GENETIC AND MORPHOMETRIC CHARACTERISTICS OF EASTERN WHITE PINE

by

Jing-Dong Yu

B.Sc. (Mathematics) 1991
University of Science and Technology of China

A PROJECT SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
in the Department of Mathematics and Statistics
of
Simon Fraser University

SIMON FRASER UNIVERSITY
January 1994

APPROVAL

Name: Jing-Dong Yu

Degree: Master of Science

Title of project: **Genetic and Morphometric Characteristics of Eastern White Pine**

Examining Committee: Dr. S. K. Thomason,
Chair

Dr. K. L. Weldon, Senior Supervisor

Dr. R. A. Lockhart, Supervisor

Dr. C. B. Dean, Supervisor

Dr. M. A. Stephens, External Examiner

Date Approved: January 11, 1994

PARTIAL COPYRIGHT LICENSE

Title of Thesis/Project/Extended Essay

Genetic and Morphometric Characteristics

of Eastern White Pine

Author: ______________________

(signature)

JINGDONG Yu

(name)

Jan, 17, '94

(date)

Abstract

A descriptive and exploratory study aimed at uncovering genetic and morphometric differences among ten geographical populations of Eastern White Pine (*Pinus strobus* L.) involves two data sets. The first data set contains morphological characteristics of 3000 Eastern White Pine cones, and the second, genetic information on 18 loci for each of 300 trees from which the 3000 cones were harvested. Using exploratory descriptive techniques, there is a suggestion of geographical population differences in cone morphometry. By a combination of multivariate techniques and ANOVA models, statistical analyses reveal that although cone morphometry of Eastern White Pine varies across geographical populations, variation from this source accounts for only a small portion of the total morphometric variability exhibited by the species. Analyses on the genetic data set show that there is very little variation attributable to geographical population differences. Finally, polychotomous logistic regression models are proposed to investigate the possible relationship between the genetic traits of the species and the morphometric measurements of its cones. The conclusion is that there exists only mild statistical evidence showing that the genotypes of certain loci studied depend on the morphometric measurements of the cones, especially on those of cone scales.

Acknowledgements

My sincere gratitude goes to my Senior Supervisor, Dr. Larry Weldon, who influenced me to appreciate applied statistical analysis. His guidance and comments directly shaped Chapter 2 of this project. It is a pleasure to express my thanks to my committee members: Drs. C. B. Dean, R. A. Lockhart and M. A. Stephens, for their encouragement and review of chapters. I am also indebted to the scientists at Canadian Forestry Service, for their experiments and data sets and especially Dr. Jean Beaulieu, for offering me additional information.

The financial support which has made my entire study here possible, is also gratefully acknowledged. This includes a Graduate Fellowship awarded by the Dean of Graduate Studies, Teaching Assistantships offered by the Department of Mathematics and Statistics, and Research Assistantships offered by Statistical Consulting Service and Dr. R. A. Lockhart. Besides, Dr. M. A. Stephens offered me several conference stipends.

Finally, I wish to thank all faculty members, graduate students and staff in the department who have made my stay in SFU enjoyable.

Dedication

To my father and mother

Contents

Chapter 1

Introduction

1.1 Background

Eastern White Pine (*Pinus strobus* L.), an important commercial species which constitutes the most valuable softwood lumber resource in eastern Canada, is a characteristic tree in the Great Lakes - St. Lawrence Forest Region, but its range also extends into the southeastern parts of the Boreal Forest Region, eastward into the Acadian Forest Region and south throughout the Deciduous Forest Region. Because of its low shrinkage and uniform texture, it is used extensively for patterns, window sashes and frames. It is also the tallest conifer in eastern Canada, commonly reaching heights of 100 ft. and diameters of 3 ft.. The crown of a mature tree growing in the open is composed of wide-spreading branches at approximately right angles to the trunk in its mid-portion. In the upper part of the tree the branches ascend, giving a broadly oval outline which often becomes irregular, or asymmetrical, owing to the effect of prevailing wind. In closed stands, the tree is often clear of branches over the lower two-thirds and the crown is columnar. It is found on many different soils[6].

In order to define the population structure of Eastern White Pine in Québec as well as in the other parts of eastern Canada so that several other studies can be conducted, research scientists at the Laurentian Forestry Centre of Forestry Canada are undertaking a study aimed at determining the nature and extent of the genetic variability of this principal coniferous forest species, creating improved synthetic varieties through the identification, selection, and hybridization of superior genotypes and populations, and refining forest tree breeding techniques for maximization of commercial yields[11].

In the natural population of living organisms, the range of genetic diversity is determined by concurrent or antagonistic forces, mainly mutation, recombination, natural selection, gene flow, and random genetic derivation. Specific environmental conditions may also play a significant role in certain cases by causing disturbances that alter the spatial uniformity of populations. Depending on the more or less dominant role played by one or more of the above elements, natural populations may exhibit greater or lesser genetic differentiation which may be observed at both the molecular and morphological levels[11].

Two types of data are therefore collected in this study from a sample of 300 Eastern White Pine trees. The first is the eight morphometric measurements of the species' cones, including cone length, cone width, number of scales on the cone, central scale length, central scale width, winged scale length, number of sound seeds and number of empty seeds. The second captures genetic characteristics of the trees by recording the genotypes of eighteen chosen loci where the solutions in the lab are available. The data was then distributed among statisticians through the Case Study section of the 1993 Annual Meeting of Statistical Society of Canada held at Acadia University, Wolfville, Nova Scotia.

This project began with participation in the Case Study organized by SSC in which participants were invited to address three clearly-defined questions (see Section 1.3). It is expansion of the two reports that I submitted to SSC before the meeting.

1.2 Data Structure

The method of data collection, or the sampling scheme, plays an important role in statistical modeling because model's validity relies on certain assumptions about the data, and different models usually prespecify different data structures. As Sir Maurice Kendall[7] noted:"... a point which is often ignored in the treatment of statistical data: the method of analysis depends on the model we have in mind; and this may depend on the way in which the data were obtained." Therefore a full understanding of sampling scheme necessitates the right choice of statistical model.

For this study, ten Eastern White Pine populations are sampled in Québec (*Figure 1.1*). Populations 1 to 4 are from the Ottawa valley, populations 5 to 8 are from the St.-Lawrence lowlands, population 9 is from Anticosti, and population 10 from Abitibi. The last two populations are sampled at the margin of the natural range of the species; Anticosti Island is furthest east and under the influence of a maritime climate, Abitibi is furthest west and under continental conditions.

In each of the ten populations, 30 trees bearing mature cones are selected so as to cover the geographical area of the population; selected trees have a well developed crown and stand at least 30 meters apart to avoid inbreeding; selection is otherwise random. On each selected tree, a cone bearing branch which protrudes from the crown is shot down with a 12 gauge deer gun, and 10 cones are randomly selected from it. For each of the 3000 cones harvested, 8 morphometric measurements are taken: [1]

[1] Cone width is measured at the fattest point before scales opened to shed their seeds. A central scale is randomly selected along the cone circumference at cone mid-length; it is removed, and its length and width are measured. Eastern white pine seeds have a wing which facilitates natural dissemination. When the winged seed is released from the scale, it leaves a scar from which the length of the winged seed can be measured directly on the removed central scale. (It is in fact the total length of the seed and its attached wing.) This measure thus is called winged scale length. Seeds from the middle of the cone are separated into sound and empty seeds with a seed blower, and examined by X-ray.

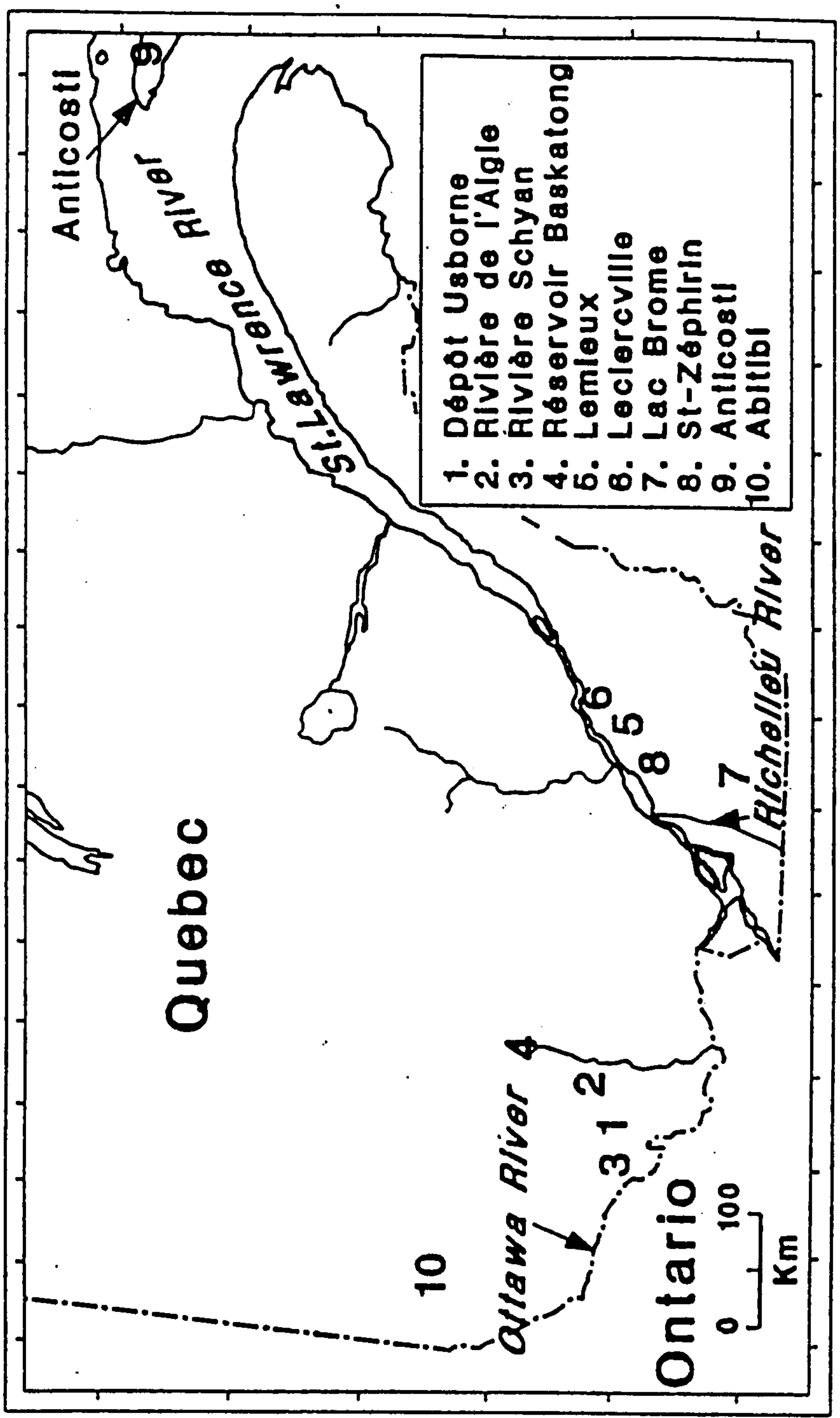

Figure 1.1: Ten Geographical Populations of Eastern White Pine

Notation	Variable	Units
X_1	Cone length	mm
X_2	Cone width	mm
X_3	Number of scales	Number
X_4	Central scale length	mm
X_5	Central scale width	mm
X_6	Winged scale length	mm
X_7	Number of sound seeds	Number
X_8	Number of empty seeds	Number

Overall, we have 3000 cones in our first data set and each cone has eight measurements.

For the second data set, we have only 300 records, one for each tree. They are the same trees as in the first data set. There are 18 variables in this data set, each standing for the genotype of a locus of the tree. (A locus is a location of enzyme on the chromosomes.) For each variable or locus, tree genotype is recorded as a pair of letters which correspond to the alleles this tree bears at the locus associated with the variable. Eastern white pine is a diploid species. Each individual has 12 pairs of homologous chromosomes on which its two alleles can be found for a given locus. In any population, there may exist several alleles. For this data set, however, we only observe three alleles A, B and C. An individual which is homozygous in a particular locus (i.e. both its alleles are of the same form in this locus) may be identified as AA, BB, or CC, depending on which form of the allele it bears. For certain loci (or variables), all 300 individuals are homozygous of the same type: all 300 trees are AA, for example. Such a locus is said to be fixed. A heterozygous individual (i.e. one which has two different alleles for a given locus) is coded AB, AC, or BC, for example, depending on which pair of alleles it has. In certain loci, some trees are homozygous, and some are heterozygous. All possible combinations of two alleles are not necessarily observed. Thus, we have 300 trees as our sampling units in this data set, and 18 loci as our variables for each unit.

Appendix A illustrates the structures of these two data sets.

1.3 Objectives

The objectives of the statistical analysis are defined by the researchers and distributed by SSC. They are:

1. To describe variability in Eastern White Pine cone morphometry within and between populations.

2. To analyze genetic variability within and among Eastern White Pine populations.

3. To investigate whether genetic characteristics depend on cone morphometry. (It is generally recognized that cone morphometry is subject to natural selection; the presence of a relationship between genetic traits and cone morphometry would suggest that the loci studied may also be the object of natural selection.)

We will try to answer the above three questions through our data analysis. In Chapter 2 we will do some preliminary analysis to get a 'feeling' about the data. In Chapter 3 we will build our model framework based on the objective and our preliminary analysis. In order to overcome some technical difficulties in modeling the data which will be presented in Chapter 5, we need to do some further preliminary analysis in Chapter 4. Finally, in Chapter 6 the results of our analysis will be summarized.

Chapter 2

Preliminary Analysis: Part I

Data analysis is a dynamic process. The first step in the process is a descriptive analysis, consisting of anomaly detection and data reduction. In this phase of the analysis, we keep in mind the ultimate objectives of the study, but we do not specify models, or attempt inferences. This step is an essential preliminary to the modelling process and inferential techniques.

2.1 Morphometric Data Set

2.1.1 Univariate Scanning

Although this is a multivariate data set, we looked at the eight morphometric variates one by one first.

Detecting Errors

Initial scanning of the data set found only one mistake in this data set, i.e., the observation of X_8, number of empty seeds, in the 3rd cone on the 17th tree of the 2nd population is recorded as −2. This is obviously a coding error. We think the most probable true value for this observation is 2, although our choice is not entirely objective. After all, it is computationally more economical than treating the observation as missing, in which case the design of experiment is unbalanced and incomplete.

Descriptive Statistics

After data correction, descriptive statistics are given in the *Table 2.1.* Note that except for X_8 (number of empty seeds), the means and medians are very close for all variates, and the skewnesses are very small. These suggest that the distributions for these variates are almost symmetric and that there are probably not many abnormal observations in the data set.

Table 2.1: Descriptive Statistics for Grand Population

Variates(unit)	Mean	Median	Range	SD	Skewness	Kurtosis
$X_1(mm)$	127	126	72-211	18.6	.428	.383
$X_2(mm)$	19.4	19.3	13.3-27.5	2.06	.208	−.003
$X_3(no.)$	61.0	60.0	33-106	10.0	.569	.595
$X_4(mm)$	27.6	27.5	18.8-38.1	2.89	.229	.140
$X_5(mm)$	14.1	14.0	9.30-19.3	1.33	.051	.262
$X_6(mm)$	23.5	23.4	14.0-32.5	2.81	.173	.207
$X_7(no.)$	41.4	39.0	0-146	22.2	.430	−.101
$X_8(no.)$	15.5	11.0	0-106	14.8	1.74	3.69

Exploring Patterns

The histograms of the eight morphometric variates were then plotted to see if there are any unusual patterns or outliers of observations. (See *Figure 2.1* and *Figure 2.2.*)

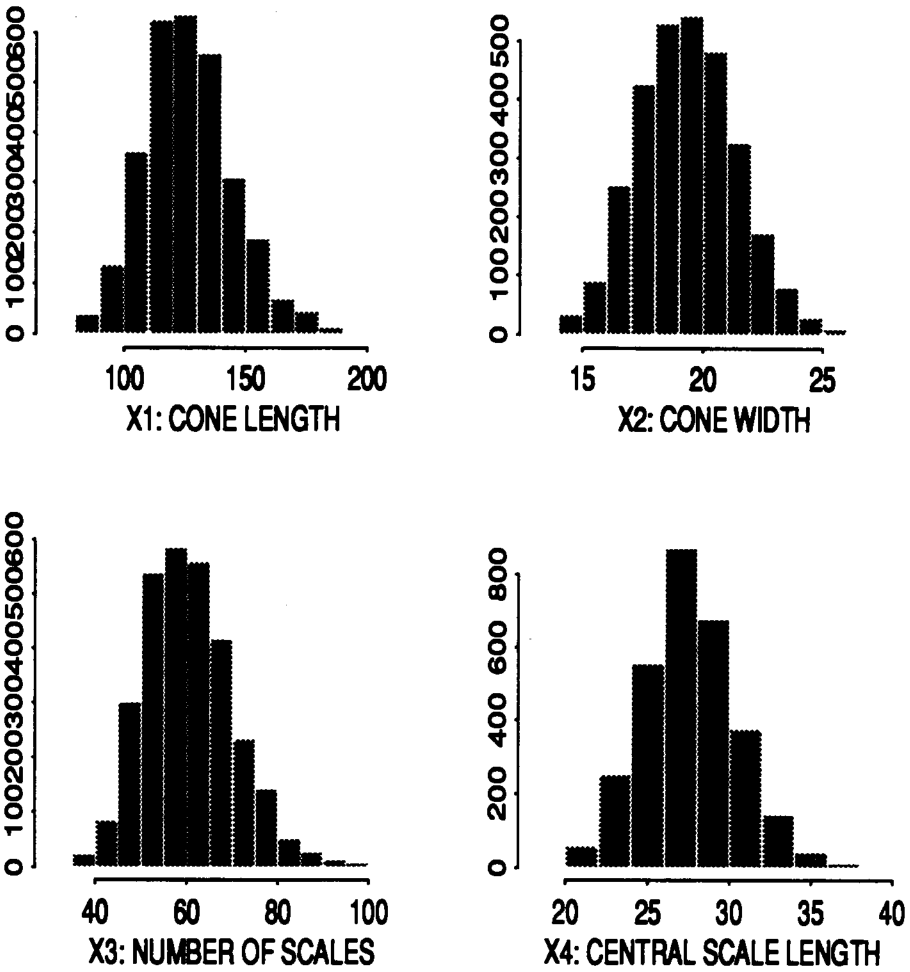

Figure 2.1: Histograms of the Morphometric Variates (Part I)

From the histograms we can see that except for X_8 (number of empty seeds), all other seven variates appear to have a bell-shaped normal distribution with different parameters. Among them, X_1 (cone length), X_3 (number of scales) and X_7 (number of sound seeds) are slightly skewed to the right. Because of the very large sample size, even negligible departure from normality may turned out to be highly significant and thus it is not necessary to test for normality formally.

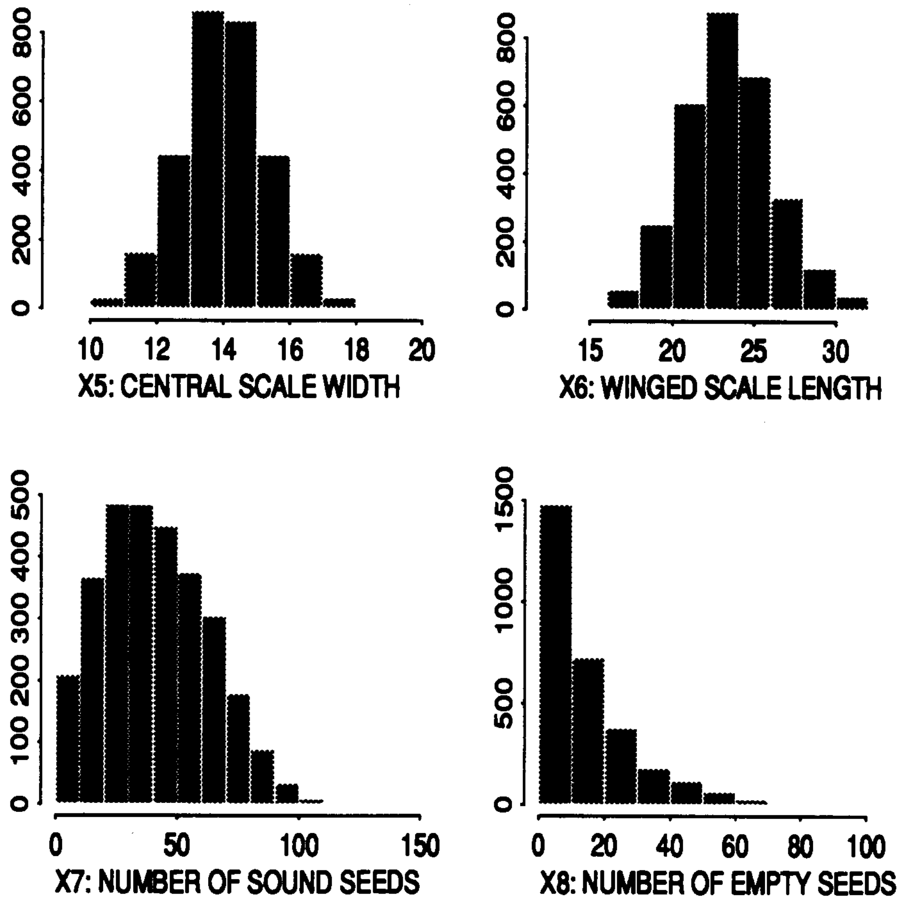

Figure 2.2: Histograms of the Morphometric Variates (Part II)

Visual Comparison Across Populations

Boxplots for each of the eight morphometric variates across populations are presented in *Figure 2.3* and *Figure 2.4*. We rearranged the ten populations according to the order for the medians of X_1 (cone length), a variate we feel the most important in characterizing cone morphometry. The order of the populations is 10, 7, 8, 4, 9, 6, 1, 3, 2, 5.

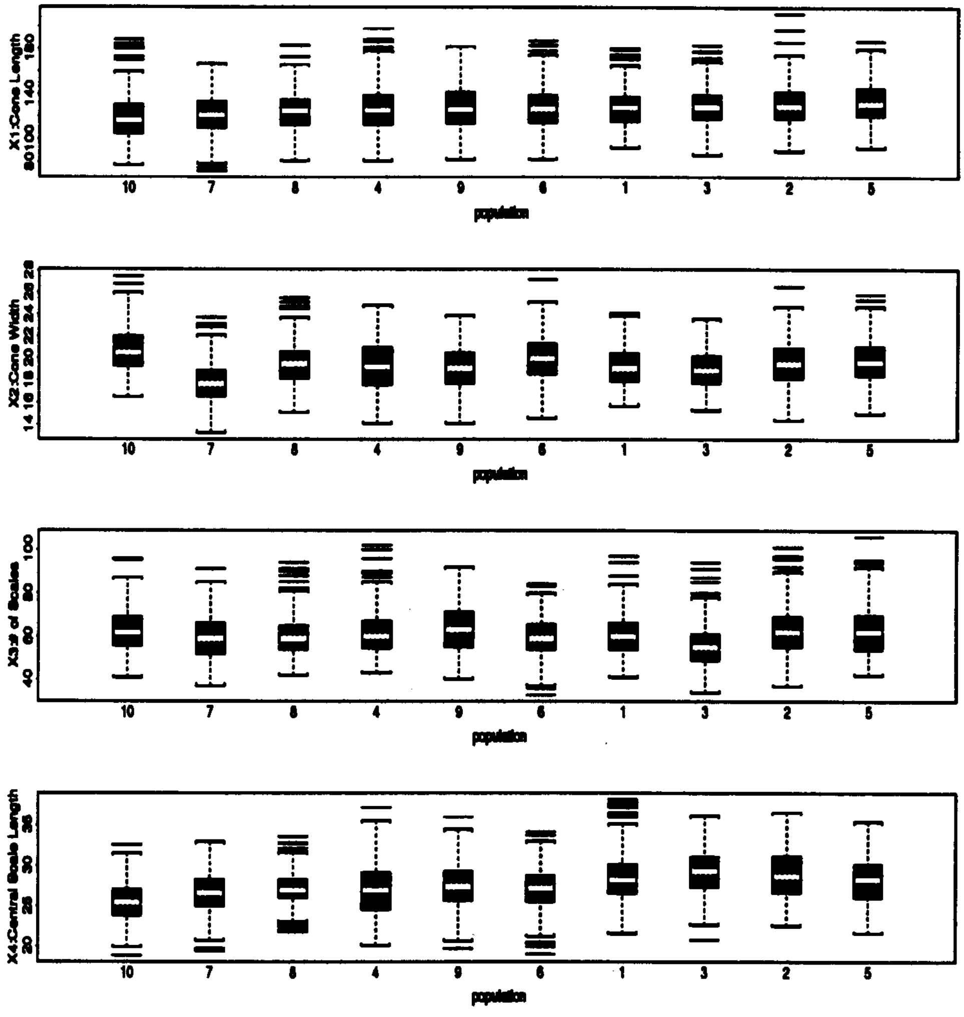

Figure 2.3: Boxplots for Morphometric Variates across Populations (Part I)

As can be seen from these boxplots, the ranking of populations on most of the variates are quite consistent with that of X_1, most notably, on X_4, X_6 and X_2, X_5. Also, population 10 differs from other populations most dramatically since its measurements on most of the morphometric variates are on the small side while its measurements on X_2 (cone width) and X_3 (number of scales on the cone) are the largest among all populations. The presence of vast amount of 'outliers' at the upper tail of the distribution of X_8 may in fact be the evidence that X_8 is not normally distributed.

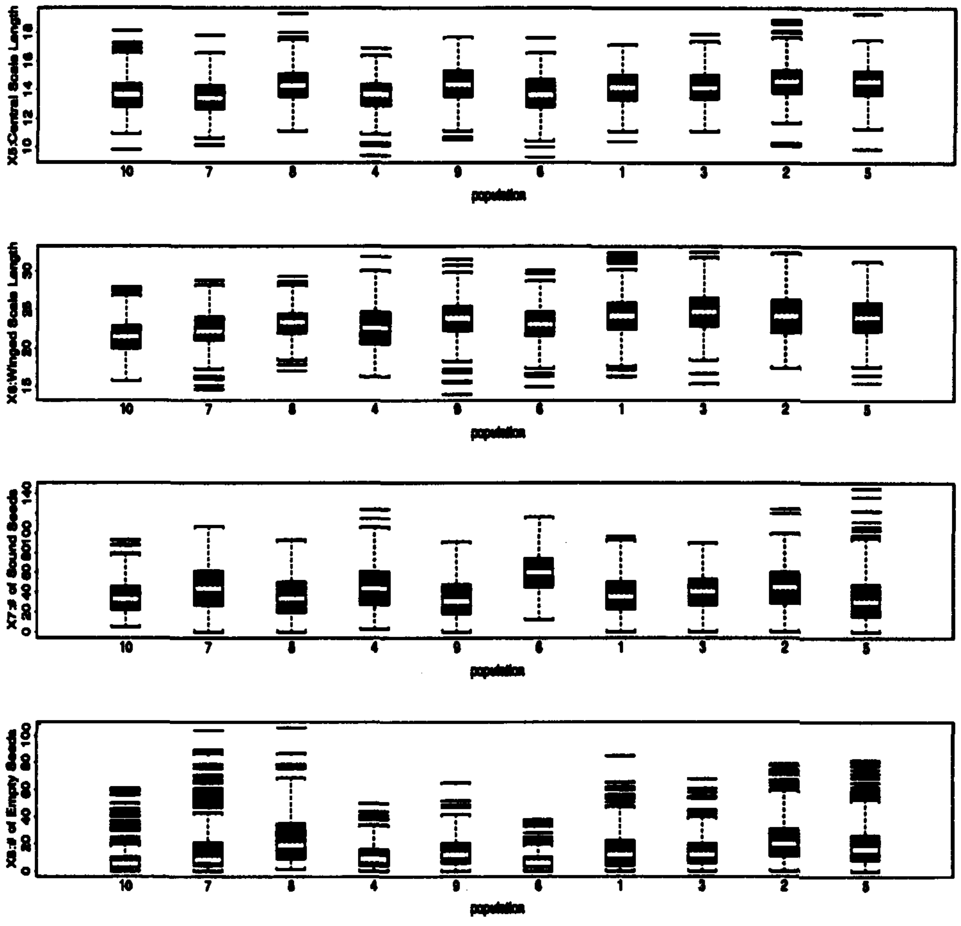

Figure 2.4: Boxplots for Morphometric Variates across Populations (Part II)

2.1.2 Multivariate Exploration

Correlation Structure and Its Implications

A first step in exploring this nearly-normal multivariate data set is to examine its correlation matrix. The original correlation matrix for the entire data set of 3000 cones is presented as follows (*Table 2.2*):

Table 2.2: Correlation Matrix (rounding the figures to the 3rd digit)

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1							
X_2	.514	1						
X_3	.570	.508	1					
X_4	.626	.401	.133	1				
X_5	.521	.603	.293	.535	1			
X_6	.597	.388	.108	.948	.532	1		
X_7	.360	.281	.345	.137	.258	.122	1	
X_8	.182	.108	.228	.130	.219	.129	−.281	1

In the case of a relatively small matrix it is often possible, by visual examination of the elements, to discover subsets of variates which correlate relatively highly with each other. From the correlation matrix in *Table 2.2*, we can see that:

1. The first six variates ($X_1 - X_6$), most of them being measurements of the cone and its bearing scale, are generally moderately correlated with each other. However, the last two variates (X_7 and X_8), number of sound seeds and number of empty seeds, are loosely correlated with the first six.

2. Among the first six, the correlation between central scale length (X_4) and winged scale length (X_6) is surprisingly high (0.948).

When we rearrange the order of the variates in the matrix, and round the figures to the first digit, the structural relationship among the eight variates becomes more apparent (*Table 2.3*).

Table 2.3: Rearranged Correlation Matrix(rounding the figures to the 1st digit)

	X_4	X_6	X_1	X_2	X_5	X_3	X_7	X_8
X_4	1							
X_6	.9	1						
X_1	.6	.6	1					
X_2	.4	.4	.5	1				
X_5	.5	.5	.5	.6	1			
X_3	.1	.1	.6	.5	.3	1		
X_7	.1	.1	.4	.3	.3	.3	1	
X_8	.1	.1	.2	.1	.2	.2	−.3	1

Moreover, we have the following observations and suggestions:

1. According to the magnitude of the correlations, the variates may be partitioned into three distinct groups namely (X_4, X_6), (X_1, X_2, X_5, X_3) and (X_7, X_8).

2. Although variate X_3 (number of scales) is in the second group, its correlations with the variates X_4 and X_6 may suggest that it could also be classified into the third group. This may be attributed to the fact that variate X_3, like X_7 and X_8, is a 'count' observation.

3. X_4 and X_6 have almost the same correlations with each of the other six variates, another piece of evidence suggesting that they are nearly perfectly linearly correlated with each other. [1]

At last, we want to add that each of the ten populations exhibits very similar patterns of correlation structures, although small variations on certain cells (correlations) do exist. See Appendix B for the ten correlation matrices.

[1] Based on the assertion:

$$\rho(X_4, X_6) = 1 \Longleftrightarrow \rho(X_4, X) = \rho(X_6, X) \quad \forall X \quad \text{r.v.}$$

Multivariate Visual Exploration

Star plots for the medians of the 10 populations on each of the eight morphometric variates are shown in *Figure 2.5.* The radii are scaled into the interval [0.2, 1] using a linear transformation, with the largest population having a radius of 1 and the smallest population having a radius of 0.2. Thus a long radius on a star would be associated with a population having a relatively large value on that variate.

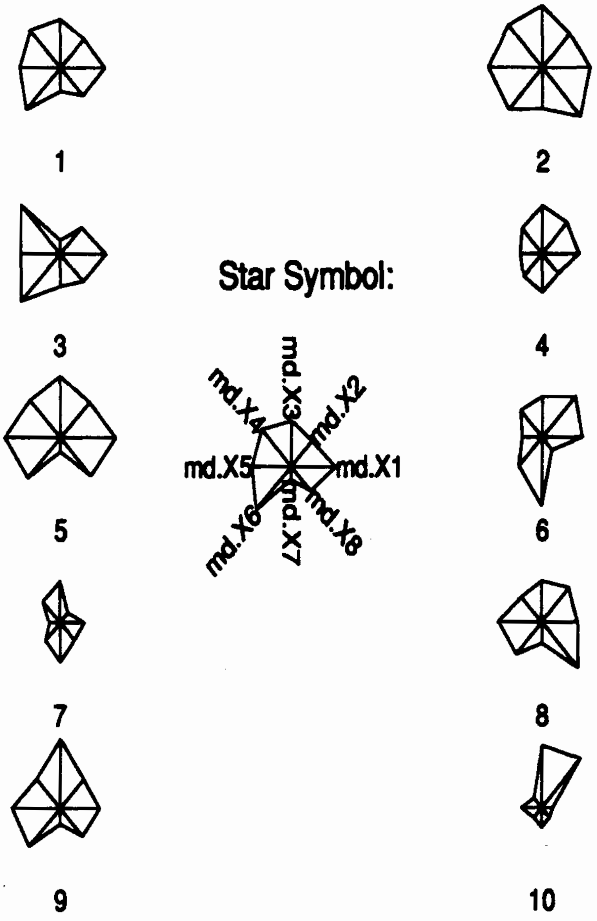

Figure 2.5: Visual Comparison of Populations (based on scaled medians across populations)

So we can envisage how populations differ from each other from the plot. What is immediately apparent is that population 2 and population 5 have relatively large morphometric measurements on most of the variates; while population 7 and population 10 have very small morphometric measurements on most of the variates.

2.2 Genetic Data Set

This data set contains the genotypes for the 300 trees on eighteen selected loci. The genotype for the first locus of tree 29 in population 3 is not given. We will treat it as missing since we could use the rest of our data set to calculate the genotypic frequency, on which most of our analysis is based.

2.2.1 Descriptive Statistics

Suppose different genotypes are considered to be different 'categories', and the number of possible genotypes on any locus is finite, then the trees sampled will essentially follow a multinomial distribution for any given locus. These distribution parameters are not necessarily the same from one locus to another. In some loci, the distributions may be reduced to binomial or degenerate distributions. The sample statistics for all 300 trees combined are calculated and presented in *table 2.4*.

As can be seen from the table, there is always an apparent dominant genotype on most of the loci studied (except for Loci 1 and 10 on which the dominant genotypes are not so apparent), such as genotype AA on Locus 2, or genotype BB on Locus 14. The trees which have the dominant genotype on a certain locus usually account for 75% to 100% of the total 300 trees sampled. The strong dominance is in fact the result of inbreeding in natural population. The fact that most dominant genotypes are homozygous (of the same allele) suggests there is not much heterogeneity among different geographical populations of trees.

Table 2.4: Frequency Table of Observed Genotypes

		Percentage of Genotypes						
Locus	Enzyme	AA	AB	AC	BB	BC	CC	Total
1	Aco	.070	.204	.090	.211	**.311**	.114	299
2	Hex	**.787**	.127		.087			300
3	Got-1	**.997**	.003					300
4	Got-2	**.960**	.040					300
5	Got-3	.200	**.527**		.273			300
6	Ak	.003	.160	.003	**.773**	.057	.003	300
7	Fum	**.980**	.020					300
8	G6p	.013	.217		**.770**			300
9	Idh	**1**						300
10	Mdh-1	.017	.070	.207	.047	.322	**.337**	300
11	Mdh-2		.007		**.923**	.070		300
12	Mdh-3	**1**						300
13	Mpi	.003	.067		**.930**			300
14	Pgm-1		.017		**.927**	.053	.003	300
15	Pgm-2	.003	.067	.030	**.543**	.320	.037	300
16	6pg	**1**						300
17	Pgi-1	**.957**	.043					300
18	Pgi-2	**.853**	.090	.053			.003	300

2.2.2 Genotypic Frequencies Across Populations

The observed genotype frequency across ten populations is then portrayed in *Table 2.5* to *Table 2.18.* These tables reveal genetic differences, if there exists any, among the ten populations for most of the loci studied. Based on these contingency tables, routinely, we should be able to perform χ^2 tests of homogeneity of genotypic distributions across ten populations. However, due to the presence of vast amount of low frequencies in certain cells, χ^2 tests may not be valid here. (In fact, the presence of vast amount of low frequency cells and their consistency across populations can be regarded as an indicator of homogeneity of genotype distributions.) There is only one exception: Locus 5 (Got-3) do not have many of such cells. The χ^2 test statistics on this locus is 25.6 with 18 degrees of freedom, yielding a p-value of 10.8%. We thus can not conclude heterogeneity in this case either.

We thus decided to test the following hypotheses: that the proportions of trees with dominant genotype are the same across ten populations. The last row in each table shows the test result.

Degenerate Case:

On Locus 9 (Idh), Locus 12 (Mdh-3) and Locus 16 (6pg), all 300 trees are homozygous of the same genotype AA. These three loci furnish no information about the genetic variability among populations. See *Table 2.4*.

Binomial Case:

On Locus 3 (Got-1), Locus 4 (Got-2), Locus 7 (Fum), and Locus 17 (Pgi), we have two genotypes AA and AB (See *Table 2.5* to *Table 2.8*). These loci are almost homozygous with genotype AA being the dominant one. Little genetic variability is expected across populations.

Table 2.5: Locus 3(Got-1): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AA	30	29	30	30	30	30	30	30	30	30
AB		1								
Total	30	30	30	30	30	30	30	30	30	30
test	$\chi^2 = 0.05 \quad p = 1.00$									

Table 2.6: Locus 4(Got-2): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AA	29	30	28	29	29	29	30	27	29	28
AB	1	1	2	1	1	1	1	3	1	2
Total	30	30	30	30	30	30	30	30	30	30
test	$\chi^2 = 0.5$ $p = 1.00$									

Table 2.7: Locus 7(Fum): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AA	28	30	29	30	30	30	30	28	30	29
AB	2		1					2		1
Total	30	30	30	30	30	30	30	30	30	30
test	$\chi^2 = 0.4$ $p = 1.00$									

Table 2.8: Locus 17(Pgi-1): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AA	28	27	27	28	30	30	30	29	28	30
AB	2	3	3	2				1	2	
Total	30	30	30	30	30	30	30	30	30	30
test	$\chi^2 = 0.86$ $p = 1.00$									

Trinomial Case:

Five loci (Hex,Got-3,G6p,Mdh-2 and Mpi) are found to have three genotypes. They are not the same three genotypes on all these loci. See *Table 2.9* to *Table 2.13*. Note for Locus 13(Mpi), we only observe one allele A in the entire 300 trees sampled. If this tree can be regarded as an outlier, the locus is degenerated into a binomial case. On the other hand, this also implies that the multinomial classification of loci based on the sample may not hold for the population due to chance error of sampling.

Table 2.9: Locus 2(Hex): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AA	27	23	27	22	21	17	28	28	21	22
AB	1	3	1	4	7	8	1	1	5	7
BB	2	4	2	4	2	5	1	1	4	1
Total	30	30	30	30	30	30	30	30	30	30
test	$\chi^2 = 4.99 \quad p = 0.835$									

Table 2.10: Locus 5(Got-3): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AA	8	9	6	7	4	9	3	4	4	6
AB	17	13	17	11	17	16	14	14	23	16
BB	5	8	7	12	9	5	13	12	3	8
Total	30	30	30	30	30	30	30	30	30	30
test	$\chi^2 = 5.81 \quad p = 0.759$									

Table 2.11: Locus 8(G6p): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AA						2	1			1
AB	6	8	7	6	9	6	4	2	8	9
BB	24	22	23	24	21	22	25	28	22	20
Total	30	30	30	30	30	30	30	30	30	30
test	$\chi^2 = 2.08 \quad p = 0.99$									

Table 2.12: Locus 11(Mdh-2): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AB		1					1			
BB	25	27	26	29	27	30	28	28	27	30
BC	5	2	4	1	3		1	2	3	
Total	30	30	30	30	30	30	30	30	30	30
test	$\chi^2 = 1.03 \quad p = 1$									

Table 2.13: Locus 13(Mpi): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AA	1									
AB	3	2	2	3	1	1	1	2	3	2
BB	26	28	28	27	29	29	29	28	27	28
Total	30	30	30	30	30	30	30	30	30	30
test	$\chi^2 = 0.35 \quad p = 1$									

4-nomial Case:

On each of Locus 14(Pgm-1) and Locus 18(Pgi-2), four genotypes are observed. The contingency tables are *Table 2.14* and *Table 2.15*.

Table 2.14: Locus 14(Pgm-1): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AB	1				2				1	1
BB	28	28	26	25	26	29	30	30	27	29
BC	1	2	3	5	2	1			2	
CC			1							
Total	30	30	30	30	30	30	30	30	30	30
test	$\chi^2 = 1.16 \quad p = 0.999$									

Table 2.15: Locus 18(Pgi-2): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AA	20	25	25	28	26	27	27	29	21	28
AB	5	4	2	2	2	3	3	1	4	1
AC	4	1	3		2				5	1
CC	1									
Total	30	30	30	30	30	30	30	30	30	30
test	$\chi^2 = 3.22 \quad p = 0.955$									

6-nomial Case:

Locus 1 (Aco), Locus 6 (Ak), Locus 10 (Mdh-1) and Locus 15 (Pgm-2) exhibit the greatest diversity of genotypes – all six possible genotypes are presented on these loci.

Table 2.16: Locus 1(Aco): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AA	2	2	2	3	1	0	3	2	2	4
AB	7	6	6	8	6	6	4	6	5	7
AC	4	1	5	4	3	1	4	2	2	1
BB	7	3	4	3	8	5	8	7	10	8
BC	8	11	9	8	7	18	6	10	7	9
CC	2	7	3	4	5	0	5	3	4	1
Total	30	30	29	30	30	30	30	30	30	30
test	$\chi^2 = 10.5 \quad p = 0.302$									

Table 2.17: Locus 6(Ak): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AA										1
AB	9	3	4	6	6	4	3	6	3	4
AC										1
BB	21	25	21	24	24	24	22	24	24	23
BC		2	5			1	5		3	1
CC						1				
Total	30	30	30	30	30	30	30	30	30	30
test	$\chi^2 = 0.76 \quad p = 1$									

Table 2.18: Locus 10(Mdh-1): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AA			1			2		1		1
AB		1	4		2	3	1	2	1	7
AC	6	5	3	7	7	8	8	4	9	5
BB			1	5	1	1	1	4	1	
BC	12	12	9	12	6	8	11	9	10	8
CC	12	12	12	6	14	9	8	10	9	9
Total	30	30	30	30	30	30	30	30	30	30
test	$\chi^2 = 4.69 \quad p = 0.861$									

Table 2.19: Locus 15(Pgm-2): Genotypic Frequency Across Populations

Genotype	Population									
	1	2	3	4	5	6	7	8	9	10
AA				1						
AB	3	2	1	2	1	4	2	1	3	1
AC	2	2			1	1		1	2	
BB	15	14	18	14	18	14	19	20	16	15
BC	10	10	10	12	9	9	8	7	8	13
CC		2	1	1	1	2	1	1	1	1
Total	30	30	30	30	30	30	30	30	30	30
test	$\chi^2 = 2.65 \quad p = 0.976$									

In short, we get a 'feeling' that the distributions of genotype are homogeneous among ten populations. However, genetic differences across populations do exist. In Chapters 4 and 5, we will reveal these differences through more detailed analysis.

Chapter 3

Model Construction

In this chapter, we intend to build the model framework necessary to answer the three questions raised by the researchers.

3.1 Models Documenting Morphometric Variability

We have several approaches to the first question in our objectives: to describe variability in Eastern White Pine cone morphometry within and among populations.

1. Given the design of the experiment and the multivariate nature of `cone morphometry`, a multivariate analysis of variance (MANOVA) model seems to be the appropriate one to first try here. Since we have not defined `cone morphometry` yet, we have at least two possible model constructions:

    - Define `cone morphometry` as a vector of the original eight morphometric variates and perform MANOVA based on this definition.

By analyzing the original data set, we utilize all the information available.

- Treat `cone morphometry` as latent variables whose indicators are the eight manifest morphometric measurements. And then perform MANOVA based on these latent variables (indices). By analyzing compressed data, we may lose some information.

2. Since all MANOVA can do is to compare the mean vector of variates across populations, we wish we would be able to state quantitatively the relative cone morphometry variability within and between populations. If the `cone morphometry` is a one-dimensional variate, an ANOVA model will yield numerically the variability within and between populations. We will use this approach as well.

3.1.1 MANOVA Model

Let $\mathbf{X} = (X^{(1)}, X^{(2)}, ..., X^{(p)})^T$ be the p-dimensional random vector characterizing cone morphometry. By subscripts, $\mathbf{X}_{ijr}$ denotes the cone morphometry for the rth cone on the jth tree in the ith population. We assume that $\mathbf{X}_{ijr}$ has the distribution of MVN$(\boldsymbol{\mu} + \boldsymbol{\alpha}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\mu} = (\mu^{(1)}, \mu^{(2)}, ..., \mu^{(p)})^T$, $\boldsymbol{\alpha}_i = (\alpha_i^{(1)}, \alpha_i^{(2)}, ..., \alpha_i^{(p)})^T$ and $\boldsymbol{\Sigma}_i$ is a positive definite $p \times p$ matrix. Then the model is written as :

$$\mathbf{X}_{ijr} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_{j(i)} + \boldsymbol{E}_{r(ij)} \qquad \sum_{i=1}^{10} \boldsymbol{\alpha}_i = \mathbf{0} \qquad \left\{ \begin{array}{rcl} i &=& 1,\ 2,\ \ldots,\ 10 \\ j &=& 1,\ 2,\ \ldots,\ 30 \\ r &=& 1,\ 2,\ \ldots,\ 10 \end{array} \right.$$

with boundary condition $\sum_{i=1}^{10} \boldsymbol{\alpha}_i = \mathbf{0}$. Because of the nested design of the experiment, $\boldsymbol{\beta}_{j(i)}$ is the random effect vector due to the the jth tree in the ith population; and $\boldsymbol{E}_{r(ij)}$ is the random effect vector due to the rth cone on the jth tree in the ith population. Then $\boldsymbol{\mu}$ can be interpreted as the overall (fixed) mean vector for the grand population, and $\boldsymbol{\alpha}_i$ as the (fixed) effect vector due to the ith population.

We want to test the simple hypothesis $H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = ... = \boldsymbol{\alpha}_{10} = \mathbf{0}$. The assumptions underlying the model are: (1) multivariate normal distributions, (2) $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = ... = \boldsymbol{\Sigma}_{10}$, i.e., the ten populations have equal covariance matries, and (3) serially uncorrelated observations. We will check the assumptions (1) and (2) in our preliminary analysis in the next chapter.

3.1.2 ANOVA Model

Let X be the random variable for any aspect of the `cone morphometry`, either as one of the eight original morphometric measurements or as one of the constructed indices. Then the model is

$$X_{ijr} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{r(ij)} \quad \sum_{i=1}^{10} \alpha_i = 0 \quad \left\{ \begin{array}{lcl} i & = & 1,\ 2,\ \ldots,\ 10 \\ j & = & 1,\ 2,\ \ldots,\ 30 \\ r & = & 1,\ 2,\ \ldots,\ 10 \end{array} \right.$$

where μ is the overall (fixed) mean of the ten populations; α_i is the fixed effect due to the population i for these ten geographical populations are not a sample (with boundary constraint $\sum_{i=1}^{10} \alpha_i = 0$); $\beta_{j(i)}$ is the random effect due to the jth tree in the ith population; $\epsilon_{r(ij)}$ is the random effect due to the rth cone on the jth tree in the ith population. The usual assumptions for the model are: (1) $\beta_{j(i)}$ has a normal distribution with mean 0 and variance σ_β^2; (2) $\epsilon_{r(ij)}$ has a normal distribution with mean 0 and variance σ^2.

This is a two-stage mixed ANOVA model. Based on the model output, we can

1. test the simple hypothesis $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_{10} = 0$;

2. test the hypothesis $H_0 : \sigma_\beta^2 = 0$;

3. estimate all the parameters in the model and thus obtain numerical comparisons for cone morphometry both between populations and between trees;

4. estimate the variance components for each variate and describe quantitatively the amount of variation exhibited either within a population or between populations.

3.2 Models Documenting Genetic Variability

In order to represent in a standardized way the amount of genetic variation, we need measures that allow us to quantify this information. The most commonly used measure of genetic variation in a population is the amount of *heterozygosity*. Because individuals in diploid species are either heterozygous or homozygous at a given locus, this measure presents intuitively a biologically useful quantity. The disadvantage of using these measures, however, is that they are not very sensitive to additional variation when the number of genotypes n is large, because the upper limit, unity, is the same for any n. [1]

The theoretical *Hardy-Weinberg heterozygosity* (will be called *heterozygosity* thereafter) of a population for a particular locus with n alleles is defined as:

$$h = \sum_{i \neq j} p_i p_j = 1 - \sum_{i=1}^{n} p_i^2$$

where p_i stands for the allelic frequency for allele i on the locus. The underlying interpretation of this definition is that heterozygosity is the probability that for a given locus, any two randomly chosen individuals from the population exhibit different alleles.

In our analysis, we feel that it is necessary to make a minor modification

[1] According to the definition, the upper limit of heterozygosity is:

$$\lim_{n \to \infty} \max_{p_i} h = \lim_{n \to \infty} [1 - \sum_i (\frac{1}{n})^2] = \lim_{n \to \infty} (1 - \frac{1}{n}) = 1$$

to the above definition. It seems to be more appropriate to use the genotypic frequency instead of the allelic frequency. (And thus n is the number of genotypes on the locus.) The reasons for the modification are:

1. Generally speaking, the information extracted from the allelic frequency is contained in the information extracted from the genotypic frequency. However, the reverse is not always right.
2. The popularity of using allelic frequency in the definition of heterozygosity is largely due to an implication of the so-called Hardy-Weinberg Principle which reveals that for a random-mating population [2], we can use n alleles at a locus to describe, without losing any information, the genetic variation instead of the $n(n+1)/2$ different genotypes formed by the n alleles. When the number of alleles in the population gets large, the computational advantage of using allelic frequency becomes apparent. In our data set, however, only 3 alleles and therefore 6 possible different genotypes are observed.
3. Biologically, it may be the combinations of the alleles, rather than the alleles themselves on the loci, that are really important in determining the traits of the species.

[2] *A random-mating population* is a group in which the probability of a mating between individuals of particular genotypes is equal to the product of their individual frequencies in the population.

3.3 Models Connecting Genetic Traits and Cone Morphometry

In this part of our analysis, we are asked to investigate whether genetic characteristics of Eastern white Pine depend on their cone morphometry. This question may seem, at first sight, nonsense as we all know that cone morphometry of a tree is controlled by its genetic structure. But in the long-term evolution process, survival of, say, AA cone is a prerequisite to observation of AA genotype in trees. So it can be considered dependence in this sense. In fact, the researchers justified their question by telling us the ultimate objective of the study: to investigate whether there exists any evidence suggesting that the loci studied may also be the object of natural selection since it is generally recognized that cone morphometry is subject to natural selection. Moreover, we feel that a dependence relationship does not necessarily imply a cause-and-effect relationship, as is the case with genetic traits and cone morphometry. The relationship between genetic traits and cone morphometry, if any being revealed through statistical analysis, is more of an association or prediction than a causation. [3]

The model we proposed here is a straightforward regression model, in which the genetic characteristics of the trees, or more specifically, the genotypes, are being treated as the response variables. Since the genotype is regarded as multinomial categorical variable, the polychotomous logistic regression model is constructed on a locus-to-locus base as follows:

$$P[y=i] = \frac{exp(\alpha_i + \beta_i \mathbf{x})}{1 + \sum_{i=1}^{I-1} exp(\alpha_i + \beta_i \mathbf{x})} \qquad i = 1, 2, ... I-1$$

and

$$P[y=I] = 1 - \sum_{i=1}^{I-1} P[y=i]$$

[3] From statistical point of view, a cause-and-effect relationship can only be confirmed through carefully designed and controlled experiments. This study, however, is virtually observational and exploratory.

where I is the total number of genotypes observed on the locus, and $\mathbf{x}$ is the vector of cone morphometry properly defined beforehand, α_i's and β_i's are the parameters to be estimated. This model is compatible with the sampling scheme since each tree is treated as an observation in the regression and the total of 300 trees are a stratified random sample from the grand population.

Several issues arise in connection with this model construction:

1. **Data Clumping:** Since every cone on the same tree has identical genotype and the each tree is being treated as one observation, we must find a representative cone for each tree. An obvious answer to this question is from the results of the ANOVA model set in Section 3.1.2. The morphometry of the representative cone for the jth tree in the ith population is estimated as:
$$\mathrm{E}(X_{ij}) \;=\; \mu + \alpha_i + \hat{\beta}_{j(i)}$$
In other words, we use estimated cone morphometry instead of the original measurements.

2. **Multicollinearity:** As revealed in the previous chapter, the correlations among some of eight original morphometric variates are quite high. If they are used as the regressors in the above model, the presence of multicollinearity may inflate the variance of the least squares estimator and possibly any predictions made, and also restrict the generality and applicability of the estimated model. We should find way to eliminate the muticollinearity of the morphometric variables while still preserve as much information as possible before we model the data. Ideally, we should regress on uncorrelated morphometric variates. This, together with the consideration of indices construction, leads to the principal factor analysis in the next chapter.

3. **Stepwise Selection:** The only statistical software available to perform polychotomous logistic regression is the BMDP **PR** program. This program produces maximum likelihood estimation of the parameters in the

above model by a stepwise forward selection procedure. My classmate, Jun Wu, gave a very crisp account for the details of the procedure in his Master's Project[14]. What we are concerned with here is the specification of the constants in the program, i.e., p-value limits controlling entry or removal of terms, convergence criterion for the likelihood function, and less importantly, number of iterations to maximize the likelihood function.

Chapter 4

Preliminary Analysis: Part II

Following the model construction, we found that we need to do some more preliminary analyses before modelling the data. These include, as a summary for the previous chapter and a preview for this chapter:

- construction of indices on the morphometric data set. The purposes of this analysis are: (1) to construct a few indices (latent variables) to characterize the `cone morphometry`; (2) to construct uncorrelated explanatory variables to be used in the logistic regression; (3) to reduce the number of the variates in the data set (dimension reduction).

- examination of the model assumptions mentioned in the previous chapter for the morphometric data set, including: multivariate normality and equal covariance matrices for the ten populations as required by the MANOVA model.

- further exploratory analysis on the genetic data set by using the measure of heterozygosity.

4.1 Construction of Indices

From the preliminary analysis in Chapter 2, we observed that among the original eight morphometric variates, X_7 (number of sound seeds) and X_8 (number of empty seeds) seem to be most different from the rest according to the correlations among the variates. Moreover, the histograms of these two variates showed that variate X_8 (number of empty seeds) has an apparent departure from normality, and variate X_7 (number of sound seeds) also departs markedly from the fit. These suggest that we should separate them out from our analysis. This is an appropriate thing to do in light of indices construction because these two variates together furnish information about the productivity of the species, an important aspect of cone morphometry.

4.1.1 The First Set of Indices

Given the first six cone measurements, the most important aspects of cone morphometry are probably the cone size and the scale size. Focusing on these two, we have the following objective approach of indices construction.

Delete variate X_6 (winged scale length) since we have obtained a lot of evidence in Chapter 2 that X_4 (central scale length) and X_6 are almost perfectly correlated with each other. [1] Then, we define:

- **cone size** as the product of cone length and cone width:

$$Y_1 = X_1 X_2$$

- **scale size** as the product of central scale length and central scale width:

$$Y_2 = X_4 X_5$$

[1] Even when we look the ten populations one by one, the correlations between these two variates are consistently high, ranging from .916 to .967 (Appendix B). This implies that, by deleting one of them, we may not lose much information regarding the cone morphometry variabilities.

- keep **number of scales**, **number of sound seeds** and **number of empty seeds** and denote them in this set of indices as:

$$Y_3 = X_3$$

$$Y_4 = X_7$$

$$Y_5 = X_8$$

Thus we have five indices in this set. The advantage of this construction is the easy interpretation of the indices. The shortcomings of this set of indices are (1) we have no idea about the amount of information lost through the construction; (2) it may fail to characterize the relevant **shape** of the cone, [2] and (3) the correlations among the indices are still quite high and thus cannot be used as the regressors in the logistic regression analysis; see *Table 4.1* for the correlation structure.

Table 4.1: Correlation Matrix for the First Set of Indices

	Y_1	Y_2	Y_3	Y_4	Y_5
Y_1	1				
Y_2	.71	1			
Y_3	.62	.24	1		
Y_4	.38	.22	.35		
Y_5	.17	.19	.23	−.28	1

4.1.2 The Second Set of Indices

Factor analysis is one of the most commonly used statistical techniques for constructing latent variables or indices. It aims at ascertaining whether the interrelations between a set of observed variables are explicable in terms of a small number of underlying, unobservable latent variables. We shall also

[2] For example, two cones may have exactly the same cone size (Y_1) index but one of them may be slimmer than another.

use this subjective way to construct our indices. The advantages of this approach are: (1) the quality of the analysis, justified by the amount of variation retained, will be known to the analyst and (2) the resulting factors are uncorrelated.

We shall perform factor analysis using principal components method. In practice, it is always important to consider the possibility, or desirability, of transforming the data before analysis. We shall use the most frequently used transformation of (natural) logarithms. The effect of the logarithmic transformation is to give measures with the same proportional variability the same variance. Moreover, since the resulting factors are all a linear combination of the variates used in the analysis, this transformation will facilitate the interpretation of the resulting components because it would be hard to interpret the meaning of the sum (or difference) of the original variates if we perform the analysis on $(X_1, X_2, X_3, X_4, X_5, X_6)^T$.

By minimum eigenvalue criterion, two factors are retained. They are, denoted as F_1 and F_2:

$$F_1 = .84(lnX_1)^t + .73(lnX_2)^t + .51(lnX_3)^t + .84(lnX_4)^t + .78(lnX_5)^t + .83(lnX_6)^t$$

$$F_2 = .16(lnX_1)^t + .38(lnX_2)^t + .76(lnX_3)^t - .48(lnX_4)^t + .02(lnX_5)^t - .50(lnX_6)^t$$

where a superscript t denotes the standardization of a random variable. The percentage of variation retained by the first factor is 58 percent, and retained by these two factors is 79 percent, which is acceptable.

These two factors are also standardized random variables. The first factor F_1 is easy to interprate because all its coefficients are positive. It is most appropriately interpreted as the latent variable of cone size. For the second factor, factor loadings are big on those of the scale morphometry, especially on the number of scales, winged scale length and central scale length. Thus this factor is mostly concerned with scale morphometry. The interpretation for this factor is not so straightforward: roughly speaking, the larger this factor is, the larger is the number of scales, or, the smaller is the scale size, or, both. (In some sense, this index is a mixture of Y_2 and Y_3 in the first set of indices.)

Using an inverse transformation of log (i.e., exponential), we obtain the following set of indices to characterize the cone morphometry:

- **cone size** defined as:

$$Z_1 = e^{F_1}$$

- **scale factor** defined as:

$$Z_2 = e^{F_2}$$

- keep **number of sound seeds** and **number of empty seeds**

$$Z_3 = X_7$$

$$Z_4 = X_8$$

So we have in this set four indices to characterize cone morphometry. The disadvantage with this set of indices is that, as we have just seen, the meaning of the indices is not very crisp. The advantages are, as opposed to the first set of indices, (1) amount of variance being kept in this set of indices is known and acceptable; (2) the correlations among the indices are small (see *Table 4.2* for the correlation matrix).

Table 4.2: Correlation Matrix for the Second Set of Indices

	Z_1	Z_2	Z_3	Z_4
Z_1	1			
Z_2	.01	1		
Z_3	.29	.18	1	
Z_4	.16	.06	−.28	1

This set of indices will be used in our logistic regression.

4.2 MANOVA Assumptions Check

4.2.1 Multivariate Normality

Many multivariate techniques rely on the assumption that the data comes from a multivariate normal distribution. Although moderate departure from normality will not cause us a serious problem in the analysis and inferences because of the very large sample size we have, the quality of our analysis will be improved if we could find suitable transformations on the variates to make them resemble normal distributions more closely.

We thus tried several common transformation on each of the original variates and constructed indices. The best transformations we found are:

- natural logarithm transformation on X_1 – X_4, X_6 and Y_1, Y_2;
- square root transformation on X_7.
- no transformation is found to be able to give a better fit for X_5.

Quantile-quantile plots were applied on both the original variates (left side of the figures) and the transformed variates (right side of the figures). They are presented in *Figure 4.1* to *Figure 4.3.* So we can see that the transformed variates generally gave a better fit to a normal distribution.

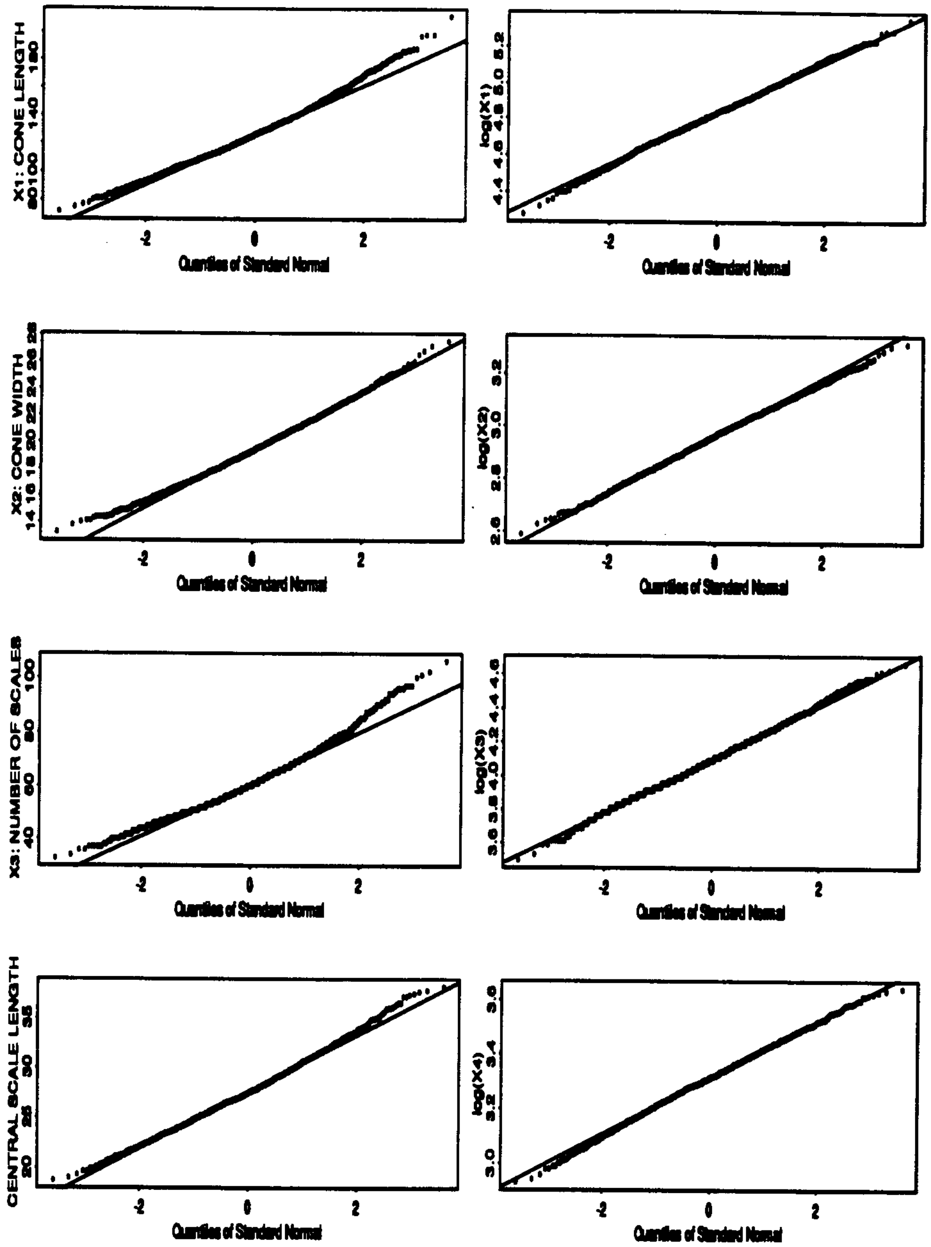

Figure 4.1: Probability Plots for Variates X_1– X_4 and Their Transformations

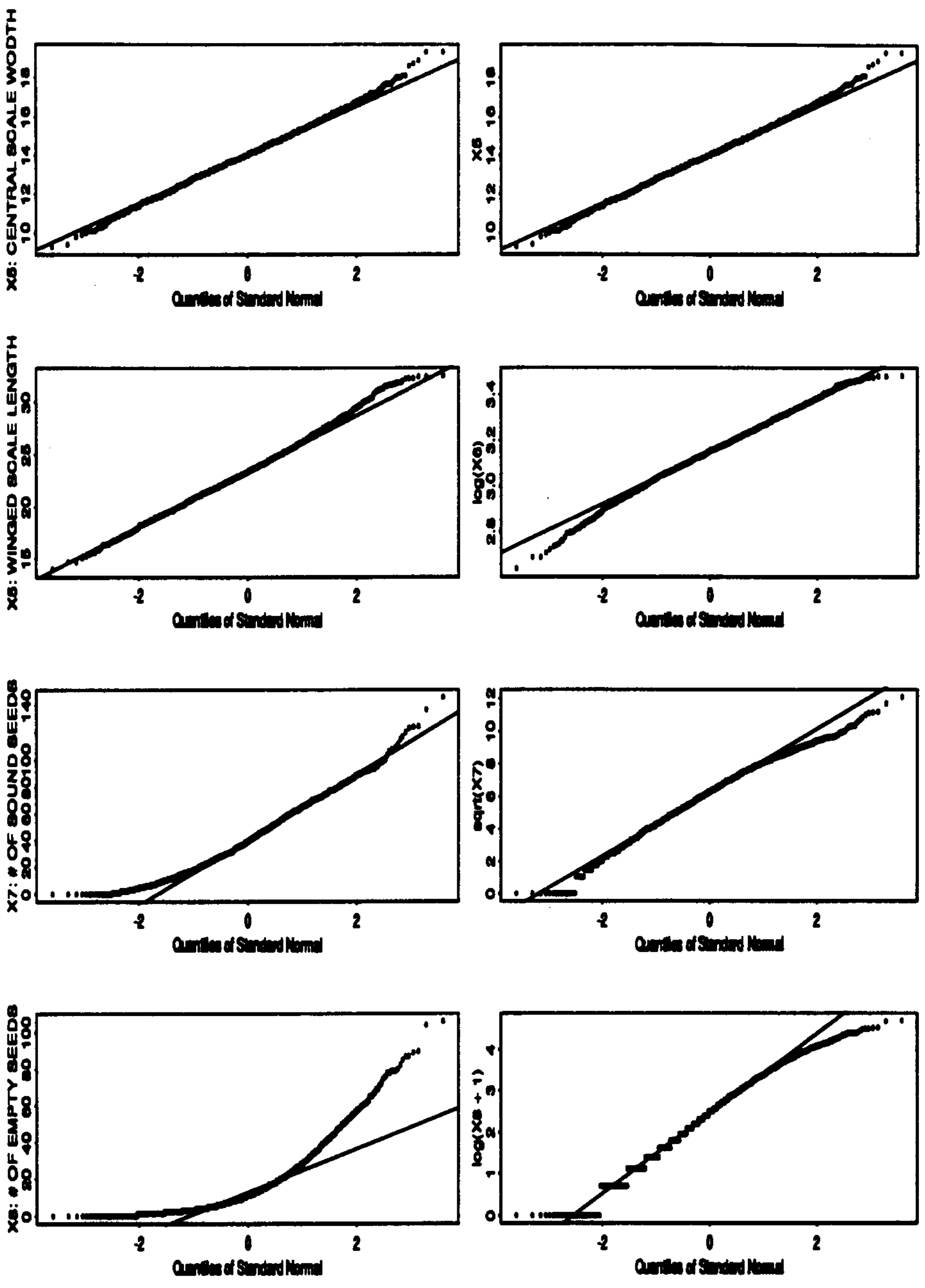

Figure 4.2: Probability Plots for Variates X_5 – X_8 and Their Transformations

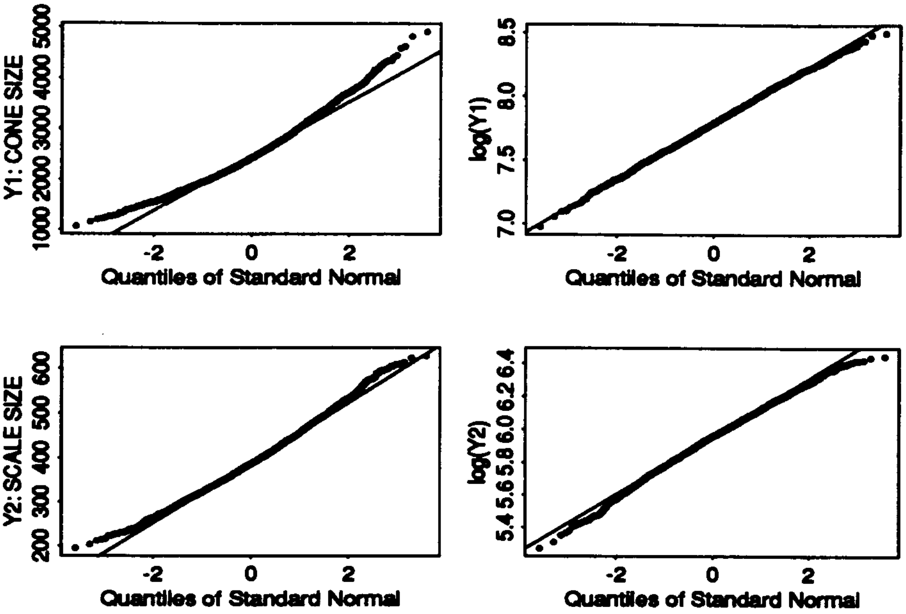

Figure 4.3: Probability Plots for Variates Y_1, Y_2 and Their Transformations

Table 4.3: Modified A^2 Statistics

Variates	test statistics		p-value	
	original	transformed	original	transformed
X_1	7.50	0.88	$< 0.5\%$	(1%, 2.5%)
X_2	1.67	1.03	$< 0.5\%$	(1%, 2.5%)
X_3	12.0	2.01	$< 0.5\%$	$< 0.5\%$
X_4	2.64	0.94	$< 0.5\%$	(1%, 2.5%)
X_6	2.30	2.12	$< 0.5\%$	$< 0.5\%$
X_7	11.3	7.60	$< 0.5\%$	$< 0.5\%$
X_8	142.0	8.06	$< 0.5\%$	$< 0.5\%$
Y_1	14.8	0.28	$< 0.5\%$	$> 50\%$
Y_2	3.27	1.80	$< 0.5\%$	$< 0.5\%$

Goodness-of-fit is a statistical techniques which is associated with the statistical testing of hypothetical models with the data. Anderson-Darling statistics is a member of the group of goodness-of-fit statistics which has come to be known as empirical distribution function (EDF) statistics [13] because the measure the discrepancy between the empirical distribution function of a given sample and the theoretical distribution to be tested. More specifically, it is defined as:

$$W = \int_{-\infty}^{\infty} [F_n(x) - F(x;\theta)]^2 \psi(x) dx$$

where $F(x;\theta)$ is the theoretical distribution under the null hypothesis,

$$\psi(x) = \frac{F(x;\theta)}{1 - F(x;\theta)}$$

is a weight function giving greater importance to observations in the tail than do other EDF statistics, and $F_n(x)$ is the empirical distribution function based on the sample

$$F_n(x) = \frac{\text{number of observations} \leq x}{n}, \quad -\infty < x < \infty$$

Practically, the numerical calculation of Anderson-Darling statistics A^2 is done by the following two steps[2].

1. Calculate $z_i = F(x_{(i)};\theta) \quad i = 1, 2, ..., n$

2. Then the statistics is given by

$$A^2 = -\frac{\sum_{i=1}^{n} (2i - 1)[ln z_i + ln(1 - z_{n+1-i})]}{n} - n$$

The improvement on the strength of fit through transformations is further being confirmed by calculating the modified A^2 statistics[2] and the corresponding p-value for testing the sample is from a normal population. The results are shown in *Table 4.3.*

4.2.2 Homogeneity of Covariance Matrices

Bartlett's modification of the likelihood ratio test is used to test another important assumption of MANOVA model, that covariance matrices are homogeneous across populations. The result shows that for each of the three set of variables, we should reject the null hypothesis at a significant level of 0.1%. This is not surprising because (1) in general, many experimental conditions which leads to higher mean value may also produce responses with larger variances; (2) our sample size is so large that it may contain a lot of evidence suggesting heteroscedasicity.

Thus the choice of a robust test is important among over half a dozen MANOVA tests that are available. Olsen[9] has made a Monte Carlo study concerning robustness of six MANOVA tests. For general protection against departures from normality and from homogeneity of covariance matrices, he has recommended the Pillai V statistics as the most robust MANOVA test. We will briefly introduce the Pillai V test before we modelling the data in the next chapter.

4.3 Heterozygosity Profiles

Based on our definition of heterozygosity, we can calculate the amount of genetic variation for each locus for all ten populations combined. The results are shown in *Table 4.4.* [3]

Table 4.4: Ordered Locus Heterozygosity

Aco(1)	Mdh-1(10)	Got-3(5)	Pgm-2(15)	Ak(6)	G6p(8)
.79	.73	.61	.60	.37	.36
Hex(2)	Pgi-2(18)	Mdh-2(11)	Pgm-1(14)	Mpi(13)	Pgi-1(17)
.36	.26	.14	.14	.13	.08
Got-2(4)	Fum(7)	Got-1(3)	Idh(9)	Mdh-3(12)	6pg(16)
.08	.04	.01	0	0	0

We can also calculate the amount of genetic variation for each locus across populations. The results are presented in *Figure 4.2*, where the profiles are arranged according the magnitude of heterozygosity in the *Table 4.4*.

On some loci, such as locus 2 (Hex) and Locus 18 (Pgi-2), locus heterozygosity varies greatly across populations; while on some other loci, all 300 sampled trees are almost homozygous, e.g., Locus 3 (Got-1) besides the obvious ones Locus 9 (Idh), Locus 12 (Mdh-3) and Locus 16 (6pg).

[3] The numbers in the brackets identify the loci in the original file.

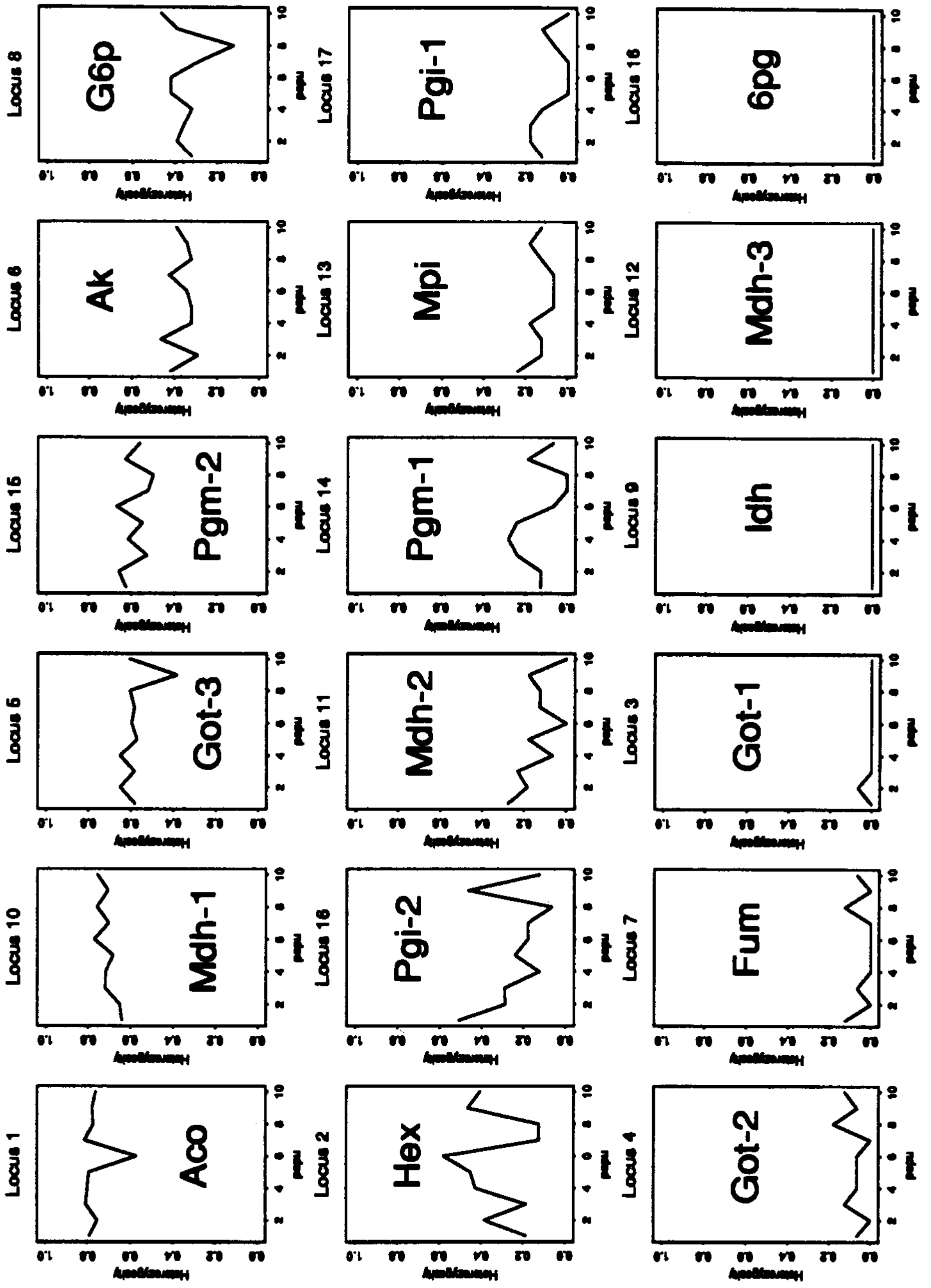

Figure 4.4: Heterozygosity Profiles

Chapter 5

Data Modelling

5.1 Morphometric Variability

In the previous chapter, we constructed two sets of indices to characterize cone morphometry, e.g., the Y-set and Z-set, from the original eight morphometric measurements, X-set. Each of the three sets of indices has its advantage and disadvantage. We will use them for different modelling purposes.

5.1.1 MANOVA

Recall the MANOVA model we specified in Chapter 3 is:

$$\mathbf{X}_{ijr} = \boldsymbol{\mu} + \boldsymbol{\alpha}_i + \boldsymbol{\beta}_{j(i)} + \boldsymbol{E}_{r(ij)} \qquad \sum_{i=1}^{10} \boldsymbol{\alpha}_i = \mathbf{0} \qquad \begin{cases} i = 1,\ 2,\ \ldots,\ 10 \\ j = 1,\ 2,\ \ldots,\ 30 \\ r = 1,\ 2,\ \ldots,\ 10 \end{cases}$$

According to this model, a vector of observations may be decomposed as shown in the following equation:

$$\mathbf{X}_{ijr} = \bar{\mathbf{X}} + (\bar{\mathbf{X}}_{\mathrm{i}} - \bar{\mathbf{X}}) + (\bar{\mathbf{X}}_{\mathrm{j(i)}} - \bar{\mathbf{X}}_{\mathrm{i}}) + (\bar{\mathbf{X}}_{\mathrm{r(ij)}} - \bar{\mathbf{X}}_{\mathrm{j(i)}})$$

i.e., any observation is an summation of the following terms: (1) overall sample mean $\bar{\mathbf{X}}$; (2) estimated population effect $(\bar{\mathbf{X}}_{\mathbf{i}} - \bar{\mathbf{X}})$; (3) estimated tree effect (inside each population) $(\bar{\mathbf{X}}_{\mathbf{j(i)}} - \bar{\mathbf{X}}_{\mathbf{i}})$; and (4) estimated cone effect or the residual term $(\bar{\mathbf{X}}_{\mathbf{r(ij)}} - \bar{\mathbf{X}}_{\mathbf{j(i)}})$.

This leads to a decomposition of the sum of squares and cross-products matrix

$$\sum_{i=1}^{10}\sum_{j=1}^{30}\sum_{r=1}^{10}(\bar{\mathbf{X}}_{\mathbf{r(ij)}} - \bar{\mathbf{X}})(\bar{\mathbf{X}}_{\mathbf{r(ij)}} - \bar{\mathbf{X}})^T$$

in the following table:

Table 5.1: MANOVA Table for Comparing Population Mean Vectors

Source of Variation	Matrix of Sum of Squares and Cross Products(SS&CP)	Degrees of Freedom
Popu	$\mathbf{P} = 300\sum_{i=1}^{10}(\bar{\mathbf{X}}_{\mathbf{i}} - \bar{\mathbf{X}})(\bar{\mathbf{X}}_{\mathbf{i}} - \bar{\mathbf{X}})^T$	9
Tree	$\mathbf{T} = 10\sum_{i=1}^{10}\sum_{j=1}^{30}(\bar{\mathbf{X}}_{\mathbf{j(i)}} - \bar{\mathbf{X}}_{\mathbf{i}})(\bar{\mathbf{X}}_{\mathbf{j(i)}} - \bar{\mathbf{X}}_{\mathbf{i}})^T$	290
Error	$\mathbf{E} = \sum_{i=1}^{10}\sum_{j=1}^{30}\sum_{r=1}^{10}(\bar{\mathbf{X}}_{\mathbf{r(ij)}} - \bar{\mathbf{X}}_{\mathbf{j(i)}})(\bar{\mathbf{X}}_{\mathbf{r(ij)}} - \bar{\mathbf{X}}_{\mathbf{j(i)}})^T$	2700
Total	$\mathbf{P}+\mathbf{T}+\mathbf{E} = \sum_{i=1}^{10}\sum_{j=1}^{30}\sum_{r=1}^{10}(\bar{\mathbf{X}}_{\mathbf{r(ij)}} - \bar{\mathbf{X}})(\bar{\mathbf{X}}_{\mathbf{r(ij)}} - \bar{\mathbf{X}})^T$	2999

One test of $H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \ldots = \boldsymbol{\alpha}_{10} = \mathbf{0}$ was proposed by Pillai[10] and known to have several optimal properties including robustness against departures from the usual population model, especially non-normality and heteroscedasicity.

The statistic, usually denoted as V, is defined to be:

$$V = tr\mathbf{P}(\mathbf{P}+\mathbf{E})^{-1}$$

Under the null hypothesis, $(N-r)V$ is distributed as a χ^2_{gu} for large sample size N. Here u is the number of variates in the random vector; and g is 9, the number of populations less 1; r is 11, the number of populations plus 1.

Consequently, we reject H_0 at significant level α if

$$(N-r)V > \chi^2_{gu}(\alpha)$$

where $\chi^2_{gu}(\alpha)$ is the upper (100α)th percentile of a chi-square distribution with gu degrees of freedom.

MANOVA Results on the Original Variates: X-set

The random vector is $(lnX_1, lnX_2, lnX_3, lnX_4, X_5, lnX_6, \sqrt{X_7}, ln(X_8+1))$.

$$\mathbf{E} = \begin{pmatrix} 30.7 & 15.0 & 28.8 & 10.9 & 152 & 11.6 & 174 & 45.3 \\ 15.0 & 13.4 & 16.9 & 7.19 & 111 & 7.93 & 124 & 31.2 \\ 28.9 & 16.9 & 43.7 & 7.95 & 131 & 8.24 & 176 & 52.3 \\ 10.9 & 7.19 & 7.95 & 10.0 & 98.9 & 11.4 & 79.0 & 18.2 \\ 152 & 111 & 131 & 98.9 & 2340 & 114 & 1550 & 343 \\ 11.6 & 7.93 & 8.24 & 11.4 & 114 & 14.4 & 89.7 & 21.9 \\ 174 & 124 & 176 & 79.0 & 1550 & 89.7 & 3820 & -13.7 \\ 45.3 & 31.2 & 52.3 & 18.2 & 343 & 21.9 & -13.7 & 940 \end{pmatrix}$$

$$\mathbf{P} = \begin{pmatrix} 3.53 & -0.143 & 0.0468 & 3.37 & 29.1 & 3.69 & -0.718 & 18.5 \\ -0.143 & 4.32 & 1.62 & -1.13 & 6.72 & -0.839 & -9.82 & -288 \\ 0.0468 & 1.62 & 4.11 & -1.45 & 11.0 & -1.15 & -22.7 & 3.41 \\ 3.37 & -1.13 & -1.45 & 4.69 & 33.2 & 4.81 & -4.12 & 25.3 \\ 29.1 & 6.72 & 11.0 & 33.2 & 493 & 39.7 & -346 & 349 \\ 3.69 & -0.839 & -1.15 & 4.81 & 39.7 & 5.27 & -12.1 & 27.8 \\ -0.718 & 0.305 & -22.7 & -4.12 & -346 & -12.1 & 1080 & -288 \\ 18.5 & -9.82 & 3.41 & 25.3 & 349 & 27.8 & -288 & 349 \end{pmatrix}$$

$$V = 1.78 \quad u = 8 \;\; g = 9 \;\; r = 11 \quad p - value = .0001$$

MANOVA Results on the First Set of Indices: Y-set

The random vector is $(lnY_1, lnY_2, lnY_3, \sqrt{Y_4}, ln(Y_5 + 1))$.

$$\mathbf{E} = \begin{pmatrix} 74.2 & 37.1 & 45.8 & 299 & 76.6 \\ 37.1 & 36.6 & 17.3 & 191 & 42.9 \\ 45.8 & 17.3 & 43.7 & 176 & 52.3 \\ 299 & 191 & 176 & 3820 & -13.7 \\ 76.6 & 42.9 & 52.3 & -13.7 & 940 \end{pmatrix}$$

$$\mathbf{P} = \begin{pmatrix} 7.57 & 4.78 & 1.67 & -0.413 & 8.75 \\ 4.78 & 12.0 & -0.681 & -29.4 & 50.5 \\ 1.67 & -0.681 & 4.11 & -22.7 & 3.41 \\ -0.413 & -29.4 & -22.7 & 1080 & -288 \\ 8.75 & 50.5 & 3.41 & -288 & 349 \end{pmatrix}$$

$$V = 1.11 \quad u = 5 \quad g = 9 \quad r = 11 \quad p - value = .0001$$

MANOVA Results on the Second Set of Indices: Z-set

The random vector is $(F_1, F_2, \sqrt{Z_3}, ln(Z_4 + 1))$.

$$\mathbf{E} = \begin{pmatrix} 1380 & 425 & 1300 & 325 \\ 425 & 1120 & 618 & 194 \\ 1300 & 618 & 3820 & -13.7 \\ 325 & 194 & -13.7 & 940 \end{pmatrix}$$

$$\mathbf{P} = \begin{pmatrix} 235 & -151 & -113 & 185 \\ -151 & 442 & -34.1 & -182 \\ -113 & -34.1 & 1080 & -288 \\ 185 & -182 & -288 & 349 \end{pmatrix}$$

$$V = .99 \quad u = 4 \quad g = 9 \quad r = 11 \quad p - value = .0001$$

Conclusions from MANOVA Analysis

For each of the three sets of indices, there is strong evidence that the population mean vectors are not the same across all ten populations. This suggests that there is a significant multivariate mean difference between at least two populations. This also suggests that there is a significant mean difference among at least one component of each set of indices and we should make the comparisons at a univariate level.

5.1.2 ANOVA

According to the model specifications in Chapter 3, the ANOVA model is essentially a part of the MANOVA model because each component of the observation $\bar{\mathbf{X}}_{\mathbf{r(ij)}}$ must satisfy the univariate model.

The following three tables show the ANOVA results from the specified model. For each variate, the population effect is highly significant. The null hypothesis $H_0 : \alpha_1 = \alpha_2 = ... = \alpha_{10} = 0$ is rejected in favor of $H_A : \alpha_i \neq 0$ for at least one $i = 1, 2, ...10$. There is at least one significant mean difference among the ten populations.

The second hypothesis of interest $H_0 : \sigma^2_\beta = 0$ is also rejected in favor of $H_A : \sigma^2_\beta > 0$ for each variate. There is significant variation on each variate among the trees within populations.

Table 5.2: ANOVA Tables for X-set Variates

Variate	Sources of Variation	Degrees of Freedom	Mean Square	F	p-value
lnX_1	Popu	9	.393	3.9	0.0001
	Tree	290	.102	8.9	0.0001
	Error	2700	.011		
lnX_2	Popu	9	.481	8.6	0.0001
	Tree	290	.056	11.2	0.0001
	Error	2700	.005		
lnX_3	Popu	9	.457	4.2	0.0001
	Tree	290	.108	6.7	0.0001
	Error	2700	.016		
lnX_4	Popu	9	.522	8.3	0.0001
	Tree	290	.063	16.8	0.0001
	Error	2700	.004		
X_5	Popu	9	54.8	6.4	0.0001
	Tree	290	8.49	9.8	0.0001
	Error	2700	.866		
lnX_6	Popu	9	.586	7.1	0.0001
	Tree	290	.083	15.4	0.0001
	Error	2700	.005		
$\sqrt{X_7}$	Popu	9	102	5.2	0.0001
	Tree	290	19.6	13.8	0.0001
	Error	2700	1.42		
$ln(X_8 + 1)$	Popu	9	38.8	8.5	0.0001
	Tree	290	4.58	13.2	0.0001
	Error	2700	.348		

Table 5.3: ANOVA Tables for Y-set Variates

Variate	Sources of Variation	Degrees of Freedom	Mean Square	F	p-value
lnY_1	Popu	9	.841	3.8	0.0001
	Tree	290	.219	8.0	0.0001
	Error	2700	.028		
lnY_2	Popu	9	1.34	8.8	0.0001
	Tree	290	.152	11.2	0.0001
	Error	2700	.014		
lnY_3	Popu	9	.457	4.2	0.0001
	Tree	290	.108	6.7	0.0001
	Error	2700	.016		
$\sqrt{Y_4}$	Popu	9	102	5.2	0.0001
	Tree	290	19.6	13.8	0.0001
	Error	2700	1.42		
$ln(Y_5 + 1)$	Popu	9	38.8	8.5	0.0001
	Tree	290	4.58	13.2	0.0001
	Error	2700	.348		

Table 5.4: ANOVA Tables for Z-set Variates

Variate	Sources of Variation	Degrees of Freedom	Mean Square	F	p-value
lnZ_1	Popu	9	26.1	5.5	0.0001
	Tree	290	4.74	9.2	0.0001
	Error	2700	.514		
lnZ_2	Popu	9	49.2	10.0	0.0001
	Tree	290	4.92	11.8	0.0001
	Error	2700	.418		
$\sqrt{Z_3}$	Popu	9	102	5.2	0.0001
	Tree	290	19.6	13.8	0.0001
	Error	2700	1.42		
$ln(Z_4 + 1)$	Popu	9	38.8	8.5	0.0001
	Tree	290	4.58	13.2	0.0001
	Error	2700	.348		

Since differences do exist between populations we proceed to find out how the ten populations differ from each other.

For each variate, based on the estimated mean, we can obtain the exact ranking for the count variates, i.e., number of scales(Y_3), number of sound seeds(Y_4) and number of empty seeds(Y_5). The results are shown in the following table:

Table 5.5: Ranking of Variates Y_3, Y_4 and Y_5

Variate	population									
	1	2	3	4	5	6	7	8	9	10
Y_3	7	3	9	8	2	4	5	1	6	10
Y_4	5	1	8	4	6	3	2	9	10	7
Y_5	5	9	4	2	6	10	8	1	3	7

For variates cone size(Y_1 and Z_1) and scale size(Y_2 and Z_2), because of two different ways of indices construction, the rankings are not entirely consistent. However, the following conclusions are consistent with both set of indices:

- **Cone Size**: population two and eight have the largest cone sizes, while population ten and four have the smallest cone size.
- **Scale Size:** population two and three have the largest scale sizes, while population ten has the smallest scale size.

As a summary, the estimated mean vector of all the indices are plotted using Chernoff's face techniques. See *Figure 5.1.* [1] Two populations, population 10 and population 2, are particularly noteworthy. Population 10, from Abitibi, has the smallest cone size, scale size, as well as number of scales. The number of sound seeds and the number of empty seeds are also small compared to

[1] Ten other features of Chernoff's Face that can be used to represent extra variates are: width of mouth; location, separation, angle, shape and width of eyes; location of pupil; location, angle and width of eyebrow. This graphical technique is most useful in grouping subjects based on multivariate observations.

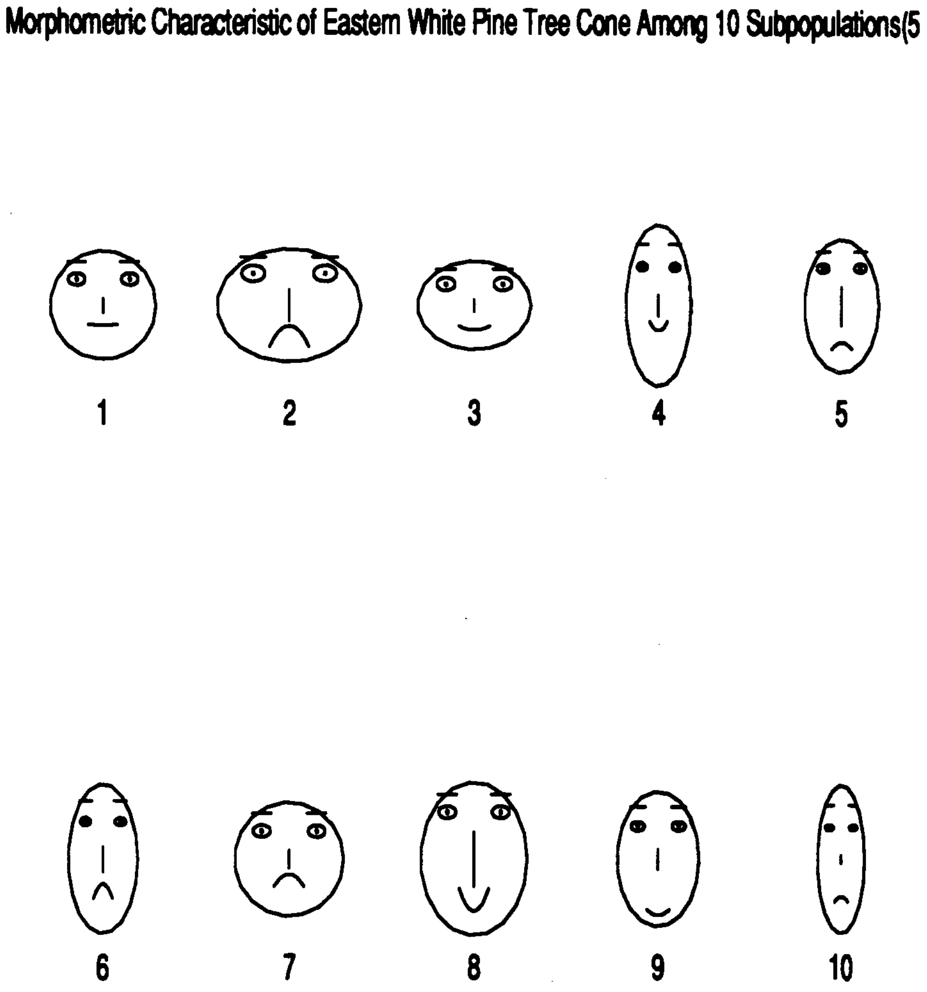

Figure 5.1: The indices are constructed subjectively. The area of face is proportional to the cone size Y_1, shape of face is for scale size Y_2 with population 2 being the largest and population 10 the smallest, length of nose is proportional to the number of scales Y_3, distance between the mouth and the nose is inversely proportional to the number of sound seeds Y_4, and curvature of smile is for the number of empty seeds Y_5 with population 8 being the largest and population 6 the smallest. Based on mean levels estimated from the ANOVA model.

that of other populations as they are both ranked at 7. Population two, from Ottawa Valley, also distinguishes itself from others by the facts that its cone has the largest cone size, scale size as well as the number of sound seeds. Its number of scales is relatively large (ranked at 3) and its number of empty seeds is the second smallest.

Furthermore, we find from this picture that populations 4, 6, 9 and 10, which are among the furthest north fringe of Eastern White Pine habitat, all appear to have small cone size as well as scale size. [2] This may suggest that temperature may be one of the environmental factors that affect the size of Eastern White Pine cone, the most important aspect of cone morphometry. However, there are no other spatial trends observed from this summary picture.

Finally, we present the estimate of variation components from each of the three sources: variability due to the cone, variability due to the tree and the variability found between populations. We shall use Z-set indices for this purpose because the correlations among variates are the smallest in this set. The result of the estimation is shown in the following table:

Table 5.6: Percentage of Variance Components

sources	variables			
	lnZ_1	lnZ_2	Z_3	Z_4
popu	7.8	14.7	10.2	13.2
tree	45.9	47.7	53.7	50.9
cone	46.3	37.6	36.1	35.9

From the table, we can conclude that roughly 10% of the total morphometric variability is found between Eastern White Pine populations. The rest is found within populations and roughly half of this variation is due to the tree variation and half is due to the cone variation.

[2] This is apparent if we could "move" the faces onto the map in *Figure 1.1*, as I did in my Defense.

5.2 Genetic Variability

5.2.1 Genetic Variability Across Populations

In the preliminary analysis of the genetic data set in last chapter, we calculated the heterozygosity of each locus across the ten populations. In this section we shall calculate the heterozygosity for each population as an index for the comparison of genetic variabilities.

We shall assume that these 18 locis studied are a simple random sample of all loci inherited in Eastern White Pine. Then we can estimate the heterozygosity for each population as the average locus heterozygosity in that population.

$$H = \frac{h_1 + h_2 + ... + h_{18}}{18}$$

Based on the above formula, heterozygosity for each of the ten populations is calculated and we find that populations 1 and 3 have the largest genetic variability, while populations 7 and 8 have the smallest genetic variability.

Table 5.7: Heterozygosity among Populations

H_{s_1}	H_{s_2}	H_{s_3}	H_{s_4}	H_{s_5}	H_{s_6}	H_{s_7}	H_{s_8}	H_{s_9}	$H_{s_{10}}$
.279	.263	.272	.260	.254	.240	.213	.216	.272	.247

From the above table, it seems that heterozygosity exhibits a spatial trend among the ten populations: the amount of genetic variation in Eastern White Pine in Ottawa valley is generally greater than that in St-Lawrence lowland. We proceed to test this hypothesis.

Viewing heterozygosity as a population parameter and again assuming that these eighteen loci studied are a simple random sample from all the loci in this species, we could combine populations 1 to 4 (from Ottawa valley) together to calculate the heterozygosity on each of the eighteen loci and use them as our observations. Similarly, we could obtain another set of observations by combining populations 5 to 8 (from St-Lawrence lowland) together. The following table shows the locus heterozygosity for populations 1 to 4 combined (Ottawa Valley), populations 5 to 8 combined (St-Lawrence lowland) and the differences between these two.

Table 5.8: Heterozygosity of Two Combined Populations

Locus	Ottawa	St-Lawrence	Difference
1	0.802	0.774	0.028
2	0.304	0.361	−0.057
3	0.017	0	0.017
4	0.064	0.080	−0.015
5	0.633	0.608	0.025
6	0.388	0.359	0.029
7	0.049	0.033	0.016
8	0.349	0.329	0.020
9	0	0	0
10	0.702	0.744	−0.042
11	0.195	0.111	0.084
12	0	0	0
13	0.168	0.080	0.088
14	0.196	0.081	0.116
15	0.612	0.568	0.045
16	0	0	0
17	0.153	0.017	0.136
18	0.317	0.169	0.148

Let h_o and h_s denote the heterozygosity of Eastern White Pine populations from Ottawa valley and St-Lawrence lowland, respectively. The hypothesis we want to test is the null $H_0 : h_o - h_s = 0$ against the alternative $H_1 : h_o - h_s > 0$. A paired t-test is an appropriate test for this purpose.

Since t-test is not very robust against departure from normality, we applied the modified A^2 statistics[2] first to test the assumption that these eighteen differences are from a normal distribution. We obtained a p-value of 12.5% (the calculated statistic is 0.590). The differences between locus heterozygosity of these two combined populations can be regarded as from a normal distribution.

The calculated t-statistic is 2.604 yielding a p-value of 0.9%. This strongly suggests that the heterozygosity of the four combined populations from Ottawa valley is higher than that from St-Lawrence lowland. The only caution in accepting this result is that we have made a somewhat unrealistic assumption that these eighteen loci are a simple random sample from all the loci inherited in Eastern White Pine.

A final speculation: based on the two spatial trends we observed on cone morphometry and heterozygosity respectively, there is a suggestion that environmental factors such as temperature may be more important in affecting cone morphometry while geographical proximity is more important in determining the amount of genetic variation in Eastern White Pine.

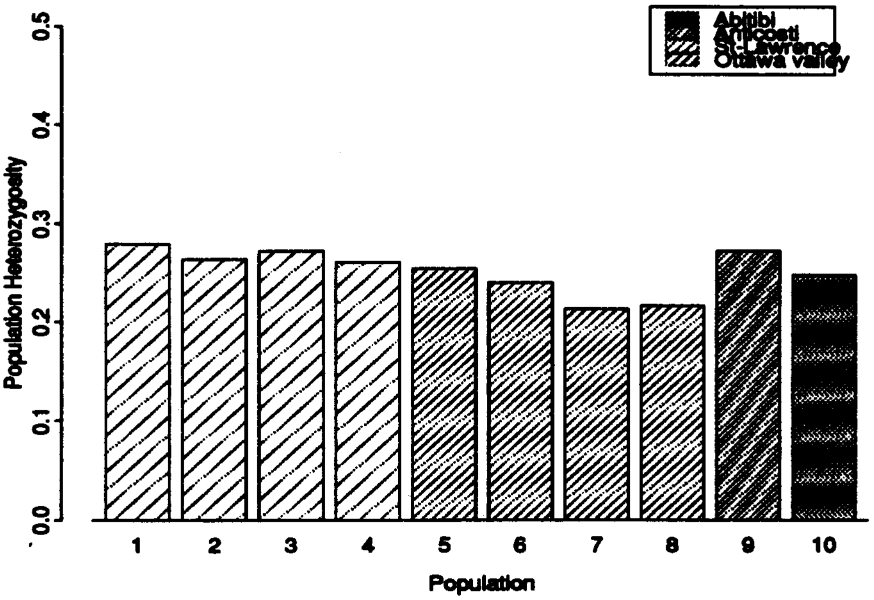

Figure 5.2: Heterozygosity across Ten Geographical Populations

5.2.2 Genetic Variability Between and Among Populations

An average genetic variability of the ten populations can be calculated as

$$\bar{H}_s = \sum_{i=1}^{10} H_{s_i}/10 = .252$$

And the total genetic variability of the whole population is calculated to be $H_t = .261$. Thus a measure of the amount of differentiation among the populations is defined and calculated to be

$$G_{st} = \frac{H_t - \bar{H}_s}{H_t} = .035$$

The interpretation is that only 3.5% of total genetic variability is found between populations in Eastern White Pine.

The amount of heterozygosity for each of the ten populations are shown in *Figure 5.2* so we can see that there is actually not much variation across populations.

5.3 Relationship Between Genetic Traits and Cone Morphometry

5.3.1 Initial Expository Analysis

From the ANOVA on the Z-set of cone morphometric indices, we obtained a typical cone morphometry for every tree, i.e.,

$$\mathrm{E}(Z_{ij}{}^{(p)}) = \mu + \alpha_i + \hat{\beta}_{j(i)} \quad \begin{cases} i = 1,\, 2,\, \ldots,\, 10 \\ j = 1,\, 2,\, \ldots,\, 30 \\ p = 1,\, 2,\, \ldots,\, 4 \end{cases}$$

As shown in Chapter 4, this set of indices retained most of cone morphometry information as contained in the original eight variates and their intercorrelations are generally small. We shall use this set of four explanatory variables in the polychotomous logistic regression model proposed in Chapter 3 (x in the model). The dependent variable y in the model is the genotype, a categorical variable which assumes values $1, 2, ..., I$, where I is the total number of genotypes observed on a particular locus. Notice that by treating it this way, we have implicitly assumed that the genotypes of the loci studied are known, i.e., all genotypes in the population are observed in the sample.

Thus the model can be re-written as:

$$P[y = n] = \frac{exp(\alpha_n + \beta_n \mathbf{z})}{1 + \sum_{n=1}^{I-1} exp(\alpha_n + \beta_n \mathbf{z})} \qquad n = 1, 2, ...I - 1$$

and

$$P[y = I] = 1 - \sum_{n=1}^{I-1} P[y = i]$$

where $\beta_n = (\beta_{n1}, \beta_{n2}, \beta_{n3}, \beta_{n4})$ and $\mathbf{z} = (z_1, z_2, z_3, z_4)^T$ as defined in Section 4.1.2.

This model is built on a locus-to-locus base. However, since we have already known there are three loci that are degenerate, i.e., Locus 9 (Idh), Locus 12

(Mdh3) and Locus 16 (6pg) are all homozygous of the same genotype, we actually need to run the regression 15 times.

For most of the 15 loci on which we performed logistic regression analysis, we found that the p-value for the likelihood ratio improvement from a null model to a one-variate model is quite high, which implies that none of the four variates can be judged as adequate explanatory variables. So there is hardly any chance that there will be a relationship between the cone morphometry and the genotypes of these loci.

However, there are three cases (Locus 5, 14 and 15) where the above p-value is slightly less than 5% which is mild evidence for selecting the variate into the model. Interestingly, the variable entering the model in all the three cases is the same, the `scale factor` Z_2. The estimated model parameters are given in the following three tables.

1. **locus 5:** the variate Z_2 (scale factor) enters the model, giving a p-value of 4.6% for testing the deviance drop.

Table 5.9: Regression Result: Locus 5 (Got-3)

genotype	code(i)	$\hat{\alpha}_i$	se($\hat{\alpha}_i$)	$\hat{\beta}_{i2}$	se($\hat{\beta}_{i2}$)
AA	1	$-.33$	(.17)	$-.54$	(.22)
AB	2	.67	(.14)	$-.18$	(.18)

2. **locus 14:** the variate Z_2 (scale factor) enters the model, giving a p-value for the deviance drop of 5.0%.

Table 5.10: Regression Result: Locus 14 (Pgm-1)

genotype	code(i)	$\hat{\alpha}_i$	se($\hat{\alpha}_i$)	$\hat{\beta}_{i2}$	se($\hat{\beta}_{i2}$)
AB	2	7.55	(5.25)	5.20	(2.82)
BB	4	11.57	(5.23)	5.03	(2.76)
BC	5	8.71	(5.24)	5.21	(2.78)

3. **locus 15:** the variate Z_2 (scale factor) enters the model, yielding a p-value of 3.9% for the deviance drop.

Table 5.11: Regression Result: Locus 15 (Pgm-2)

genotype	code(i)	$\hat{\alpha}_i$	se($\hat{\alpha}_i$)	$\hat{\beta}_{i2}$	se($\hat{\beta}_{i2}$)
AA	1	−4.16	(2.36)	3.10	(1.56)
AB	2	.55	(.41)	−.12	(.49)
AC	3	−.11	(.47)	.51	(.58)
BB	4	2.79	(.34)	.64	(.41)
BC	5	2.26	(.35)	.63	(.42)

5.3.2 Signal Amplification

Since Z_2 is mostly concerned with morphometric measurements of the scales on the cone, the above finding suggests that these three loci would most probably depend on the scale morphometry. However, the signal is very weak. This could be due to the masked effect of the principal components, e.g., we may lose some important information by using the principal components instead of the original variates as the regressors; or, if the hypothesis that these loci are depend on the scale measurements rather than other aspects of cone morphometry is true, using principal components could diminish this relationship because they are a mixture of all aspects of cone morphometry.

Based on this suggestion, we performed further logistic regression analysis by using the original scale measurements X_3, X_4, and X_5. We deleted X_6 again because it has a correlation of almost 1 with X_4. The model used here is the same polychotomous logistic regression model. We re-write it as:

$$P[y = i] = \frac{exp(\alpha_i + \beta_i \mathbf{x})}{1 + \sum_{i=1}^{I-1} exp(\alpha_i + \beta_i \mathbf{x})} \qquad i = 1, 2, ...I - 1$$

and

$$P[y = I] = 1 - \sum_{i=1}^{I-1} P[y = i]$$

where $\mathbf{x} = (X_3, X_4, X_5)^T$.

Not surprisingly, our logistic regression results have been improved significantly. For each of the three cases, one scale measurement is able to enter the model. The following table shows the amplification of signal as justified by the improvement in p-values.

Table 5.12: Summary of Signal Amplifications

Locus	Z-set		$(X_3, X_4, X_5)^T$	
	variate entered	p-value	variate entered	p-value
5	Z_2	4.6%	X_3	3.8%
14	Z_2	5.0%	X_3	2.0%
15	Z_2	3.9%	X_4	0.6%

1. **Locus 5:** the variate, number of scales, enters the model with the corresponding p-value for the the likelihood ratio improvement from a null model to a one-variate model being 3.8%.

Table 5.13: Regression Result: Locus 5 (Got-3)

genotype	code(i)	$\hat{\alpha}_i$	se($\hat{\alpha}_i$)	$\hat{\beta}_{i1}$	se($\hat{\beta}_{i1}$)
AA	1	3.73	(1.62)	−.067	(.027)
AB	2	2.03	(1.28)	−.022	(.021)

2. **Locus 14:** the variate, number of scales, enters the model with the corresponding p-value for the the likelihood ratio improvement from a null model to a one-variate model being 2.0%.

Table 5.14: Regression Result: Locus 14 (Pgm-1)

genotype	code(i)	$\hat{\alpha}_i$	se($\hat{\alpha}_i$)	$\hat{\beta}_{i1}$	se($\hat{\beta}_{i1}$)
AB	2	−35.39	(33.5)	0.76	(0.74)
BB	4	−37.14	(33.3)	0.86	(0.74)
BC	5	−40.81	(33.4)	0.87	(0.74)

3. **Locus 15:** the variate, central scale length, enters the model with the corresponding p-value for the the likelihood ratio improvement from a null model to a one-variate model being 0.6%.

Table 5.15: Regression Result: Locus 15 (Pgm-2)

genotype	code(i)	$\hat{\alpha}_i$	se($\hat{\alpha}_i$)	$\hat{\beta}_{i2}$	se($\hat{\beta}_{i2}$)
AA	1	43.96	(25.8)	−1.90	(1.15)
AB	2	−2.43	(4.27)	0.11	(0.15)
AC	3	−1.47	(5.12)	0.045	(0.18)
BB	4	6.97	(3.59)	−0.15	(0.12)
BC	5	5.11	(3.66)	−0.11	(0.12)

5.3.3 Conclusion and Discussion

From the above analyses, we could conclude that, although most of the loci studied have no relationship with the cone morphometry, there is some evidence suggesting that the genotypes on Locus 5 (Got-3), Locus 14 (Pgm-1) and Locus 15 (Pgm-2) might depend on the cone morphometry with the most important aspects being the scale measurements of the cone.

However, in drawing the above conclusion, several cautionary notes should be sound.

1. The signal is too weak and it prevent us from drawing any more concrete conclusions. Although the smallest p-value we got is 0.6%, it was obtained only after running 15 logistic regressions. By the Bonferroni Inequality, the overall significant level will lie somewhere between 0.6% and $15 \times 0.6\% = 9.0\%$.

2. From previous analysis, we know that there is not much variation in the genetic structure of Eastern White Pine. Also, the morphometric variability among populations is also small. The homogeneous nature of

these two data sets may mask any relationship between these two and thus make statistical inferences more difficult.

3. We still feel our analysis is worthwhile because this study is exploratory. Forward selection regression is an appropriate technique to use at this stage. The fact that in our initial logistic regressions on the Z-set, the variable entered the models on all these three loci is the same may help to reinforce that our marginally significant regression results are not just happened purely by chance.

4. We feel our conclusion is meaningful and useful provided we regard it as expository and temporary and use it as a guidance and reference for future studies.

Chapter 6

Summary of Results

6.1 Morphometric Variability

Objective

To describe variability in Eastern White Pine cone morphometry within and between populations.

Results

1. All the ten populations show a statistically significant difference for each of the cone morphometric measurements.

2. In terms of indices cone size and scale size, all the ten populations also show a statistically significant difference. Eastern White Pine situated at the northern brink of its natural population appears to have a relatively small cone.

3. Although cone morphometry of Eastern White Pine varies from population to population, variation from this source only accounts for a small

portion (roughly 10%) of the total morphometric variability; i.e., most of the cone morphometric variation is from within populations.

4. For the within population cone morphometric variation, the amount of variability due to tree-to-tree variation is roughly equal to that due to cone-to-cone variation.

5. In some sense, the cone morphometric difference between Eastern White Pine trees are greater than the average cone morphometric difference between geographical populations.

6. Eastern White Pine cones of population 2 (from Ottawa Valley) are noteworthy in that they have the largest average cone size and and scale size and the most number of sound seeds. The average number of scales on these cones is relatively large (ranks 3 among all ten populations) and the average number of empty seeds is the second smallest.

7. Population 10 (from Abitibi) also distinguishes itself by the fact that its cone has the smallest average cone size, scale size and number of scales. Both its number of sound seeds and number of empty seeds are also on the small side (rank 7) among all populations.

6.2 Genetic Variability

Objective

To analyze genetic variability within and among Eastern White Pine populations.

Results

1. Genotypic dominance is a feature of most loci studied.

2. Distributions of genotype on the loci studied are homogeneous across populations.

3. On some loci studied, there is little or no genetic variation; however, some other loci studied exhibit very large genetic variation.

4. There is a spatial trend in genetic variation: the amount of genetic variation in populations 1 to 4 (from Ottawa Valley) is generally greater than that in populations 5 to 8 (from the St.-Lawrence lowlands).

5. Only 3.5% of the total genetic variation is due to the population differences of Eastern White Pine, i.e., the vast amount of genetic variation is found within populations.

6.3 Relationship between Genetic Traits and Cone Morphometry

Objective

To investigate whether genetic characteristics depend on cone morphometry.

Results

1. On most loci studied, there is no statistical evidence for the presence of a relationship between genetic traits and cone morphometry.

2. On locus 5 (Aco), Locus 14 (Pgm-1) and Locus 15 (Pgm-2), however, there exists mild evidence showing that the genotype distributions do depend on some characteristics of the cone morphometry, with the most important aspects being the morphometric measurements of the scale on the cone.

3. Our investigation is intended to be expository rather than authoratative. The result should be useful as a reference and guidance for further study.

Appendix A

Data Structure

A.1 Morphometric Data Set

popu	tree	cone	X1	X2	X3	X4	X5	X6	X7	X8
1	1	1	148	19.6	67	29.3	14.3	24.8	42	34
1	1	2	137	19.3	60	28.1	14.5	24.0	45	16
1	1	3	126	17.5	53	27.1	12.9	23.5	53	4
1	1	4	143	19.0	59	28.3	14.7	24.2	16	47
1	1	5	149	19.6	65	27.7	13.3	23.9	51	20
1	1	6	118	16.6	50	25.2	13.1	22.0	36	17
1	1	7	129	18.5	51	27.6	13.7	23.1	23	27
1	1	8	133	16.6	53	25.3	12.8	21.5	26	22
1	1	9	124	18.3	52	27.1	13.9	22.6	50	0
1	1	10	116	18.2	46	27.9	13.8	23.9	55	4
1	2	1	150	20.5	97	27.7	14.5	24.4	17	30
1	2	2	123	20.0	70	29.6	14.6	26.0	11	39
1	2	3	136	20.4	81	28.2	17.2	24.1	28	48
1	2	4	128	18.4	62	28.6	15.4	25.5	9	41
1	2	5	135	19.6	68	27.6	16.1	24.2	29	41
1	2	6	149	20.8	88	29.0	15.9	26.0	46	3
1	2	7	114	17.1	60	28.9	14.5	25.1	18	15
1	2	8	116	17.8	54	28.1	15.9	25.0	11	43
1	2	9	122	18.7	62	27.8	14.1	24.2	16	21
1	2	10	119	19.6	64	28.2	15.0	25.2	27	23
1	3	1	111	19.7	51	27.4	15.3	23.5	34	6
1	3	2	143	20.9	65	27.6	14.4	23.9	20	10
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.

A.2 Genetic Data Set

		LOCI																	
P	T			G	G	G					M	M	M		P	P		P	P
O	R	A	H	O	O	O	A	F	G	I	D	D	D	M	G	G	6	G	G
P	E	C	E	T	T	T	K	U	6	D	H	H	H	P	M	M	P	I	I
.	E	O	X	1	2	3		M	P	H	1	2	3	I	1	2	G	1	2
1	1	BB	BB	AA	AA	BB	BB	AA	BB	AA	BC	BB	AA	BB	BB	BC	AA	AA	AC
1	2	BB	AA	AA	AA	BB	BB	AA	BB	AA	CC	BC	AA	BB	BB	BC	AA	AA	AA
1	3	AB	AA	AA	AA	AB	AB	AA	BB	AA	AC	BC	AA	BB	BB	BB	AA	AA	AA
1	4	BC	AA	AA	AA	AA	AB	AA	BB	AA	BC	BB	AA	BB	BB	BB	AA	AA	AA
1	5	AC	AA	AA	AA	AA	BB	AA	BB	AA	CC	BB	AA	BB	BB	BB	AA	AA	AA
1	6	BC	AA	AA	AA	AB	BB	AA	AB	AA	CC	BB	AA	AB	BB	BC	AA	AB	AA
1	7	BB	AA	AA	AA	BB	BB	AA	BB	AA	BC	BB	AA	BB	BB	BB	AA	AA	AA
1	8	CC	AA	AA	AA	AB	BB	AB	BB	AA	BC	BB	AA	BB	BB	BC	AA	AA	AB
1	9	BB	AA	AA	AA	AB	BB	AA	BB	AA	CC	BB	AA	BB	BB	AB	AA	AA	AA
1	10	AA	AA	AA	AA	AB	BB	AA	BB	AA	CC	BB	AA	BB	BB	BB	AA	AA	AA
1	11	BC	AA	AA	AA	AB	AB	AA	BB	AA	AC	BB	AA	BB	BB	AC	AA	AA	AA
1	12	BB	AA	AA	AA	AA	BB	AB	BB	AA	CC	BB	AA	BB	BB	BC	AA	AA	AA
1	13	BC	AA	AA	AA	AB	BB	AA	AB	AA	CC	BB	AA	BB	BB	BB	AA	AA	AA
1	14	BC	AA	AA	AA	BB	BB	AA	AB	AA	CC	BB	AA	BB	BB	BB	AA	AA	AA
1	15	CC	AB	AA	AA	AA	AB	AA	BB	AA	AC	BB	AA	BB	BB	AC	AA	AA	AC
1	16	BC	AA	AA	AA	AB	BB	AA	BB	AA	BC	BB	AA	AA	BB	BB	AA	AA	AA
1	17	BC	AA	AA	AB	AB	BB	AA	BB	AA	BC	BB	AA	BB	AB	BB	AA	AA	AA
1	18	BB	AA	AA	AA	BB	AB	AA	BB	AA	AC	BB	AA	BB	BB	BB	AA	AB	CC
1	19	AC	AA	AA	AA	AA	AB	AA	BB	AA	CC	BB	AA	BB	BB	BB	AA	AA	AA
1	20	BC	AA	AA	AA	AB	AB	AA	AB	AA	BC	BB	AA	BB	BB	BB	AA	AA	AA
1	21	AB	AA	AA	AA	AB	AB	AA	AB	AA	BC	BC	AA	BB	BB	BB	AA	AA	AA
1	22	BB	AA	AA	AA	AA	BB	AA	BB	AA	BC	BB	AA	AB	BB	BC	AA	AA	AA
1	23	AB	AA	AA	AA	AA	BB	AA	AB	AA	BC	BB	AA	BB	BC	AB	AA	AA	AB
1	24	AB	BB	AA	AA	AB	AB	AA	BB	AA	AC	BB	AA	BB	BB	BC	AA	AA	AA
1	25	AB	AA	AA	AA	AB	BB	AA	BB	AA	AC	BC	AA	BB	BB	BB	AA	AA	AB
1	26	AB	AA	AA	AA	AB	BB	AA	BB	AA	CC	BB	AA	BB	BB	BC	AA	AA	AB
1	27	AA	AA	AA	AA	AB	BB	AA	BB	AA	CC	BB	AA	BB	BB	BC	AA	AA	AB
1	28	AC	AA	AA	AA	AA	BB	AA	BB	AA	BC	BB	AA	AB	BB	BC	AA	AA	AA
1	29	AC	AA	AA	AA	AB	BB	AA	BB	AA	BC	BB	AA	BB	BB	BB	AA	AA	AC
1	30	AB	AA	AA	AA	AB	BB	AA	BB	AA	CC	BC	AA	BB	BB	AB	AA	AA	AC
2	1	BC	AA	AA	AA	AA	BB	AA	BB	AA	BC	BB	AA	BB	BB	BC	AA	AA	AA
2	2	BC	AA	AA	AA	AB	AB	AA	AB	AA	AC	BC	AA	BB	BB	AC	AA	AB	AA
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Appendix B

Correlation Matrices

Table B.1: Correlation Matrix for Population 1

	X_4	X_6	X_1	X_2	X_5	X_3	X_7	X_8
X_4	1							
X_6	.956	1						
X_1	.652	.625	1					
X_2	.547	.519	.626	1				
X_5	.464	.450	.418	.551	1			
X_3	.174	.161	.599	.436	.122	1		
X_7	.257	.270	.313	.274	.359	.212	1	
X_8	.055	.120	.194	.093	.050	.263	−.201	1

Table B.2: Correlation Matrix for Population 2

	X_4	X_6	X_1	X_2	X_5	X_3	X_7	X_8
X_4	1							
X_6	.953	1						
X_1	.623	.600	1					
X_2	.515	.469	.625	1				
X_5	.537	.524	.579	.699	1			
X_3	.158	.119	.652	.579	.368	1		
X_7	.301	.308	.587	.490	.499	.467	1	
X_8	.016	−.009	.055	.011	.044	.227	−.358	1

Table B.3: Correlation Matrix for Population 3

	X_4	X_6	X_1	X_2	X_5	X_3	X_7	X_8
X_4	1							
X_6	.938	1						
X_1	.507	.474	1					
X_2	.498	.490	.562	1				
X_5	.646	.616	.393	.560	1			
X_3	−.061	−.048	.373	.514	.091	1		
X_7	.028	.017	.307	.325	.166	.434	1	
X_8	.039	.018	.046	.202	.202	.105	−.420	1

Table B.4: Correlation Matrix for Population 4

	X_4	X_6	X_1	X_2	X_5	X_3	X_7	X_8
X_4	1							
X_6	.965	1						
X_1	.725	.708	1					
X_2	.603	.568	.569	1				
X_5	.546	.549	.579	.675	1			
X_3	.197	.196	.617	.380	.259	1		
X_7	.136	.146	.488	.301	.292	.572	1	
X_8	−.068	−.040	.042	.128	.058	.231	.087	1

Table B.5: Correlation Matrix for Population 5

	X_4	X_6	X_1	X_2	X_5	X_3	X_7	X_8
X_4	1							
X_6	.944	1						
X_1	.648	.620	1					
X_2	.456	.416	.656	1				
X_5	.507	.509	.524	.674	1			
X_3	.242	.183	.687	.572	.335	1		
X_7	.150	.142	.453	.419	.392	.419	1	
X_8	.156	.184	.238	.112	.109	.257	−.370	1

Table B.6: Correlation Matrix for Population 6

	X_4	X_6	X_1	X_2	X_5	X_3	X_7	X_8
X_4	1							
X_6	.967	1						
X_1	.590	.565	1					
X_2	.531	.498	.475	1				
X_5	.592	.564	.493	.776	1			
X_3	.224	.175	.618	.542	.341	1		
X_7	.143	.099	.457	.363	.311	.624	1	
X_8	−.006	.004	.066	.398	.294	.202	.018	1

Table B.7: Correlation Matrix for Population 7

	X_4	X_6	X_1	X_2	X_5	X_3	X_7	X_8
X_4	1							
X_6	.916	1						
X_1	.600	.564	1					
X_2	.527	.486	.612	1				
X_5	.454	.442	.585	.616	1			
X_3	.252	.211	.661	.523	.394	1		
X_7	.232	.189	.241	.170	.212	.325	1	
X_8	−.010	.031	.248	.091	.213	.290	−.480	1

Table B.8: Correlation Matrix for Population 8

	X_4	X_6	X_1	X_2	X_5	X_3	X_7	X_8
X_4	1							
X_6	.932	1						
X_1	.542	.530	1					
X_2	.491	.477	.521	1				
X_5	.561	.494	.451	.671	1			
X_3	.173	.150	.577	.558	.290	1		
X_7	.044	.003	.272	.208	.293	.316	1	
X_8	.209	.194	.265	.270	.174	.356	−.405	1

Table B.9: Correlation Matrix for Population 9

	X_4	X_6	X_1	X_2	X_5	X_3	X_7	X_8
X_4	1							
X_6	.920	1						
X_1	.603	.492	1					
X_2	.386	.340	.522	1				
X_5	.404	.415	.462	.644	1			
X_3	.149	.028	.575	.649	.416	1		
X_7	−.001	.060	.247	.283	.407	.319	1	
X_8	.327	.281	.309	.315	.213	.292	−.065	1

Table B.10: Correlation Matrix for Population 10

	X_4	X_6	X_1	X_2	X_5	X_3	X_7	X_8
X_4	1							
X_6	.960	1						
X_1	.676	.656	1					
X_2	.496	.456	.543	1				
X_5	.506	.489	.604	.713	1			
X_3	.273	.245	.609	.511	.322	1		
X_7	.262	.231	.486	.331	.446	.346	1	
X_8	.027	−.052	.194	.286	.361	.169	.082	1

Table B.11: Correlation Matrix for Grand Population

	X_4	X_6	X_1	X_2	X_5	X_3	X_7	X_8
X_4	1							
X_6	.948	1						
X_1	.626	.597	1					
X_2	.401	.388	.514	1				
X_5	.535	.532	.521	.603	1			
X_3	.133	.108	.570	.508	.293	1		
X_7	.137	.122	.360	.281	.258	.345	1	
X_8	.130	.129	.182	.108	.219	.228	−.281	1

Appendix C

Covariance Matrices

Table C.1: Covariance Matrices for Population 1

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	234							
X_2	16.9	3.11						
X_3	82.7	6.94	81.5					
X_4	30.1	2.91	4.74	9.11				
X_5	7.73	1.18	1.33	1.69	1.46			
X_6	27.4	2.62	4.16	8.28	1.56	8.23		
X_7	101	10.2	40.6	16.4	9.18	16.4	448	
X_8	46.6	2.58	37.3	2.60	.945	5.44	−66.9	247

Table C.2: Covariance Matrices for Population 2

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	319							
X_2	21.7	3.77						
X_3	125	12.0	115					
X_4	32.9	2.96	5.00	8.72				
X_5	13.6	1.78	5.17	2.08	1.72			
X_6	32.7	2.79	3.89	8.61	2.10	9.34		
X_7	235	21.4	112	19.9	14.7	21.5	505	
X_8	16.3	.343	40.3	.802	.961	−.467	−132	273

Table C.3: Covariance Matrices for Population 3

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	255							
X_2	15.3	2.90						
X_3	60.0	8.82	101					
X_4	21.5	2.25	−1.63	7.05				
X_5	8.07	1.27	1.17	2.21	1.65			
X_6	21.6	2.39	−1.38	7.13	2.26	8.18		
X_7	93.9	10.6	83.9	1.43	4.08	.918	368	
X_8	8.34	3.91	12.1	1.17	2.95	.560	−91.5	129

Table C.4: Covariance Matrices for Population 4

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	464							
X_2	27.4	5.01						
X_3	131	8.37	97.1					
X_4	47.1	4.08	5.89	9.10				
X_5	15.0	1.82	3.08	1.99	8.62			
X_6	45.1	3.76	5.73	8.62	1.96	8.77		
X_7	220	14.1	118	8.63	7.39	9.08	440	
X_8	7.80	2.45	19.5	−1.75	.601	−1.03	15.7	73.6

Table C.5: Covariance Matrices for Population 5

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	318							
X_2	22.9	3.84						
X_3	134	12.3	120					
X_4	30.9	2.39	7.08	7.17				
X_5	11.6	1.63	4.54	1.68	1.53			
X_6	29.0	2.14	5.24	6.26	1.65	6.88		
X_7	202	20.5	115	10.0	12.1	9.33	624	
X_8	67.5	3.47	44.6	6.64	2.13	7.67	−147	252

Table C.6: Covariance Matrices for Population 6

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	366							
X_2	18.9	4.35						
X_3	101	9.62	72.3					
X_4	29.5	2.90	4.98	6.83				
X_5	12.9	2.21	3.97	2.12	1.87			
X_6	27.2	2.61	3.74	6.35	1.94	6.32		
X_7	174	15.1	106	7.45	8.48	4.96	398	
X_8	8.92	5.88	12.2	−.111	2.84	.079	2.53	50.2

Table C.7: Covariance Matrices for Population 7

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	312							
X_2	19.6	3.29						
X_3	113	9.20	2.38					
X_4	26.4	2.38	6.08	4.51				
X_5	12.2	1.32	4.51	1.33	5.53			
X_6	24.2	2.14	4.97	5.53	1.27	5.90		
X_7	98.2	7.12	72.7	13.4	5.77	10.6	532	
X_8	82.0	3.08	52.6	−.476	4.72	1.43	−207	351

Table C.8: Covariance Matrices for Population 8

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	273							
X_2	15.3	3.15						
X_3	83.2	8.64	76.0					
X_4	17.6	1.71	2.96	3.87				
X_5	8.96	1.43	3.03	1.33	1.45			
X_6	17.2	1.67	2.58	3.61	1.17	3.87		
X_7	91.7	7.50	56.0	1.75	7.16	.134	413	
X_8	79.1	8.65	56.0	7.44	3.77	6.88	−149	326

Table C.9: Covariance Matrices for Population 9

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	382							
X_2	19.5	3.66						
X_3	116	12.9	107					
X_4	32.9	2.06	4.31	7.79				
X_5	12.0	1.64	5.76	1.50	1.78			
X_6	24.7	1.67	.751	6.59	1.42	6.58		
X_7	90.6	10.2	62.1	−.075	10.2	2.90	352	
X_8	67.7	6.75	34.0	10.2	3.19	8.08	−13.6	126

Table C.10: Covariance Matrices for Population 10

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	364							
X_2	20.6	3.94						
X_3	115	20.0	97.8					
X_4	31.3	2.40	6.58	5.93				
X_5	15.3	1.88	4.22	1.63	1.76			
X_6	29.0	2.10	5.64	5.43	1.51	5.39		
X_7	170	12.0	62.8	11.7	10.8	9.84	336	
X_8	39.8	6.11	18.0	.719	5.15	−1.30	16.3	116

Table C.11: Covariance Matrices for Grand Population

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	345							
X_2	19.6	4.22						
X_3	106	10.5	101					
X_4	33.6	2.38	3.86	8.34				
X_5	12.9	1.65	3.91	2.05	1.77			
X_6	31.1	2.23	3.04	7.69	1.98	7.88		
X_7	149	12.8	77.1	8.77	7.62	7.59	494	
X_8	50.3	3.28	34.0	5.56	4.32	5.39	−92.7	220

Bibliography

[1] Becker, R. A., Chambers, J. M. and Wilks, A. R. (1985), *The New S Language.* Pacific Grove, CA: Wadsworth & Brooks/Cole

[2] D'Agostino, R. B. and Stephens, M. A. (1986), *Goodness-of-fit Techniques*, New York: Marcel Dekker, Inc.

[3] Dixon, W. J. (1990), *BMDP Statistical Software Manual.* vol.2, Berkeley: University of California Press

[4] Green, P. J. (1984), Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives(with discussion), *J. Royal Statist. Soc.* **46**(2): 149-192

[5] Hedrick, P. W. (1985), *Genetics of Populations.* Boston: Jones and Bartlett

[6] Hosie, R. C. (1979), *Native Trees of Canada.* Don Mills, Ontario: Fitzhenry & Whiteside

[7] Kendall, M. (1975) *Multivariate Analysis.* London: Charles Griffin

[8] McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models.* London: Chapman & Hall

[9] Olsen, C. L. (1974) Comparative robustness of six tests in multivariate analysis of variance, *J. Amer. Stat. Ass.*, **69**, 894-908

[10] Pillai, E. J. C. (1955), Some new test criteria in multivariate analysis, *Ann. Math. Stat.*, **26**, 117-121

[11] Research Directorate (1991) *Forest Research at the LFC: 1991-1992 Overview*, Sainte-Foy, Quebec: Forestry Canada, Quebec Region

[12] SAS Institute Inc. (1990), *SAS/STAT User's Guide.* Cary, NC: SAS Institute

[13] Stephens, M. A. (1974), EDF statistics for goodness-of-it and some comparisons, *J. Amer. Stat. Ass.*, **69**, 730-737

[14] Winer, B. J., Brown, D. R. and Michels, K. M. (1991), *Statistical Principles in Experimental Design.* New York: McGraw-Hill

[15] Wu, Jun (1992) *Proportional Logistic Regression Analysis of a Forestry Data Set*, M.Sc Project, Department of Mathematics and Statistics, Simon Fraser University, Burnaby, B.C.