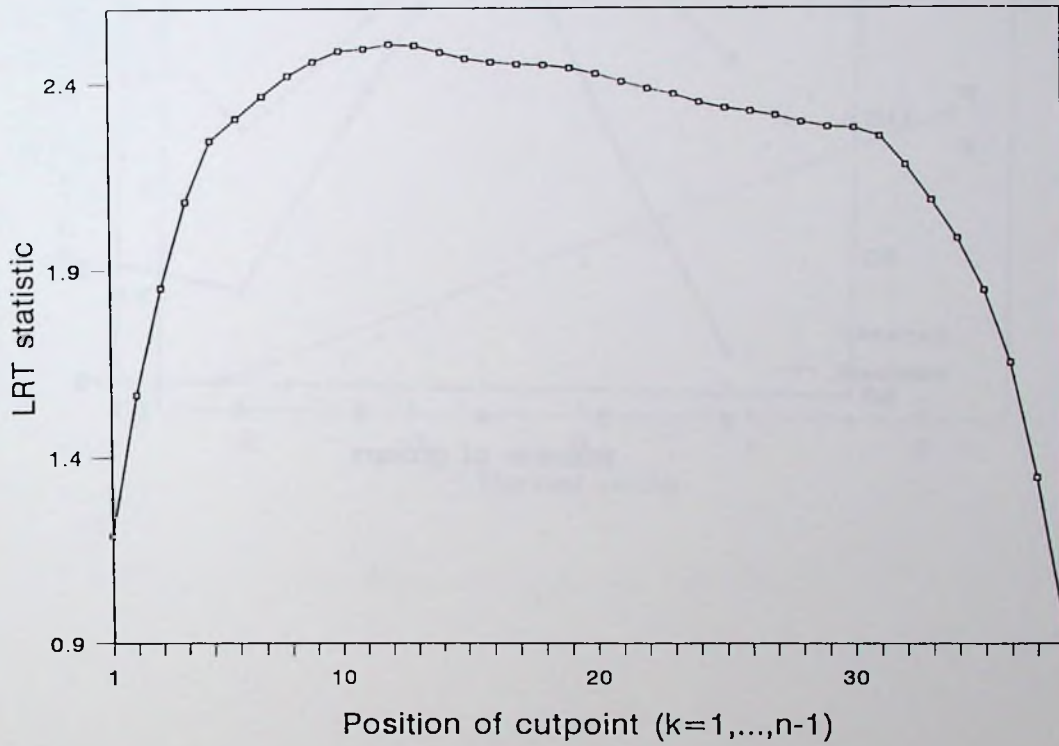


Figure 5.7b Plot of LRT statistic for the 2-group model against hypothetical cutpoints for C-LF.

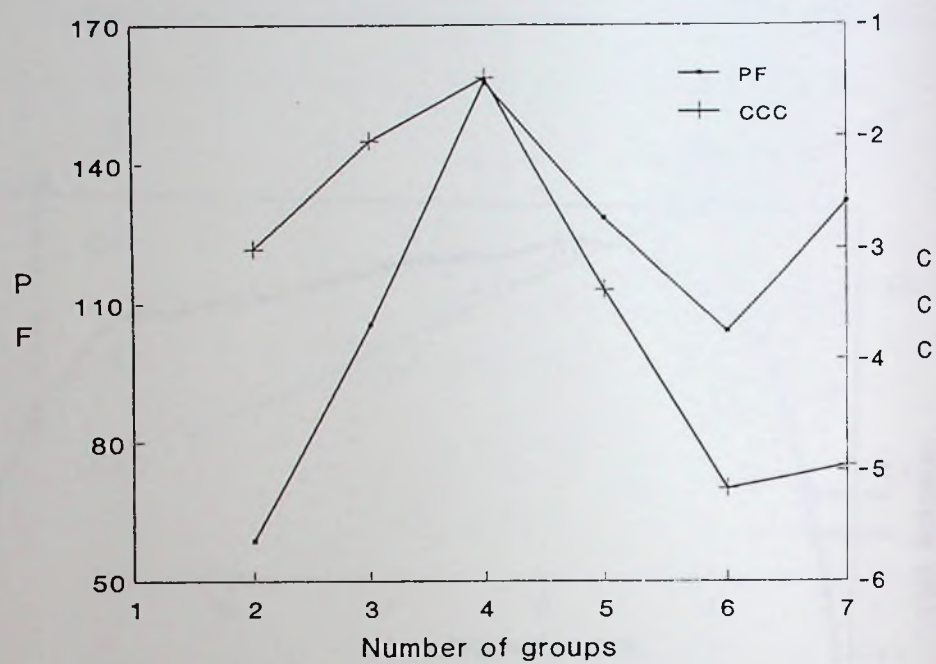


Figure 5.7c Plot of PF and CCC for optimum classifications groups with respect to C-LF.

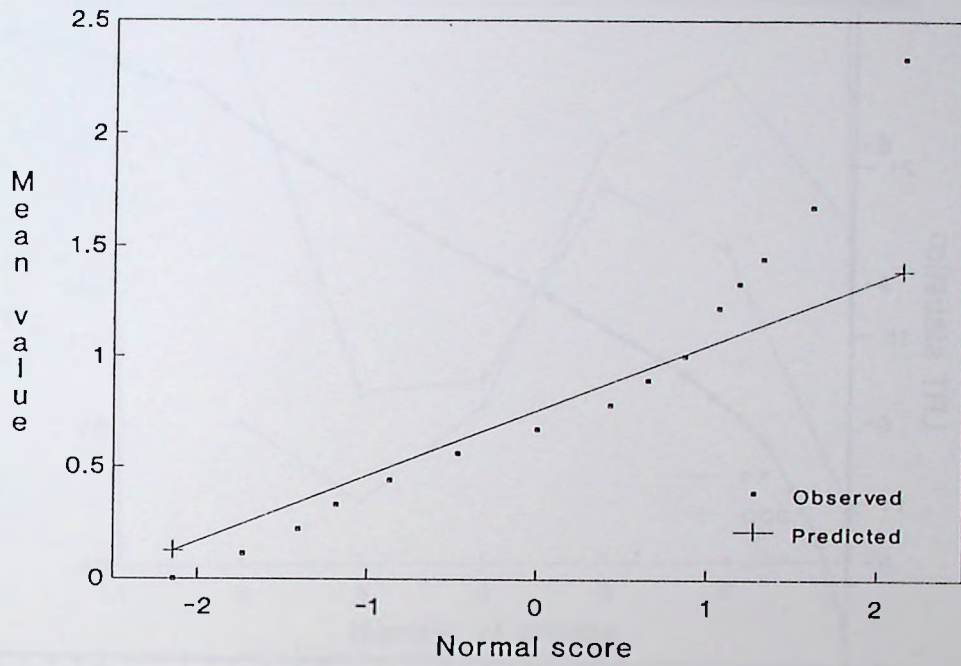


Figure 5.8a Plot of C-DH for the 39 entries against their respective normal scores.

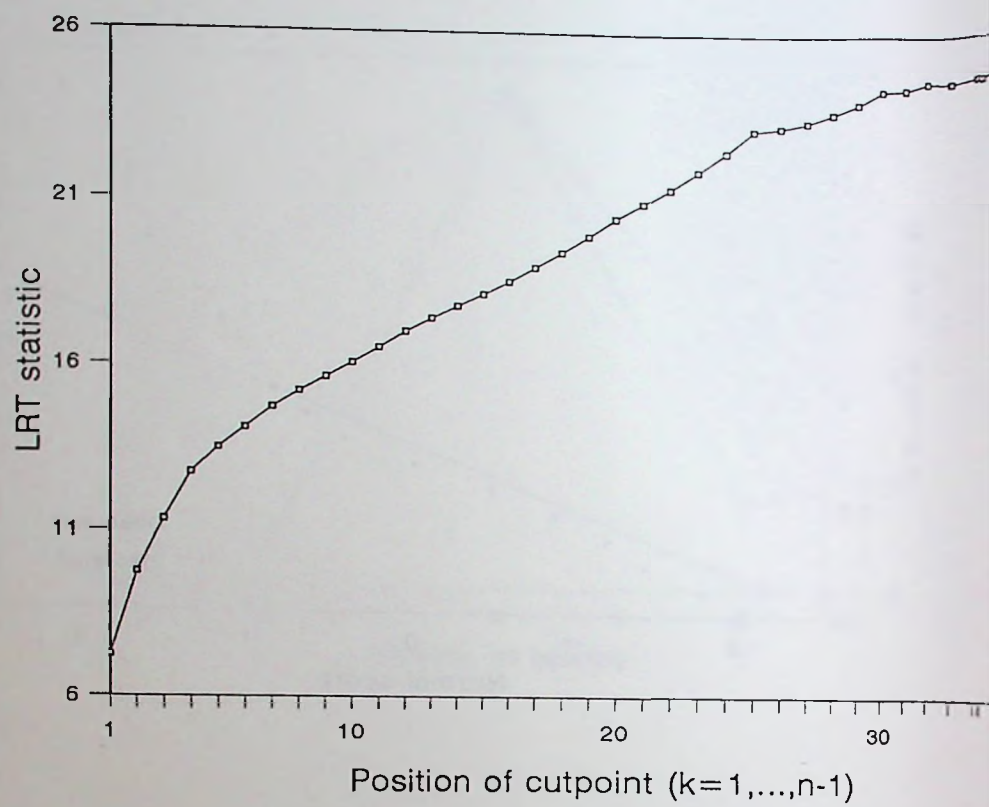


Figure 5.8b Plot of LRT statistic for the 2-group model against hypothetical cutpoints for C-DH.

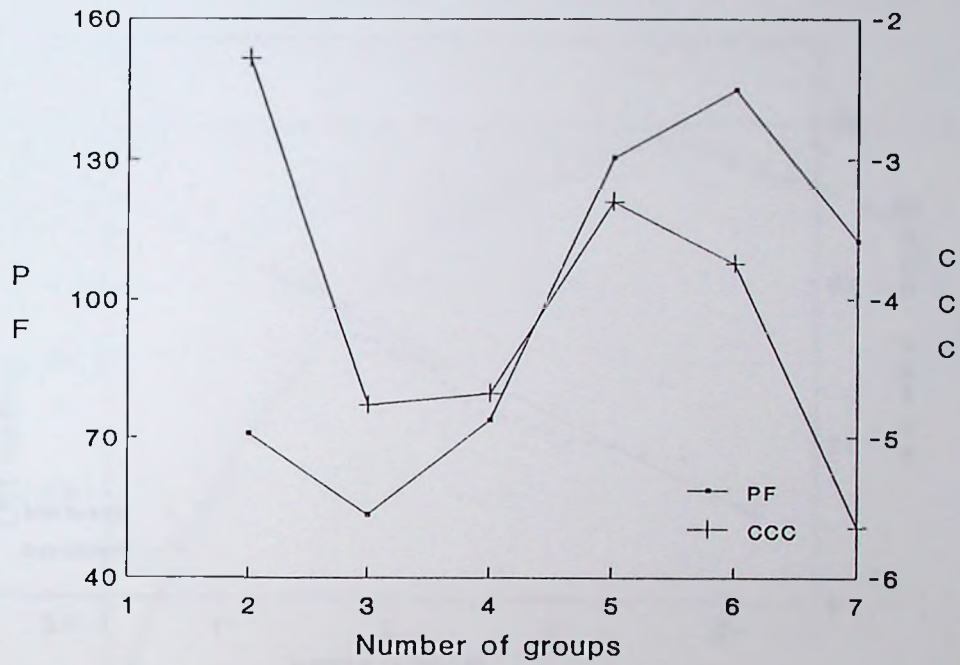


Figure 5.8c Plot of PF and CCC for optimum classifications into g groups with respect to C-DH.

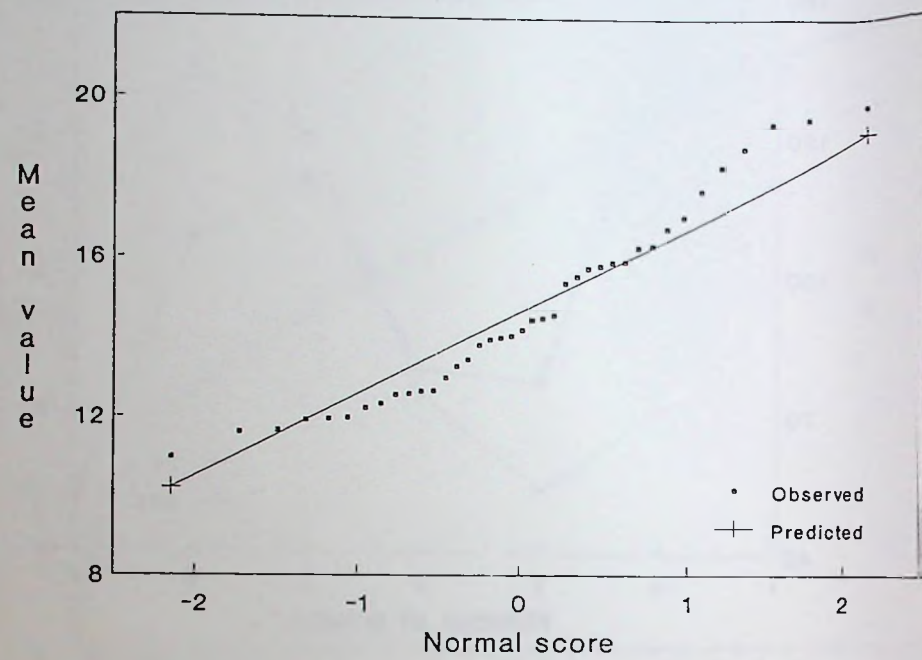


Figure 5.9a Plot of C-ST for the 39 entries against their res
normal scores.

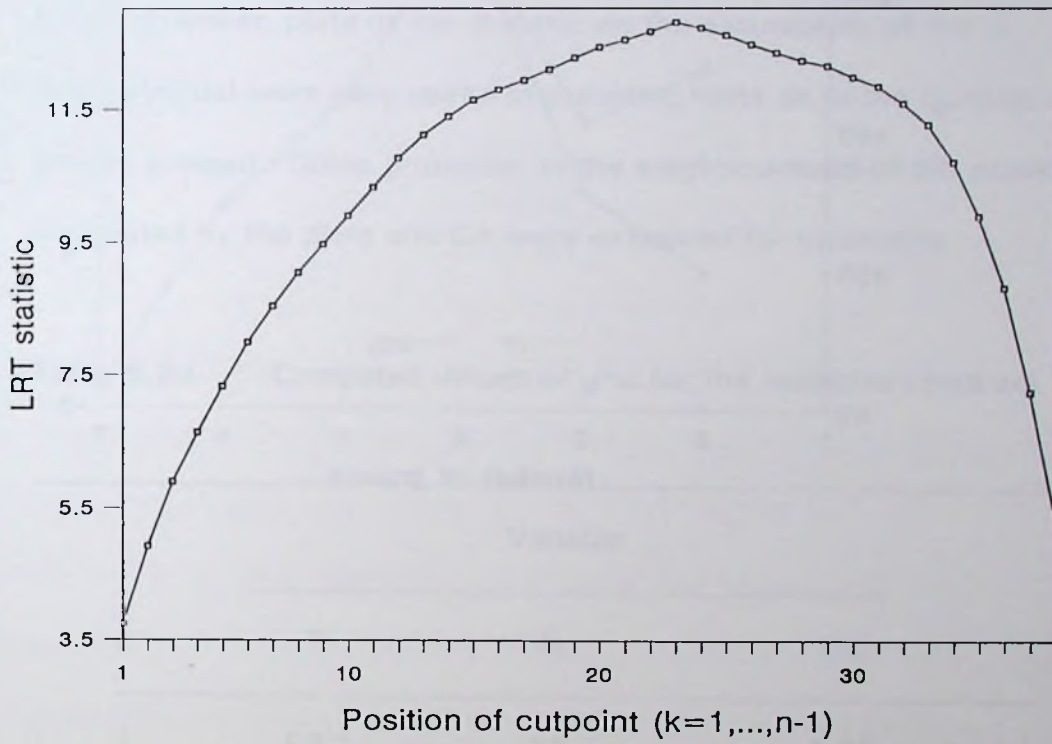


Figure 5.9b Plot of LRT statistic for the 2-group model against hypothetical cutpoints for C-ST.

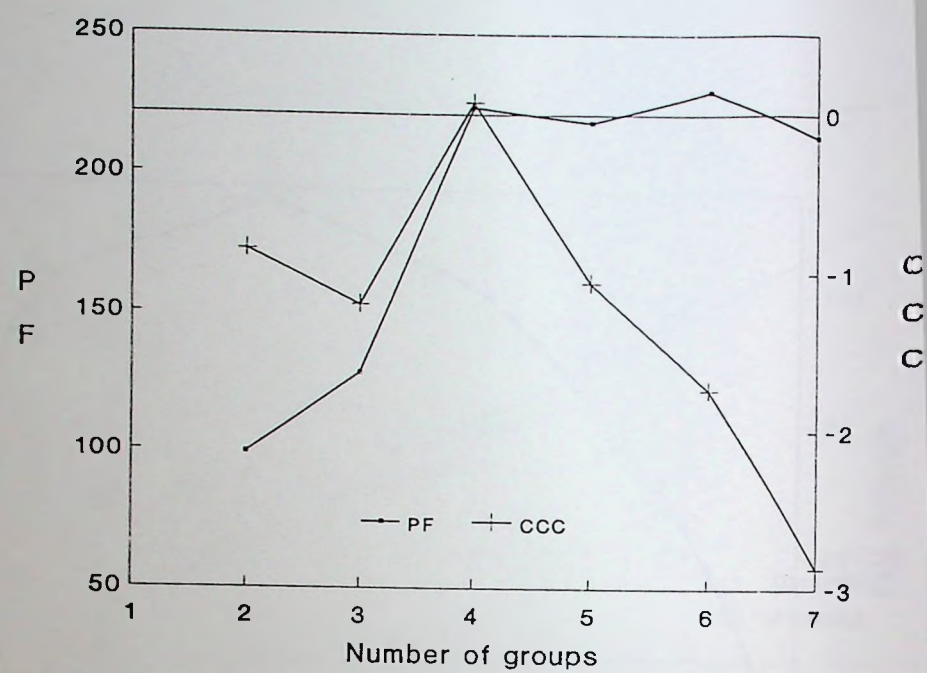


Figure 5.9c Plot of PF and CCC for optimum classification groups with respect to C-ST.

5.2.4 DATA SET D

For this data set, a maximum of 4 groups were considered for the LRT statistics due to limitations of computer memory (see table 5.1). However, plots of the statistic on the assumption of the 2-groups model were very useful in providing hints as to the number of groups present. Some groupings in the neighbourhood of the partition suggested by the plots and CA were examined for optimality.

Table 5.2d Computed values of g^2w for the variables (data set D).

g	Variable		
	T	N	WL
1	53.2	89.7	1.05
2	39.1	98.0	1.11
3	51.2	106.2	1.28
4	55.6	110.9	1.34

5.2.4.1 Time to 50% emergence (D - T)

The NS plot (Fig 5.10a) suggests two groups. The classification into two groups is further supported by virtually all the other procedures. The LRT statistic plot (Fig 5.10b) exhibits no other cutpoint from the one at point 47. Determination of optimum number of classes gave the minimum value of g^2w at $g = 2$ (Table 5.2d). A

plot of CCC and PF using CA across 2-10 groups shows no peak other than at 2 groups (Fig 5.10c). The descriptions of classes are:

j	n_j	mean (s.e.)
1	47	40.54 (0.47)
2	30	53.96 (0.56)

QBT classified the entries into groups of 43 and 34 entries. Ten percent of the number of entries in this data set gives 8 entries. The wide margin between this number and the 30 supported by procedures reveals the need for compulsory statistical methods in the selection process. A decision to divide the entries into 2 equal groups, which is a usually employed conventional method, would lead to loss of useful resources. Using the variable overall mean and standard deviation of 45.77 and 7.37 respectively, another popular method (e.g. Bhumannavar, 1989) of classification would have resulted into three classes of 11, 50 & 16 given by $(\min, y - sd)$, $(y - sd, y + sd)$; and $(y + s.d, \max)$. Thus 16 entries would be selected leaving out some 14 the other procedures included in the relatively resistant class.

5.2.4.2 Number of adult insects (D - N)

Each procedure suggests a different partition. The normal score plot (Fig 5.11a) suggests 2-4 groups. The LRT plot (Fig 5.11b) curves smoothly without showing any major angle (which marks the cutpoints). CA results are somehow inconsistent. While CCC suggests no more than 2 groups, PF suggests no less than 7 groups (Fig 5.11c). The g^2w statistic (Table 5.2d) increased with increasing values of g (at least within the range of g values investigated). QBT classifies the varieties into groups of 42 & 35 entries. The best classification into 2 groups by LRT is into 47 & 30 entries while CA gives 49 & 28 entries.

Results of all the procedures are rather inconsistent and inconclusive. This is probably a consequence of the structure of the data resulting into gross violation of the distributional assumptions. Such data should not be relied on for classification.

5.2.4.3 Weight loss (D - WL)

The normal score plot suggests 5 groups (Fig 5.12a). QBT gave a grouping into 43 & 34, while LRT and CA suggest two classes of 50 & 27 entries. However, the LRT plot (Fig 5.12b) reveals no distinct cutpoints. CA gives 5 groups (Fig 5.12c) of 11, 27, 17, 12, 10 entries. Determination of optimum number of groups using the g^2w statistic did not indicate any optimum number of groups within the range of g investigated (Table 5.2d). However, the computed value

(1.049) for the 5 groups model identified by CA would suggest a similar classification. The description of the 5-group model is

j	n_j	mean (s.e.)
1	11	0.24 (0.06)
2	27	1.03 (0.04)
3	17	1.78 (0.06)
4	12	2.66 (0.06)
5	10	3.44 (0.08)

Group 1 entries look very distinct from the others. The group constitutes 14% of the study population. The conventional sales intensity would have left out a good chunk of this distinct group.

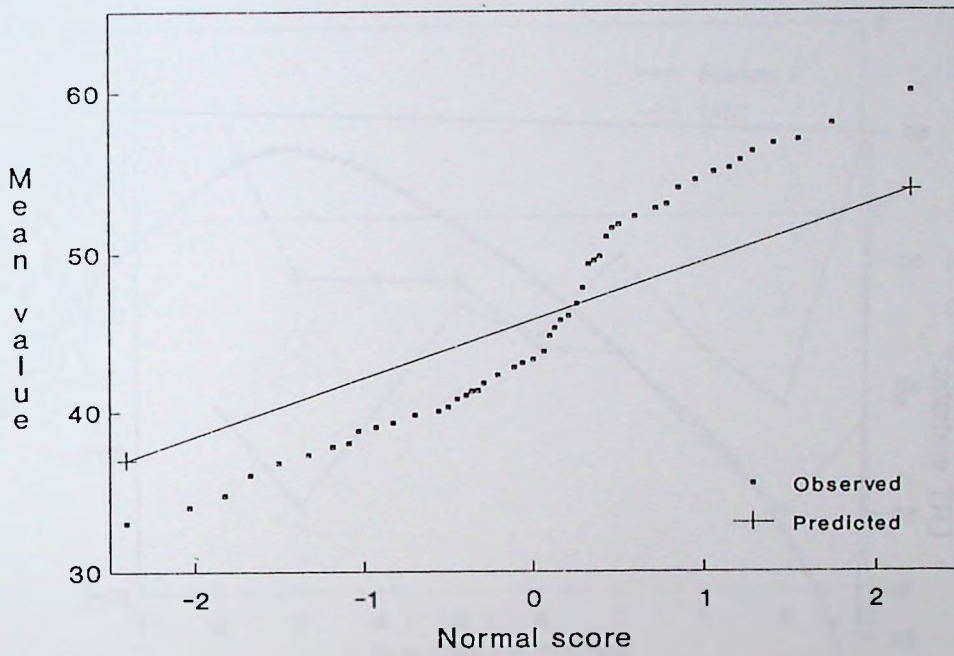


Figure 5.10a Plot of D-T for the 77 entries against their respective normal scores.

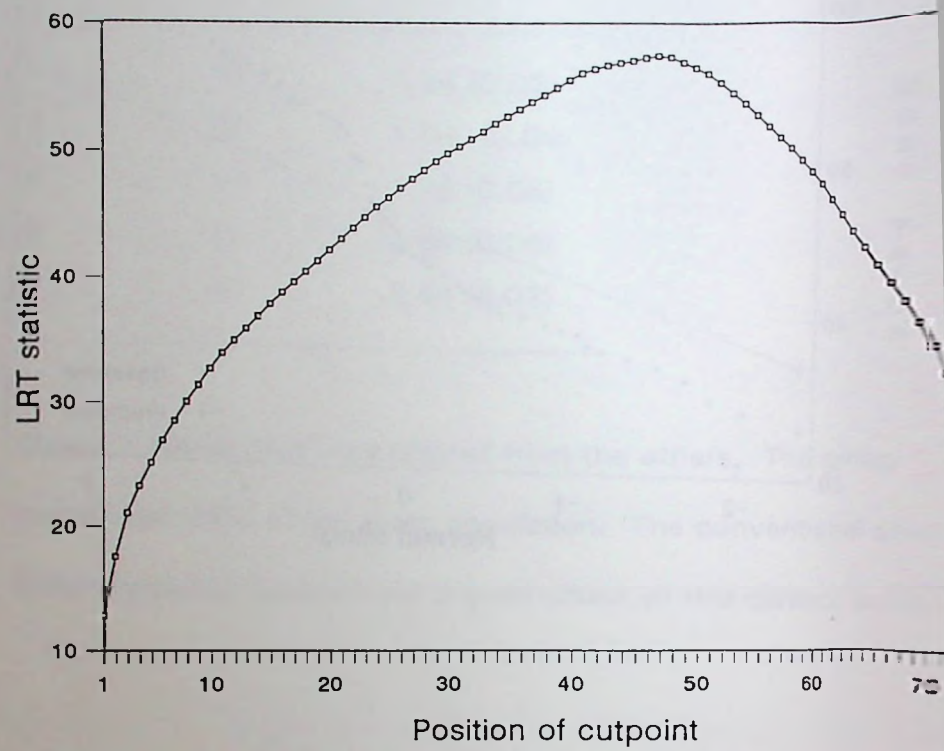


Figure 5.10b Plot of the LRT statistic for the 2-group model with hypothetical cutpoints for D-T.

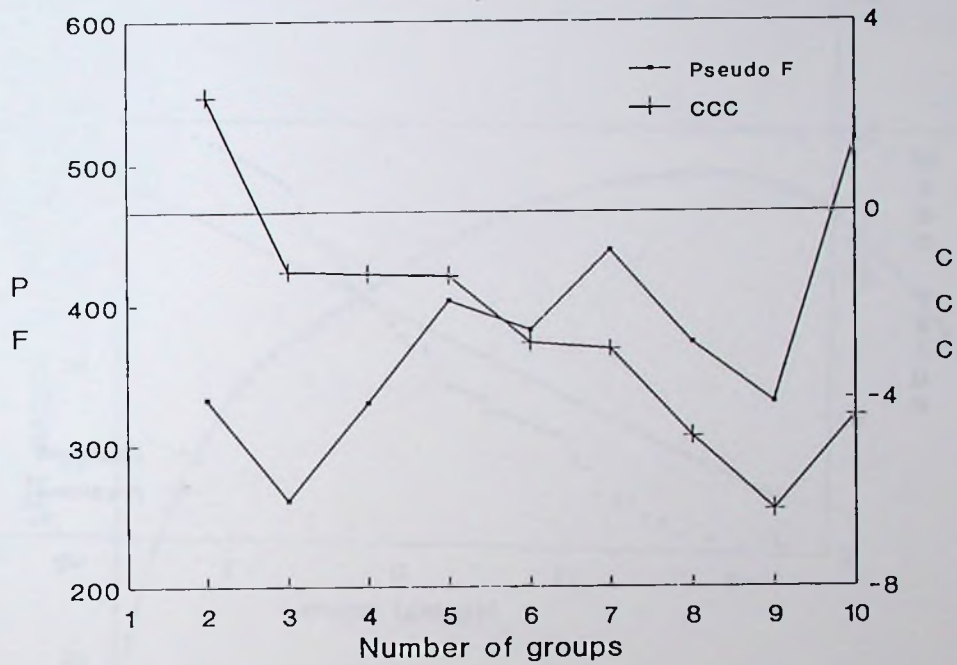


Figure 5.10c Plot of PF and CCC for optimum classifications into g groups with respect to D-T.

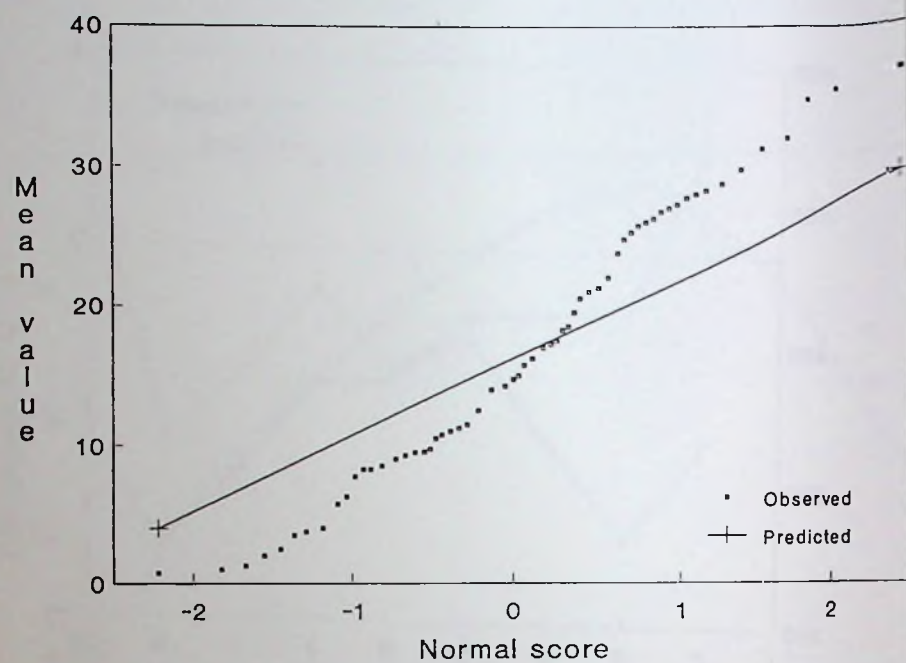


Figure 5.11a Plot of D-N for the 77 entries against their normal scores.

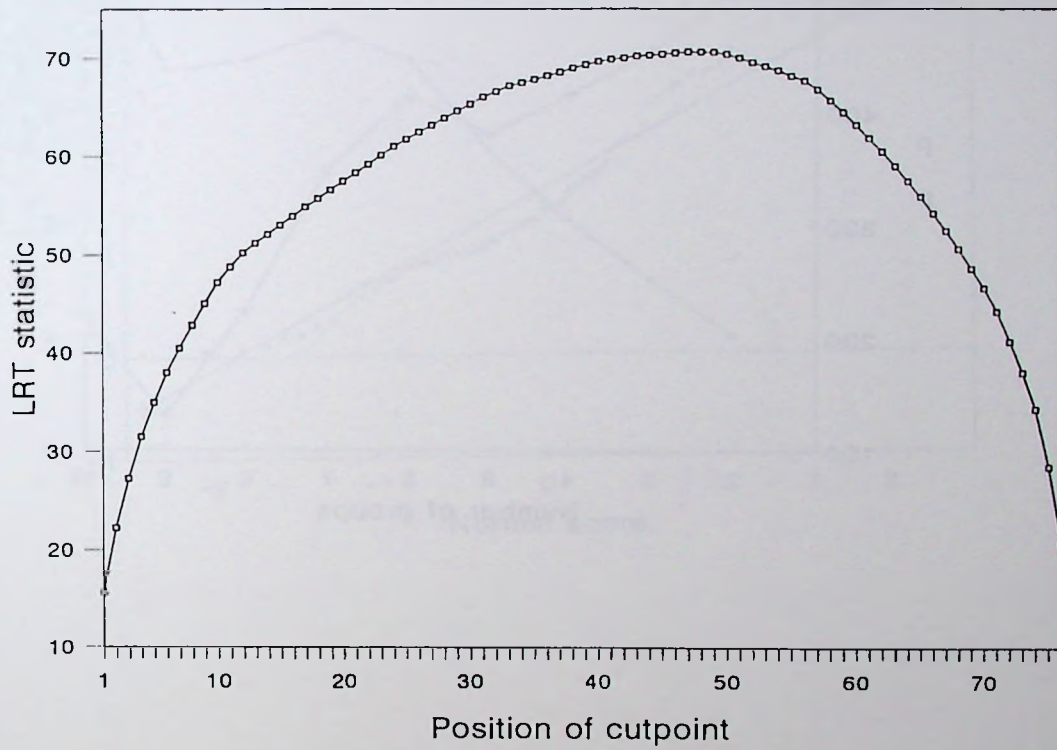


Figure 5.11b Plot of the LRT statistic for the 2-group model against hypothetical cutpoints for D-N.

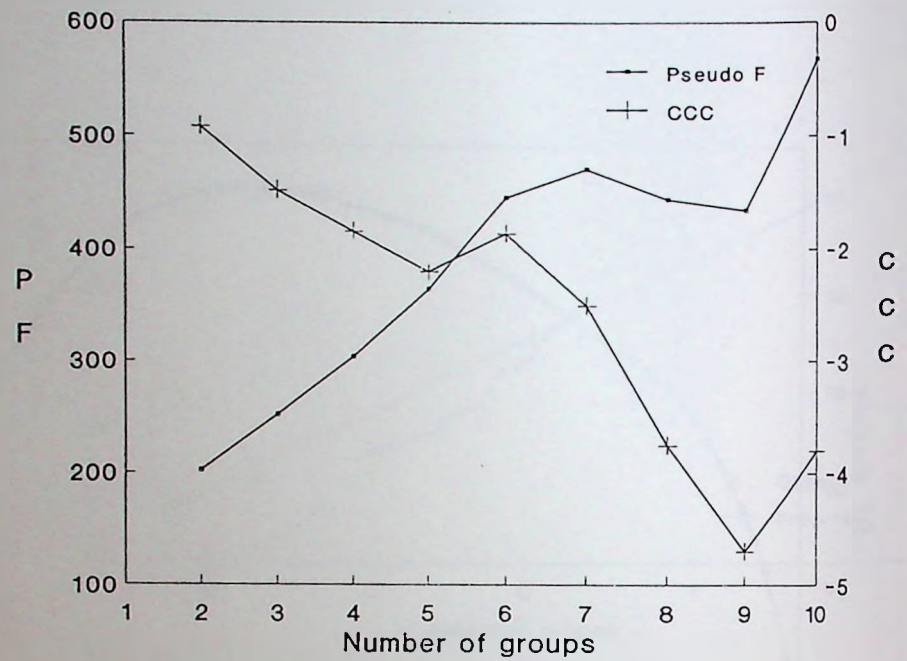


Figure 5.11c Plot of PF and CCC for optimum classification groups with respect to D-N.

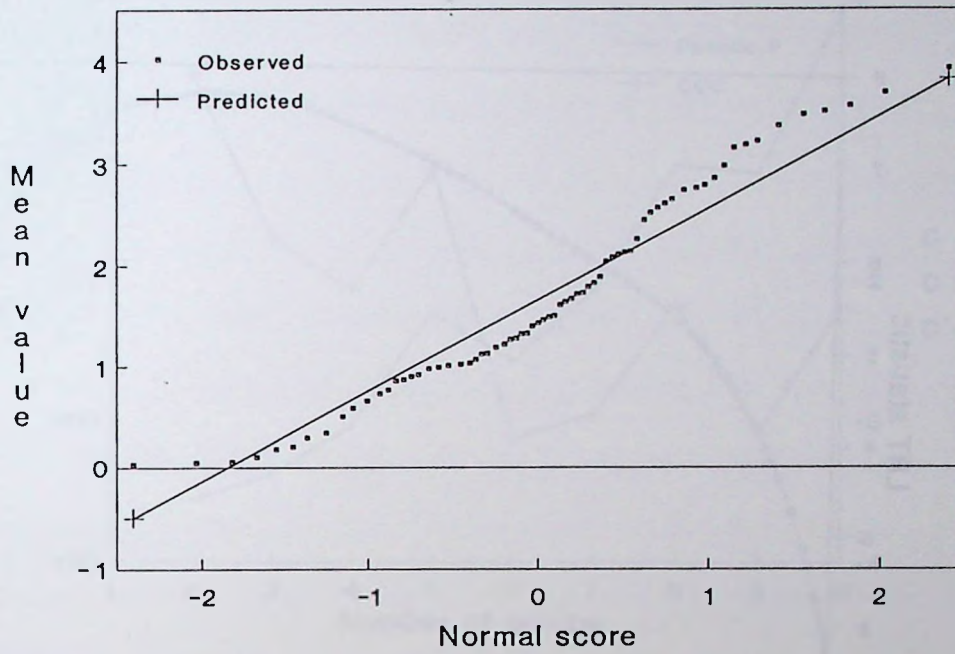


Figure 5.12a Plot of D-WL for the 77 entries against their respective normal scores.

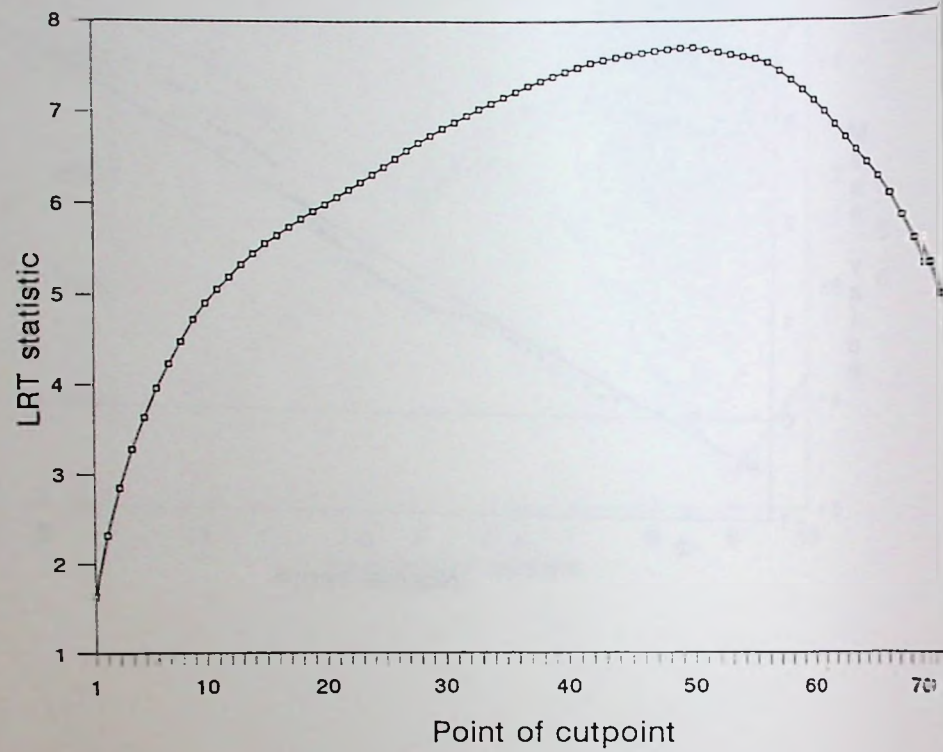


Figure 5.12b Plot of the LRT statistic for the 2-group model $\hat{\alpha}_{D-WL}$ hypothetical cutpoints for D-WL.

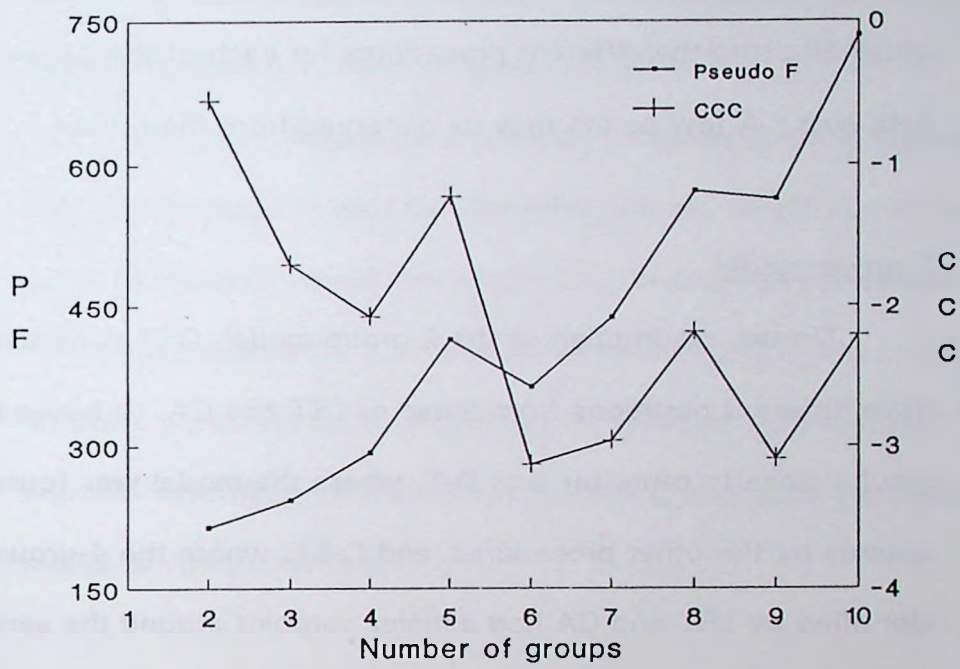


Figure 5.12c Plot of PF and CCC for optimum classifications into g groups with respect to D-WL.

5.2.5 Summary of results for univariate analyses

Table 5.3 gives a summary of the optimum classifications obtained using the different procedures for each of the 12 univariate data sets. A few points may be observed from the results.

2-group model

On the assumption of the 2-group model, QBT consistently gave different partitions from those of LRT and CA. It however gave results close to other for sets D-T, where the model was found optimal by the other procedures, and C-ST, where the 4-group model identified by LRT and CA had a major cutpoint around the same position. From this observation it appears QBT gets close to the optimal partition when the 2-group model is adequate. Otherwise, the use of the statistic tended to partition the entries into near (or close to) equal sized groups.

Optimum model

In 6 data sets, CA and LRT gave similar classifications and for these sets, NS plots gave the exact number of groups as identified by these procedures. These data sets included A-LF, A-DH, A-ST, C-LF, C-ST and D-T. Apparently the assumptions for the application of the techniques were satisfied, including normality, equal variance among groups and existence of distinct groups which the techniques sought for.

Among the other 6 sets, no solution was obtained for D-N. Other than lack of distinct groups, the data set may have had a non-normal distribution. Perhaps the same applied for D-WL.

The major reason for inconsistent results for B-LF seems to be lack of groupings among the observed means. ANOVA results (Table 4.1b) showed no significant differences among the tested entries. The other variables in this data set (B-DH and B-ST) also resulted into different partitions being suggested by different procedures. The CA and LRT also differed markedly in partitions based on the 2-group model for these variables.

Table 5.3 Summary of optimal groupings by different procedures for the different data sets.

Variable	NS plot	Two groups model			Optimum model	
		QBT	LRT	CA	LRT	CA
A-LF	3 groups	28, 23	36, 15	36, 15	3: 14,22,15	3: 14,22,15
A-DH	3 groups	34, 17	40, 11	40, 11	3: 38,11,2	3: 38,11,2
A-ST	≥ 5 groups	30, 21	40, 11	40, 11	6: 4,19,9,9,7,3	6: 4,19,9,9,7,3
B-LF	≈ 2 groups	28, 27	36, 19	36, 19	1 or 2 groups	3 or 4 groups
B-DH	≥ 4 groups	35, 20	49, 6	51, 4	5: 20,16,13,4,2	4: 36,13,4,2
B-ST	≥ 5 groups	27, 28	37, 18	44, 11	5: 8,20,18,7,2	6: 6,18,20,7,2,2
C-LF	4 groups	17, 22	12, 27	13, 26	4: 5,8,18,8	4: 5,8,18,8
C-DH	≥ 3 groups	24, 15	35, 4	35, 4	3: 25,10,4	2: 35,4
C-ST	4 groups	22, 17	23, 16	23, 16	4: 13,10,10,6	4: 13,10,10,6
D-T	2 groups	43, 34	47, 30	47, 30	2: 47,30	2: 47,30
D-N	≥ 4 groups	42, 35	47, 30	49, 28	-	-
D-WL	≥ 5 groups	43, 34	50, 27	50, 27	-	5:11,27,17,12,10

5.3 MULTIVARIATE ANALYSES

Data arrays and computation of indices

The linear relationships between traits were described in Chapter 4 (see Table 4.2) and partitions based on these were made using multivariate techniques. PCA was carried out on the correlation matrices using PROC PRINCOMP of SAS.

Only PCs that cumulatively accounted for at least 80% of the total variance were retained for subsequent analyses. The classifications of the entries into resistance groups based on the computed PCs or the observed damage variables, as appropriate, were compared with those based on appropriate indices for some selected data sets. This included RMI for data sets A, B and C; SHI for data sets A and B; and I for data set D.

Proc PLOT, FASTCLUS and GLM (or CANDISC) in SAS were used to construct bivariate scatter plots, perform CA and LRT procedures. Proc GLM (or CANDISC) performs a one-way MANOVA when the hypothetical classes are specified in the model statement and provides various test statistics (e.g. Wilk's λ). Since the LRT procedure is iterative, requiring examination of all possible groupings, some programs were written in BASIC (analogues of the univariate) to generate the groupings, compute the required matrices and their determinants (e.g. Appendix A3).

Ordering (ranking) of the entries, which is vital for the LRT procedure, was done with respect to PC1, this being the major source of variation or the variable with the greatest variation.

Computation of indices

Rank mean index (RMI)

RMI values were computed for each entry using PROC RANK in SAS and obtaining the average of the rankings with respect to each variable (see equation (2.7)).

Smith Hazel index (SHI)

Variance component analysis was performed to obtain the genotypic and phenotypic variances-covariance matrices that were used to obtain values of the b vector (see equations (2.1) & (2.2)). PROC VARCOMP in SAS and FACTOR (with covariance) in MSTAT were employed to generate combined ANOVA for individual variables and ANOCOVA tables for pairs of variables as were required.

The computed b vectors for data sets A and B are [3.04, -0.119] (see Appendix A4) and [0.07, -0.44 -0.63] (see Appendix A) respectively. The nonsignificance of LF in ANOVA in data set B seems to have an association with the size of the weight placed on LF compared to the others.

Dobie's index (I)

See Sections 2.1.2.1 & 4.1.1. It should be noted that the index is only appropriate for storage pests studies. Its use will therefore be evaluated in comparison with use of PCs for data set

5.3.1 RESULTS AND DISCUSSION

5.3.1.1 DATA SET A

Results of PCA performed on the correlation matrix are summarised in Table 5.4a.

Table 5.4a A summary of results of PCA of data set A damage variables.

Variable	Eigenvectors		
	PC1	PC2	PC3
LF	0.569	-0.641	0.515
DH	0.552	0.762	0.338
ST	0.612	-0.092	-0.788
Eigenvalue	2.370	0.440	0.190
Cumulative proportion	0.790	0.937	1.000

The first PC accounts for 79% of the multivariate variation among the entries. It is basically a measure of the average damage with respect to the three measures, low values of which characterise a resistant entry. PC2 contrasts LF with DH. Almost 94% of the variability is explained by the first two PCs. A plot of entries according to their first two PC scores is given in Fig 5.13a.

Classification based on PCS

CA suggested optimal classifications with $g = 3$ & 7 (Fig 5.13b). The 7-group model was found optimal by the LRT method when the CA solutions for $g = 2-10$ were used to compute the statistics. Group estimates are given below.

j	n_j	Mean (s.e.)		
		LF	DH	ST
1	3	5.66 (0.39)	17.09 (0.86)	24.29 (1.3)
2	24	3.73 (0.11)	2.72 (0.35)	14.26 (0.4)
3	2	7.20 (0.40)	16.67 (2.56)	30.06 (2.9)
4	3	5.82 (0.20)	6.41 (0.01)	24.98 (0.3)
5	8	3.66 (0.20)	12.18 (1.32)	16.63 (0.8)
6	2	7.11 (0.03)	34.62 (5.12)	33.40 (1.2)
7	9	5.26 (0.22)	4.79 (0.58)	18.72 (0.9)

The various groups plotted along the first two components are shown in Fig 5.13c. Entries in group 2 are the most resistant while groups 6 & 3 are the most susceptible.

Classifications based on RMI and SHI

Application of the univariate classification procedures gave the following results.

RMI - At least 3 groups are suggested by the NS plot (Fig 5.14a), while about 5 are suggested by the LRT plot (Fig 5.14b). However, CA gives 7 groups as the optimal classification (Fig 5.14c) which is also supported by the LRT. Assuming a 2-group model, QBT, LRT and CA give groups of 28 & 23, 27 & 24 and 28 & 23, respectively.

Group estimates for the 7-group model are,

j	n_j	Mean (s.e.)		
		LF	DH	ST
1	8	3.66 (0.20)	12.18 (1.32)	16.63 (0.87)
2	20	3.65 (0.12)	2.82 (0.35)	13.68 (0.39)
3	1	4.89 -	15.38 -	21.56 -
4	10	4.72 (0.24)	3.20 (0.44)	17.96 (1.02)
5	4	6.63 (0.37)	17.31 (1.11)	27.86 (1.75)
6	2	7.11 (0.03)	34.62 (5.12)	33.40 (1.28)
7	6	5.72 (0.10)	6.54 (0.24)	22.10 (1.30)

SHI - The NS and LRT plots (Figs 5.15a & b) suggest at least 6 groups. CA (Fig 5.15c) gives an optimum 7-group model. QBT, LRT and CA on the assumption of a 2-groups model give groups of 32 & 19, 35 & 16 and 37 & 14, respectively. Group estimates for the 7-group model are,

j	n_j	Mean (s.e.)		
		LF	DH	ST
1	9	5.66 (0.12)	5.50 (0.55)	20.42 (1.1)
2	4	6.63 (0.37)	17.82 (1.11)	27.86 (1.1)
3	2	4.52 (0.37)	10.26 (5.12)	23.19 (1.6)
4	8	3.66 (0.22)	12.18 (1.33)	16.63 (0.8)
5	10	3.23 (0.12)	2.18 (0.60)	12.95 (0.5)
6	2	7.11 (0.03)	34.62 (5.13)	33.40 (1.2)
7	16	4.19 (0.09)	3.20 (0.39)	15.48 (0.5)

High correlations ($p < 0.001$) were observed among the indices (and the damage variables) as indicated below:

	PC1	RMI	SHI
PC1	-		
RMI	0.93	-	
SHI	0.92	0.90	-
LF	0.87	0.86	0.98
DH	0.85	0.73	0.60
ST	0.98	0.88	0.87

Perhaps the high correlations explain why 7 groups were indicated as optimal using all the indices. However the sizes and composition of groups differ. The distributions of groups along the first two PCs based on RMI and SHI, respectively, are given in Figs

5.14d and 5.15d (cf. Fig 5.13c). A list of entries showing group membership based on the indices is included in Table 5.4b.

A close study of the graphs and the table indicates similarities of placement of entries in various groups. Apparently similar information, to some extent, is carried by these indices. However, the slight differences between them probably result in the observed differences in the classifications.

MANOVA, used to compare the efficiencies of indices (differences in variability accounted for by the classifications based on indices), gave the results summarised below:

Index	Wilk's λ	$F_{18, 119}$	P
PCs	0.0124	24.69	< 0.001
RMI	0.0138	23.54	< 0.001
SHI	0.0111	25.88	< 0.001

The similarity in classifications obtained with different indices is further indicated by the relative reduction in the within groups variation which is represented by the Wilk's statistic.

Table 5.4b Composition of groups based on the indices (data set

ENTRY	LF	DH	ST	PCGP	RMIGP	SHGP
49	6.03	17.95	25.38	1	5	2
46	6.07	17.95	25.95	1	5	2
36	4.89	15.38	21.56	1	3	3
12	2.47	1.28	15.22	2	2	5
23	2.93	0.00	10.68	2	2	5
10	3.11	2.56	11.27	2	2	5
29	3.31	0.00	10.06	2	2	5
8	3.11	3.85	13.45	2	2	5
17	3.20	2.56	13.97	2	2	5
24	3.28	2.56	13.79	2	2	5
22	3.56	6.41	13.48	2	2	5
13	3.61	1.28	13.98	2	2	5
37	3.75	1.28	13.61	2	2	5
30	3.91	5.13	13.51	2	2	7
18	3.88	2.56	14.40	2	2	7
15	4.08	5.13	12.65	2	2	7
2	4.14	2.56	12.17	2	2	7
33	3.72	2.56	17.41	2	2	7
1	4.18	6.41	13.98	2	2	7
16	4.01	2.56	15.97	2	2	7
14	4.17	2.56	14.35	2	2	7
7	3.75	1.28	20.03	2	4	7
28	4.25	3.85	14.77	2	2	7
40	4.06	2.56	17.10	2	4	7
26	4.36	1.28	14.82	2	2	7
9	4.14	1.28	18.04	2	4	7
44	4.73	3.85	13.64	2	4	7
51	6.80	19.23	33.01	3	5	2
50	7.61	14.10	27.11	3	5	2
43	5.45	6.41	24.47	4	7	1
6	5.88	6.41	24.82	4	7	1
48	6.14	6.41	25.66	4	7	1
25	3.03	11.54	13.59	5	1	4
27	3.28	10.26	16.32	5	1	4
3	3.16	16.67	20.13	5	1	4
42	3.36	7.69	17.18	5	1	4

Table 5.4b continued

ENTRY	LF	DH	ST	PCGP	RMIGP	SHIGP
38	3.83	12.82	16.43	5	1	4
32	4.19	17.95	14.26	5	1	4
31	3.69	7.69	20.09	5	1	4
45	4.72	12.82	15.03	5	1	4
5	7.08	39.74	32.13	6	6	6
4	7.14	29.49	34.68	6	6	6
21	4.89	3.85	20.14	7	4	1
39	5.53	7.69	19.57	7	7	1
35	5.92	3.85	14.62	7	4	1
47	5.60	5.90	18.68	7	7	1
41	5.89	2.56	16.37	7	4	1
20	5.69	6.41	19.42	7	7	1
19	4.14	5.13	24.83	7	4	3
34	4.39	5.13	18.73	7	4	7
11	5.28	2.56	16.08	7	4	7

Comment on results for data set A

Based on PCs, the 7-group model suggests selecting 24 entries, while 20 and 10 entries will be selected based on RMI and SHI, respectively. Depending on requirements for further studies, the selection by SHI may be preferred in case of limited resources when recombinations are to be made. However, some useful materials may be left out. The conventional selection intensity would have resulted into 5 entries, far much fewer than suggested by CA and LRT based on either index.

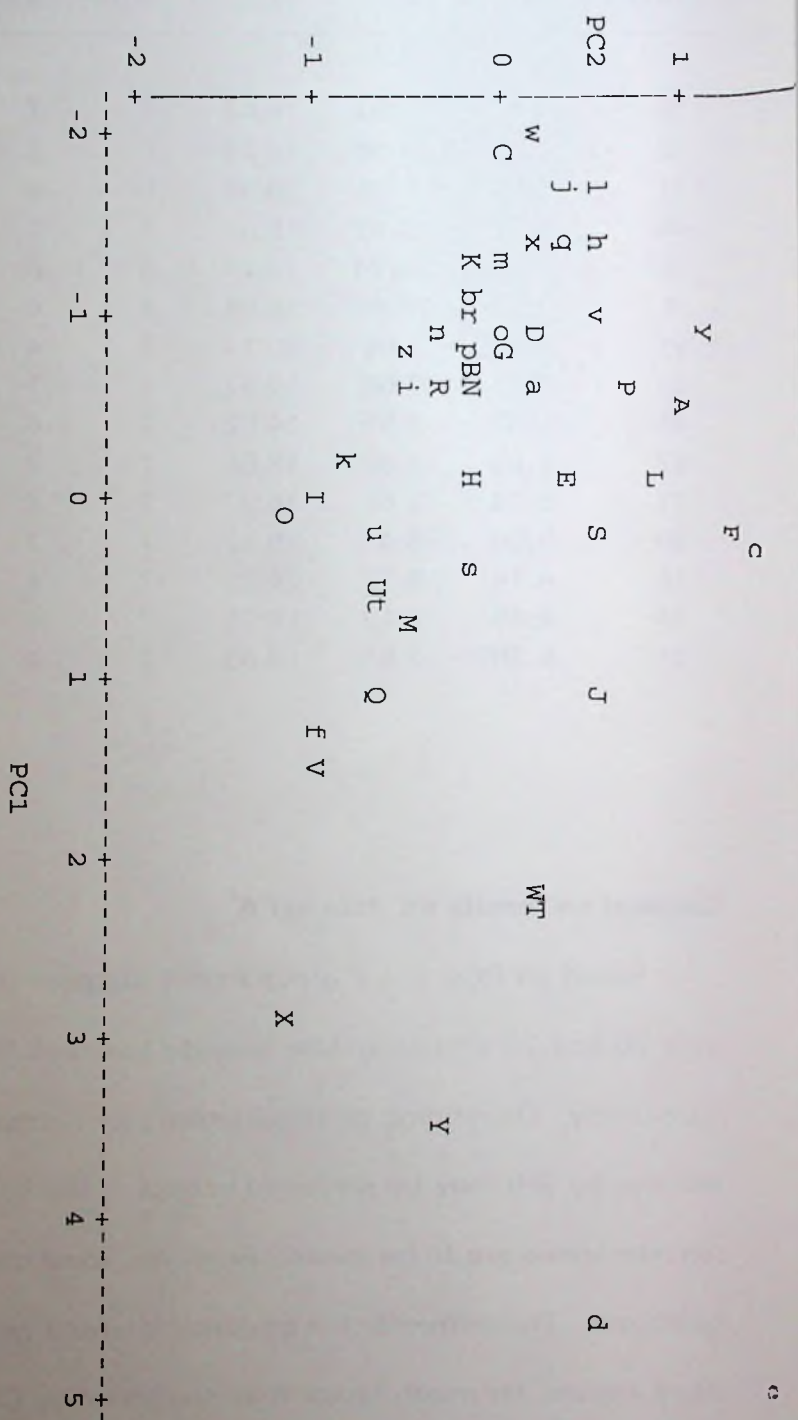


Figure 5.13a Plot of the 51 entries against their values for PC1 and PC2 (letters represent the entry codes).

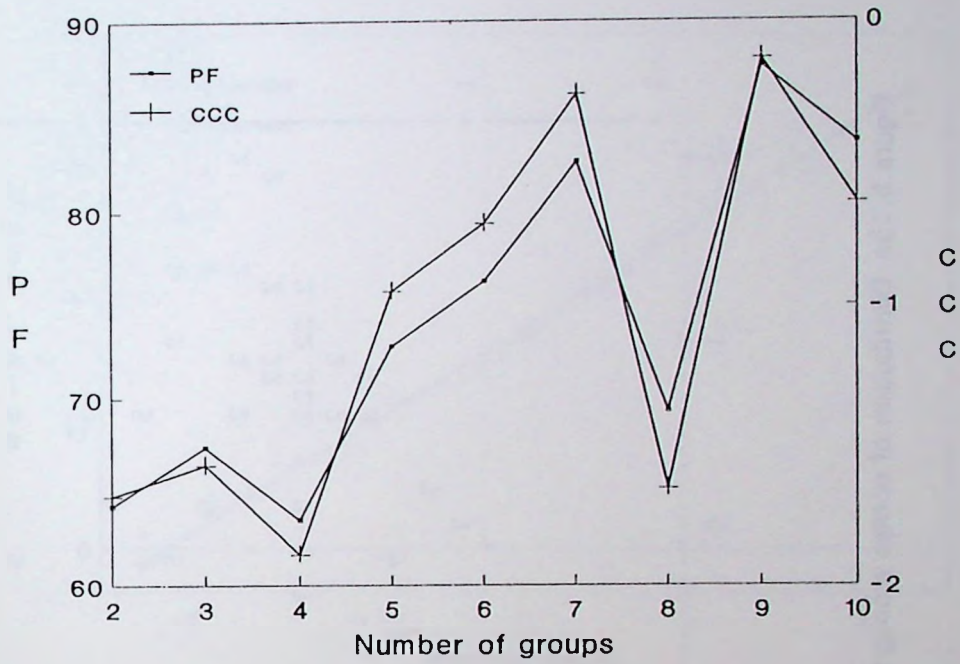


Figure 5.13b Plot of PF and CCC for optimum classifications into g groups using the first two PCs for data set A.

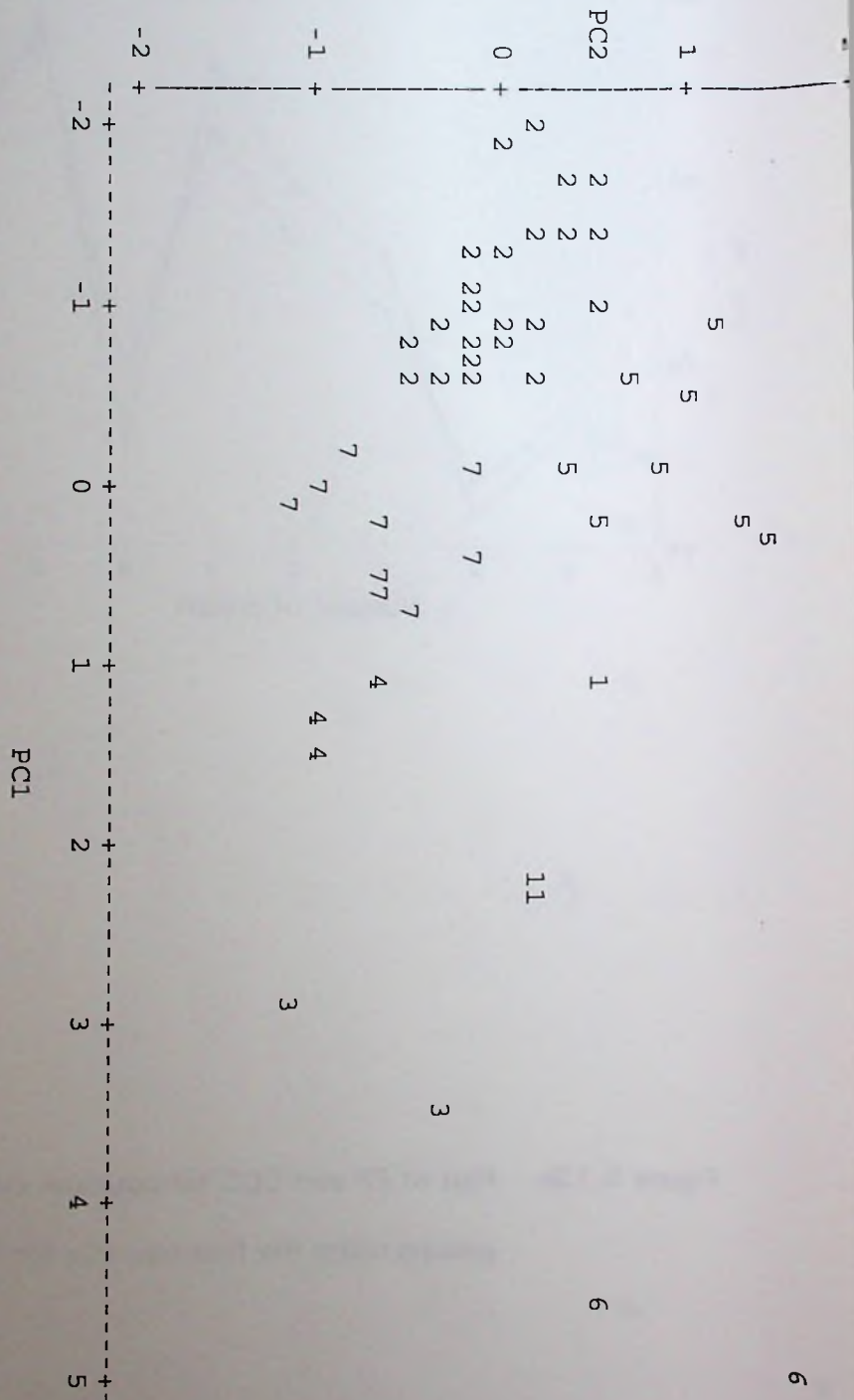


Figure 5.13c Distribution of entries among the 7 classes along the PC1 and PC2 axes.

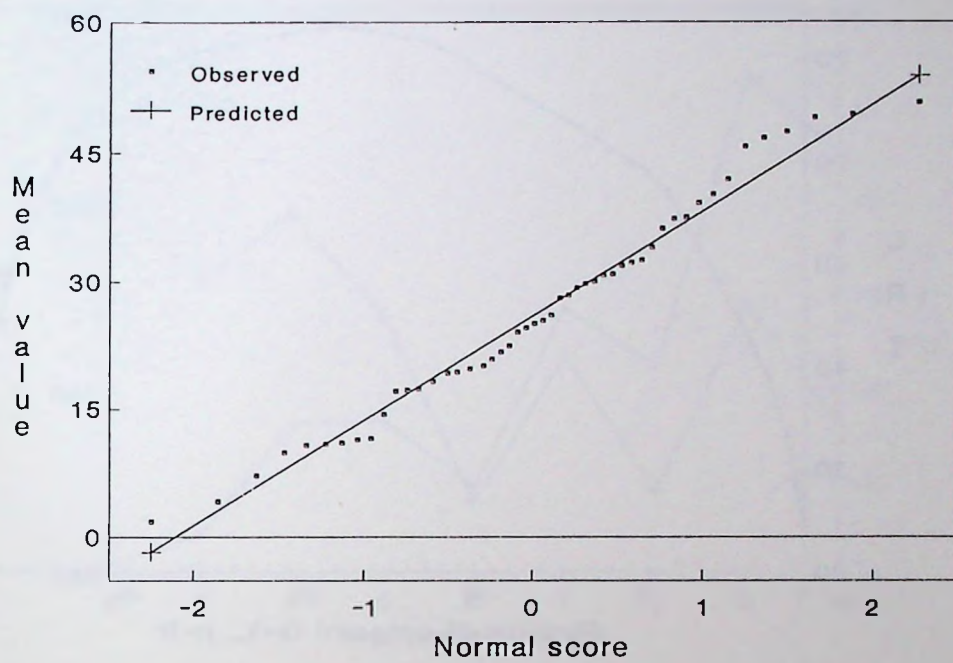


Figure 5.14a Plot of A-RMI for the 51 entries against their respective normal scores.

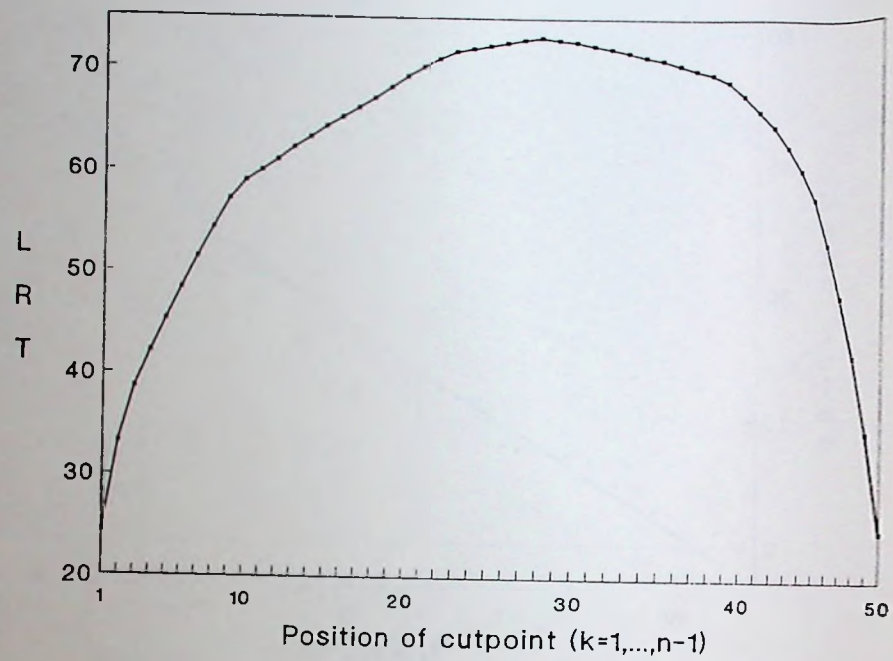


Figure 5.14b Plot of the LRT statistic for the 2-group model as
hypothetical cutpoints for A-RMI.

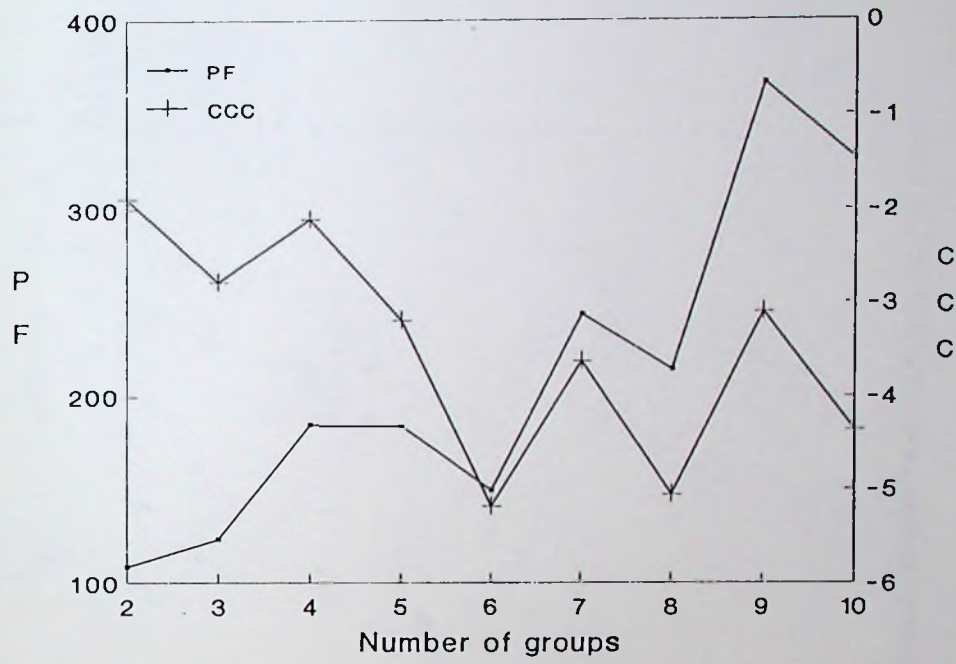


Figure 5.14c Plot of PF and CCC for optimum classifications into g groups with respect to A-RMI.

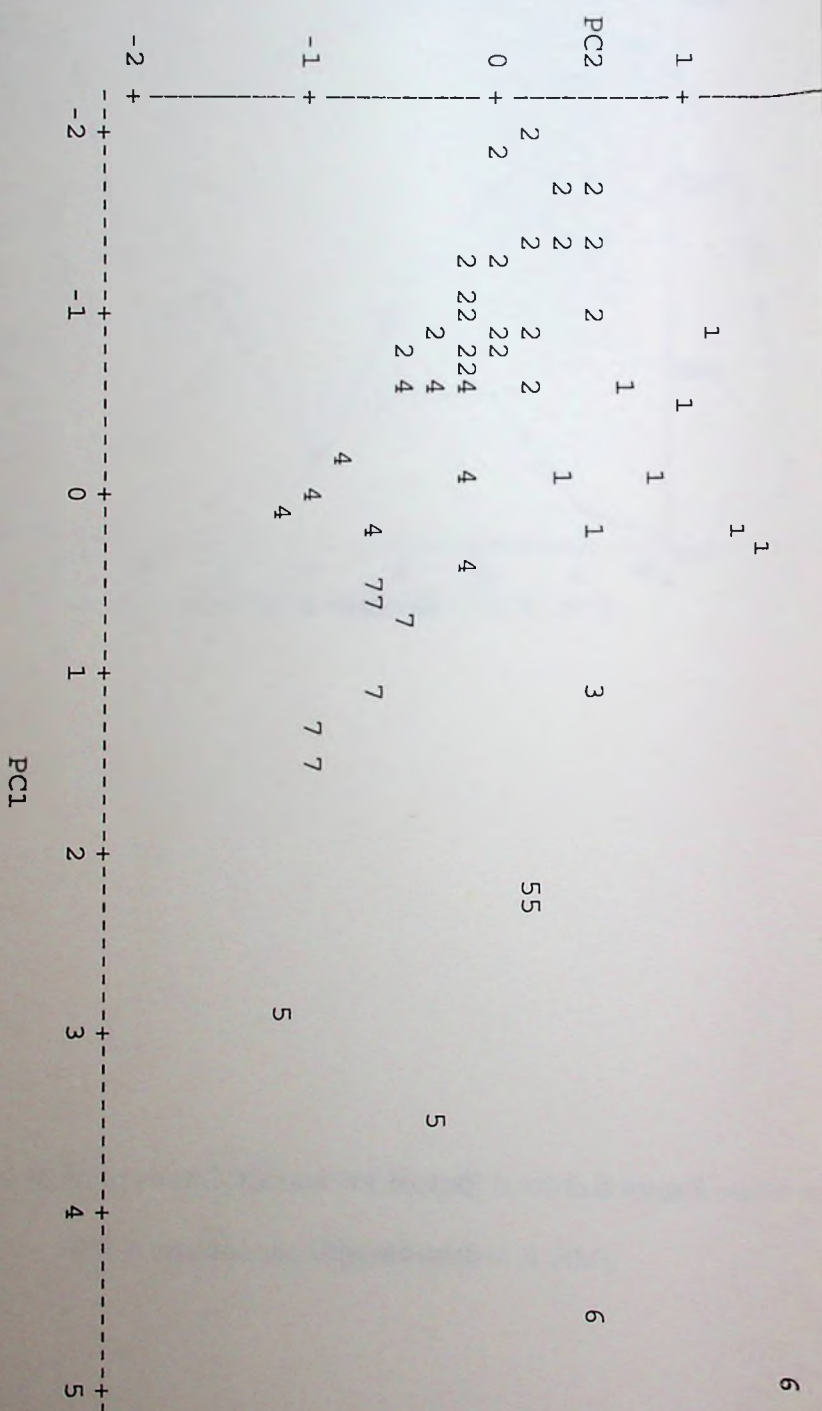


Figure 5.14d Distribution of entries in various groups (based on RMI) along the PC axes.

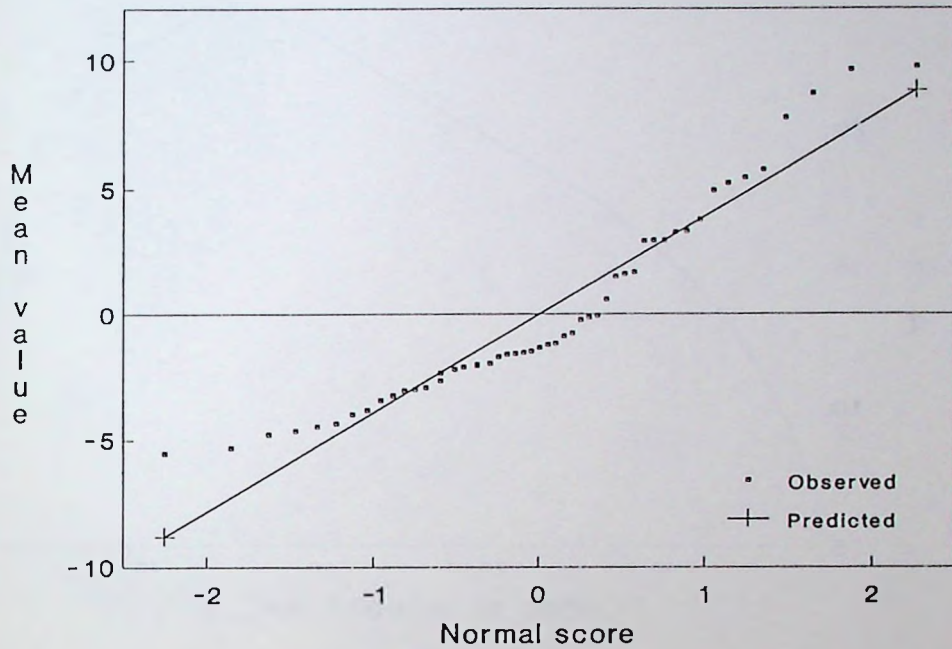


Figure 5.15a Plot of A-SHI for the 51 entries against their respective normal scores.

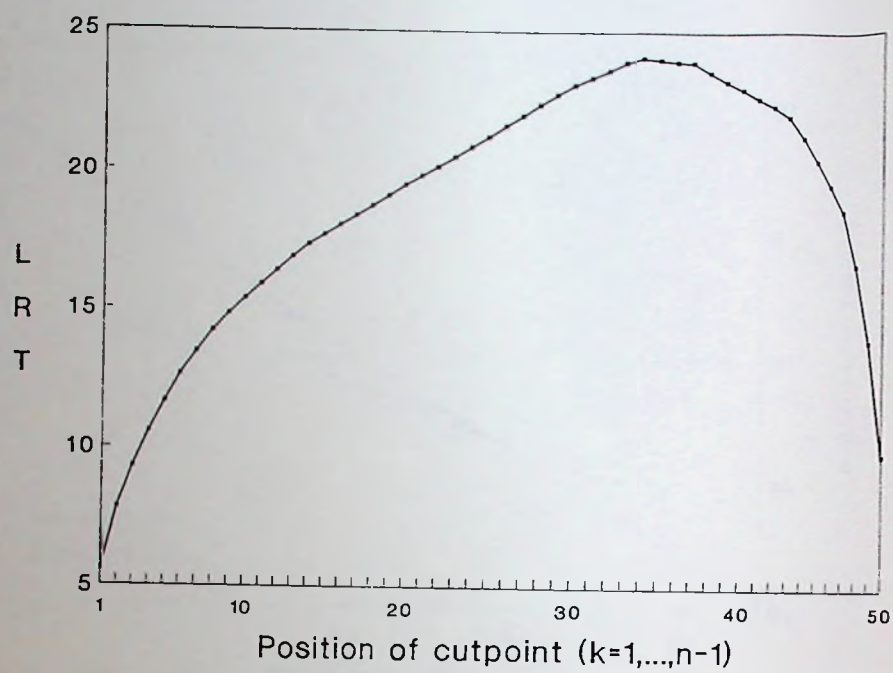


Figure 5.15b Plot of LRT statistic for the 2-group model against hypothetical cutpoints for A-SHI.

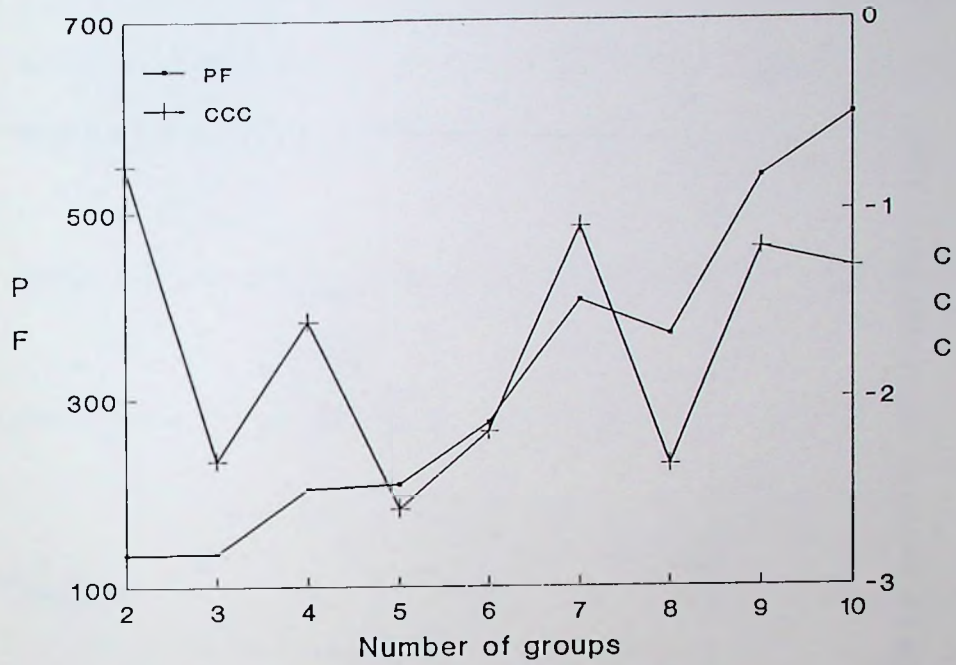


Figure 5.15c Plot of PF and CCC for optimum classifications into g groups with respect to A-SHI.

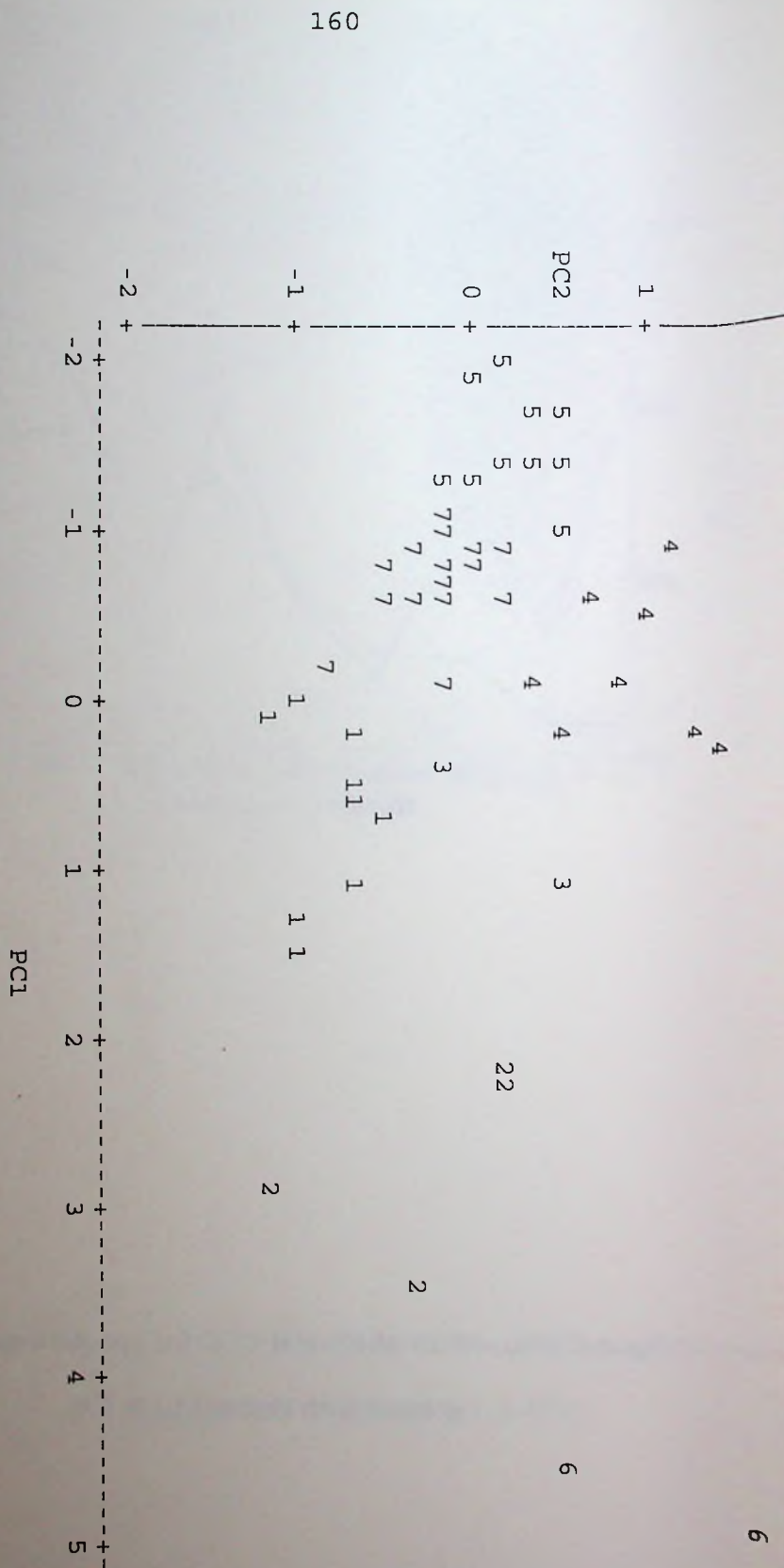


Figure 5.15d Distribution of entries in various groups (based on SHI) along the PC axes.

5.3.1.2 DATA SET B

There were no significant correlations among the damage measures (Table 4.2). An attempt to generate indices for use in subsequent analyse by PCA gave results summarised in Table 5.5a.

Table 5.5a A summary of results of PCA of data set B damage variables.

Variable	Eigenvectors		
	PC1	PC2	PC3
LF	0.031	0.999	-0.041
DH	-0.706	0.051	0.706
ST	0.707	0.007	0.707
Eigenvalue	1.101	1.000	0.899
Cumulative proportion	0.367	0.700	1.000

It is evident from the results that PCA was not able to capture much variation of the original measures among the 55 entries, due to lack of linear relationships among the measures. The three PCs, individually account for 37%, 33% and 30% of the total variation. The magnitude of each of the 3 eigenvalues is almost 1 (each accounting for just as much variation as accounted for by the original measures independently). There is therefore no advantage in using

PCA in this particular case. The resistance measures (standardised) were used to determine the appropriate classification for the entries.

CA gave 7 groups (Fig 5.16). LRT, however, did not provide any meaningful classification as values of the statistic increased with increasing values of g ($=2-10$), using CA solutions. Generation of reasonable groupings was ineffective as there were no useful graphs to suggest possible groupings or PCS on which to base ordering. The most variable measure, ST, was used for the ranking of entries, but still the method could not provide meaningful results. The estimates for the 7-groups model identified by CA are

j	n_j	Mean (s.e.)		
		LF	DH	ST
1	2	5.10 (0.18)	0.00 (0.00)	26.44 (2.9)
2	4	3.90 (0.13)	0.84 (0.15)	17.19 (1.6)
3	13	3.74 (0.07)	0.14 (0.04)	18.66 (0.3)
4	18	4.15 (0.05)	0.14 (0.04)	14.63 (0.4)
5	1	3.79 -	0.00 -	30.86 -
6	2	4.36 (0.29)	2.00 (0.17)	15.32 (0.15)
7	15	4.69 (0.07)	0.23 (0.06)	17.28 (0.5)

Classifications based on RMI and SHI

RMI - The NS plot (Fig 5.17a) suggests about 6 groups while the IL plot (Fig 5.17b) is not very informative.

SHI - The NS and LRT plots (Fig 5.18 a & b) suggest 6 and ≥ 4 groups, respectively.

CA classifies entries into 6 groups based on each of the indices (Fig 5.17c & Fig 5.18c). However, the allocation of entries to the groups varies as indicated by the group estimates based on the two indices given below.

Based on RMI:

j	n_j	Mean (s.e.)		
		LF	DH	ST
1	10	3.98 (0.11)	0.05 (0.04)	14.74 (0.64)
2	14	4.22 (0.09)	0.24 (0.07)	16.72 (0.49)
3	7	4.69 (0.16)	0.40 (0.17)	21.90 (1.44)
4	1	3.89 -	0.00 -	10.18 -
5	12	4.39 (0.13)	0.60 (0.19)	16.84 (0.47)
6	11	3.99 (0.15)	0.12 (0.06)	18.29 (1.45)

Based on SHI:

j	n_j	Mean (s.e.)		
		LF	DH	ST
1	17	4.29 (0.08)	0.15 (0.03)	15.42 (0.28)
2	20	4.04 (0.09)	0.26 (0.06)	17.31 (0.34)
3	3	4.41 (0.44)	1.11 (0.63)	22.34 (3.90)
4	1	3.79 -	0.00 -	30.86 -
5	9	4.26 (0.19)	0.44 (0.19)	20.00 (0.82)
6	5	4.58 (0.24)	0.03 (0.03)	12.40 (0.74)

Comment on results for data set B

The results from this data set are inconsistent (Table 5.5b). Possible reasons could be the lack of correlation among variables as well as absence of distinct groups among entries studied. Apart from RMI, use of the other indices would result into selecting entries that are resistant for one damage variable and susceptible for the other(s). Based on SHI, for example, the lone entry in group 4 (same as group based on standardised variables) was resistant to LF and DH yet highly susceptible to ST. RMI would lead to selection of the lone entry in group 4, which consistently performed best, and perhaps the 10 in group 1 for that model. Selections based on SHI and the standardised variables would require introduction of economic weights (relative importance) for each variable.

Table 5.5b Composition of the groups based on indices (data set B).

ENTRY	LF	DH	ST	STDGP	RMIGP	SHIGP
1	4.73	0.17	18.34	7	4	3
2	4.39	0.17	14.65	4	6	2
3	3.85	0.17	19.43	3	3	3
4	4.02	0.17	13.39	4	2	2
5	5.01	0.50	14.66	7	5	2
6	4.16	1.00	14.04	2	3	3
7	4.76	0.33	19.82	7	4	5
8	3.70	0.17	17.70	3	6	3
9	3.92	0.00	15.33	4	2	2
10	4.68	0.00	11.09	4	2	6
11	4.83	0.17	16.07	7	5	2
12	4.18	0.00	13.72	4	2	6
13	4.86	0.17	16.47	7	5	2
14	4.74	0.32	22.04	7	4	5
15	4.24	0.00	15.73	4	6	2
16	5.29	0.00	29.34	1	4	4
17	4.19	0.33	15.60	4	3	2
18	4.30	0.33	17.26	7	5	3
19	5.01	0.17	13.63	7	3	6
20	4.63	0.33	16.36	7	5	3
21	3.76	0.00	17.49	3	2	3
22	4.04	0.00	14.89	4	2	2
23	4.54	0.00	17.78	7	3	2
24	5.14	0.00	13.37	7	6	6
25	4.54	0.17	17.45	7	5	3
26	3.57	0.67	15.59	2	6	3
27	3.87	1.17	21.83	2	4	4
28	4.37	0.00	19.12	7	3	3
29	4.05	0.50	14.77	4	3	3
30	3.54	0.33	17.50	3	3	3
31	4.31	0.17	16.44	4	3	2
32	4.22	0.00	15.19	4	2	2
33	3.88	0.00	15.59	4	2	2
34	3.41	0.00	20.99	3	6	5
35	4.00	0.00	20.94	3	6	5
36	4.11	0.17	17.78	3	3	3
37	3.33	0.00	17.74	3	2	3

Table 5.5b continued

ENTRY	LF	DH	ST	STDGP	RMIGP	SHI
38	4.35	0.00	18.45	7	3	3
39	4.02	0.50	17.29	2	5	3
40	4.24	0.17	16.07	4	3	2
41	3.89	0.00	10.18	4	1	6
42	3.82	0.33	12.94	4	2	2
43	3.55	0.17	17.45	3	6	3
44	4.02	0.17	17.93	3	3	3
45	4.92	0.00	23.53	1	4	5
46	4.28	0.17	15.61	4	3	2
47	4.13	0.17	15.73	4	6	2
48	3.58	0.33	20.09	3	5	5
49	3.79	0.00	30.86	5	6	1
50	3.79	0.33	19.38	3	5	5
51	4.26	0.33	16.37	4	5	3
52	4.65	1.83	14.80	6	5	5
53	4.07	2.17	15.84	6	5	4
54	4.02	0.00	18.22	3	6	3
55	4.53	0.83	18.43	7	4	5

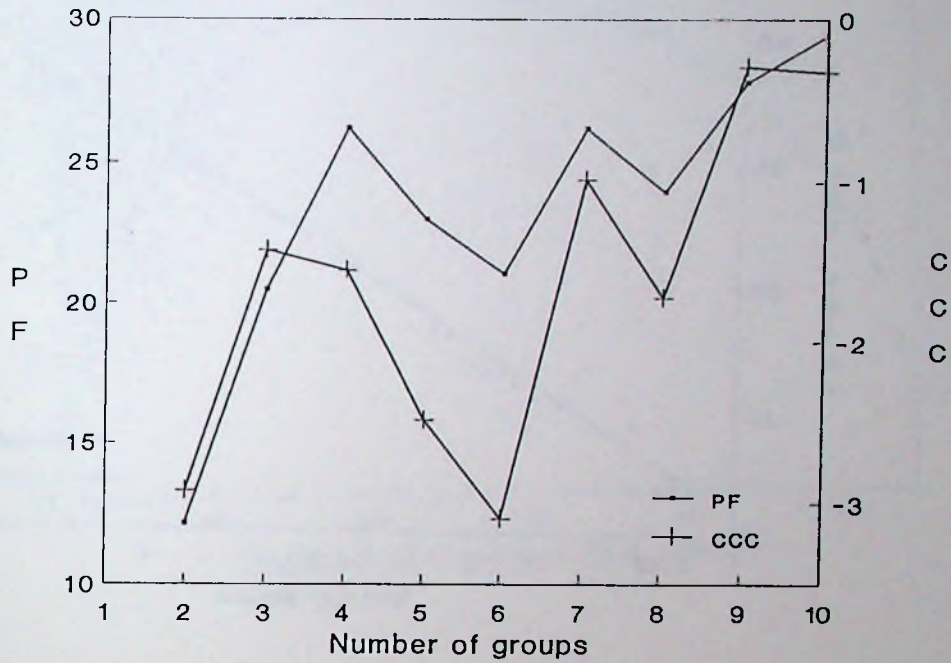


Figure 5.16 Plot of PF and CCC for optimum classifications into g groups based on damage variables (standardised).

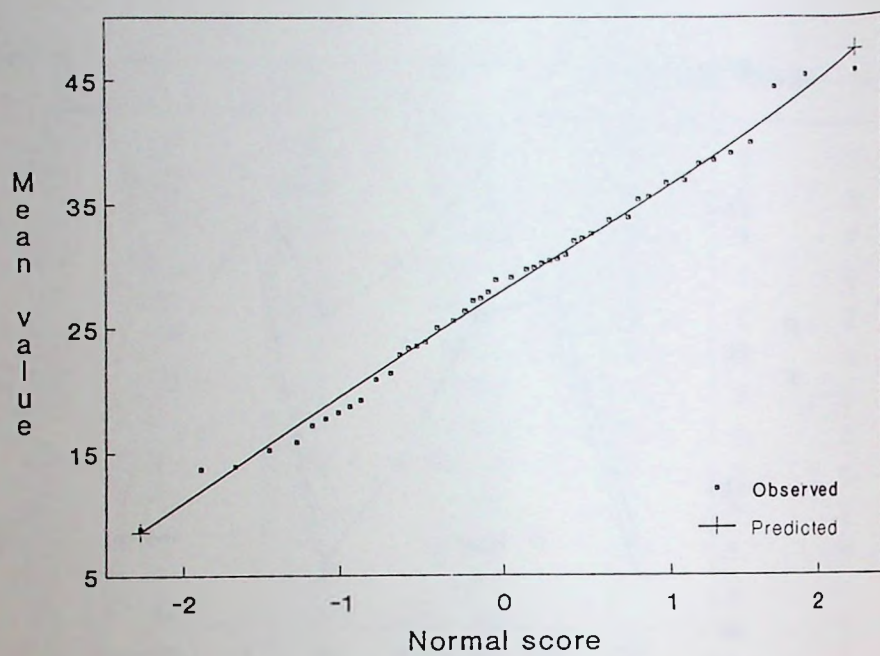


Figure 5.17a Plot of B-RMI for the 55 entries against their respective normal scores.

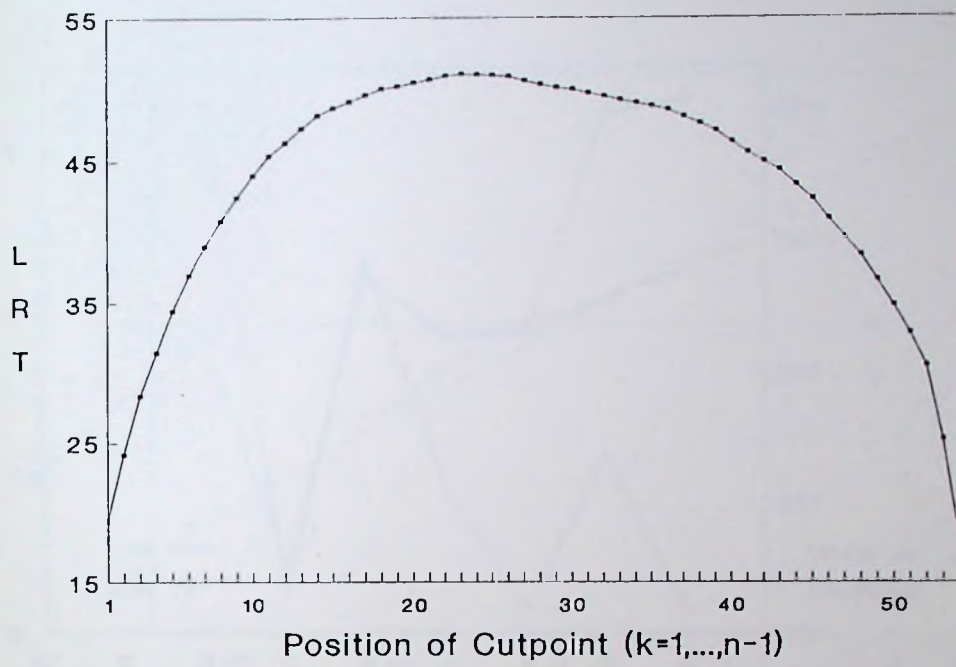


Figure 5.17b Plot of LRT statistic for the 2-group model against hypothetical cutpoints for B-RMI.

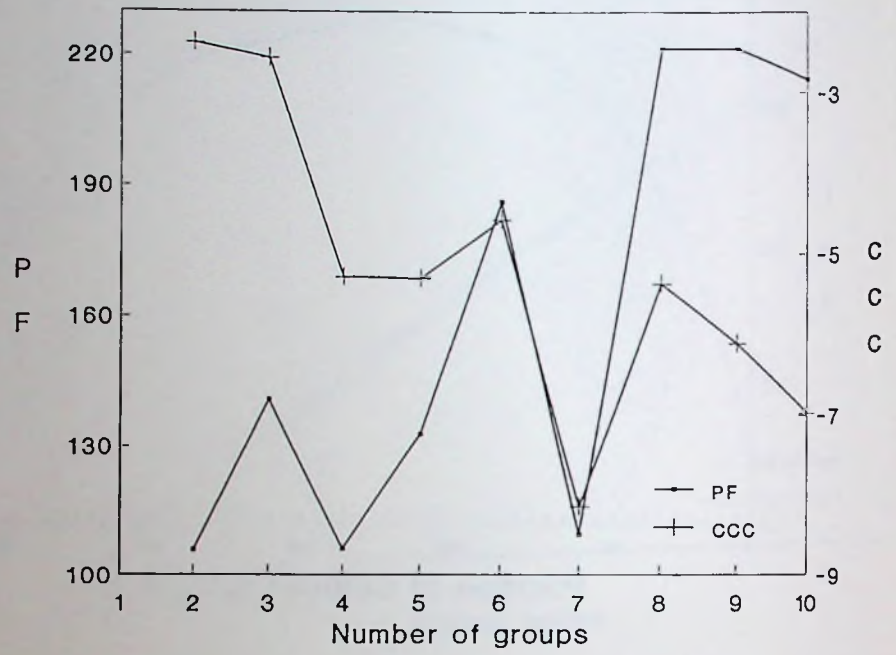


Figure 5.17c Plot of PF and CCC for optimum classifications into groups with respect to B-RMI.

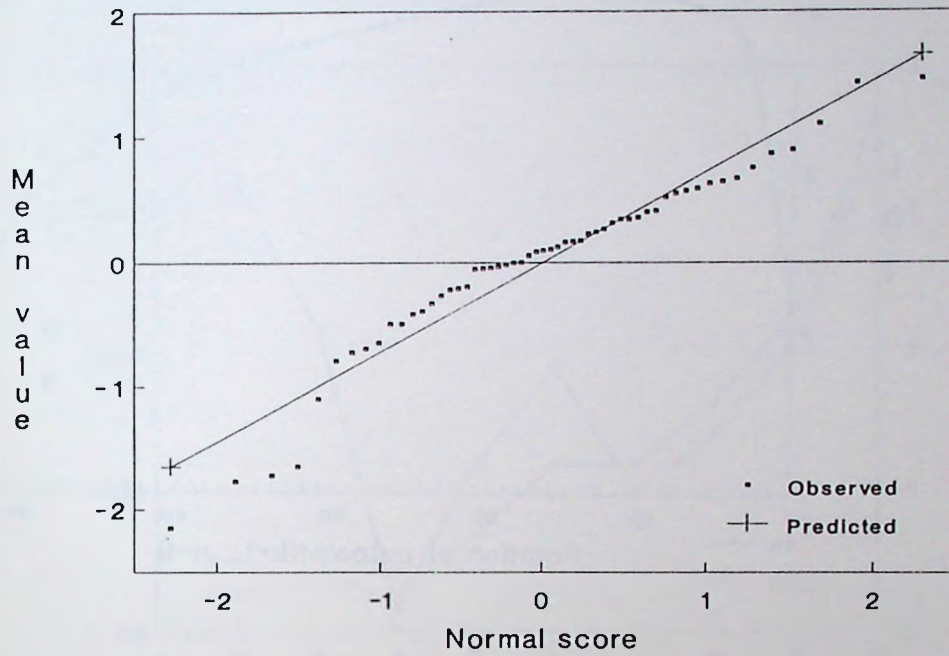


Figure 5.18a Plot of B-SHI for the 55 entries against their respective normal scores.

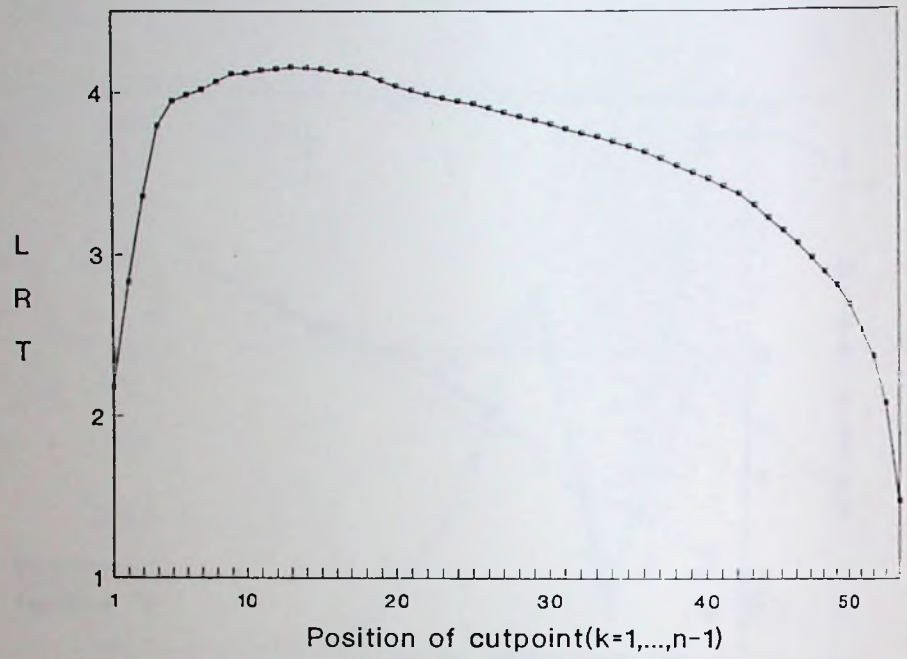


Figure 5.18b Plot of LRT statistic for the 2-group model against hypothetical cutpoints for B-SHI.

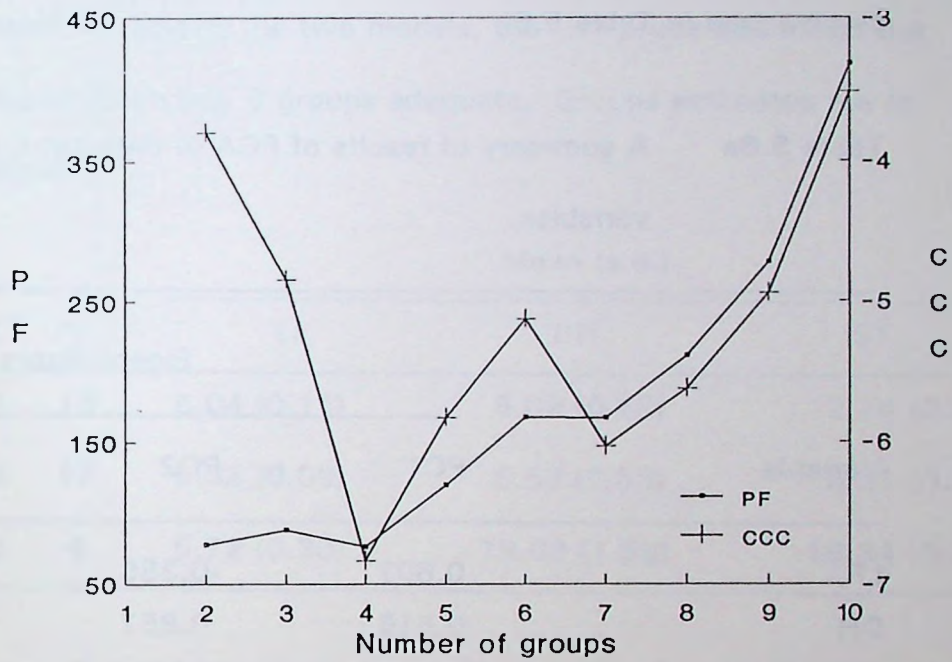


Figure 5.18c Plot of PF and CCC for optimum classifications into g groups with respect to B-SHl.

5.3.1.3 DATA SET C

Results of PCA performed on the correlation matrix (Table 4.2) are summarised in Table 5.6.

Table 5.6a A summary of results of PCA of data set C damage variables.

Variable	Eigenvectors		
	PC1	PC2	PC3
LF	0.607	-0.355	-0.711
DH	0.515	0.857	0.012
ST	0.605	-0.373	0.703
Eigenvalue	1.713	0.742	0.544
Cumulative proportion	0.571	0.818	1.000

As implied by the eigenvectors (Table 5.6), PC1 is a measure of the average contribution of each damage variable and PC2 is a contrast between DH and the other two variables. The two PCs accounted for 82% of the total variation in the data. The distribution of entries along the first two PCs is given in Fig 5.19a.

Classification based on PCS

CA indicated optimal classifications at $g = 3$ & 5 (Fig 5.19b).

However, among the two models, the LRT procedure found the classification into 3 groups adequate. Groups estimates are as follows:

j	n_j	Mean (s.e.)		
		LF	DH	ST
1	18	5.04 (0.11)	5.09 (0.69)	12.74 (0.27)
2	17	5.62 (0.09)	6.53 (0.59)	16.31 (0.46)
3	4	5.72 (0.30)	19.03 (1.59)	16.34 (0.31)

Plots of groups along the PCs are shown in Fig 5.19c. As can be seen, the 18 entries in group 1 are the most resistant, followed by the 17 in group 2, while the most susceptible are the 4 in group 3.

Classification by RMI

The NS and LRT plots suggest at least 3 groups (Figs 5.20a & b). CA (Fig 5.20c) gives 2 - 3 classes. The estimates for the 3-group model are:

j	n_j	Mean (s.e.)		
		LF	DH	ST
1	11	4.82 (0.12)	3.06 (0.72)	12.78 (0.38)
2	15	5.44 (0.08)	6.54 (0.64)	14.14 (0.48)
3	13	5.73 (0.11)	10.81 (1.72)	16.87 (0.59)

A comparison of the classifications obtained using the two index forms by MANOVA gave the following results:

Index	Wilk's λ	$F_{6, 68}$	P
PCs	0.1245	20.78	< 0.001
RMI	0.2374	11.80	< 0.001

From the values of the test statistics, it would appear the reduction in total variation by the classification using PCs was much higher than using the RMI.

Comment on results for data set C

Based on PCs and RMI, using LRT or CA would select 18 and 11 entries (Table 5.6b), respectively, while 4 entries would result from setting the selection intensity at 10% or 8 if it were increased to 20%.

Table 5.6b Composition of the groups based on indices (set C).

ENTRY	LF	DH	ST	PCGP	RMIGP
12	4.19	0.85	11.64	1	1
38	4.62	1.85	12.21	1	1
39	4.38	8.12	10.97	1	1
26	4.44	4.89	12.65	1	1
14	4.89	4.49	12.64	1	1
9	5.37	0.00	12.32	1	1
29	5.05	6.04	11.93	1	1
28	4.50	4.07	15.73	1	1
35	5.00	3.56	13.81	1	1
19	5.29	1.71	13.27	1	1
32	4.86	10.97	12.55	1	2
7	5.54	5.88	11.59	1	2
4	5.32	4.49	13.43	1	1
1	5.59	5.13	11.91	1	2
10	5.59	6.12	11.97	1	2
5	5.41	5.34	13.98	1	2
24	5.22	8.33	14.18	1	2
34	5.46	9.71	12.58	1	2
17	5.15	6.04	15.80	2	2
15	5.31	5.41	15.53	2	2
27	5.14	8.12	15.36	2	2
21	5.51	3.63	15.90	2	2
13	5.83	5.30	14.03	2	2
11	5.45	10.72	13.94	2	2
30	5.51	9.70	14.49	2	3
8	6.01	5.48	14.44	2	3
20	5.33	4.34	18.29	2	2
23	6.26	3.13	14.57	2	2
3	5.31	8.42	17.04	2	3
33	5.58	16.89	12.98	3	3
37	5.89	6.12	16.32	2	3
22	5.47	5.44	18.74	2	3
36	4.93	18.59	16.27	3	3
18	5.51	5.34	19.81	2	3
2	6.31	8.07	15.89	2	3
6	5.67	11.32	17.70	2	3
16	5.92	4.58	19.49	2	3
25	6.11	16.95	16.75	3	3
31	6.27	23.63	19.36	3	3

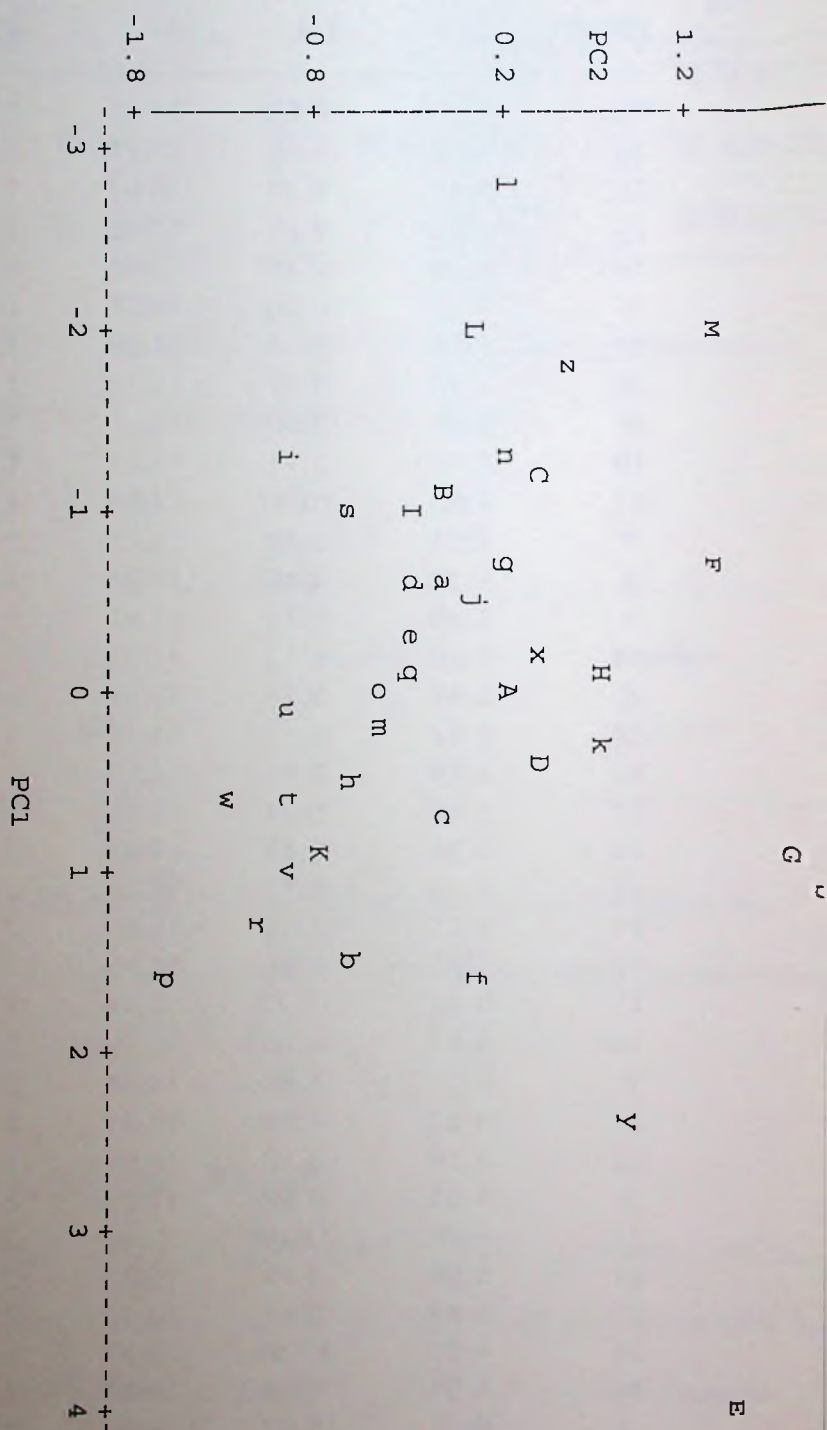


Figure 5.19a Plot of the 39 entries against their values for PC1 and PC2 (letters represent the entry codes).

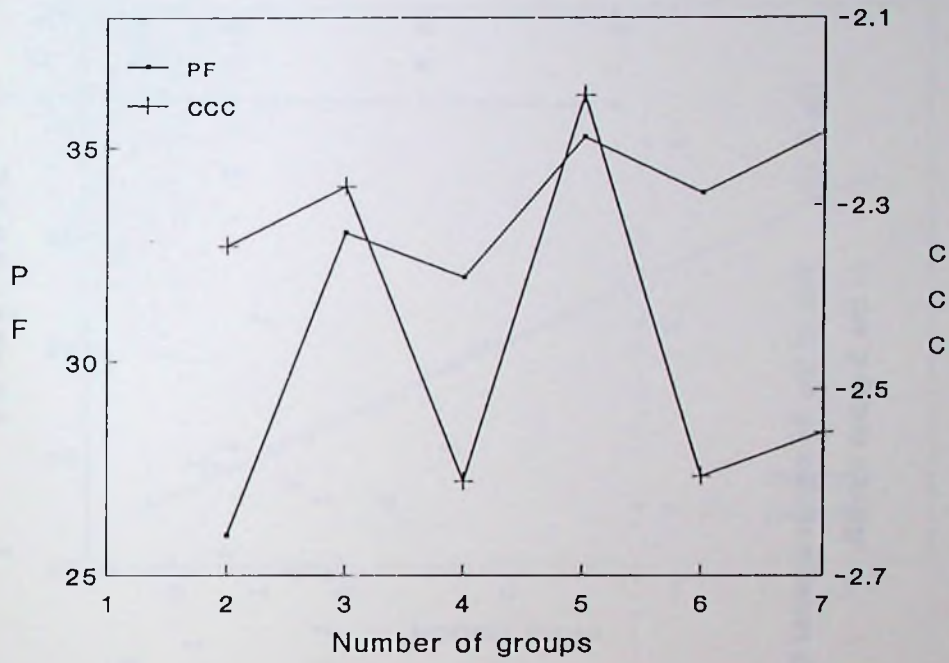


Figure 5.19b Plot of PF and CCC for optimum classifications into g groups using the first two PCs for data set C.

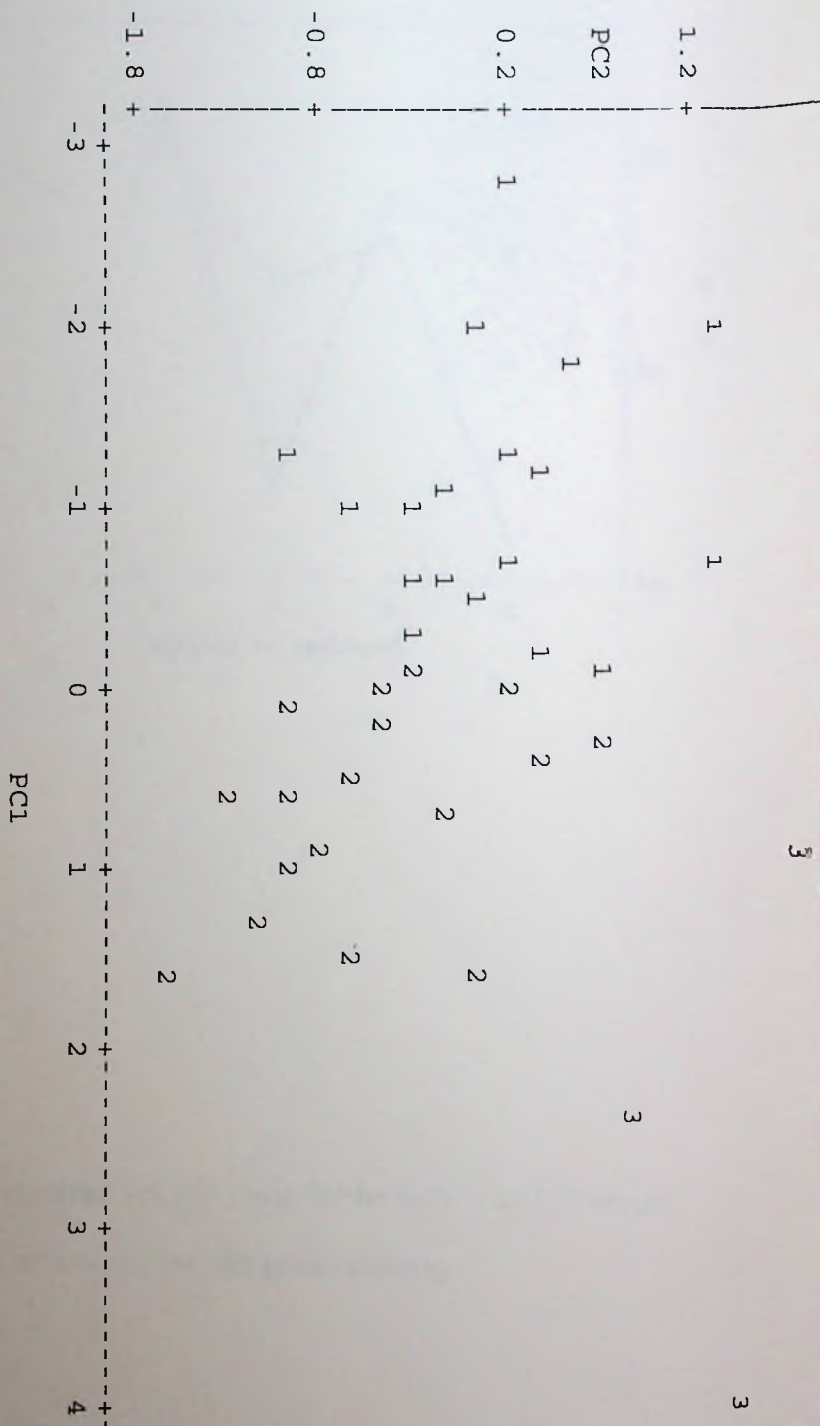


Fig 5.19c Plot of the 39 entries against their values for PC1 and PC2 (figures indicate their respective classes in the 3-group model).

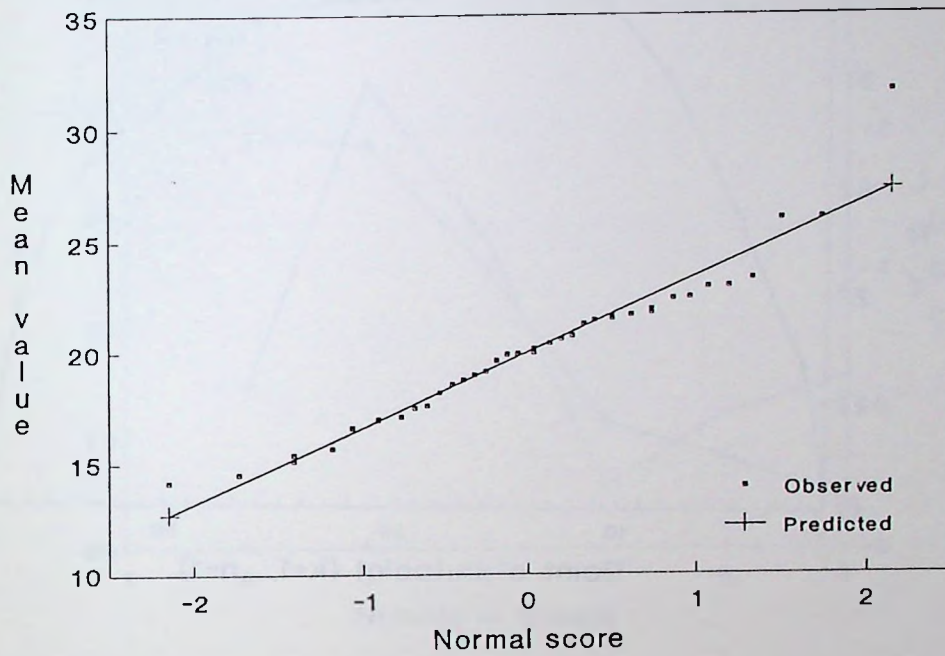


Figure 5.20a Plot of C-RMI for the 39 entries against their respective normal scores.

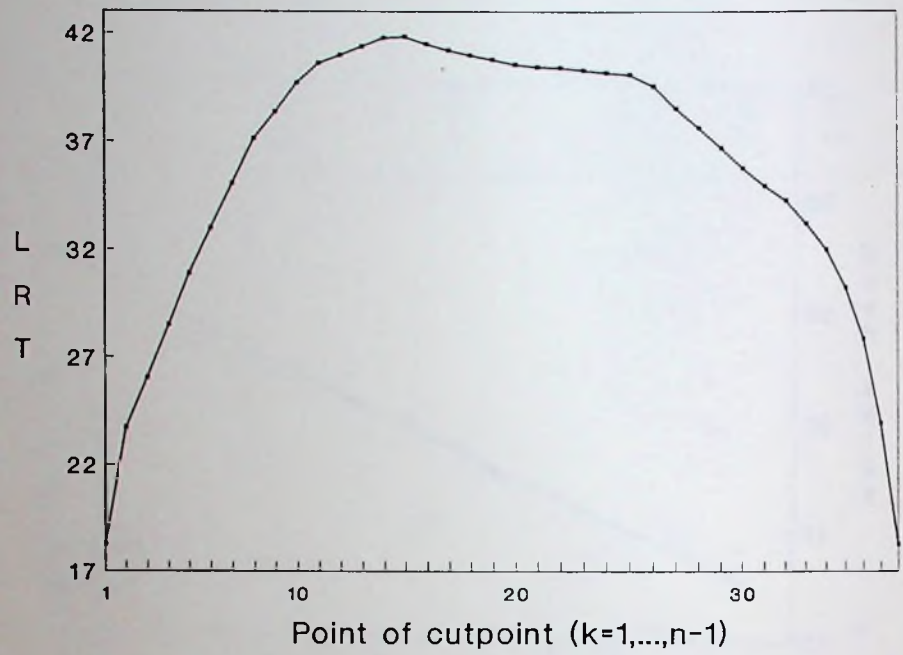


Figure 5.20b Plot of LRT statistic for the 2-group model against hypothetical cutpoints for C-RMI.

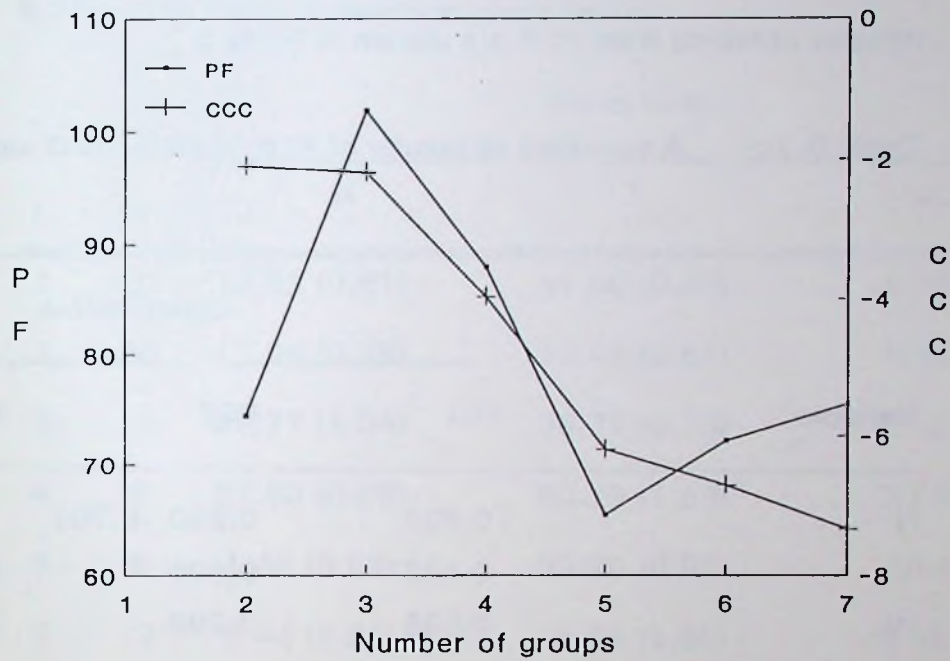


Figure 5.20c Plot of PF and CCC for optimum classifications into g groups with respect to C-RMI.

5.3.1.4 DATA SET D

Results obtained from PCA are shown in Table 5.7.

Table 5.7c A summary of results of PCA of data set D variables

Variable	Eigenvectors		
	PC1	PC2	PC3
N	0.633	0.330	-0.701
T	-0.438	0.898	0.027
WL	0.638	0.299	0.713
Eigenvalue	2.237	0.706	0.056
Cumulative proportion	0.746	0.981	1.000

PC1 and PC2 account for 75% and 24% of the variation in the original variables, respectively. Variables N and WL, which are highly correlated (Table 4.2), contribute equally to PC1 while T contributes a relatively smaller negative weight. The distribution of entries along the two PC axes is shown in Fig 5.21a.

Classification based on PCS

CA gave optimal classifications with $g = 3$ and 6 (Fig 5.21b). LRT supported the 6-groups model as the optimum solution. The

distribution of the resulting classes along the two PCs is shown in Fig 5.21c. The table of means is given below:

j	n_j	Mean (s.e.)		
		N	T	WL
1	20	17.62 (0.81)	41.08 (0.57)	1.76 (0.09)
2	20	11.64 (0.75)	53.49 (0.51)	1.14 (0.06)
3	14	29.77 (1.04)	38.71 (0.78)	3.16 (0.12)
4	6	27.50 (0.65)	50.79 (1.63)	2.85 (0.08)
5	5	1.45 (0.52)	58.00 (0.95)	0.14 (0.06)
6	12	6.44 (0.91)	41.34 (1.06)	0.58 (0.09)

Classification based on I

The normal score plot (Fig 5.22a) does not clearly indicate a definite grouping in the data, as most of the points lie quite close to the line, but at least 4 groups are suggested. QBT partitions the entries into groups of 35, 42 entries. With the 2-group model assumption, LRT and CA partition the entries into groups of 34, 43 entries. However, further splitting into 4 groups is suggested as the optimum classification by both the CA and LRT (see Fig 5.22b & c). The table below shows the classes description.

Mean (s.e.)

j	n_j	I	N	T
1	7	0.89 (0.23)	1.68 (0.39)	52.14 (3.91)
2	27	4.27 (0.14)	9.70 (0.66)	49.66 (1.19)
3	30	6.73 (0.11)	19.66 (1.01)	44.02 (0.96)
4	13	8.89 (0.23)	29.87 (1.14)	38.29 (0.74)

Comment on results for data set D

Results of LRT and CA procedures based on the first two PCs suggest selection of 5 entries. Using I results into selection of 7 entries. Since N and WL are highly correlated ($r = 0.943$, Table 4.2) exclusion of WL in the computation of I leads to somewhat similar selections (Table 5.7b) as those based on PCs where appropriate weights are assigned to the variables based on existing correlations. These selections are close to the 10% selection intensity often employed. This is probably due to chance as stated earlier.

Table 5.7b Composition of the groups based on indices (set D).

ENTRY	N	T	WL	I	IGP	PCGP
77	0.75	60.00	0.10	0.29	1	5
76	0.75	58.00	0.05	0.30	1	5
75	1.00	57.00	0.04	0.48	1	5
74	1.25	55.00	0.18	0.73	1	5
73	2.00	41.00	0.03	1.04	1	6
72	2.50	34.00	0.34	1.58	1	6
71	3.50	60.00	0.34	1.83	1	5
70	3.75	36.25	0.20	2.73	3	6
69	4.00	41.29	0.73	2.83	3	6
68	6.29	54.50	0.86	3.16	3	2
67	4.00	39.75	0.50	3.20	3	6
66	8.25	56.25	0.66	3.73	3	2
65	9.50	58.00	1.02	3.80	3	2
64	8.28	54.00	0.87	3.86	3	2
63	9.00	54.50	1.02	3.92	3	2
62	5.75	43.25	0.66	3.97	3	6
61	9.75	55.75	1.01	4.03	3	2
60	7.75	44.75	0.98	4.11	3	6
59	9.25	52.25	0.98	4.22	3	2
58	9.55	52.75	0.90	4.23	3	2
57	11.00	54.00	1.01	4.37	3	2
56	11.50	55.25	1.49	4.39	3	2
55	10.75	52.75	1.19	4.41	3	2
54	9.50	49.50	1.07	4.50	3	2
53	11.25	52.25	1.40	4.54	3	2
52	9.00	46.00	0.92	4.75	3	6
51	12.50	51.75	1.03	4.82	3	2
50	15.75	56.75	1.27	4.84	3	2
49	8.50	42.25	0.59	5.02	3	6
48	16.25	55.00	1.13	5.04	3	2
47	8.50	42.25	0.77	5.05	3	6
46	11.00	46.00	0.29	5.16	3	6
45	15.00	51.50	1.19	5.21	3	2
44	16.25	42.25	1.12	5.31	3	1
43	18.50	51.00	1.67	5.60	4	2
42	17.00	49.75	1.64	5.69	4	2
41	28.00	56.75	2.86	5.87	4	4
40	10.50	39.25	0.99	5.99	4	6
39	14.00	52.25	1.46	6.17	4	2

Table 5.7b continued

ENTRY	N	T	WL	I	IGP	PCGP
38	14.00	42.75	1.50	6.18	4	1
37	25.75	52.25	2.57	6.21	4	4
36	11.50	39.00	1.43	6.23	4	1
35	28.75	53.00	2.79	6.33	4	4
34	19.50	46.50	1.79	6.33	4	1
33	12.50	39.75	1.22	6.34	4	1
32	18.25	45.25	1.28	6.39	4	1
31	14.25	42.75	1.33	6.52	4	1
30	17.50	43.75	1.32	6.53	4	1
29	14.25	39.75	1.83	6.62	4	1
28	26.75	49.25	2.98	6.67	4	4
27	14.00	36.75	2.11	6.72	4	1
26	21.25	43.25	2.26	7.04	4	1
25	17.25	39.75	1.61	7.09	4	1
24	29.75	47.75	3.18	7.10	4	4
23	26.00	45.75	2.74	7.12	4	4
22	14.75	37.25	1.73	7.16	4	1
21	21.00	41.75	1.72	7.28	4	1
20	22.00	42.25	1.89	7.29	4	1
19	17.00	38.75	2.13	7.30	4	1
18	21.00	40.75	2.65	7.40	4	1
17	27.00	43.00	2.76	7.65	4	3
16	22.00	40.25	2.14	7.66	4	1
15	20.50	39.00	2.04	7.74	4	1
14	25.25	41.25	3.15	7.81	4	3
13	23.75	40.00	2.08	7.88	2	1
12	28.28	40.75	3.22	8.18	2	3
11	34.75	43.25	3.51	8.19	2	3
10	27.75	40.00	2.74	8.31	2	3
9	24.75	37.75	2.52	8.48	2	3
8	28.75	39.00	3.48	8.60	2	3
7	32.00	39.75	2.45	8.72	2	3
6	26.25	37.25	3.37	8.79	2	3
5	27.25	36.75	3.37	8.99	2	3
4	31.25	37.75	3.69	9.09	2	3
3	36.75	38.00	2.61	9.48	2	3
2	31.25	34.50	3.92	10.01	2	3
1	35.50	33.00	3.56	10.82	2	3

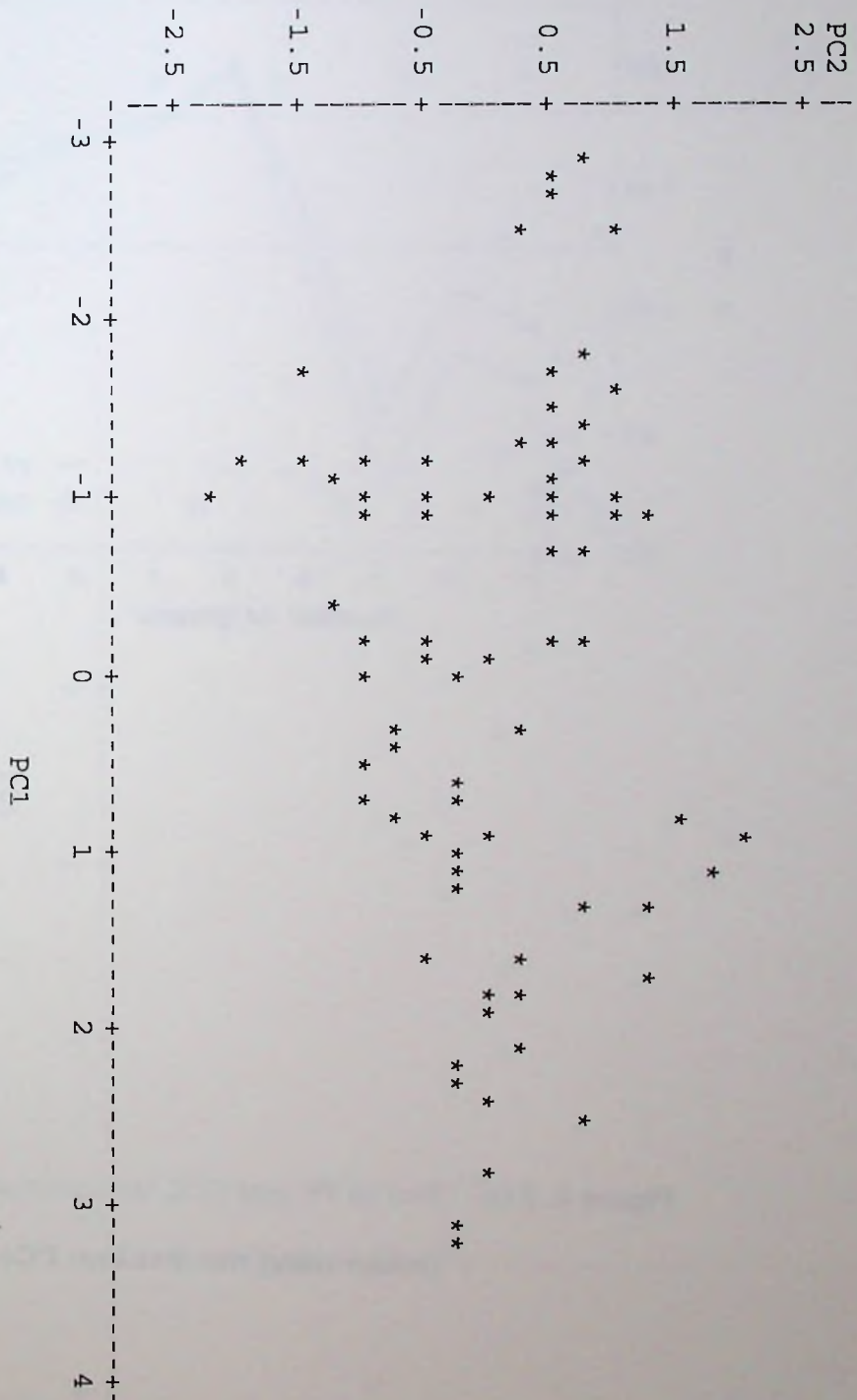


Figure 5.21a Plot of the 77 entries against their values for PC1 and PC2.

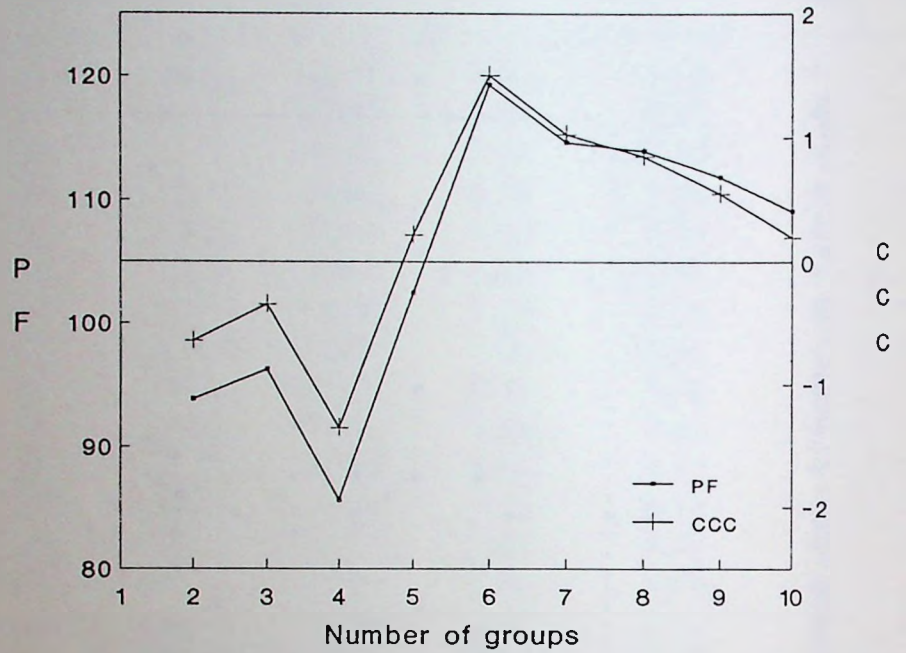


Figure 5.21b Plot of PF and CCC for optimum classifications into groups using the first two PCs for data set D.

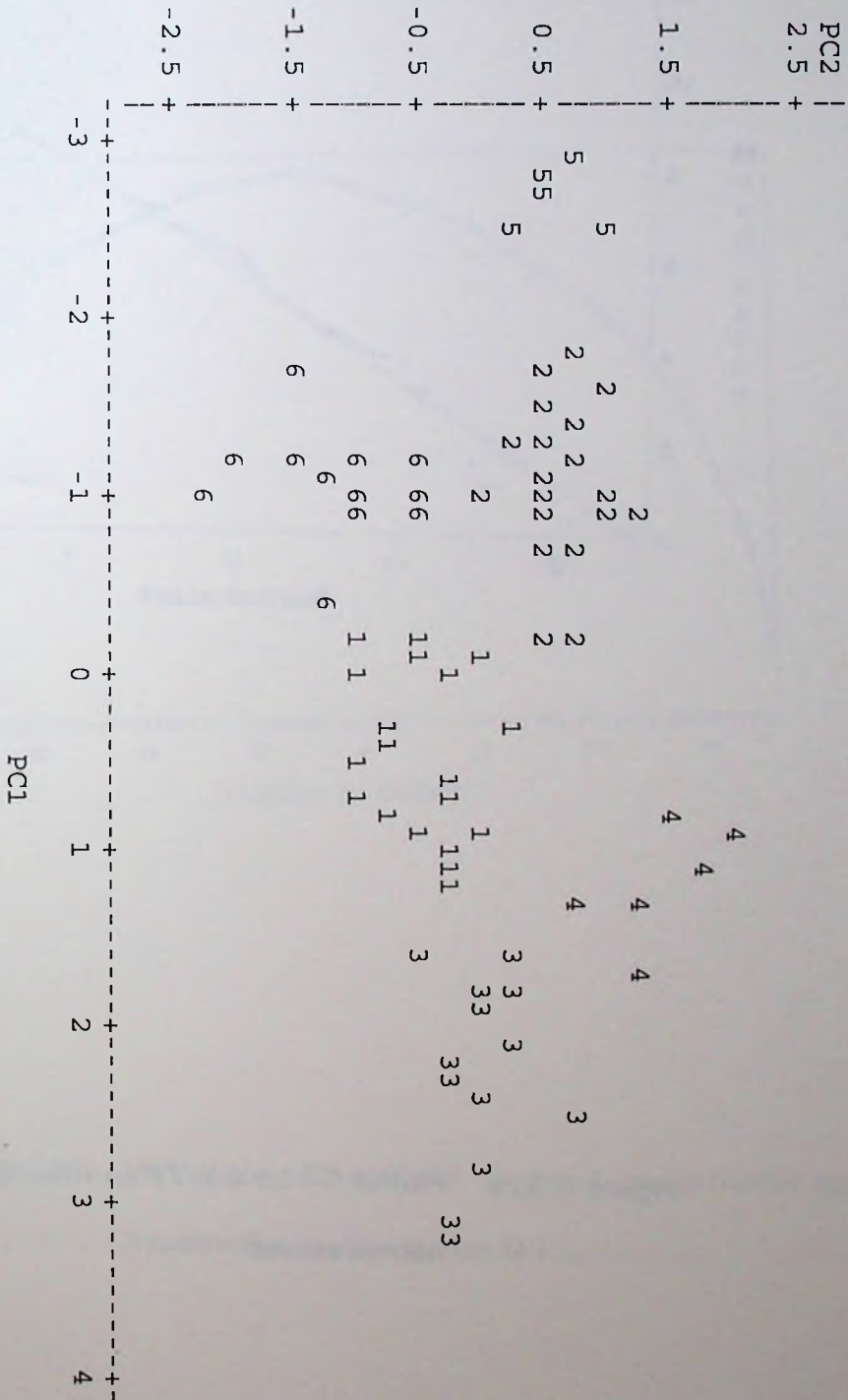


Figure 5.21c Plot of the 77 entries against their values for PC1 and PC2 (figures indicate their respective classes in the 6-group model).

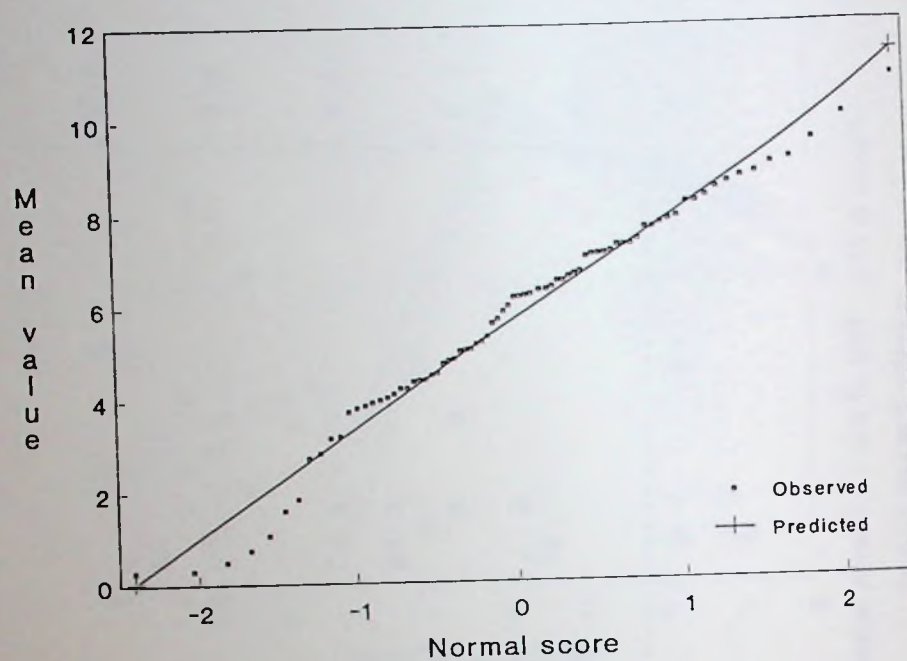


Figure 5.22a Plot of D-I for the 77 entries against their respective normal scores.

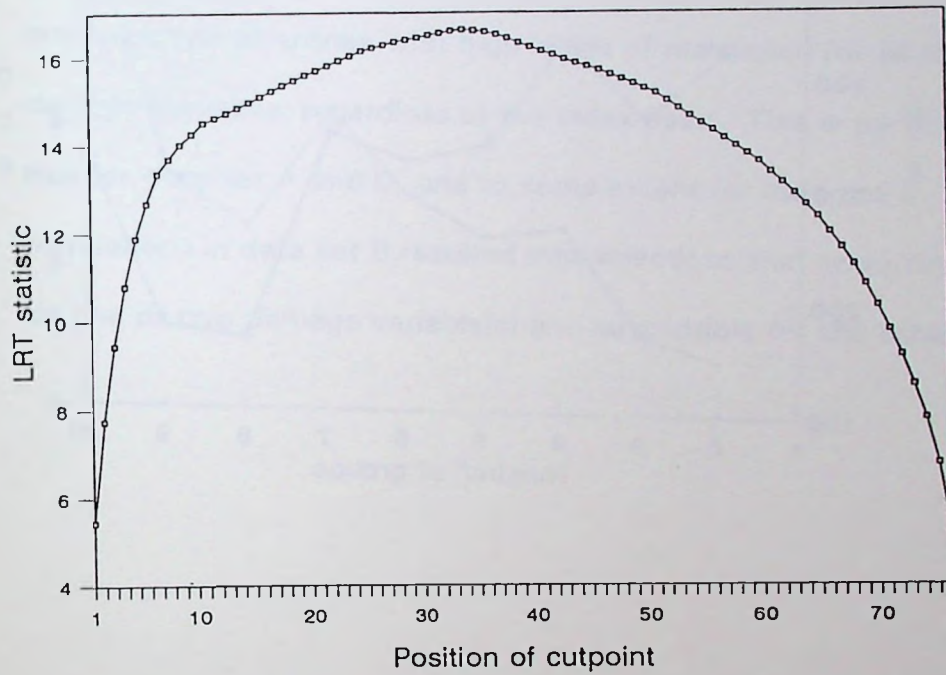


Figure 5.22b Plot of the LRT statistic for the 2-group model against hypothetical cutpoints for D-I.

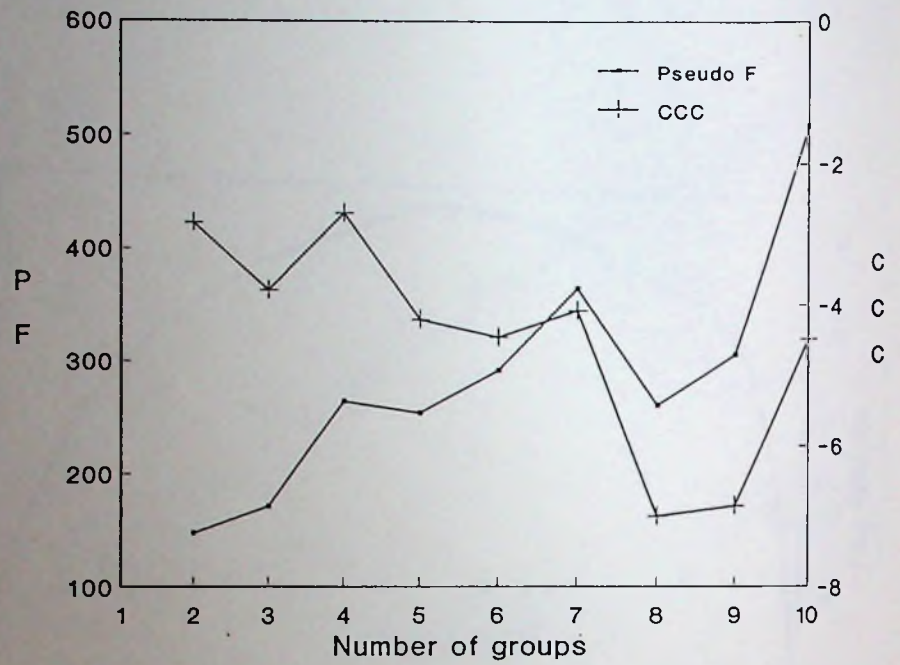


Figure 5.22c PF and CCC for optimum classifications into g groups with respect to D-l.

5.3.2 Summary of results for multivariate analyses

Table 5.8 gives a summary of the optimum classifications obtained using the different procedures for each of the 4 data sets.

Significant intercorrelations among damage measures resulted into selection of entries with high levels of resistance for all the damage variables, regardless of the index used. This is particularly true for data set A and D, and to some extent for data set C. Lack of correlations in data set B resulted into selections that were resistant for one or two damage variable(s) and susceptible for the other(s).

Table 5.8 Summary of optimal groupings by different indices.

Data set	Index			
	PCs	RMI	SHI	I
A	7: 3,24,2,3,8,2,9	7:8,20,1,10,4,2,6	7: 9,4,2,8,10,2,16	na
B	7:2,4,13,18,1,2,15	6:10,14,7,1,12,11	6: 17,20,3,1,9,5	na
C	3: 18,17,4	3: 11,15,13	na	na
D	6: 20,20,14,6,5,12	na	na	4: 7,27,30,13

5.4 COMPARISON OF PROCEDURES USING SIMULATED DATA

In Chapter 3, it was indicated that it is necessary to develop a criterion for comparing the proposed procedures as well as the performance of the indices included in the study. In this section results of the application of the procedures to simulated data, described in Chapter 4, are used for the comparisons. For this purpose the procedures are applied to known data structures and comparisons made on the number (or proportion) of groupings recovered, which represents the ability of the procedures to identify the optimum partitioning of n entries into g groups. Conditions for optimal performance of the procedures are also identified.

5.4.1 Number of groupings

For the different simulated data sets described above, the number of possible groupings to consider is shown in Table 5.9b.

Table 5.9 Number of groupings examined using LRT and QBT procedures.

n	Number of groups (g)			
	2	3	4	5
30	29	406	3654	23751
60	59	1711	32509	455126
120	119	7021	273819	7940751 ¹
240	239	28441	2246839 ¹	132563501 ¹
600	599	179101	35641099 ¹	5310523751 ¹
1200	1199	718201 ¹	286562199 ¹	85682097501 ¹

¹ Not all groupings were investigated due to limitations of computer capacity. For these cases, the study concentrated on a few groupings in the neighbourhood of the correct cutpoints.

5.4.2 Results

5.4.2.1 *Univariate classifications*

An inspection of the normal score (NS) plots revealed that they tended to give a clear indication of the number of groups, but actual location of the cutpoint depended on how distinct the groups were as well as the number of data points.

For a given classification method, the frequency of correct recovery of the original grouping in the data sets is presented in Table 5.10. A correct classification in this case refers to a situation when the correct number of groups was determined and every observation was correctly placed in the known group it belonged. Otherwise, the classification was considered not correct.

Analysis of frequencies

Chi-square tests were carried out (using PROC FREQ of SAS) on the observed frequencies. The tests indicated the following:

- (a) Size of groups influenced the recovery rate of the procedures.

The probability of making a correct classification depends on the composition of the groups ($\chi^2_2 = 19.7$, $p < 0.001$). It is evident from Table 5.10 that in general, regardless of method, equal sized groups were easier to recover than unequal ones. The overall recovery rate was least for data sets that had the largest range between group sizes (s_2).

Table 5.10 Frequency of indication of correct and incorrect classifications of univariate data sets* (C and N represent correct and incorrect classifications, respectively).

Number of groups (g)		Method of classification					
		LRT		CA		QBT	
		C	N	C	N	C	N
g = 1		20	4	-	-	-	-
g = 2	s ₁	24	0	24	0	18	6
	s ₂	22	2	20	4	10	14
g = 3	s ₁	24	0	22	2	-	-
	s ₂	22	2	18	6	-	-
	s ₃	20	4	18	6	-	-
g = 4	s ₁	22	2	22	2	-	-
	s ₂	18	6	16	8	-	-
	s ₃	20	4	20	4	-	-

* Data sets include 4 replicates for each of the 54 factor combinations shown in Table 4.3.

Table 5.10a Total frequencies of correct (C) and incorrect (N) classifications for different levels of s .

	s_1	s_2	s_3	Total
C	156	126	78	360
N	12	42	18	18
Total	168	168	96	432

(b) There was, however, no association between relative group sizes and method of determination. That is, the influence of relative size was consistent among the procedures. In other words, each procedure was exposed to the same influence of relative group sizes. Even QBT recovered more equally sized groups in the case of $g = 2$.

(c) No significant difference was observed in recovery rates of LRT and CA procedures. In absolute terms, LRT appeared to have achieved more correct determinations than CA (90% and 83%, respectively), but the observed difference was not significant ($\chi^2_1 = 3.2$, $p = 0.07$).

(d) Number of groups (g) influenced their recovery with 94%, 88% and 82% for $g = 2, 3$ and 4 , respectively ($\chi^2_2 = 6.88$, $p = 0.03$).

There was, however, no association between number of groups and their relative sizes ($\chi^2_2 = 0.43$, $p = 0.81$).

(e) There was no association between number of groups ($g > 1$) and method of determination ($\chi^2_2 = 0.16$, $p = 0.92$). Neither was the three factor interaction (gxsmethod) significant.

(f) Number of observations (entries) did not affect the performance of the procedures. The relative proportions of correct and incorrect determinations were maintained at various values of n . However, LRT required tremendous amounts of time and computer memory, especially with large data sets, to be able to go through the entire set of iterations.

(g) For the 2-groups model, where QBT was included among the determinant procedures, the recovery rate for QBT was much lower than those of the others ($\chi^2_2 = 27.4$, $p < 0.001$) (see Table 5.10b).

Table 5.10b Total frequencies of correct (C) and incorrect (N) classifications by different procedures for the 2-group model.

	CA	LRT	QBT	Total
C	44	46	28	118
N	4	2	20	26
Total	48	48	48	144

While CA and LRT recovered 92% and 96%, respectively, QBT recovered 58%.

(h) One advantage of LRT is its ability to almost correctly identify the 1-group model. This property has no obvious counterpart with other procedures studied, except with NS plots.

(i) It was further observed that among the incorrectly classified data sets with a large difference in group sizes (s_2), there was a tendency for the procedures to misclassify them into nearly equally-sized groups. In some few cases, more groups were created by splitting some bigger groups. This could have arisen partly from some outliers in the simulated groups.

5.4.2.2 *Multivariate classifications*

CA and LRT procedures were applied to the simulated multivariate data sets. For $r \approx 0$, their comparison included use of RMI and standardised variables, while for $r \approx 0.6$ & 0.9 , use of RMI was compared to using PCs. Combinations of $p = 2$ & 3 with $n = 30$ & 60 and $g = 1, 2, 3, 4$ with the respective levels of s , were used for the comparisons (see section 4.2).

Tables 5.11, 5.12, 5.13 and 5.14 show the frequencies of correct determinations for $r \approx 0.9$, 0.6 , and 0 , respectively.

Highly correlated variables ($r \geq 0.8$)

The results are summarised in Table 5.11.

Table 5.11 Frequency of correct and incorrect classifications when $r \geq 0.8$.

Number of groups* (g)		Method of classification							
		LRT (PC1)		LRT (RMI)		CA (PC1)		CA (RMI)	
		C	N	C	N	C	N	C	N
g = 1		3	1	2	2	-		-	
p = 2	s ₁	12	0	10	2	12	0	9	3
	s ₂	11	1	9	3	10	2	9	3
	s ₃	7	1	7	1	7	1	5	3
g = 1		2	2	2	2	-		-	
p = 3	s ₁	11	1	10	2	11	1	10	2
	s ₂	10	2	6	6	9	3	7	5
	s ₃	8	0	6	2	8	0	7	1

* Results for g = 2, 3 & 4 are combined.

Frequencies of correct and incorrect determinations for the three levels of s are given in Table 5.11a.

Table 5.11a Total frequencies of correct (C) and incorrect (N) classifications for different levels of s .

	s_1	s_2	s_3	Total
C	85	71	55	211
N	11	25	9	45
	96	96	64	256

There were significant differences in the overall recovery rates for different levels of s ($\chi^2_2 = 7.8$, $p = 0.021$). Frequencies of correct determinations for each group size (s) by different methods are given in Table 5.11b.

Table 5.11b Two-way table of method by level of s .

Size	Method of determination				Total
	CA (PC1)	CA (RMI)	LRT (PC1)	LRT (RMI)	
s_1	23	19	23	20	85
s_2	19	16	21	15	71
s_3	15	12	15	13	55
Total	57	47	59	48	211

No significant association was observed between relative group size and method of determination. That is, the influence of relative group size on power of procedure was consistent across the procedure-index combinations. However, there were differences between procedures ($\chi^2_3 = 12.16, p = 0.007$), with both LRT and CA using PC giving more correct determinations than when RMI was used (89% & 73%, respectively). No difference was indicated between the overall performance of CA and LRT (regardless of index adopted) procedures.

Moderately correlated variables ($0.5 \leq r \leq 0.7$)

Results are given in Table 5.12.

Table 5.12 Frequency of correct and incorrect classifications when
 $0.5 < r < 0.7$.

Number of groups* (g)	Method of classification								
	LRT (PC1&2)		LRT (RMI)		CA (PC1&2)		CA (RMI)		
	C	N	C	N	C	N	C	N	
g = 1	2	2	1	3	-		-		
p = 2	s ₁	10	2	6	6	11	1	5	7
	s ₂	7	5	3	9	7	5	3	9
	s ₃	6	2	5	3	6	2	4	4
g = 1	3	1	1	3	-		-		
p = 3	s ₁	10	2	9	3	12	0	6	6
	s ₂	9	3	3	9	5	7	4	8
	s ₃	6	2	5	3	5	3	3	5

* Results for g = 2, 3 & 4 are combined.

Table 5.12a Total frequencies of correct (C) and incorrect (N) classifications for different levels of s.

	s ₁	s ₂	s ₃	Total
C	69	41	50	150
N	27	55	20	106
	96	96	64	256

Table 5.12a gives frequencies of recoveries for different levels of s. There is a significant effect of relative sizes of groups on the recovery rate ($\chi^2_2 = 17.37, p < 0.001$). There is no association between method and level of s (Table 5.12b).

Table 5.12b A two-way table of method by levels of s.

Size	Method of determination				Total
	CA (PC1&2)	CA (RMI)	LRT (PC1&2)	LRT (RMI)	
s ₁	23	11	20	15	69
s ₂	19	7	16	12	54
s ₃	11	7	12	10	40
Total	53	25	48	37	163

No differences, either between CA and LRT, but both recovered significantly more classifications with PC(s) (73%) than with RMI (44%) ($\chi^2_1 = 12.16$, $p < 0.001$). In fact, as implied by 44%, use of RMI made more incorrect determinations than it recovered.

Uncorrelated variables ($r \leq 0.3$)

As indicated in Section 4.2, values of $p = 2$ and 3 were considered with $n = 30$ and 60. Recovery frequencies for $p = 2$ and $p = 3$ are given in Tables 5.13 and 5.14.

Table 5.13 Frequency of correct and incorrect classifications when $r \leq 0.3$ and $p = 2$.

Number of groups (g)		Method of classification							
		LRT*		LRT (RMI)		CA*		CA (RMI)	
		C	N	C	N	C	N	C	N
g = 1		6	6	2	10	-	-	-	-
g = 2	s ₁	11	1	7	5	9	3	10	2
	s ₂	10	2	3	9	10	2	4	8
g = 3	s ₁	12	0	7	5	12	0	9	3
	s ₂	9	3	4	8	8	4	7	5
	s ₃	8	4	5	7	11	1	7	5
g = 4	s ₁	10	2	8	4	11	1	10	2
	s ₂	2	10	2	10	9	3	4	8
	s ₃	9	3	7	5	12	0	8	4

* The procedures were applied to standardised variables.

Table 5.14 Frequency of correct and incorrect classifications when
 $r \leq 0.3$ and $p = 3$.

Number of groups (g)	Method of classification								
	LRT*		LRT (RMI)		CA*		CA (RMI)		
	C	N	C	N	C	N	C	N	
g = 1	7	5	2	10	-		-		
g = 2	s ₁	10	2	6	6	11	1	6	6
	s ₂	8	4	7	5	10	2	4	8
g = 3	s ₁	3	3	3	3	9	3	5	7
	s ₂	2	4	2	4	6	6	6	6
	s ₃	4	2	3	3	8	4	7	5
g = 4 ⁺	s ₁	-	-	-	-	11	1	4	8
	s ₂	-	-	-	-	11	1	6	6
	s ₃	-	-	-	-	8	4	6	6

- * The procedures were applied to standardised variables.
- + Results of LRT were not obtained due to limitations of computer memory and time. Subsequent analyses which make comparisons between methods (LRT & CA) will leave out the corresponding results of CA.

There were no major differences in performance of procedures when $p = 2$ or $p = 3$. Therefore the following discussion uses the combined results. Different groups sizes gave the following frequencies:

Table 5.14a Total frequencies of correct (C) and incorrect (N) classifications for different levels of s .

	s_1	s_2	s_3	Total
C	184	134	103	421
N	168	118	53	239
	252	252	156	660

73%, 53% and 66% correct determinations were obtained for s_1 , s_2 and s_3 , indicating significantly different recovery rates ($\chi^2_2 = 21.9$, $p < 0.001$). Like in the case of correlated variables, this effect was consistent among the two procedures considered, with either the RMI index or the standardised variables.

Major differences occurred in methods as indicated in the Table 5.14b ($\chi^2_3 = 38.8$, $p < 0.001$).

Table 5.14b Frequencies of correct and incorrect determinations for different method-index combinations.

	Method of determination				Total
	CA	CA (RMI)	LRT	LRT (RMI)	
Correct	126	87	98	75	386
Incorrect	30	69	40	74	213
Total	156	156	138	149	599

It would appear CA and LRT, using RMI, gave similar results (56% and 50% correct decisions, respectively), while a significant difference between CA and LRT (regardless of index) was indicated ($\chi^2_1 = 4.2$, $p = 0.04$). Therefore, CA gave significantly more correct determinations (78%) than LRT (60%). Apparently, standardised forms of the variable, which recovered 76% of the original groupings, retained more structure than RMI (53%) could reveal. The underlying distribution may have contributed to this difference.

Effect of intercorrelations among variables

The overall recovery rates with different correlations among variables are summarised in Table 5.15.

Table 5.15 Effect of intercorrelations among variables on performance of procedures.

	Value of correlation coefficient		
	$r < 0.3$	$0.5 < r < 0.7$	$r > 0.8$
Correct	386	150	211
Incorrect	213	106	45
	599	256	256

That is, 64%, 58% and 82% recovery rates for $r < 0.3$, $0.5 < 0.7$ and $r > 0.8$, respectively ($\chi^2_2 = 37.8$, $p < 0.001$). This suggests that the best rate of recovery in multivariate data was when the variables under study were highly correlated.

5.4.3 Discussion of results from the simulations

A major point derived from this study is that the performance of various procedures was affected by the size of the groups (underlying structure of the data), without any differential advantage for any particular procedure. That is, the effect was relatively similar among the procedures. Equally sized groups were easier to detect (recover)

than when a wide range of group sizes existed. This was true for both univariate and multivariate determinations.

A similar observation was reported by Marriott (1971), Scott and Symons (1971), Sen and Srivastava (1975) and Milligan and Cooper (1985), among others in relation to other uses of the procedures.

Our results seem to suggest that groups among highly correlated variables are easier to distinguish than those existing among uncorrelated variables. And in this case, using PCs to effect classifications gave better results than using the RMI. Standardised variables also gave better results than RMI in uncorrelated cases. However, RMI gave best results for highly correlated variables. Its use has been associated with the assumption of independent and equally important variables. On average, its use recovered 52%, 49% and 74% cases when correlations among variables were approximately 0, 0.6 and 0.9, respectively.

No differences between the groups recovery powers of CA and LRT were indicated for correlated variables. However, CA gave more correct determinations for uncorrelated variables than LRT.

In general, recovery rate decreased with increasing number of groups ($g > 1$). However, when $g = 1$ the LRT and NS were the only reliable procedures. The other factors - sample size (n) and number of variables (p) - showed no major effects on performance of procedures.

There is probably no direct way of quantifying ease of use of these procedures, but it must be noted that use of LRT involves heavy computations that required large amounts of (or imposed excessive demands on) time and computer memory, especially when large values of n were involved. However, the present surge in computer technology is likely to result in very powerful micro-chips and micro-processors which, in addition to more efficient software, could remove this computing drudgery. Moreover, its efficiency for the 1-group model is worth the effort, especially if grouping is doubtful.

CHAPTER 6

GENERAL DISCUSSION AND CONCLUSIONS

The primary approach to selection, in our study, required the development of classification procedures that would maximise differences in resistance levels among the groups while minimising the differences within the selected group.

In Chapter 3, the procedures which could achieve these goals were developed. Three categories of procedures were considered. For single-trait selection, one procedure considered is a graphical method in which observed data was plotted against its expected values under the null hypothesis when certain distributional assumptions hold. Visual inspection of the resulting normal score plots suggested possible partitions for the genotypes. The second procedure used cluster analysis techniques, while the third used a hypothesis testing approach to identify optimum classification models using appropriately selected test statistics. The multivariate approach focussed on intercorrelations among resistance measures. PCA generated indices (PCs) whose potential use for classification was compared with some commonly used indices, including RSI, SHI and Dobie's index of susceptibility (I).

In Chapter 5, the practical application of the procedures was investigated using two categories of real data described in Chapter 4.

Further investigations and comparison of the procedures were made through simulation studies. This involved evaluation of the procedures using sets of data simulated with known groupings and determination of the ability of each procedure to detect the groupings.

In this Chapter, we provide a general discussion of the results and draw plausible conclusions from the study. Further investigations will also be suggested.

6.1 DISTRIBUTIONAL ASSUMPTIONS

The results of the studies for various resistance measures demonstrated the value of studying the distribution of the data prior to other analyses. The commonly assumed distributions (for example, the normal distribution for B-LF and C-LF) were found not to hold in some situations. Transformations based on theoretical assumptions could as well cause more errors than not transforming at all, while empirical determinations would suggest appropriate ones, if required.

6.2 ANALYSES OF VARIANCE

The ANOVA provided a general overview of variability among genotypes, and perhaps the best estimate (the residual mean square) of the background variation in the experiment. The pooled standard error played a major role in the graphical method. However, the other classification procedures did not make much use of the results.

6.3 CLASSIFICATION PROCEDURES

6.3.1 Single-trait

6.3.1.1 *Normal score plots*

The results amply demonstrated the value of the graphical method in classification for resistance. The normal score plots, in most cases, were able to provide an indication of the optimum number of groups in the data. However, the cutpoints (class limits) determined from visual inspection could be imprecise, having no measure of precision themselves. However, the plots gave satisfactory and useful results when used with other procedures.

One advantage of this method over the others is that it utilises the variability in the data (pooled s.e.) for the reference line. In fact if variability within groups were the same, then the points, representing entries, would all lie parallel to the line.

While other procedures were greatly affected by the relative group sizes, NS plots were not. The groups would show up without much interference as to whether some groups were large, even when groups of 1 entry were involved. Further investigations would be required to extensively quantify the usefulness of this procedure.

One limitation to the method, is that it would appear to be sensitive to distributional assumptions. The method, and indeed the others, assumes normality and equal variance. Failure of the assumptions seems to have caused dramatic effects on the results.

This was demonstrated by the real problem data, especially variable DH, where parallelism of data points and the computed line was hard to achieve.

Otherwise, the method was found easy, simple and convenient to use. Its previous application, as indicated by Plackett (1971) and Perry (1986), depended on use of normal probability paper and tables as those of Fisher and Yates (1963). These are quite restrictive; for example, the tables only allow for a sample size of not more than 51. An approximation adopted in this study allows us to handle larger numbers of entries as is often the case with crop breeding trials.

6.3.1.2 *Tests statistic*

The LRT statistics plots, originally, in this study, meant for locating the optimum cutpoint in the 2-groups model, proved more useful as it provided information on the actual number of groups in the data. In some cases, when distinct groups existed, the plot of the test statistic values would suffice for the classification. However, as with NS plots, the location of cutpoints tended to be fairly imprecise, especially when large sets of data and non-distinct groups were involved. A 2-groups model (e.g. Fig 5.10b) was characterised by a sharp rise and fall off, almost at the same rate, at the cutpoint, while more groups were indicated by changes in the slope at various positions, or angular shapes at the respective cutpoints (e.g. Fig 5.6b). This is a rather interesting result, because the primary use of

these plots was to locate the optimal cutpoint for the 2-groups model. It appears the procedure has an inherent (implicit) detection device, for if the plot suggested 3 groups, say, the same number was eventually confirmed by the other procedures. In some cases the entire classification would be indicated by the plot.

The ability of LRT (using g^2w) to detect a 1-group model can only be compared to the normal score plot procedure, which as noted above, may not be as precise.

The QBT statistic has proved not very useful in this study. It has no measure of adequacy of the 2-group model. Previous use of the statistic was in locating a single change point in various processes. More studies would be required to find out the conditions required for the statistic to be optimally useful.

One other point to note, which indeed generally applies to other procedures, is that this procedure has a tendency to place the cutpoint around half-way between entries it is dividing. Apparently the method may be useful in subdividing entries into groups which are close to equal in size. This may only apply when interest is in just reducing the number of entries to about half the entire set. In such situations its use may provide a better division than just ranking and selecting the first half of the entries. This may require further investigation.

6.3.1.3 *Cluster analysis*

The procedure tended to give results similar to those of the LRT procedure. One major limitation is its inability to consider the 1-group model (null hypothesis) in its classification process. To this effect, Arnold (1979) indicated that the procedure will proceed to group observations even when there are no meaningful groupings. However, inconsistency in the two statistics considered in this study seemed to suggest the 1-group model. But, the same 'symptom' could also suggest other alternative models from those considered (e.g. $g > 10$).

The method was easy to use, and fairly robust to distributional assumptions.

6.3.2 *Multi-trait selections*

6.3.2.1 *Procedures*

The points made on LRT and CA with respect to the univariate procedures apply to the multivariate procedures as well.

Results indicated that the appropriate index or procedure to adopt depended heavily on correlations between various measures.

Classification by LRT relied heavily on the data arrangement. CA does too, to some extent, but the iterative nature of the algorithm(s) allows for swapping and transfer of entries until convergence is achieved. This may be one of the reasons why, when considering uncorrelated variables where there was no natural basis for ordering data, the LRT method was not as good as CA.

6.3.2.2 *Indices*

Use of PCs as the basic variables for classification gave promising results. Results of simulations indicated that for correlated variables, using PCs recovered more classifications than when RMI was used.

For data sets examined in the study, it was observed that resistance measures could be highly correlated. As a result of these relationships, substantial variability could be extracted by the PCs to the effect that a few of them were enough to produce the necessary classifications. This is an efficient way of utilising much of the information about the entries. RMI, on the other hand, makes no use of the underlying intercorrelations among variables and actual variability among the entries.

The high correlations observed among indices (PC1, RMI and SHI) in data set A resulted into similar classifications and selection of entries that were largely resistant to all the damage measure. Lack of correlations among variables (e.g. data set B), on the other hand resulted into large differences in the classifications based on different indices. In this case, use of RMI gave more meaningful results than the standardised variables and SHI.

6.4 **CONCLUSIONS**

The fundamental point that came out of the simulations is that the best power of a given procedure was realised when group sizes

were about equal. An examination of wrongly classified results showed that there was a tendency to give solutions with about equally sized groups. This was most prevalent with the use of QBT. This situation is more likely to arise when the groups are not distinct/disparate. To this effect, Scott and Symons (1971) suggestion that some care should be taken in interpreting an analysis if the resulting clusters were about the same size, is supported by our results.

With the 2-group model, as was the case for data D-T, for example, all procedures, including QBT tended to converge to the same cutpoint, while they differed if the 2-group model was inadequate.

While univariate procedures were sensitive to the underlying distribution of the classification variable, multivariate procedures were sensitive to correlations among the classification variables. Univariate-based classifications may therefore result into misleading selections when such correlations are not considered. With highly correlated variables, use of PCs, RMI and SHI resulted into comparable selections. With uncorrelated variables, the best results were obtained using RMI.

Results obtained using LRT and CA were very similar. However, LRT seems to be more precise, while CA is fairly robust to distributional assumptions.

Much as the procedures proposed in this study were meant to be alternatives for solving the classification problem, they would better serve as complimentary procedures. As noted from results of the simulations, each would serve as a verification check for the others. It is therefore recommended that, where possible, more than one technique be applied before deciding on the optimal solution. Since the procedures were distributional dependent, the empirical determination of the underlying distribution of the classification variables should be a procedural first step to classification.

Finally, the results obtained from the real data have demonstrated that most conventional methods employed by researchers in selecting genotypes could lead to enormous loss of information and vital resources.

6.5 RELEVANCE OF RESULTS FROM THE STUDY

Use of the classification procedures that utilise the inherent variability among entries, instead of rankings (even when the difference numerically observed may not be real), is expected to reduce subjectivity and thus improve objectivity in the selection process. Maximising the variability between selected group and the other groups is likely to ensure maximal and faster gain/progress from selection since the primary objective, in the long run, is to improve the level of resistance.

Use of these procedures will reduce dependence on theoretical values (e.g. fixed selection intensity, etc.) The selection intensity may now be determined empirically. Perhaps this will reduce the gap between the predicted and actual gains from selection.

Many previous selections have relied on standard varieties for selection of new ones. As noted earlier (Section 1.5), this may be a limiting factor to selection under varying conditions. The procedures included herein will reduce dependence on standard varieties.

Use of PCA will provide an alternative to construction of indices for use in the classification process. Its use is likely to solve the difficulty in choice of, and dangers of using irrational, weights for various resistance measures. This is likely to increase the effectiveness of the classifications and thus the selections based on them.

The results of this study, though developed in the host plant resistance context, are indeed applicable to other fields.

6.6 SUGGESTIONS FOR FURTHER WORK

In this study we considered resistance measures to classify entries. Crop yield, which is the ultimate goal, is of paramount importance especially at later stages of selection programmes. In addition to classifying genotypes for yield potential, the investigator will wish to select genotypes with some tolerance to the pest of interest. Presently, there are no effective criteria for quantifying

tolerance (Alexander & Bramel-Cox, 1991; Ajala, 1994). One approach is to examine a function of loss in yield and level of damage as a possible measure of the level of tolerance. This would facilitate classification of the genotypes into tolerance groups and the subsequent selection for tolerance. Future study in this direction is necessary. Further investigations should also consider the effect of environmental factors.

A combination the potential yield, infestation and damage parameters as well as yield loss to come up with an index or indices would be very appropriate. This could be effected by canonical correlation analysis where yield parameters make one group and the infestation and damage parameters the other. This approach may result into better classification indices.

The comparison of various classification procedures may be carried out by constructing an index that will determine how 'good' a classification obtained by a given method is. This could take the form of discriminant analysis function which may be used to determine the probability of misclassification. This approach will require some studies to develop appropriate functions. Comparisons based on actual gains from repeated selections made using different procedures would further improve the validity of results obtained in this study.

Use of power functions would be very useful in establishing appropriate data transformations for resistance measures. Future work should include this approach.

REFERENCES

- Agrawal, B.L. and Taneja, S.L. (1989). Breeding for resistance to stem borer (Chilo partellus Swinhoe) in sorghum. *ICRISAT conference paper* no. CP 491, 159-168.
- Agrawal, B.L., Taneja, S.L., House, L.R. and K. Leuschner. (1990). Breeding for resistance to Chilo partellus Swinhoe in sorghum. *Insect Science and its Applications* 11 (4/5), 671-682.
- Ajala, S.O. (1993). Population cross diallel among maize genotypes with varying levels of resistance to the spotted stem-borer Chilo partellus (Swinhoe). *Maydica* 38, 39-45.
- Ajala, S.O. (1994). Maize (Zea mays L.) stem borer (Chilo partellus Swinhoe) infestation/damage and plant resistance. *Maydica* 39, 203-205.
- Ajala, S.O. and Saxena, K.N. (1994). Interrelationships among Chilo partellus (Swinhoe) damage parameters and their contribution to grain yield reduction in maize (Zea mays L.). *Applied Entomology and Zoology* 29 (4), 469-474.
- Ajala, S.O., Saxena, K.N. and Chiliswa, P. (1995). Selection in maize (Zea mays L.) for resistance to the spotted stem borer (Chilo partellus (Swinhoe)). *Maydica* (In press).
- Alexander, H.M. and Bramel-Cox, P.J. (1991). Sustainability of genetic resistance. In *Plant breeding and sustainable agriculture: considerations for objectives and methods*. Ed. D.A. Sleper, T.C. Barker, P.J. Bramel-Cox. CSSA Special publication no. 18, pp 11-27.

- Alghali, A.M. (1983). Relative susceptibility of some rice varieties to stalk-eyed fly Diopsis thoracica West. *Insect Science and its Application* 4, 135-140.
- Amaro, A. (1994). A classification methodology. A case study. *Proceedings of XVIIth International Biometric Conference, vol 2: Contributed papers*, pp 242. Hamilton, Canada.
- Ampofo, J.K.O., Saxena, K.N., Kibuka, J.G. and Nyangiri, E.M. (1986). Evaluation of some maize cultivars for resistance to the stem borer Chilo partellus (Swinhoe) in Western Kenya. *Maydica* 31, 379-389.
- Anscombe, F.J. (1949). The statistical analysis of insect counts based on the negative binomial distribution. *Biometrics* 5, 165-175.
- Anscombe, F.J. (1950). Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika* 37, 358-382.
- Arnold, S.J. (1979). A test for clusters. *Journal of Marketing Research* 16, 545-55.
- Ashraf, M. and Fatima, B. (1990). Breeding for resistance to Chilo spp. in Sugar-cane. *Insect Science and its Applications* 11 (4/5), 683-687.
- Barrow, M.R. (1985). The effect of different maize genotypes on the maize stalk-borer, Busseola fusca (Fuller) (Lepidoptera: Noctuidae), feeding in whorl tissue. *Journal of Entomological Society of South Africa* 48 (1), 113-119.

- Barrow, M.R. (1987). The effect of first generation maize stalkborer Busseola fusca (Fuller) (Lepidoptera: Noctuidae), on yield of different maize genotypes. *Journal of Entomological Society of South Africa* 50 (2), 291-298.
- Beall, G. (1940). The fit and significance of contagious distribution when applied to observations on larval insects. *Ecology* 21, 460-471.
- Bell, R.L. and Stuart, L.C. (1990). Resistance in East European Pyrus Germplasm to Pear Psylla Nymphal Feeding. *HortScience* 25 (7), 789-791.
- Bhumannavar, B.S., Singh, S.P. and Sulladmath, V.V. (1989). Field evaluation of citrus germplasm for resistance to the black aphid, Toxoptera aurantii (Boy.) under tropical humid south Indian conditions. *Insect Science and its Applications* 10 (1), 81-88.
- Binelli, G., Gianfranceschi, L., Angelini, A., Cammusi, A. and Pe, M.E. (1992). Multivariate analysis for the evaluation of stalk-rot resistance in twenty seven maize inbreds. *Maydica* 37, 339-342.
- Blashfield, R.K. and Aldenderfer, M.S. (1978). The literature on cluster analysis. *Multivariate Behavioral Research* 13, 271-295.
- Bliss, C.I. and Fisher, R.A. (1953). Fitting the negative binomial distribution to biological data and note on the efficient fitting of the negative binomial. *Biometrics* 9, 176-200.

- Brown, D. and Rothery, P. (1993). *Models in Biology: Mathematics, Statistics and Computing*. John Wiley & Sons, New York. 688 pp.
- Calinski, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics* 3, 1-27.
- Cate, R.B. Jr and Nelson, L.A. (1971). A simple statistical procedure for partitioning soil test correlation data into two classes. *Soil Science Society of America Proceedings* 35, 658-660.
- Chatfield, C. and Collins, A.J. (1980). *Introduction to multivariate analysis*. Chapman & Hall, London. 246 pp.
- Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subject to changes in time. *Annals of Mathematical Statistics* 35, 999-1028.
- Chheda, H.R. and Fatokun, C.A. (1982). Numerical analysis of variation patterns in Okra (Abelmoschus esculatus (L.) Moench). *Bot. Gaz.* 143 (2), 253-261.
- Chiang, H.C. and Hodson, A.C. (1959). Distribution of the first generation egg masses of the European corn borer (Pyrausta nubilalis (Hbn.)) in North Central states. *Minnesota Agricultural Experiment station Technical Bulletin* 229. 63 pp.
- Compton, W.A. and Lonquist, J.H. (1981). A Multiplicative Selection Index Applied to Four Cycles of Full-Sib Recurrent Selection in Maize. *Crop Science* 22, 981-983.

- Cornelius, P.L., Sanford, D.A.V. and Seyedsadr, M.S. (1993). Clustering Cultivars into Groups without Rank-Change Interactions. *Crop Science* 33, 1193-1200.
- Crosbie, T.M., Mock, J.J. and Smith, O.S. (1980). Comparison of gains predicted by several selection methods for cold tolerance traits. *Crop Science* 20, 649-655.
- Crump, S.L. (1951). The present status of variance component analysis. *Biometrics* 7 (1), 1-16.
- Curnow, R.N. (1961). Optimal Programmes for Varietal Selection. *Journal of Royal Statistical Society B* 23, 282-318.
- Dagneillie, P. and Merckx, A. (1991). Using generalised distances in classification of groups. *Biometrics Journal* 33 (6), 683-695.
- Davis, F.M. (1985). Entomological techniques and methodologies used in Research Programmes on Plant Resistance to Insects. *Insect Science and its Application* 6 (3), 391-400.
- Davis, F.M., Williams, W.P., Mihm, J.A., Barry, B.D., Overman, J.L., Wiseman, B.R. and Riley, T.J. (1988). *Resistance to Multiple Lepidopterous Species in Tropical Derived Corn Germplasm. Mississippi Agricultural & Forestry Experiment Station. Technical Bulletin* 157.
- Dent, D. (1992). *Insect Pest Management*. CAB International, U.K. 604 pp.

- Dhooria, M.S. (1981). Differential response of citrus species/varieties to infestation of Eutetranychus orientalis. *Indian Journal of Acarology* 7 (1), 11-14.
- Digby, P.G.N. and Kempton, R.A. (1987). *Multivariate analysis of ecological communities*. Chapman and Hall, London.
- Dobie, P. (1974). The laboratory assessment of the inherent susceptibility of maize varieties to post-harvest infestation by Sitophilus zeamais Motsch. (Coleoptera, Curculionidae). *Journal of Stored Product Research* 10, 183-197.
- Dobson, A. (1983). *An introduction to statistical modelling*. Chapman and Hall, London.
- Elliott, J.M. (1977). Some methods for the statistical analysis of Benthic Invertebrates. *Freshwater Biological Association, Scientific publication* no. 25.
- Ellis, P.R., Pink, D.A.C. and Ramsey, A.D. (1994). Inheritance of resistance to lettuce root aphid in the lettuce cultivars 'Avoncrisp' and 'Lakeland'. *Annals of Applied Biology* 124 (1), 141-151.
- Elston, R.C. (1963). A weight-free index for the purpose of ranking or selection with respect to several traits at a time. *Biometrics* 19, 85-97.
- Everitt, B.S. (1994). An introduction to finite mixtures. *Proceedings of the XVIIth International Biometric Conference, vol 1: Invited papers* 139-150. Hamilton, Canada.

- Falconer, D.S. (1989). *Introduction to quantitative genetics*. Longman Group Ltd. 3rd edn. 483 pp.
- Fisher, R.A. and Yates, F. (1963). *Statistical tables for biological, agricultural and medical research*. Longman.
- Flury, B.N. (1984). Common principal components in k groups. *Journal of the American Statistical Association* 79, 892-898.
- Friedman, H.P. and Rubin, J. (1967). On some invariant criteria for grouping data". *Journal of the American Statistical Association* 62, 1159-78.
- Gail, M.H. and Green, S.B. (1976). A generalization of the one-sided two-sample Kolmogorov-Smirnov statistic for evaluating diagnostic tests. *Biometrics* 32, 561-570.
- Gardner, L.A. Jr. (1969). On detecting Changes in the mean of normal variates. *The Annals of Mathematical Statistics* 40 (1), 116-126.
- Giga, D.P. and Mazarura, V.W. (1991). Levels of resistance to maize weevil, Sitophilus zeamais (Motsch.) in exotic, local open-pollinated and hybrid maize germplasm. *Insect Science and its Application* 12, 159-169.
- Gomez, K.A. and Gomez, A.A. (1984). *Statistical procedures for agricultural research*. John Wiley & Sons. New York.
- Gordon, A.D. (1973). Classification in the presence of constraints. *Biometrics* 29 (4), 821-828.

- Gower, 1967. A comparison of some methods of cluster analysis. *Biometrics* 23, 623-637.
- Green, P.E. (1978). *Mathematical tools for Applied Multivariate Analysis. Student edition*. Academic Press, Inc., London.
- Gupta, S.C. (1992). *Fundamentals of statistics*. Himalaya Publishing House, New Dheli. 503 pp
- Hallauer, A.R. and Miranda, J.B. (1988). *Quantitative Genetics in Maize Breeding*. Second edition. Iowa State University Press, Ames.
- Harcourt, D.G. (1963). Population Dynamics of Leptinotarsa decemlineata (Say) in Eastern Ontario. 1. Spatial Pattern and Transformation of field counts. *Canadian Entomology* 95, 813-820.
- Harris, D.L. (1964). Expected and predicted progress from index selection involving estimates of population parameters. *Biometrics* 20 (1), 46-72.
- Hazel, L.N. (1943). The genetic basis for constructing selection indexes. *Genetics* 28, 476-490.
- Horber, E. (1972). Plant resistance to insects. *Agricultural Science Review* 18, 1-10.

- Horber, E. (1980). Methods to detect and evaluate resistance in maize to grain insects in the field and in storage. In *Toward insect resistant maize for the third world: Proceedings of the International symposium on methodologies for developing host pest resistance to maize insects, D.F. CIMMYT*. pp 140-150.
- Horn, D.J. (1988). *Ecological Approach to Pest Management*. Elsevier Applied Science Publishers Ltd., London. pp 186-194.
- James, B., James, K.L. and Siegmund, D. (1987). Tests for a Change-point. *Biometrika* 74 (1), 71-83.
- Kamali, K., Dickie, F.F. and Guthrie, W.D. (1989). Resistance-Susceptibility of Maize Genotypes to Artificial Infestations by Two spotted Spider Mites. *Crop Science* 29, 936-938.
- Kaufman, L. and Rousseeuw, P.J. (1990). *Finding groups in data. An introduction to cluster analysis*. John Wiley & Sons, Inc. New York. 342 pp.
- Kogan, M. and Ortman, E.E. (1978). Antixenosis - a new term proposed to define Painter's 'non-preference' modality of resistance. *Bulletin of Entomological Society of America* 24 (2), 175-6.
- Krzanowski, W.J. (1979). Between-groups comparison of principal components. *Journal of the American Statistical Association* 74, 703-707.

- Lausen, B. and Schumacher, M. (1992). Maximally Selected Rank Statistics. *Biometrics* 48, 73-85.
- Lynch, R.E., Robinson, J.F. and Berry, E.C. (1980). European Corn Borer: yield losses and damage resulting from a simulated natural infestation. *Journal of Economical Entomology* 73, 141-144.
- Manly, B.F.J. (1986). *Multivariate Statistical Methods A Primer*. Chapman and Hall, London.
- Mansour, F. and Bar-Zur, A. (1992). Resistance of maize inbred lines to the carmine spider mite, Tetranychus cinnabarius (Acari: Tetranychidae). *Maydica* 37, 343-345.
- Mansur, L.M., Carriquiry, A.L. and Rao-Arelli, A.P. (1993). Generation mean analysis of resistance to race 3 of soybean cyst nematode. *Crop Science* 33 (6), 1249-1252.
- Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979). *Multivariate analysis*. Academic Press, London.
- Marriott, F.H.C. (1970). A problem of optimum stratification. *Biometrics* 26 (4), 845-847.
- Marriott, F.H. (1971). Practical problems in a method of cluster analysis. *Biometrics* 27, 501-514.
- Marriott, F.H.C. (1974). *The interpretation of multiple observations*. Academic Press, London.

- McGuire, J.U., T.A. Bridley and Bancroft, T.A. (1957). The distribution of European Corn Borer larvae Pyrausta nubilalis (Hbn.) in field corn. *Biometrics* 13, 65-78.
- Mendell, N.R., Thode, H.C. Jr. and Finch, S.J. (1991). The likelihood ratio test for the two-component normal mixture problem: Power and sample size analysis. *Biometrics* 47 (3), 1143-1148.
- Mendell, N.R., Finch, S.J. and H.C. Thode, H.C Jr. (1993). The consultant's forum. Where is the likelihood ratio test powerful for detecting two component normal mixtures? *Biometrics* 49 (3), 907-915.
- Mendes, A.P., Farina, M.P.W., Channon, P. and Smith, M. (1985). A field evaluation of the differential tolerance to soil acidity of forty-eight South African maize cultivars. *South African Journal of Plant Soil* 2 (4), 215-220.
- Milligan G.W. and Cooper, M.C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50, 159-179.
- Mudholkar, G.S. and McDermott, M.P. (1989). A class of tests for equality of ordered means. *Biometrika* 76 (1), 161-168.
- Mulamba, N.N. and Mock, J.J. (1978). Improvement of yield potential of the Eto Blanco Maize (*Zea mays* L.) population by breeding for plant traits. *Egyptian J. Genetics and Cytology* 7, 40-51.

Nabasiye, M., Odulaja, A. and Ajala, S.O. (1993a). Critical resistance levels (CRL): thoughts and views. *African Association of Insect Science 10th conference: Pre-conference papers proceedings*. Mombasa, Kenya. pp 87-91.

Nabasiye, M., Odulaja, A. and Luboobi, L.S. (1993b). Application of cutpoint statistics to crops varietal selection. In *Biometry for sustainable food productio. Proceedings of the 3rd meeting of the East, Central & Southern African Biometric Network*. Eds. Luboobi, L.S., Ssewanyana, S. and Nabugoomu, F. Uganda Biometric Society, Uganda. pp 90-98.

Neave, H.R. (1981). *Elementary statistical tables for all users of statistical techniques*. Chapman and Hall, New York. 49 pp.

Odulaja, A. and Nokoe, S. (1993). A maxmin-minimax approach for classifying crop varieties into resistance groups based on yield potential and loss. *International Journal of Pest Management* 39 (1), 64-66.

Oghiake, S. and Odulaja, A. (1993a). A multivariate analysis of growth and development parameters of the legume pod borer, Maruca testulalis on variably resistant cowpea cultivars. *Entomol. exp. appl.* 66, 275-282.

Oghiake, S. and Odulaja, A. (1993b). Classification of cowpea cultivars for field resistance to the legume pod borer, Maruca testulalis in Nigeria using cluster analysis. *Annals of Applied Biology* 122, 69-77.

- Painter, R.H. (1951). *Insect Resistance in Crop Plants*. MacMillan, New York. 520 pp.
- Painter, R.H. (1958). Resistance of plants to insects. *Annual Review of Entomology* 3, 267-290.
- Panda, N. and Heinrichs, E.A. (1983). Levels of tolerance and antibiosis in rice varieties having moderate resistance to the brown planthopper, Nilaparvata lugens. *Environmental Entomology* 12, 1204-1204.
- Pathak, M.D. and Saxena, R.C. (1979). Insect resistance. pp 270-92. In *Plant Breeding Perspectives*. Eds J. Snee and A.J.T. Hendriksen. Centre for Agricultural Publishing Documentation, Wageningen. 460 pp.
- Perry, J.N. (1986). Multiple-comparison procedures: a dissenting view. *Journal of Economical Entomology* 79, 1149-1155.
- Pesek, J. and Baker, R.J (1969). Desired improvement in relation to selection indices. *Canadian Journal of Plant Science* 49, 773-781.
- Pla, S. (1991) Determining stratum boundaries with multivariate real data. *Biometrics* 47(4), 1409-1422.
- Plackett, R.L. (1971). Contribution to discussion on paper by O'Niell and Wetheril. *Journal of Royal Statistical Society, B* 36, 242-244.

- Quenouille, M.H. (1949). A relation between the Logarithmic, Poisson and Negative Binomial series. *Biometrics* 5, 162-164.
- Rana, B.S. and Murty, B.R. (1971). Genetic analysis of resistance to stem borer in sorghum. *Indian Journal of Genetics & Plant Breeding* 31 (3), 521-529.
- Rao, C.R. (1994). Statistics: an essential technology in environmental research and management. *Environmental and Ecological Statistics* 1, 7-19.
- Reddy, S.K.V. (1985a). Relative susceptibility and resistance of some sorghum lines to stem-bores in Western Kenya. *Insect Science and its Application* 6 (3), 401-404.
- Reddy, S.K.V. (1985b). Screening maize genotypes for multiple resistance to stem-borers. *Insect Science and its Application* 6 (3), 405-408.
- Sanford, D.A.V., Pfeiffer, T.W. and Cornelius, P.L. (1993). Selection Index Based on Genetic Correlations Among Environments. *Crop Science* 33, 1244-1248.
- Sarle, W.S. (1983). Cubic clustering criterion. *SAS technical report A-108*. Cary, NC: SAS Institute Inc. pp 52.
- SAS Institute Inc. *SAS Procedures Guide for Personal Computers, Version 6 Edition*. Cary, NC: SAS Institute Inc., 1985. 373 pp.

- SAS Institute Inc. *SAS/STAT™ Guide for Personal Computers, Version 6 Edition*. Cary, NC:SAS Institute Inc., 1987. 1028 pp.
- Saxena, K.N. (1989). Sorghum resistance to stem borers. In *ICIPE Seventeenth Annual Report*. Nairobi, Kenya. pp 8-9.
- Scott, A.J. and Symons, M.J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics* 27, 387-397.
- Sen, A. and Srivastava, M.S. (1975). On tests for detecting change in mean. *Annals of Statistics* 3 (1), 98-108.
- Shapiro, S.S. and Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika* 52, 591-611.
- Shapiro, S.S., Wilk, M.B. and Chen, H.J. (1968). A comparative study of various tests for normality. *Journal of American Statistical Association* 63, 1343-1372.
- Singh, B.U. and Rana, B.S. (1989). Varietal resistance in sorghum to spotted stem borer, Chilo partellus (Swinhoe). *Insect Science and its Applications* 10 (1), 3-27.
- Singh, I.D. and Allen, D.J. (1980). Pests, diseases, and protection in cowpeas, pp. 419-43. In *Advances in Legume Science*. Eds R. Summerfield and H. Bunting. Ministry of Agr. For. Fish. Her Majesty's Govt. and Kew Gardens. 667 pp.
- Smith, H.F. (1936). A discriminant function for plant selection. *Annals of Eugenics* 7, 240-250.

- Smith M.F. (1990). A multivariate approach to a specific problem of grouping maize cultivars. *South African Journal of Plant Soil* 7 (3), 167-171.
- Smith, O.S., Hallauer, A.R. and Russel, W.A. (1981). Use of index selection in recurrent selection programs in maize. *Euphytica* 30, 611-618.
- Smith, C.M., Khan, Z.R. and Pathak, M.D. (1994). *Techniques for evaluating insect resistance in crop plants*. CRP Press, Inc. (Lewis publishers) - Florida.
- Sneath, P.H.A. and Sokal, R.R. (1973). *Numerical Taxonomy*. W.H. Freeman & Co., San Francisco.
- Snedecor, G.W. and W.G. Cochran (1980). *Statistical methods*. 7th edition. The Iowa State University Press, USA. 507 pp.
- Sokal, R.R. and Rolf, F.J. (1981). *Biometry. The principles and practice of statistics in biological research*. 2nd edition. W.H. Freeman and company, New York. 859 pp.
- Southwood, T.R.E. (1978). *Ecological methods with particular reference to the study of insect populations*. 2nd edition. 524 pp.
- Souza, E. and Sorrells, M.E. (1991). Relationships among 70 North American oat germplasm: I. Cluster analysis using quantitative characters. *Crop Science* 31, 599-605.

- Subandi, W., Compton, A. and Emping, L.T. (1973). Comparisons of efficiencies of selection indices for three traits in two variety crosses of corn. *Crop Science* 13, 184-186.
- Taylor, L.R. (1961). Aggregation, variance, and the mean. *Nature* 189, 732-735.
- Taylor, L.R. (1984). Assessing and interpreting the spatial distributions of insect populations. *Annual Review of Entomology* 29, 321-57.
- Thorne, C.R., Smith, M.E. and Mihm, J.M. (1994). Yield reduction in a maize diallel under infestation with South Western Cornborer. *Crop Science* 34 (6), 1431-1435.
- Widstrom, N.W. (1967). An Evaluation of Methods for Measuring Corn Earworm Injury. *Journal of Economic Entomology* 60 (3), 792-794.
- Widstrom, N.W., Redlinger, L.M. and Wiser, W.J. (1971). Appraisal of Methods for Measuring Corn Kernel Resistance to Sitophilus zeamais. *Journal of Economic Entomology* 65 (3), 790-792.
- Widstrom, N.W. (1974). Selection Indexes for Resistance to Corn Earworm Based on Realized Gains in Corn. *Crop Science* 14, 673-675.
- Williams, J.S. (1962). The evaluation of a selection index. *Biometrics* 18, 375-393.

Worsley, K.J. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of American Statistical Association* 74, 365-367.

Zar, J.H. *Biostatistical analysis*. (1984). 2nd edition. Prentice-Hall, Inc., New Jersey. 718 pp.

Zeinali-Khanghah, H., Dreen, D.E. and Shibles, (1993). Use of morphological, developmental and plant nitrogen traits in a selection scheme in soybean. *Crop Science* 33 (6), 1121-1127.

APPENDICES

Appendix A1 Generation of bivariate normal variables using SAS.

```
OPTIONS LS = 78 PS = 30 NODATE NONUMBER;

DATA A;

KEEP X Y ZX ZY;

SA = 1.8; SA2 = SA*SA; RX = 0.9; RX2 = RX*RX;

SB = 1.5; SB2 = SB*SB; RY = 0.6; RY2 = RY*RY;

N = 30 MX = 10; MY = 5; LINK GENERATE;

N = 60 MX = 30; MY = 10; LINK GENERATE;

N = 30 MX = 20; MY = 8; LINK GENERATE;

STOP;

GENERATE:

DO I = 1 TO N;

  X = RANNOR(1)*SA + MX;

  Y = RANNOR(1)*SB + MY;

  ZX = RANNOR(1)*SA2*(1-RX2) + MX + RX*SB*(Y-MY)/SA;

  ZY = RANNOR(1)*SB2*(1-RY2) + MY + RY*SA*(X-MX)/SB;

OUTPUT;

END;

RETURN;
```

Appendix A2 Iterative generation of groupings of n entries into 3 groups and computation of univariate LRT statistics.

```

10  CLS : CLEAR
20  REM sets generation for 3 groups
30  INPUT "No. of entries (varieties) "; EN
40  GR = 3: RR = 0: NN = 0
50  DIM Y(EN)
60  INPUT "Data file name "; DF$
70  OPEN "i", #2, DF$
80  INPUT "Output file name "; FL$
90  OPEN "o", #1, FL$
100 FOR I = 1 TO EN: INPUT #2, Y(I): SU = SU + Y(I):
    SST = SST + Y(I) ^ 2: NEXT I
110 CT = SU ^ 2 / EN: TOSS = SST - CT
120 PRINT "Total ss = "; TOSS: PRINT
130 PRINT "No. of entries = "; EN: PRINT
140 PRINT "No. of groups = "; GR: PRINT
150 PRINT "Total d.f. = "; EN - 1: PRINT
160 PRINT "Trt. d.f. = "; GR - 1: PRINT
170 PRINT "Error d.f. = "; EN - GR: PRINT
180 PRINT #1, "Total ss = "; TOSS: PRINT #1,
190 PRINT #1, "No. of entries = "; EN: PRINT #1,
200 PRINT #1, "No. of groups = "; GR: PRINT #1,
210 PRINT #1, "Total d.f. = "; EN - 1: PRINT #1,
220 PRINT #1, "Trt. d.f. = "; GR - 1: PRINT #1,
230 PRINT #1, "Error d.f. = "; EN - GR: PRINT #1,
240 PRINT "iter Grp1 Grp2 Grp3 ";
250 PRINT "trt ss", : PRINT "Error ss", : PRINT "F-Ratio", :
    PRINT "R-square"

```

```

260 PRINT #1, "iter Grp1 Grp2 Grp3 ";
270 PRINT #1, "trt ss", : PRINT #1, "Error ss", :
    PRINT #1, "F-Ratio", : PRINT #1, "R-square"
280 DD = EN - GR + 1
290 DIM SZ(DD)
300 FOR CL = 1 TO DD
310 SZ(CL) = EN + 2 - GR - CL
320 FOR SET = 1 TO SZ(CL): NN = NN + 1
330 GT1 = 0: GT2 = 0: GT3 = 0: TRSS = 0
340 GZ(1) = CL: GZ(3) = EN - SET - CL:
    GZ(2) = EN - GZ(1) - GZ(3)
350 FOR I = 1 TO GZ(1): GT1 = GT1 + Y(I): NEXT I
360 FOR I = GZ(1) + 1 TO GZ(1) + GZ(2):
    GT2 = GT2 + Y(I): NEXT I
370 FOR I = GZ(1) + GZ(2) + 1 TO EN:
    GT3 = GT3 + Y(I): NEXT I
380 TRSS = GT1 ^ 2 / GZ(1) + GT2 ^ 2 / GZ(2) + GT3 ^ 2 / GZ(3) -
    CT: ERSS = TOSS - TRSS
390 TRMS = TRSS / (GR - 1): ERMS = ERSS / (EN - GR)
400 PRINT USING "####"; NN; GZ(1); GZ(2); GZ(3); :
    PRINT TRSS, ERSS, TRMS / ERMS, 100 * TRSS / TOSS
410 PRINT #1, USING "#####"; NN; GZ(1); GZ(2); GZ(3); :
    PRINT #1, TRSS, ERSS, TRMS / ERMS, 100 * TRSS / TOSS
    NEXT SET
430 NEXT CL

```

Appendix A3 Iterative generation of groupings of n entries into 3 groups and computation of multivariate LRT statistics.

```

10  CLS : CLEAR
20  REM sets generation for 3 groups
30  INPUT "No. of entries (varieties) "; EN
40  INPUT "No. of variables (traits) "; TR
50  DIM Y(EN, TR)
60  INPUT "Data file name "; DF$
70  OPEN "i", #2, DF$
80  INPUT "Output file name "; FL$90      OPEN "o", #1, FL$
100 FOR I = 1 TO EN: INPUT #2, II: FOR J = 1 TO TR:
    INPUT #2, Y(I, J)
110  SU(J) = SU(J) + Y(I, J): SST(J) = SST(J) + Y(I, J) ^ 2
120  NEXT J: NEXT I
130  FOR J = 1 TO TR: ME(J) = SU(J) / EN: CT(J) = SU(J) ^ 2 /
    EN: TOSS(J) = SST(J) - CT(J): NEXT J
140  PRINT "iter Grp 1 Grp 2 Grp 3 ";
150  PRINT "det(T) det(W) det(w)/det(t) phi"
160  PRINT #1, "iter Grp 1 Grp 2 Grp 3 ";
170  PRINT #1, "det(T) det(W) det(w)/det(t) phi"
180  GR = 3: RR = 0: NN = 0
190  DD = EN - GR + 1
200  DIM SZ(DD)
210  FOR CL = 1 TO DD
220  SZ(CL) = EN + 2 - GR - CL
230  FOR SET = 1 TO SZ(CL): NN = NN + 1
240  GT1 = 0: GT2 = 0: GT3 = 0: TRSS = 0
250  FOR I = 1 TO GR: FOR J = 1 TO TR: GT(I, J) = 0:
    FOR K = 1 TO TR
260  T(I, J, K) = 0: W(I, J, K) = 0: TRC(J, K) = 0:
    WRC(J, K) = 0: NEXT K: NEXT J: NEXT I

```

```

270  GZ(1) = CL: GZ(3) = EN - SET - CL: GZ(2) = EN - GZ(1) - GZ(3)
280  FOR I = 1 TO TR: FOR J = 1 TO GZ(1): GT(1, I) = GT(1, I)
      + Y(J, I): NEXT J: GM(1, I) = GT(1, I) / GZ(1): NEXT I
290  FOR I = 1 TO GZ(1): FOR J = 1 TO TR: FOR K = 1 TO TR
300  T(1, J, K) = T(1, J, K) + ((Y(I, J) - ME(J)) * (Y(I, K) - ME(K)))
310  W(1, J, K) = W(1, J, K) +
      ((Y(I, J) - GM(1, J)) * (Y(I, K) - GM(1, K)))
320  NEXT K: NEXT J: NEXT I
330  FOR I = 1 TO TR: FOR J = GZ(1) + 1 TO GZ(1) + GZ(2):
      GT(2, I) = GT(2, I) + Y(J, I): NEXT J: GM(2, I) = GT(2, I) / GZ(2):
      NEXT I
340  FOR I = GZ(1) + 1 TO GZ(1) + GZ(2): FOR J = 1 TO TR:
      FOR K = 1 TO TR
350  T(2, J, K) = T(2, J, K) + ((Y(I, J) - ME(J)) * (Y(I, K) - ME(K)))
360  W(2, J, K) = W(2, J, K) +
      ((Y(I, J) - GM(2, J)) * (Y(I, K) - GM(2, K)))
370  NEXT K: NEXT J: NEXT I
380  FOR I = 1 TO TR: FOR J = GZ(1) + GZ(2) + 1 TO EN:
      GT(3, I) = GT(3, I) + Y(J, I): NEXT J: GM(3, I) = GT(3, I) / GZ(3):
      NEXT I
390  FOR I = GZ(1) + GZ(2) + 1 TO EN: FOR J = 1 TO TR:
      FOR K = 1 TO TR
400  T(3, J, K) = T(3, J, K) + ((Y(I, J) - ME(J)) * (Y(I, K) - ME(K)))
410  W(3, J, K) = W(3, J, K) + ((Y(I, J) - GM(3, J)) *
      (Y(I, K) - GM(3, K)))
420  NEXT K: NEXT J: NEXT I
430  FOR I = 1 TO GR: FOR J = 1 TO TR: FOR K = 1 TO TR
440  TRC(J, K) = TRC(J, K) + T(I, J, K)
450  WRC(J, K) = WRC(J, K) + W(I, J, K)
460  NEXT K: NEXT J: NEXT I
470  'determinant of a matrix
480  FOR I = 1 TO TR: FOR J = 1 TO TR
490  IF I = J THEN BT(I, J) = 1 ELSE BT(I, J) = 0

```

```

500  IF I = J THEN BW(I, J) = 1 ELSE BW(I, J) = 0
510  NEXT J: NEXT I
520  FOR J = 1 TO TR: FOR I = J TO TR
530  IF TRC(I, J) <> 0 THEN 550
540  NEXT I: PRINT "MATRIX IS SINGULAR": PRINT : STOP
550  FOR L = 1 TO TR
560  WT = TRC(J, L): TRC(J, L) = TRC(I, L): TRC(I, L) = WT
570  WT1 = BT(J, L): BT(J, L) = BT(I, L): BT(I, L) = WT1
580  NEXT L
590  QT = 1 / TRC(J, J): QQT(J) = TRC(J, J)
600  QT1 = 1 / BT(J, J): QQT1(J) = BT(J, J)
610  FOR L = 1 TO TR
620  TRC(J, L) = QT * TRC(J, L)
630  BT(J, L) = QT * BT(J, L)
640  NEXT L
650  FOR I = 1 TO TR: IF I = J THEN 710
660  QT = -TRC(I, J): QT1 = -TRC(I, J)
670  FOR L = 1 TO TR
680  TRC(I, L) = TRC(I, L) + QT * TRC(J, L)
690  BT(I, L) = BT(I, L) + QT * BT(J, L)
700  NEXT L
710  NEXT I: NEXT J
720  FOR J = 1 TO TR: FOR I = J TO TR
730  IF WRC(I, J) <> 0 THEN 750
740  NEXT I: PRINT "MATRIX IS SINGULAR": PRINT : STOP
750  FOR L = 1 TO TR
760  WW = WRC(J, L): WRC(J, L) = WRC(I, L): WRC(I, L) = WW
770  WW1 = BW(J, L): BW(J, L) = BW(I, L): BW(I, L) = WW1
780  NEXT L
790  QW = 1 / WRC(J, J): QQW(J) = WRC(J, J)
800  QW1 = 1 / BW(J, J): QQW1(J) = BW(J, J)
810  FOR L = 1 TO TR
820  WRC(J, L) = QW * WRC(J, L)

```

```
830  BW(J, L) = QW * BW(J, L)
840  NEXT L
850  FOR I = 1 TO TR: IF I = J THEN 910
860  QW = -WRC(I, J): QW1 = -WRC(I, J)
870  FOR L = 1 TO TR
880  WRC(I, L) = WRC(I, L) + QW * WRC(J, L)
890  BW(I, L) = BW(I, L) + QW * BW(J, L)
900  NEXT L
910  NEXT I: NEXT J
920  DET = QQT(1): FOR I = 2 TO TR: DET = DET * QQT(I): NEXT I
930  DEW = QQW(1): FOR I = 2 TO TR: DEW = DEW * QQW(I): NEXT I
940  PHI = ((EN - 1) - (.5 * (TR - GR) * LOG(DET / DEW)))
950  PRINT NN; TAB(6); GZ(1); TAB(12); GZ(2); TAB(18); GZ(3);
      TAB(24); DET; TAB(36); DEW; TAB(48); DEW / DET;
      TAB(64); PHI
960  PRINT #1, NN; TAB(6); GZ(1); TAB(12); GZ(2); TAB(18); GZ(3);
      TAB(24); DET; TAB(36); DEW; TAB(48); DEW / DET;
      TAB(64); PHI
970  NEXT SET
980  NEXT CL
```

Appendix A4 Computation of coefficients b_j for the Smith-Hazel Index (SHI) for data set A.

P	P^{-1}
1.469 4.509 5.305	1.678 -0.043 -0.240
4.509 46.563 23.233	-0.043 0.034 -0.017
5.305 23.233 32.988	-0.240 -0.017 0.081

G	$P^{-1} G$
1.124 4.649 4.509	0.606 1.431 1.004
4.649 5.944 25.519	0.033 -0.435 0.287
4.509 25.519 22.813	0.016 0.850 0.328

a^{-1}	b_j	h^2
1	3.0406	0.765
1	-0.1140	0.128
1	1.1933	0.692

Matrices P , G are the variance-covariance matrices for the phenotypic and genotypic values. a is a vector of economic weights and b_j is the computed coefficient for variable j . Heritability estimates (h^2) for each variable are also computed.

Appendix A5 Computation of coefficients b_j for the Smith-Hazel
Index (SHI) for data set B.

P	P^{-1}
0.2106 0.0003 0.0054	4.7646 -0.1408 -0.6278
0.0003 0.0204 -0.0023	-0.1408 49.3320 2.7524
0.0054 -0.0023 0.0415	-0.6278 2.7524 24.3306

G	$P^{-1} G$
0.0538 -0.0126 -0.0294	0.2766 -0.0579 -0.1458
-0.0126 0.0110 -0.0058	-0.7101 0.5285 -0.2534
-0.0294 -0.0058 0.0104	-0.7838 -0.1029 0.2555

a^{-1}	b_j	h^2
1	0.0728	0.255
1	-0.4350	0.539
1	-0.6312	0.251

See appendix A3 for notation.

