



**Addis Ababa University
Postgraduate Program**

**ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES**

**THE APPLICATION OF D^2 STATISTIC TO MEASURE GENETIC DIVERGENCE IN
DIFFERENT BREAD WHEAT GENOTYPES**

LEUL MEKONNEN

A Thesis Submitted to

The Department of Statistics

Presented in Partial Fulfillment of the Requirements for the

Degree of Master of Science in Statistics

Addis Ababa University

Addis Ababa, Ethiopia

June, 2014

Addis Ababa University
School of Graduate Studies

This is to certify that the thesis prepared by Leul Mekonnen, entitled: **The Application of D^2 Statistic to Measure Genetic Divergence in Different Bread Wheat Genotypes** and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Statistics complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

Examiner _____ Signature _____ Date _____

Examiner _____ Signature _____ Date _____

Advisor _____ Signature _____ Date _____

Chair of Department or Graduate Program Coordinator

ABSTRACT

The Application of D^2 Statistic to Measure Genetic Divergence in Different Bread Wheat Genotypes

Leul Mekonnen

Addis Ababa University, 2014

Wheat is one of the most important cereal crop in Ethiopia, ranking third in total production next to maize and teff. Wheat covers a total arable land of 110,434 ha with average productivity of about 8.4 qt /ha, which is below the national production average (14.4 qt /ha). The objective of this research was to study the genetic dissimilarity among twenty bread wheat genotypes. The study is made based on twenty bread wheat genotypes which was obtained in 2007 and 2008 from trials conducted by the Ethiopian Institute of Agricultural Research (EIAR) and were evaluated for their diversity to estimate the genetic divergence and clustering them into homogenous groups for the hybridization program in 8 environments using Mahalanobis D^2 statistic. The D^2 value data can be used in cluster analysis to identify groups of related cultivars. Using UPGMA clustering technique all the genotypes were grouped into six cluster on the basis of D^2 value using Tocher's optimization technique. The intra- and inter-cluster D^2 values suggested that within cluster genetic diversity is narrow, but the genetic diversity among clusters is greater. Therefore, The presence of significant genetic variability among the evaluated bread wheat genotypes suggests an opportunity for improvement of grain yield through hybridization of genotypes from different clusters and subsequent selection from the segregating generations would broaden the genetic base of bread wheat breeding populations. Thus, those analysis based on multivariate methods using D^2 statistic is useful in providing information for more efficient variety development programmes.

ACKNOWLEDGEMENT

Praise be to God, the Cherisher and Sustainer of the worlds, Most Gracious, Most Merciful! He has made the completion of this study a reality.

I owe the deepest gratitude to Professor M. K. Sharma, my thesis advisor for his exemplary guidance, monitoring and constant encouragement throughout the course of this thesis work. Without his professional assistance and guidance this study would not be realized. His stimulating supports helped me in preparing my thesis within the time-frame.

I record with appreciation the help rendered by my dear classmates encouraging me to work hard. My best wishes to all who helped directly or indirectly in ensuring a better thesis write-up.

A large degree of emotional support and encouragement has been received from the many friends I have made during my stay at the University. My deep appreciation and thanks goes to my friends Amare Wubishet, Birhane Zelalem and Sebsib Muanenda for the help they gave me. Also I am grateful to all of those who supported me in any respect during the completion of my study.

Finally, and most importantly, this might be a good opportunity to acknowledge all members of my family, relatives and friends for all their encouragement and support throughout my study.

Table of Contents

ABSTRACT.....	i
ACKNOWLEDGEMENT.....	ii
LIST OF FIGURES.....	v
LIST OF TABLES.....	vii
ACRONYMS	viii
INTRODUCTION.....	1
1.1 Background of the Study.....	1
1.2 Statement of the Problem.....	4
1.3 Objective of the Study.....	5
LITERATURE REVIEW.....	6
2.1 Concepts and Definitions of Genetic Diversity.....	6
2.2 Analysis of Genetic Divergence Based on Multivariate Methods Using D ² Statistic.....	8
2.3 Analysis of Genetic Divergence Based on Multivariate Methods Using D ² Statistic in Bread Wheat.....	11
2.4 Analysis of Genetic Divergence Based on Multivariate Methods Using D ² Statistic in Other Crops.....	14
MATERIALS AND METHODS.....	19
3.1 Materials.....	19
3.2 The Methodology.....	21
3.2.1 Test of Significance of Differences Among Genotypes/Populations.....	21
3.2.1.1 Univariate Analysis of Variance(ANOVA).....	21
3.2.1.2 Wilk's Lambda Criterion (Λ) (Wilk's, 1932).....	22
3.2.2 Analysis of Genetic Divergence.....	23
3.2.2.1 Multivariate Analysis Using D ² Statistic.....	23
3.2.2.2 Test of Significance of D ² Values.....	33
3.2.2.3 Contribution of Individual Environments for Divergence.....	33
3.2.2.4 Cluster Analysis (Group Constellation).....	33
3.2.2.4.1 Aims of Clustering Analysis.....	36
3.2.2.4.2 General Guide-Lines for Cluster Analysis.....	37
3.2.2.4.3 Criterion for Clustering.....	37
3.2.2.5 Intra and Inter Cluster Divergence.....	38
3.2.2.6 Intra Cluster Divergence.....	38

3.2.2.7 Inter Cluster Divergence.....	38
3.2.2.8 Cluster Diagram.....	39
3.2.2.9 Cluster Mean Analysis.....	39
3.2.3 Canonical Variate Analysis.....	39
3.2.4 Graphical Method for Assessing Multivariate Normality Using Mahalanobis Distance.....	43
3.2.6 Multivariate Outliers.....	44
3.2.6.1 Mahalanobis Distance Method.....	44
RESULT AND DISCUSSION.....	46
4.1 Variation (Significance of Differences) Among Genotypes.....	46
4.1.1 Univariate Analysis of Variance(ANOVA).....	46
4.1.2 Wilk's Criterion of Aggregate Variation.....	47
4.2 D ² Analysis.....	48
4.2.1 Computation of D ² Values.....	55
4.3 Cluster Analysis.....	58
4.3.1 Intra-and Inter-Cluster Distances.....	60
4.3.2 The Nearest and Distant Clusters.....	61
4.3.3 Diagrammatic Representation of Cluster Divergence.....	68
4.3.4 Cluster Mean Analysis.....	69
4.4 Mahalanobis Distance for Assessing Multivariate Normality and Outlier.....	73
4.5 Discussion.....	74
CONCLUSION AND RECOMMENDATION.....	78
REFERENCE.....	80
APPENDIXES.....	89

LIST OF FIGURES

Figure 4.1	Scatter plot of the six cluster groupings in relation to the two representative axes of the first two canonical variates (CV1 and CV2) in 8 environments evaluated.....	60
Figure 4.2	Dendrogram depicting the hierarchical clustering of 20 genotypes of bread wheat tested in 8 environments ,o - (D ²) as dissimilarity measure.....	63
Figure 4.3	Dendrogram depicting grouping by drawing boxes into six of the hierarchical clustering of 20 genotypes of bread wheat tested in 8 environments ,o - distance (D ²) as dissimilarity measure.....	65
Figure 4.4	Cophenetic distance defined by UPGMA.....	67
Figure 4.5	Generalized distance - inter-cluster divergence ($\sqrt{D^2}$) among six clusters of 20 bread wheat genotypes.....	68
Figure 4.6	Dendrogram depicting the hierarchical clustering of four localities used in bread wheat multi-environment trials ;Cophenetic correlation=0.89,o - ized distance (D ²) as dissimilarity measure.....	71
Figure 4.7	Dendrogram depicting the hierarchical clustering of 8 environments (location × year) used in testing 20 genotypes of bread wheat at four locations, 2007–08. Cophenetic correlation=0.78,o - (D ²) as dissimilarity measure.....	72

Figure 4.8 Mahalanobis distance plotted versus chi-square quantiles for checking multivariate normality.....	73
Figure 4.9 Plot of Mahalanobis distance for checking multivariate outlier.....	74
Figure A.1 Normal probability plot of residuals for each environment from the separate ANOVA.....	91
Figure A.2 Plot of residuals against the predicted value in each of the separate ANOVA model.....	92

LIST OF TABLES

Table 1.1	Production of wheat world-wide and in Ethiopia during the period 2007- 2011 and an estimated value for 2011-2012 (Sources: USDA, 2011; GMR, 2011; CSA, 2010).....	2
Table 3.1	Genotype codes of 20 wheat genotypes.....	19
Table 3.2	Description of the experimental sites and environments.....	20
Table 3.3	The general separate ANOVA structure for each environment.....	21
Table 4.1	Mean Sum of Square for 20 genotypes in bread wheat.....	47
Table 4.2	variance and covariance matrix at different levels.....	49
Table 4.3	Reduction of common dispersion (error) matrix by pivotal condensation method.....	51
Table 4.4	Divergence as measured by D^2 statistic among 20 genotypes of bread wheat.....	57
Table 4.5	Contribution of each environment to the total divergence.....	58
Table 4.6	Cluster number with their respective genotypes.....	59
Table 4.7	Intra- (main diagonal) and inter-cluster D^2 values along with their $\sqrt{D^2} = D$ value in parenthesis.....	62
Table 4.8	The nearest and distant clusters from each cluster based on D value.....	62
Table 4.9	Cluster Means for 8 quantitative environments in 20 bread wheat genotypes	69
Table A.1	Reduction of common dispersion (error + genotype) matrix by pivotal condensation method.....	89
Table A.2	Individual (separate) analyses of variance (RCB design) for a trial with 20 genotypes and four replications, by environment.....	90
Table A.3	Bartlett's test for homogeneity of residual variance using genotype as a group in each of the separate ANOVA.....	91
Table A.4	Test for Normality of residuals in each of the separate ANOVA model using the Shapiro-Wilk (W) statistic.....	91

ACRONYMS

ANOVA	Analysis of Variance
CCC	Cophenetic Correlation Coefficient
CSA	Central Statistical Agency
CV	Coefficient of Variation
CVA	Canonical Variate Analysis
DF	Degrees of Freedom
DNA	Deoxyribonucleic Acid
EIAR	Ethiopian Institute of Agricultural Research
ESE	Ethiopian Seed Enterprise
<i>et al.</i>	And others
FAO	Food and Agriculture Organization
GE ($G \times E$)	Genotype by Environment
GEI	Genotype by Environment Interaction
GMR	Grain Market Report
kg/ha	Kilogram per Hectar
LHS	Left Hand Side
m.a.s.l	Meter Above Sea Level
MANOVA	Multivariate Analysis of Variance
MD	Mahalanobis Distance

MSD	Mahalanobis Squared Distance
MSE	Mean Square Error
q/ha	Quintal per Hectar
RCBD	Randomized Complete Block Design
RD	Robust Distances
RHS	Right Hand Side
SAS	Statistical Analysis System
SSCP	Sum of Squares and Cross Products
t/ha	Tone per Hectar
UPGMA	Unweighted Pair - Group Method with Arithmetic averages
UPGMC	Unweighted Pair - Group Method using Centroid
USDA	United States Department of Agriculture

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND OF THE STUDY

Wheat is one of the most important cereals world-wide and it is grown in many areas (Curtis, 2002; Briggie & Curtis, 1987). During the past four decades, wheat has made a significant contribution to the increase in global food production. This is due to the use of higher-yielding, water and fertilizer responsive, and disease resistant cultivars, combined with strengthened input delivery systems, tailored management practices and improved marketing (Ortiz *et al.*, 2008; Dixon *et al.*, 2006). A rapid increase in global wheat production has taken place during the last five decades, mainly due to increased productivity rather than an expansion of the cultivated area, with average global yields having risen from 1 t/ha in the 1950s to about 2.5 t/ha at the turn of the century (Curtis, 2002). According to estimates produced by USDA (2011), GMR (2011) and CSA (2010), wheat is one of the major cereal crops grown world-wide in terms of total production tonnage used for food. Wheat is currently second to rice and ahead of maize as the main human food crop. However, maize is more extensively used in animal feed.

Wheat is one of the major cereal crops principally grown in the highlands of Ethiopia, basically in the south east, central and North West parts. Considerable amount is also produced in the northern and southern regions (CSA, 2010). It is the third most grown crop, after teff and sorghum, both in terms of acreage and production volumes (Table 1.1). In global terms, wheat is the leading source of cereal proteins in human food, having higher protein content than maize or rice. In total, 16% and 26% of total dietary calories in developing and developed countries, respectively, come from wheat (Ortiz *et al.*, 2008). Wheat is the most traded food crop internationally and is used as an emergency food in aid for developing countries. Wheat grain is a staple food used to make flour for leavened, flat and steamed breads, biscuits, cookies, breakfast cereal, pasta, noodles, bio-fuel, and for fermentation to make alcoholic beverages such as beer and liquors (Tsegaye & Berg, 2007a). Its importance may vary in different countries, partly as a result of the diversity in different species of wheat. For example, tetraploid wheat is widely grown by farmers on the heavy black clay soils (vertisols) of the Ethiopian Highlands, at

altitudes of 1500-3000 meter above sea level (m.a.s.l.) during the main cropping season. Most of the tetraploid wheat under cultivation consists of a mixture of types (Tesemma & Belay, 1991). However, the most suitable range of altitude is 1900-2700 m.a.s.l., where the annual rainfall range is between 600 and 2000 mm (Gebre-Mariam, 1991).

Table 1.1 Production of wheat world-wide and in Ethiopia during the period 2007- 2011 and an estimated value for 2011-2012 (Sources: USDA, 2011; GMR, 2011; CSA, 2010).

Year	World-wide			Ethiopia		
	Area (million hectares)	Yield (metric tons per hectare)	Production (million metric tons)	Area (million hectares)	Yield (metric tons per hectare)	Production (million metric tons)
2011/12*	227.31	3.01	684.40	**	**	**
2010/11	222.48	2.91	648.20	1.61	1.71	2.75
2009/10	222.63	3.05	678.12	1.68	1.83	3.08
2008/09	222.76	3.06	683.41	1.45	1.75	2.54
2007/08	222.55	3.02	607.00	1.43	1.73	2.20

*Predicted production on September 2011; ** Actual value not yet available

Within the total area of wheat under cultivation in Ethiopia, it has been reported tetraploid and hexaploid species each occupy approximately 50% of the area (Gorfu *et al.*, 2001). However, a change in the relative proportions of wheat types grown has been reported more recently, with *e.g.* tetraploid and hexaploid species occupying approximately 30% and 70%, respectively, of the total wheat area under cultivation (CSA, 2010). Due to the introduction of productive semi-dwarf cultivars, improved varieties, the tetraploid wheat landraces that were widely cultivated in the early 20th century are now increasingly being replaced. This has resulted in the loss of genetically diverse, locally well-adapted but unimproved landraces and in an extinction of on-farm genetic variability (Royo *et al.*, 2009). Thus, the level of genetic diversity in tetraploid wheat is being affected by the high selection pressure applied in breeding programmes and the relatively small number of varieties currently in cultivation (Skovmand *et al.*, 2005).

For a successful genetic improvement program for any crop, conservation of germplasm resources is a basic prerequisite, but to understand the usable variability, grouping or classification of genetic stocks and choice of genetically divergent parents for hybridization is extremely essential. Quantitative classification based environmental measurements offers a quantifiable degree of divergence among genotypes and population. The models of multivariate analysis are D^2 statistic and canonical analysis. Plant breeders have often used D^2 statistic for measuring divergence quantitatively and canonical analysis to determine spatial divergence; both methods have been used to define the relative contribution of different environments to differentiate among genotypes. For determining the divergence among populations in terms of Mahalanobis D^2 statistic, which has been successfully used widely in psychometry and anthropometry (Rao 1952). Later on, Chandrashekhariah et al. (1969), and Murty et al. (1967) extended it to biology. Nowadays, divergence studies are being made on the basis of molecular markers, which provide precise information about phylogenetic hierarchy or descriptive account of individuals or groups of individuals but lack in quantification of continuous variation, the backbone of practical plant breeding. Mahalanobis D^2 statistic (Rao 1952) has been widely used in crop plants for rational selection of parents for cross hybridization (Ramanujam et al. 1974; Arunachalam 1981; Rabindra Babu et al. 2006).

Genetic diversity plays an important role in plant breeding because hybrids between lines of diverse origin generally display a greater heterosis than those between closely related parents. Mahalanobis D^2 analysis is a powerful biometrical techniques for measuring genetic divergence in plant breeding which has been widely used in crops like maize, sorghum, pearl millet, wheat, linseed, cotton, tobacco, alfalfa, brassica, etc. (Murty and Arunachalan, 1966). D^2 statistic measure the forces of differentiation at intra and inter cluster levels and determine the relative contribution of each component trait to the total divergence. Its estimates are free from genetic assumptions and help to identify suitable germplasm for incorporation into plant breeding stocks. The choice of parents is of paramount importance in any breeding program. Assessment of a large number of inbreeds for genetic diversity is of utmost importance. The Mahalanobis D^2 statistic enables one to discriminate between different cultivars according to the diversity present (Mahalanobis, 1936). It provides a clear idea about the diverse nature of populations. Clusters

formed according to Tochers method (Rao, 1952) were used to determine the distance between and within clusters.

1.2 STATEMENT OF THE PROBLEM

Ethiopia is known for its diverse/heterogeneous agro-ecology ranging from 100m below sea level in the Danakil depression to 4620m above sea level at mount Ras Dashan that contributes further to the problem of selecting high yielding genotypes of bread wheat varieties for particular environmental conditions . Even, Wheat is the most important cereal crop in the mid and high land areas of the Ethiopia, but its productivity remained unsatisfactory because of lack of improved varieties of high yielding genotypes. Some of the improved varieties tested were found to be adaptable to the agro ecology and are still under cultivation, but majority of these have showed lower yield performance. Even the performance of the improved varieties being cultivated in the Ethiopia is low and this could be due to the genotype divergence across environment.

Various studies have been conducted to identify divergent genotypes for hybridization purposes on the Ethiopian wheat varieties using Mahalanobis D^2 statistic based on physiological, morphological and agricultural characters of a group of genotypes. Interestingly, it was not done in identifying divergent genotypes of wheat varieties based on different environmental conditions (location-year combination) using Mahalanobis D^2 statistic in Ethiopia. In addition, the changing environmental conditions of Ethiopia, the expansion of wheat to new agro-ecologies coupled with inadequate wheat varieties available for the different environments necessitate a rigorous and continuous study of genotypic divergence for dynamic crop improvement program.

1.3 OBJECTIVE OF THE STUDY

General objective:

To assess genetic variability/diversity between bread wheat genotypes and to classify them based on their genetic distance.

Specific objective:

1. To identify the genetic diversity of bread wheat genotypes in the future bread wheat breeding program.
2. To assess the magnitude of genetic diversity and attain meaningful classification or grouping of genotype collections based on the genetic distance .
3. To determine relative contribution of each environment to the total divergence.
4. To identify those cluster of genotypes which are suitable for a particular environment that can produce a high yield capacity.
5. To evaluate the degree of divergence of performance of 20 bread wheat genotypes grown in different environments.

CHAPTER TWO

LITERATURE REVIEW

2.1 CONCEPTS AND DEFINITIONS OF GENETIC DIVERSITY

A knowledge of the genetic diversity of a plant population helps a plant breeder in the selection of desirable parents for a breeding program. Geographic diversity is normally used as an index of genetic variability in crop plants. Therefore, the relationship between geographic and genetic divergence, and the relative role of each character in contributing to maximum divergence should be evaluated prior to starting a breeding program.

Genotype : Although the term has a formal definition to geneticists, the term 'genotype' refers to an identifiably specific subspecies. This variety, line, or cultivar is usually identified by a commercial name or code.

Environment: This term refers to a set of growing conditions under which genotypes are examined. These conditions are dependent on both the physical location and the timing of the trial, and include, climatic, geographic and edaphic factors.

Genetic diversity refers to the variation of genes within populations/species, making it possible to develop new breeds of crop plants and domestic animals and allowing species in the wild to adapt to the changing conditions. In crop plants, genetic diversity arises as consequences of interplay of evolutionary forces (mutation, selection and random genetic drift) and the influence of human through domestication and selection (Bhatt, G.M. 1970). Study of genetic diversity is the process by which variation among individuals or groups of individuals or populations is analyzed by a specific method or a combination of methods. The data often involve numerical measurements and in many cases, combinations of different types of variables. Diverse data sets have been used by researchers to analyze genetic diversity in crop plants; most important among such data sets are pedigree data, morphological data and biochemical data obtained by analysis of isozymes and storage proteins and, recently, DNA-based marker data.

Genetic relationships in crop species is an important component of crop improvement programs, as it serves to provide information about genetic diversity, and is a platform for stratified sampling of breeding populations. Accurate assessment of the levels and patterns of genetic diversity can be invaluable in crop breeding for diverse applications including (i) analysis of genetic variability in cultivars, (ii) identifying diverse parental combinations to create segregating progenies with maximum genetic variability for further selection, and (iii) introgressing desirable genes from diverse germplasm into the available genetic base. An understanding of genetic relationships among inbred lines or pure lines can be particularly useful in planning crosses, and in assigning lines to specific heterotic groups. Analysis of genetic diversity in germplasm collections can facilitate reliable classification of accessions and identification of subsets of core accessions with possible utility for specific breeding purposes.

Genetic improvement programs are based on the genetic variability of the populations, which may be changed by introducing new genotypes in the population. Crosses are also frequently used in these programs, and the success of this practice is dependent on the genetic divergence between the parents. Parents showing high productivity indexes and great genetic diversity may produce a progeny that is more productive and shows great genetic variability (Piassi et al., 1995b). Studies on the genetic diversity of plants and livestock have increased after the 70's, simultaneously to the sudden increase in the variability of informatics resources (Sakaguti et al., 1996).

Genetic divergence studies may be used to evaluate the behavior of genotypes in different environments, to evaluate the superiority of some genotypes over others, to identify divergent genotypes that may be used as parents in breeding programs and to relate genetic divergence with heterosis (Piassi, 1994).

Genetic diversity among parentals is considered an important factor for obtaining heterotic hybrids. This diversity is one of the restraining tools for breeding programs based on hybridization, because it generates parameters for identifying superior parentals. According to Cruz and Carneiro. (2003), this distance is essential to increase the chance of recovering superior genotypes. To understand the usable variability grouping or classification of genetic stocks based

on suitable scale is quite imperative. Similarly, choice of genetically divergent parents for hybridization under transgressive breeding program is also dependent upon categorization of breeding materials on the basis of appropriate criteria. On the other hand, quantitative classification offers a quantified degree of divergence among genotypes or populations. This serves as a sound basis of grouping any two or more genotypes based on minimum divergence or resemblance between them.

2.2 ANALYSIS OF GENETIC DIVERGENCE BASED ON MULTIVARIATE METHODS USING D^2 STATISTIC

With increases in the sample sizes of germplasm accessions used in crop improvement programs, methods to classify and order genetic variability are assuming considerable significance. The use of established multivariate statistical methods is an important strategy for classifying germplasm, ordering variability for a large number of accessions, or analyzing genetic relationships among breeding materials. Multivariate statistical techniques, which simultaneously analyze multiple measurements on each individual under investigation, are widely used in analysis of genetic diversity irrespective of the data set (morphological, biochemical, or molecular marker data) (Mohammadi and Prasanna, 2003). Quantitative classification based on trait measurements offers a quantifiable degree of divergence among genotypes and population. The models of multivariate analysis are D^2 statistic and canonical analysis. Plant breeders have often used D^2 statistic for measuring divergence quantitatively and canonical analysis to determine spatial divergence; both methods have been used to define the relative contribution of different characters to differentiate among genotypes. A method utilizing multiple measurements, which are subject to multivariate analysis based on generalized distance as, indicated by D^2 statistic proved to be worth in quantifying the degree of divergence between genotypes or populations (Rao, 1952).

Different approaches to measurement of genetic distance have been proposed to suite various objectives (Mahalanobis, 1930, 1936) and (Rao, 1952, 1960). Mahalanobis D^2 statistic provided measure of the generalized distance in case of multiple measurements. D^2 statistic helped in the identification of genetically divergent genotypes that facilitated grouping and characterization using morphological and agronomic characteristics. The

multivariate D^2 analysis has been favored as a tool in estimating genetic divergence for use in plant breeding, since it helps in the choice of parental combinations of the greatest promise.

Selection of parents in hybridization programme based on Mahalanobis D^2 statistic is more reliable as the requisite knowledge of parents in respect of many characters is available prior to crossing. Nair and Mukherjee (1960) were the pioneers to use D^2 statistic as a measure of genetic divergence in the field of plant breeding for classification of teak. Genetic divergence analysis was done based on multivariate analysis of variance. D^2 statistic is useful to assess the extent and pattern of variation of the germplasm with respect to environments, regions, altitude classes and species. It also helps in identification of major traits contributing to the diversity which provides opportunities to be utilized for genetic improvement (Faris, *et al.*, 2006). Significant heterotic effects for yield and yield components were observed in the hybrids of the parents belonging to different groups but not in the same group. Genetic divergence between the parents had a positive relationship with heterosis effects of the hybrids were also observed (Shamsuddin, 1985).

Techniques of multivariate analysis such as clustering analysis have been successfully employed as a means to identify divergent genotypes and better utilize the advantages provided by heterosis. The objective of this model is to join the parents (or any other kind of sample objects) in many clusters based on a classification criterion, so that there is within - group homogeneity and between - group heterogeneity. Alternatively, techniques of clustering analysis divide a group of objects into many subgroups, according to a criterion of similarity or dissimilarity. Such process involves basically two steps. The first one is related to estimating a measure of similarity (or dissimilarity) between parents, whereas the second is related to clustering the objects into subgroups based on a joining technique (Cruz *et al.*, 2004).

Cluster analysis refers to a group of multivariate techniques whose primary purpose is to group individuals or objects based on the characteristics they possess, so that individuals with similar descriptions are mathematically gathered into the same cluster (Hair *et al.*, 1995). The resulting clusters of individuals should then exhibit high internal (within cluster) homogeneity and high external (between clusters) heterogeneity. Thus, if the classification is

successful, individuals within a cluster shall be closer when plotted geometrically and different clusters shall be farther apart (Hair *et al.*, 1995).

Cluster analysis based on algorithms such as UPGMA(Unweighted Pair Group Method with Arithmetic Averages), UPGMC has drawbacks. For instance, these algorithms do not provide an objective definition of what constitutes an optimal tree or dendrogram, and systemic errors are likely to be introduced during cluster analysis reconstructions. However, Mahalanobis distance (D^2) between centroids (vectors of means) of the groups can be used to identify the best clustering algorithm (Franco *et al.*, 1997). The best clustering method produces the largest distance, D^2 , among groups or clusters; this method may be particularly appropriate for quantitative data. Determining the number of acceptable clusters involves in deciding the true or natural groups. An acceptable cluster is defined as a group of two or more genotypes with a within-cluster genetic distance less than the overall mean genetic distance and between cluster distances greater than their within cluster distance of the two clusters involved. One of the relatively simple ways of finding optimal number of clusters is the D^2 . On the basis of D^2 , the best point for cutting a cluster is the one that shows the largest D^2 between centroids of the groups created at that point (Franco *et al.*, 1997).

Studies on genetic divergence have used the mean Euclidian distance or the standardized squared mahalanobis distance as dissimilarity measures (Cruz *et al.*, 2004). Although the latter has been preferred , it can be estimated only when the residual covariance matrix is available. Therefore, the advantage of the standardized squared Mahalanobis distance (D^2) over the Euclidian distance is that it considers the correlations among the chosen traits. The second step of the clustering analysis consists in choosing a joining method. The most used methods in studies of genetic diversity are the hierarchical and optimization methods. The Tocher's method is the most used among the optimization methods (Piassi, 1994).

Estimation of genetic diversity is an important step for a breeding program, but not the last one. Another helpful issue to be evaluated is the relative importance of the characters. The method to estimate their relative importance was proposed by Singh (1981). In this procedure (based on Mahalanobis D^2 statistic distance), each character i.e. the environment has a percentage which

refers to its contribution to the whole divergence. The characteristic with the least percentage is able to be omitted from the genotype evaluations (Cruz and Carneiro, 2003).

Literatures on multivariate statistical methods regarding the bread wheat and some other crops have been reviewed as below.

2.3 ANALYSIS OF GENETIC DIVERGENCE BASED ON MULTIVARIATE METHODS USING D^2 STATISTIC IN BREAD WHEAT

The assessment of genetic diversity using different environments has been of prime importance in many contexts particularly in differentiating well defined populations. The germplasm in a self pollinated crop can be considered as heterogeneous sets of groups, since each group being homozygous within it. Selecting the parents for breeding program in such crops is critical because, the success of such program depends on the segregates of hybrid derivatives between the parents, particularly when the aim is to improve the quantitative characters like yield. To help the breeder in the process to identify the parents that nick better, several methods of divergence analysis based on quantitative traits have been proposed to suit various objectives. Among them

’ D^2 statistic (Mahalanobis, 1936) has been put to good use for discriminating divergent populations (Murthy and Quadri, 1966) rather than inferring from earlier indices like morphological similarity, eco geographic diversity, phylogenetic relationship and similar other criteria. Estimation of degree of divergence between biological population and computation of different components to the total divergence is done completely by Mahalanobis generalized distance estimated by D^2 statistic (Nair and Mukherjee, 1960).

The Mahalanobis D^2 statistic has been used to quantify the degree of divergence in different crops (Das and Das Gupa, 1984; Ariyo, 1987a and Nair et al., 1998). The technique gave insight into the most genetically divergent parents that could be used for hybridization purpose. Das and Das Gupa (1984) and Ariyo (1987a) noted earlier that, geographical diversity was not always related to genetic diversity and therefore not an adequate index of genetic diversity. Genotypes within clusters often showed great geographical diversity. Successful establishment of germplasm collections and plant introduction for crop improvement as well as for germplasm conservation require studies in genetic variability within plant populations.

In a species, the pattern of variation can be measured in terms of levels of diversity in different populations and genetic distances between different populations. Vavilov (1951) reported the presence of high genetic diversity in cultivated tetraploid wheat and considered Ethiopia to be a centre of diversity and the site of origin of tetraploid wheat. Most of the studies to date on Ethiopian wheat landraces have been based on the diversity of agro-morphological characters, which are highly heritable (Eticha *et al.*, 2005; Bechere *et al.*, 1996; Pecetti & Damania, 1996; Abebe & Giorgis, 1991; Tesemma *et al.*, 1991; Tesemma & Belay, 1991; Perrino & Porceddu, 1991; Spagnoletti-Zeuli & Qualset, 1987; Negassa, 1986; Jain *et al.*, 1975; Porceddu *et al.*, 1973), on isozymes (Tsegaye *et al.*, 1994) and on cytological markers (Belay & Merker, 1999). There are specific endemic characters for some of the tetraploid wheat species in Ethiopia, such as violet grain and beardless or half-bearded hard durum wheat, which shows great variation from place to place (Teklu & Hammer, 2006; Belay *et al.*, 1993; Tesemma *et al.*, 1991; Jain *et al.*, 1975). The variation in the Ethiopian landraces includes some significant values such as resistance to diseases, pests and various stress conditions. For wheat landraces collected from the central highlands of Ethiopia, Tesemma *et al.* (1991) have found regional diversity, whereas Bechere *et al.* (1996) have reported the absence of such differences between seeds collected from the north central highlands. Messele (2001) has also shown the presence of high diversity across regions and altitudes using morphological, chromosomal and molecular markers. However, in most of these previous studies samples from all regions were not included, since some studies examined samples from the central highlands, whereas others studied samples from the north central highlands. Furthermore, the research was conducted outside the natural habitat of the landraces and qualitative and quantitative traits and different wheat types were lumped together. The studies mentioned above have contributed to the identification of useful materials for plant breeding purposes and to a better understanding of the crop in designing appropriate collection and conservation strategies. However, despite these studies, knowledge of the variation in Ethiopian wheat landraces for biochemical and molecular markers remains limited, which indicates there is still a need for more information on population structure and on potential inputs for breeding. Harlan (1975) indicated that genetic variation can be attributed to human activities, environmental factors and the dynamic process of hybridization, segregation and selection. Hence, the variation may be the result of the interaction between abiotic and biotic factors and

population characteristics. Physical and biotic environmental factors exert selection pressure on variants, leading to the establishment of specific types, which indicates that the genetic variation is different from place to place depending on differences in the environment. For instance, Porceddu *et al.* (1988) observed a significant differentiation in agro-morphological traits among Ethiopian tetraploid wheat landraces collected from the same regions, but from different places which varied in their edaphic and climatic conditions.

The genetic divergence among indigenous durum wheat (*Triticum durum* Desf.) genotypes from northeastern Ethiopia for diverse origin and effort was made to cluster them into homogenous groups of ten clusters for the hybridization program. And it was found that there was no correspondence between geographic and genetic distances, i.e., germplasms, collected from the same geographic area were placed into different cluster groups and those collected from different geographic regions were placed into the same cluster (Arega, *et al.*, 2007). The presence of significant genetic variability among the genotypes provide an opportunity for improvement of grain yield through hybridization of genotypes from different clusters and subsequent selection from the segregating generations.

Genetic divergence study among accessions also help in knowing the secondary centers of diversification. Durum wheat (*Triticum durum* Desf.) collection with 111 accessions represents a large part of the collection of Triticeae species in Tunisia. Evaluation showed an average of diversity index about 0.77, hence presence of a large diversity of this collection. Germplasm has a relatively important diversity and confirming that Tunisia is one of the principal durum wheat secondary centers of diversification (Ayed and Slim-Amara, 2009). This can also be used in special breeding programs like abiotic stress conditions. With this regard a set of eight differentially heat tolerant wheat genotypes were used to study divergence for grain yield and its components. The eight genotypes were grouped into four clusters on account of diversity among them.

In bread wheat breeding program, several character of environments are simultaneously measured, then it is possible to estimate the genetic divergence by using multivariate methods. Choice of divergent parent(s) can be made for hybridization purposes on the basis of Mahalanobis D^2 statistic - values between two genotypes or two clusters. Members from most

divergent clusters can be chosen taking into account their desirable or complementary characters. This divergence is based on physiological, morphological and agricultural characters of a group of genotypes (Machado *et al.*, 2002). The efficacy of the genetic divergence as a criterion for choosing parents to be crossed has been reported by several authors (Machado *et al.*, 2002). Moreover, the Cluster Analysis and the Canonical Variable Analysis have analogous efficacy to determine the most suitable combinations for a breeding program (Machado *et al.*, 2002). Somayajulu *et al.* (1970) employed Mahalanobis' distances to measure the genetic diversity among 67 wheat strains. The estimates were obtained for three different environments representing different levels of fertility. When the 67 populations were grouped into clusters for each environment, the number of clusters varied from environment to environment, as did the populations within clusters. However, while instability of grouping was evident when the strains involved exhibit a low magnitude of genetic divergence, it did not occur for clusters representing the highest order of divergence.

2.4 ANALYSIS OF GENETIC DIVERGENCE BASED ON MULTIVARIATE METHODS USING D² STATISTIC IN OTHER CROPS

Goodman (1967) applied Mahalanobis generalized distances to measure morphological similarities among 15 races of maize from the subtropical region of southeastern South America; the conclusions coincided with those from conventional taxonomic results. Multivariate analysis of variance on the above data showed little variation from the results obtained by using Mahalanobis generalized technique (Goodman, 1968). Euclidean distances between race means in the form of standardized principal components were calculated from 20 ear characters for 219 Latin American races of corn by Goodman and Bird (1977). From an unweighted cluster analysis, 14 groups of races were delimited. Goodman and Bird (1977) reported that they used principal components analysis, factor analysis, canonical variate analysis and cluster analysis for grouping maize races on the basis of ear morphology and that principal components analysis had shown more general utility.

Sohoo and Gill (1978), using Mahalanobis generalized distance on nine characters of 100 strains of cluster bean, *Cyamopsis tetragonotoba*, determined that the maximum contribution for

divergence was due to days to maturity, closely followed by clusters per plant. The minimum was due to grain gum content. They grouped the strains into 17 clusters.

Cluster analysis was applied to 16 genotypes of mung bean, *Vigna raahata*, to provide an index of similarity of the genotypes for four characters: yield, pods per plant, seeds per pod and seed weight. No relationship between geographic distribution and genetic diversity was obtained (Ghaderi *et al.*, 1979). Multivariate analysis of variance, canonical analysis and canonical distances (generalized distance between groups based on the significant canonical variates) were used to discriminate between tetraploid-diploid hybrid families of potatoes and their relationship to potato cultivars (Tai and De Jong, 1980). The same analysis was applied to discriminate between TUBerosun-Andigena hybrid families and their relationships to their parents by Tai and Tarn (1980). Five significant canonical variates were determined, accounting for 96.6% of the total variation. Denton and Del Moral (1976) evaluated three cluster strategies and stepwise discriminant analysis for taxonomic data of *Oxalis* section *Inoxalis*. The cluster methods were isodata (a polythetic, divisive, iterative and non-hierarchical method which used Euclidean distances between centroids (group means) as a fundamental metric); cluster (an agglomerative, hierarchical procedure which uses the unweighted pair group method with the inter individual correlation coefficient) and MINFO (a polythetic, agglomerative, hierarchical procedure which uses 18 the within group dispersion method). A canonical analysis revealed that MINFO segregated discrete clusters better than the other two. They suggested that a polythetic hierarchical clustering method using an information statistic produces the most efficient and reliable phenetic classification.

Rhodes *et al.* (1970) measured variability for 73 characters of 40 cultivars of mango (*Mangifera indica*) and specimens of *M. odorato* and *M. Zeylanica* using the unweighted pair group method of cluster analysis with distance coefficients. Most of the cultivars were clustered in four groups. Using similar techniques and principal components analysis, Rhodes *et al.* (1971) conducted a numerical taxonomic study of the avocado (*Persea americana*), based on 67 characters of 38 cultivars. Correlation and distance phenograms from cluster analysis seemed more useful than principal components analysis, but the authors concluded that the use of both methods were complementary in studying patterns of variation in this species.

The genetic divergence among 36 genotypes of bitter gourd (*Momordica charantia* L.) was estimated using D^2 analysis (Kundu *et al.*, 2012). Through multivariate analysis based on 22 characters 36 genotypes were grouped into six distant clusters. Cluster VI includes maximum genotypes (12) followed by cluster I (6) and cluster II (6). Cluster V, cluster III and cluster I comprised 5, 4 and 3 genotypes respectively. The inter-cluster distances were higher than the intra-cluster distances. The inter-cluster distance was maximum between cluster III and IV (28.71) followed by the distance between cluster I and cluster IV (23.61). The intra-cluster distances in all the 6 clusters were more or less low indicating the closeness of genotypes within the same cluster. The highest intra-cluster distance was observed for cluster III (1.84) followed by the cluster I (1.38). The genotypes within the same clusters were collected from different places and genotypes collected in the same place fall in different cluster, which indicated that genetic divergence are not dependent on its geographical position from where the genotypes were collected (Kundu *et al.*, 2012).

Genetic divergence among 22 genotypes of lentil was estimated using D^2 and principal component analysis (A.A. Jewel, A. K. Chowdhury, A. K. M. M. Alam, M. A. Latif, and M. M. Hassan, 2010). The genotypes under study fell into four clusters. The distribution indicated that the maximum number of genotypes were included in cluster III (12) followed by cluster IV (7), while cluster I and II included only two and single genotype, respectively. D^2 values for all possible combinations were computed between pair of genotypes and the value ranged from 0.936 (between genotypes BLX-02009-17-4 and BLX-02009-04-1) to 0.105 (between genotypes BARI masur-5 and BLX-02009-21-5). Statistical distances represent the index of genetic diversity among the clusters. The inter-cluster distance in all of the cases was larger than intra-cluster distances suggesting wider genetic diversity among the genotypes of different groups. The intra-cluster distances in all the four clusters were low. This indicated that the genotypes within the same cluster were closely related. The highest inter-cluster distance (18.175) was observed between cluster I and II. The intra cluster distance was the highest (0.531) in cluster I and the minimum in cluster II. The lowest inter-cluster distance was observed between cluster III and IV (2.923) followed by cluster I and IV (7.600) and between II and III (8.574). Moderate or intermediate distance was found between cluster II and IV (11.414), cluster I and III (10.353).

The inter cluster distances were higher than the intra cluster distances suggesting wider genetic diversity among the genotype of different groups. The genotypes belonging to the distant clusters could be used in hybridization program for obtaining a wide spectrum of variation among the segregant. Similar reports were made by Jagadev and Samal (1991) in niger. The inter-cluster distance was larger than the intra-cluster distance in all the cases suggesting the wider genetic diversity among the genotypes of different groups. The inter-cluster value indicated maximum distance between cluster I and II than cluster III and IV. Thus the genetically diverged genotypes of the clusters could be used as parent in hybridization program to get desirable genotypes. Cluster I had the highest cluster mean. Genotypes with those characters in respective groups would, therefore, offer a good scope of improvement of lentil through selection (A.A. Jewel, A. K. Chowdhury, A. K. M. M. Alam, M. A. Latif, and M. M. Hassan , 2010).

Malhotra *et al.* (1973) studied the relationship between geographic diversity and genetic divergence and the relative role of each character towards maximizing divergence in green gram. They used sixty strains of indigenous and exotic green gram and observed 7 quantitative characters for each strain ANOVA revealed high significant differences among the strains for all studied characters. Pooled differences for the characters were also highly significant when tested by Wilk's criterion ($\chi^2 = 1702.47$ at 413 df). Based on the magnitude of the D^2 values, they found that the 60 strains that are originated in different geographic areas were grouped into 14 clusters with the consideration that strains within the clusters have smaller variation in D^2 values than strains between the clusters. The clustering pattern of the strains showed that genetic diversity was not directly related to geographic diversity because strains from one geographic area grouped into different or similar cluster. Thus it could be concluded that geographic diversity cannot be always used as an index of genetic diversity.

Sarma and Roy (1994) classified 42 early maturing pigeon pea genotypes on the basis of D^2 analysis. The analysis of variance revealed significant differences among the 42 genotypes for all the characters under study, indicating considerable variation among the genotypes. The D^2 values ranged from 11.5 to 2658.6, reflecting wide diversity among the genotypes. Based on these values, they grouped the 42 genotypes into eight clusters. Cluster means for branches/plant

Pods/plant, harvest index and yield/plant were conspicuous and contributing more to the total genetic divergence, which was also reflected by their high coefficient of variation.

Jaylal (1994) carried out genetic divergence study in forty genotypes of soybean using Mahalanobis D^2 statistic. The Wilk's test revealed highly significant differences ($\chi^2 = 242.31$) for all the characters, and he grouped the forty genotypes into 9 and 7 clusters respectively, based on physiological and yield attributes. The analysis for estimating the contribution of characters to the divergence indicated carotene content and total chlorophyll in the case of physiological and pods/cluster, branches/plant and seed yield/plant in the case of yield attributes contributed maximum to the total genetic divergence.

Singh *et al.* (1997) studied genetic divergence in 100 rice bean (*Vigna umbellata*) cultivars of different geographic origin using D^2 statistic. The analysis grouped the cultivars into 15 clusters. The random distribution of cultivars into the different clusters indicated the weak relationship between genetic distance and geographic diversity. Cluster mean of characters indicated the importance of a character within the different clusters and its significance for improvement.

CHAPTER THREE

MATERIALS AND METHODS

3.1 MATERIALS

Twenty wheat genotypes, listed in Table 3.1, were evaluated over a period of two years from 2007 to 2008 in 4 locations (Adet, Holeta, Kulumsa, and Sinana) under irrigated condition. The experimental layout was a randomized complete block design (RCBD) with four replications. Planting method was on 30 cm apart at a seed rate of about 120 kg/ha. Plots were managed conventionally and followed the established local practices but usually the plot area ranged from 10 to 15m².

Table 3.1 Genotype codes of 20 wheat genotypes

No.	Genotype code	No.	Genotype code	No.	Genotype code
1.	G1	8.	G8	15.	G15
2.	G2	9.	G9	16.	G16
3.	G3	10.	G10	17.	G17
4.	G4	11.	G11	18.	G18
5.	G5	12.	G12	19.	G19
6.	G6	13.	G13	20.	G20
7.	G7	14.	G14		

These wheat hybrids were selected based on their relative yield performance among the different experimental hybrids developed by the Ethiopian Seed Enterprise (ESE). These hybrids were released varieties adapted to the medium altitude wheat growing areas of Ethiopia. All the hybrids are categorized under the medium maturity group (between 140 and 145 days) and their broad adaptation zone is mid-altitude sub-humid which includes areas with an elevation range of 1000-2000m above sea level and an annual rainfall between 1000-1200mm.

The trials were conducted under irrigated conditions and fertilization at each site and other management activities were done according to the practices of each farmer (co-operator) for his

farm and the specific field. The whole plot was harvested to estimate grain yields and to reduce border effects, data were recorded from the two central rows of each plot. Grain yields are expressed in kg/ha at 12.5 moisture content.

The locations where the experiment was conducted were different in soil type, altitude, mean annual temperature and rainfall and considered as individual environment (Table3.2). Also the years differentiated in terms of mean seasonal rainfall. Therefore, locations in each year were considered as different environments. Hence, an environment is defined here as a location-year combination. Consequently, combinations of seasons (Year 1 and Year 2) and four locations were treated as eight environments (E1-E8). Therefore, the trial consists of the yield of g genotypes(G1, G2, ... , G20) is measured in e environments(E1, E2, ... , E8) each with r replicates(R1, R2, ... , R4).

Table 3.2 Description of the experimental sites and environments

Year	Locations	Codes	Altitude (masl)*	Soil texture	Agro-ecologies
2007	Adet	E1	2240	Nitosol	Tepid Moist Mid Highlands
	Holeta	E2	2400	Silt clay loam	Tepid Moist Mid Highlands
	Kulumsa	E3	2200	Clay loam	Tepid Sub-Moist Mid Highlands
	Sinana	E4	2400	Sandy loam	Cool Sub-humid Mid Highlands
2008	Adet	E5	2240	Nitosol	Tepid Moist Mid Highlands
	Holeta	E6	2400	Silt clay loam	Tepid Moist Mid Highlands
	Kulumsa	E7	2200	Clay loam	Tepid Sub-Moist Mid Highlands
	Sinana	E8	2400	Sandy loam	Cool Sub-humid Mid Highlands

* masl = meter above sea level

The data being considered here are obtained from trials conducted by the Ethiopian Institute of Agricultural Research (EIAR).

3.2.THE METHODOLOGY

3.2.1. Test of significance of differences among genotypes/populations: This can be done in two ways:

3.2.1.1 Univariate Analysis of Variance(ANOVA):

Analysis of variance was done separately for each environment. If the individual experiments are laid out as RCBD, separate ANOVA structure for each environment is:

Table 3.3 The general separate ANOVA structure for each environment

Sources of variation	Degrees of freedom	Sum of squares	Mean squares	Expected mean squares	"F" ratio
Replicates(blocks)	r-1	A	P	$\sigma_e^2 + \frac{r-1}{r} g \sigma_b^2$	P/R
Genotypes(entries)	(g-1)	B	Q	$r \sigma_g^2 + \sigma_e^2$	Q/R
Error	(r-1)(g-1)	C	R	σ_e^2	

Where, r = number of replications

g = number of genotypes

A = replication sum of squares

B = sum of square for genotypes(entries)

C = error sum of squares

P, Q, and R = mean sum of squares replications, Genotypes(entries) and error, respectively.

The statistical model for this design is

$$y_{ij} = \mu + G_i + B_j + \varepsilon_{ij} \begin{cases} i = 1, 2, \dots, g \\ j = 1, 2, \dots, r \end{cases}$$

Where y_{ij} is the observed yield response of the i^{th} genotype at the j^{th} block (replication), μ is an overall mean, G_i is the effect of the i^{th} genotype, β_j is the effect of the j^{th} block (replication) and ε_{ij} is NID $(0, \sigma^2)$ random error term. Genotype and blocks are considered as

fixed factors. Furthermore, the genotype and block effects are defined as deviations from the overall mean so that $\sum G_i = 0$ and $\sum_j = 0$.

By simple ANOVA as explained above, significance of difference among genotypes/populations for all or majority of environments would justify further calculation of D^2 .

3.2.1.2 Wilk's lambda criterion (Λ) (Wilk's, 1932)

A simultaneous test of significance of difference in the mean value of a number of correlated variables (Wilk's, 1932). For this purpose, sum of squares and sum of products of the experimental error were used and the significance of the genotypes was tested. The Wilk's lambda criterion was estimated using the following relationship.

$$\Lambda = \frac{|E|}{|T|} = \frac{|E|}{|E+G|} = \prod_{i=1}^n \frac{1}{1 + \lambda_i},$$

where λ_i denotes the i^{th} eigen value of the matrix $E^{-1}G$ $i = 1, 2, \dots, n$; $n = \min(p, q - 1)$, g is number of genotypes and p is the number groups / environments.

$|E| = \left| \sum_{i=1}^g \sum_{j=1}^{e_i} (X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i) \right|$ = determinant of the error dispersion (error sum of squares and products) matrix,

$|T| = |E + G| = \left| \sum_{i=1}^g \sum_{j=1}^{e_i} (X_{ij} - \bar{X})(X_{ij} - \bar{X}) \right|$ = (determinant of the error dispersion matrix plus genotype dispersion which is population sum of squares and sum of products) matrix.

$|G| = \left| \sum_{i=1}^g \sum_{j=1}^{e_i} (\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X}) \right|$ = determinant between group sum of squares and cross products (SSCP) matrix of the genotypic variables.

If there is no difference between genotype; i.e. $|G|=0$, then $\Lambda = 1$, whereas if $|G|$ is very large relative to $|E|$ then Λ tends to 0. The quantity $(1 - \Lambda)$ is often interpreted as the proportion of variance in the dependent variables explained by the model. However, this quantity is not unbiased and can be quite misleading in small samples. The sampling distribution of this is very complicated and generally an approximation is used.

The significance of Λ was tested by V-statistics (Bartlett, 1947) representing central χ^2 distributions at pq degree of freedom.

$$V\text{-stat} = -m \ln \Lambda = -m \log_e \Lambda$$

$$\text{Where, } m = t - \frac{(p + q + 1)}{2}$$

t = total degree of freedom (degree of freedoms for error plus genotypes).

p = number of environments

q = d.f. for genotypes

If the significance of V-stat shows that the difference between the means in respect of the pooled environments between different genotypes/populations is significant. Hence, further analysis can be made to estimate D^2 values.

3.2.2 Analysis of Genetic divergence

3.2.2.1 Multivariate analysis using D^2 statistic

Definition of D^2 statistic and Explanation

The Mahalanobis distance, also called generalized Euclidean distance, was introduced by P.C. Mahalanobis in 1936 as one of the most well known distance measure used in multivariate statistics which corrects the correlation in the variables due to the fact that it is calculated using the inverse of the covariance matrix.

A major problem with averaging univariate distances to estimate genetic distances is that it ignores the intercorrelations among variables. Ignoring intercorrelations among characters will cause the average univariate distance estimate to be biased (Atchley 1980). The Mahalanobis (1936) generalized distance was developed to resolve the problem of intercorrelated characters. This is a measure which overcomes the drawback of the Euclidean distance in correlated data where in uncorrelated data the Mahalanobis distance is equal to the Euclidean distance.

Mahalanobis squared distance is defined between two group populations G_1 and G_2 as

$$\Delta^2 = (\bar{\mu}_1 - \bar{\mu}_2)^T \Sigma^{-1} (\bar{\mu}_1 - \bar{\mu}_2),$$

where the super fix T denotes matrix transpose, Σ denotes the common (nonsingular) covariance matrix of vector of the measurements made in each group populations G_1 & G_2 and $\bar{\mu}_1$ & $\bar{\mu}_2$ are mean vectors of both populations. It can be seen that since Σ is a (nonsingular) covariance matrix, it is positive-definite and hence Δ is a metric.

Uses of Δ^2

Mahalanobis Δ^2 is mainly used in classification problems, where there are several groups and the investigation concerns the affinities between groups. The object of the study may be to form clusters of members who are similar to each other, perhaps in a hierarchical scheme. Another situation in which Mahalanobis Δ^2 is relevant is in the problem of pattern recognition or discriminant analysis, where a formula is developed on the basis of knowledge of μ_1 , μ_2 and Σ in order that a new element can be classified (assigned, identified or recognized) into one of these two groups with as little chance of error as possible under the circumstances.

Sample Version of the Mahalanobis Distance

In practice, the means μ_1 and μ_2 and the common covariance matrix Σ of the two groups G_1 and G_2 are generally unknown and must be estimated from random samples of sizes n_1 and n_2 from G_1 and G_2 , yielding sample means $\bar{x}_1$ and $\bar{x}_2$ and (bias corrected) sample covariance matrices S_1 and S_2 . The common covariance matrix Σ can then be estimated by the pooled estimate,

$$S = \{(n_1 - 1)S_1 + (n_2 - 1)S_2\} / N,$$

Where $N = n_1 + n_2 - 2$. The sample version of the Δ^2 is denoted by D^2 and is given by

$$D^2 = (\bar{x}_1 - \bar{x}_2)^T S^{-1} (\bar{x}_1 - \bar{x}_2).$$

Although D^2 is the sample Mahalanobis distance, it is usually referred to simply as the Mahalanobis distance, with Δ being referred to as the population or true Mahalanobis distance. The Mahalanobis distance D^2 is known to overestimate its population counterpart Δ^2 .

The method of derivation of the sampling distribution of D^2 is the same as that of what is known as Hotelling's T^2 since $T^2 = kD^2$, where $k = \frac{(n_1 n_2)}{(n_1 + n_2)}$, Rao (1973a). It follows under the assumption of normality that cD^2 is distributed as F-distribution with p and $N-p+1$ degrees of freedom and non centrality parameter $c\Delta^2$, where $c = \frac{k(N-p+1)}{(pN)}$. It can be used

obviously to test the null hypothesis that $\Delta^2 = 0$, or equivalently, $\mu_1 = \mu_2$. It also forms the basis of tests for the equality of two group means given that some covariates have also been observed on the individuals.

The Mahalanobis distance is thus a very useful statistic in multivariate analysis. It can be used, also for example, to test that an observed random sample $\mathbf{x}_1, \dots, \mathbf{x}_n$ is from a multivariate normal distribution. We can proceed to test this null hypothesis by forming the Mahalanobis (squared) distances $D_1^2, \dots, D_n^2$, where $D_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})^T \mathbf{S}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}})$, and $\bar{\mathbf{x}}$ and $\mathbf{S}$ denote respectively the sample mean and the (bias-corrected) sample covariance matrix of the n observations in the observed sample. Then under the null hypothesis the D_j^2 should be distributed independently (approximately), with a common distribution that can be approximated by a chi-squared distribution with p degrees of freedom. Alternatively, we can form the modified Mahalanobis distances $d_1^2, \dots, d_n^2$, where

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}}_{(j)})^T \mathbf{S}_{(j)}^{-1} (\mathbf{x}_j - \bar{\mathbf{x}}_{(j)}),$$

where $\bar{\mathbf{x}}_{(j)}$ and $\mathbf{S}_{(j)}$ denote respectively the sample mean and (bias-corrected) sample covariance matrix of the $n-1$ observations after the deletion of $\mathbf{x}_j$, ($j = 1, \dots, n$). In this case, the d_j^2 can be taken to be approximately independent with the common distribution of $q d_j^2$ given exactly by a F-distribution with p and $n-p-1$ degrees of freedom, where $q = \frac{(n-1)(n-p-1)}{\{(pn)(n-2)\}}$.

The values of D^2 or d_j^2 can be used also to assess whether $\mathbf{x}_j$ is an outlier.

The concept of distance between populations is frequently used in statistics. When the variables measured on individuals are continuous, a satisfactory distance measure between populations is the Mahalanobis D^2 [1936].

Genetic divergence analysis in general helps in assessing nature of diversity in order to identify specific parents for realizing higher heterosis for yield and other economic characters in hybrid breeding programme.

The genotypes are discriminated by employing the environment by the use of Mahalanobis D^2 statistic which measures the generalized distance between the two genotypes implicitly taking not only the distance between the two population means, but also the intrinsic variances and covariance's among these character. The squared Mahalanobis' generalized distance between genotype i and j is defined by:

Since

$$\mu_1 = (\mu_{11}, \mu_{12}, \dots, \mu_{1n})$$

is the mean column vectors for the first and second genotypes and the variance and covariance matrix Σ , assumed to be common for both, be denoted by;

$$\Sigma = n \times n \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \sigma_{13} & \dots & \sigma_{1j} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22}^2 & \dots & \dots & \sigma_{2j} & \dots & \sigma_{2n} \\ \sigma_{i1}^2 & \sigma_{i2} & \dots & \dots & \sigma_{ij} & \dots & \sigma_{in} \\ \sigma_{n1} & \sigma_{n2} & \dots & \dots & \sigma_{nj} & \dots & \sigma_{nn}^2 \end{pmatrix}$$

Then, the square of the generalized distance between the two genotypes is defined as Δ^2 .

$$\text{where, } \Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$$

Equivalently the generalized distance between any populations or genotype is defined as,

$$D_x^2 = \sum_i^p \sum_j^p \lambda^{ij} d_i d_j$$

Where,

x = number of different environments in point.

p = no of population/genotypes.

λ^{ij} = The reciprocal matrix to the common dispersion matrix (i.e., λ_{ij}).

d_i = The difference between the two mean values of the two populations for i^{th} environment $(\mu_{i1} - \mu_{i2})$.

$d_j = (\mu_{j1} - \mu_{j2})$ The difference between the mean values of the two populations for the j^{th} environment $(\mu_{i1} - \mu_{i2})$.

μ = Vector of mean values for all the environments.

Since, mean vectors μ_1 and μ_2 and the variance, covariance matrix Σ are not known, the sample estimate of D^2 is obtained by substituting the sample estimates for the population values viz., μ_1 , μ_2 and Σ . Thus, the sample estimate of Δ^2 , given by D^2 is defined as;

$$D^2 = (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2)$$

Where,

$$\bar{x}_i = (X_{i1}, X_{i2}, X_{i3}, \dots, X_{in}) \quad (i = 1, 2)$$

is sample mean vector of $\bar{x}$ and S is the pooled variance-covariance matrix computed from the sample estimates. Thus, S is defined as,

$$S = \begin{pmatrix} S_{11}^2 & S_{12} & S_{13} & \dots & S_{1j} & \dots & S_{1n} \\ S_{21} & S_{22}^2 & \dots & \dots & S_{2j} & \dots & S_{2n} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ S_{n1} & S_{n2} & \dots & \dots & S_{nj} & \dots & S_{nn}^2 \end{pmatrix}$$

In which,

$$S_i^2 = \frac{\sum_{j=1}^n (X_{ij} - \bar{X}_i)^2}{n-1} \quad \text{and} \quad S_{ij} = \frac{\sum_{k=1}^n (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{n-1}$$

are used as unbiased estimates of sample variances and covariances, respectively on n observations.

Formula for computation of D^2 values, which requires inversion of higher order determinant of the matrix, becomes complicated especially when the number of variables under consideration are large. Therefore, the original correlated unstandardised variables (X_i) was transformed to standardized uncorrelated variables (y_j), so that the computation of D^2 values reduce to simple summation of squares of the difference between values of transformed variables of the two populations i, e D_i^2 . This transformation is done by pivotal condensation method.

Pivotal Condensation Method: The method of pivotal condensation is useful for transformation of correlated unstandardised variables into standardized uncorrelated variables and convenient for practical computations of determining the value of a determinant. Only the relevant computations need be retained in any problem. In most of the statistical computations symmetric matrices are met with. This results in a certain amount of reduction in computations, and the layout of a simplified procedure in such cases is as follows.

- i. For a given $n \times n$ symmetric matrix put an equivalent unit matrix ($I_{n \times n}$) at the right hand side.
- ii. Obtain the first row of the reduction matrix from the first row of a given matrix by reducing the first element to unity. This is called a the first pivotal row of the reduction matrix and the first element from the first row of a given square matrix is the first pivotal element.
- iii. Multiply the first pivotal row by the first pivotal element and subtract from the second row of a given square matrix, the second row of the reduction matrix i.e. the row following from the first pivotal row is obtained in which the first element is zero.

- iv. Similarly, by eliminating the first element from all the remaining rows of the given square matrix, the remaining rows following from the first pivotal row are obtained.
- v. Starting with the reduced matrix in rows beginning from the first pivotal row the whole operation is repeated. The product of all pivotal elements is the value of the determinant of the given square matrix. If only the value of the determinant is needed, the calculation beyond the last column of the given square matrix i.e. for reduction of the unit matrix should be omitted.
- vi. A sum check provided in the last column ensures the accuracy of all the calculations. For subsequent operations this is treated as an extra column. Any derived row has the property that the last element is the sum of the other elements. This may be checked whenever the new row is obtained either by reducing the first coefficient to unity or by sweeping out a first column.

From the newly transformed uncorrelated variables, the square of the distance was computed using the following formula

$$D^2 = \sum_i^x d_i^2 = \sum_i^x (y_i^j - y_i^k)^2, \quad j \neq k$$

Where,

y = Uncorrelated variable (environment) which varies from i=1 to x, i.e., number of environments.

The super scripts j and k to y represent a pair of any two genotypes, eg., 1-2, 1-3, 1-4 ... 1-n; 2-3, 2-4 2-n; 3-4, 3-5; (n-1)-n, (where n is number of populations).

Thus, D^2 is the sum of squares of differences between any two populations for each of the uncorrelated variables [(Ys), obtained by transforming correlated variables (Es) through pivotal condensation method]. The D^2 values were obtained as the corresponding uncorrelated (Ys) values of any two genotypes (Rao, 1952).

The squared Mahalanobis' generalized distance between genotype i and j is defined by

$$D^2_{ij} = (\underline{\mu}_1 - \underline{\mu}_2)^T \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$$

where $\underline{\mu}_i$ is the vector of environmental means for genotype i and Σ is the common within population covariance matrix (Mahalanobis, 1936). The squared Mahalanobis' distance between genotype i and j can be expressed equally well as follows:

If $\tilde{x}_1$ and $\tilde{x}_2$ are two independent random vectors with $E(\tilde{x}_i) = \mu_i$, $i = 1, 2$ and $\text{cov}(\tilde{x}_i) = \Sigma = \{\sigma_{kl}\}$, $k, l = 1, \dots, P$. Then a formal definition of the Mahalanobis generalized distance is $D^2(\mu_1, \mu_2, \Sigma) = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$

FACT: The Mahalanobis distance is invariant under the following linear transformation:

$$\tilde{y}_i = A\tilde{x}_i - \tilde{c}$$

where,

$$\tilde{c} = \text{vector of known constants}$$

$$A = \text{Diagonal } (\sigma_{11}, \sigma_{22}, \sigma_{33}, \dots, \sigma_{pp})$$

$$\text{then, } E(\tilde{y}_i) = A\tilde{x}_i - \tilde{c} = \mu_i^*$$

$$\text{var}(\tilde{y}_i) = \text{var}(A\tilde{x}_i - \tilde{c})$$

$$= \text{var}(A\tilde{x}_i)$$

$$= \Sigma A \Sigma$$

$$= R$$

$$\begin{aligned} \text{Hence, } D^2(\mu_1^*, \mu_2^*, R) &= (A(\mu_1 - \mu_2))' R^{-1} (A(\mu_1 - \mu_2)) \\ &= (\mu_1 - \mu_2)' A (A \Sigma A)^{-1} A (\mu_1 - \mu_2) \\ &= (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2) \\ &= D^2(\mu_1, \mu_2, \Sigma) \end{aligned}$$

Therefore, the squared Mahalanobis' distance between genotype i and j can be expressed equivalently as:

$$D^2_{ij} = (\mu_i^* - \mu_j^*)' R^{-1} (\mu_i^* - \mu_j^*)$$

Where μ_i^* is the standardized vector of means for genotype i, μ_j^* is the standardized vector of means for genotype j and R is the common within population correlation matrix.

Then, the squared Mahalanobis distance is usually estimated by

$$\hat{D}^2_{ij} = (\bar{z}_i - \bar{z}_j)' R^{-1} (\bar{z}_i - \bar{z}_j)$$

where $\bar{z}_i$ is the estimated standardized vector of means for genotype i, $\bar{z}_j$ is the estimated standardized vector of means for genotype j and R the sample pooled within population correlation matrix.

Relationship Between Principal Components And Mahalanobis' Distance

Principal components: The method of principal component analysis consists in transforming the set of variables $x_1, x_2, \dots, x_p$ to a new set $y_1, y_2, \dots, y_p$ satisfying the following conditions:

1. Each y_i is a linear combination of the x 's, say

$$y_i = a_1x_1 + a_2x_2 + \dots + a_px_p$$

2. If y_j is another linear combination of the x 's, say

$$y_j = b_1x_1 + b_2x_2 + \dots + b_px_p$$

then we want $\underline{a}'\underline{a} = 1$, $\underline{b}'\underline{b} = 1$, $\underline{a}'\underline{b} = 0$. This last condition leads to the y 's being uncorrelated.

3. Of all possible combinations of this type, y_1 has the greatest variance, y_2 , the second highest variance and so on (Morrison, 1967).

A geometrical interpretation is that principal components result in rotation of the coordinate axes to a new coordinate system with certain statistical properties such as the new set of variables are uncorrelated, and their variances are the nonzero characteristic roots of the covariance matrix of the original set of variables.

The technique of principal components has been introduced here for several reasons. It can be shown that the squared Mahalanobis' distance between genotype i and j is also expressed by based upon the new set of variables obtained by the principal components technique as follows:

Let $x_1, x_2, \dots, x_p$ be the elements of the random vector $\underline{x}$ with mean $\underline{\mu}$ and covariance matrix Σ positive semidefinite. Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ be the eigenvalues of Σ . From matrix theory, we can find an orthonormal matrix P of order $p \times p$ whose columns are the associate orthonormal characteristic vectors of Σ such that

$$\Sigma P = P \Lambda \quad \Sigma = P \Lambda'$$

where

$$\Lambda = D \quad \{ \lambda_1, \lambda_2, \dots, \lambda_p \}$$

Consider the orthogonal transformation

$$\underline{y} = P'(\underline{x})$$

then, $y_1, y_2, \dots, y_p$, the elements of $\underline{y}$, are called the principal components of $\underline{x}$. The first principal component of $\underline{x}$: which correspond to the maximum eigen-value λ_1 is y_1 , y_2 is the second, etc.

Note,

$$E(\underline{y}) = p'(\underline{\mu}) \quad \text{and} \quad \text{Var}(\underline{y}) = p'\Sigma p = \Lambda$$

Now, consider $\underline{x}_1, \underline{x}_2$ be two random vectors with $E(\underline{x}_i) = \underline{\mu}_i$ and $\text{var}(\underline{x}_i) = \Sigma$, $i = 1, 2$.

Let P be defined as above, then the transformation

$$\underline{y}_i = p'(\underline{x}_i)$$

implies, $E(\underline{y}_i) = p'(\underline{\mu}_i) = \underline{\xi}_i$, $\text{Var}(\underline{y}_i) = \Lambda$. F formal definition of the Mahalanobis generalized distance,

$$\begin{aligned} D^2(\underline{\xi}_1, \underline{\xi}_2, \Lambda) &= (p'(\underline{\mu}_1) - p'(\underline{\mu}_2))' \Lambda^{-1} (p'(\underline{\mu}_1) - p'(\underline{\mu}_2)) \\ &= (\underline{\xi}_1 - \underline{\xi}_2)' \Lambda^{-1} (\underline{\xi}_1 - \underline{\xi}_2) \\ &= \sum_{k=1}^p \left[(\xi_{1k} - \xi_{2k})^2 / \lambda_k \right] \end{aligned}$$

Therefore, the squared Mahalanobis' distance between genotype i and j is also expressed by

$$D^2_{ij} = (\underline{\xi}_i - \underline{\xi}_j)' \Lambda^{-1} (\underline{\xi}_i - \underline{\xi}_j)$$

where $\underline{\xi}_i$ is the vector of principal component means for genotype i; i. e., $\underline{\xi}_i$ is the mean vector for genotype i based upon the new set of variables obtained by the principal components

$$q \quad ; \quad \Lambda \quad \quad \quad x \quad \quad \quad f \quad \quad \quad x \quad \Sigma.$$

Thus the relation between principal components analysis and the Mahalanobis' generalized distance analysis is direct.

Now let $\underline{\xi}_i = (\xi_{i1}, \xi_{i2}, \dots, \xi_{ip})'$ then the squared Mahalanobis' generalized distance can also be expressed by

$$D^2_{ij} = \sum_{k=1}^p \left[(\xi_{ik} - \xi_{jk})^2 / \lambda_k \right]$$

where λ_k are the characteristic roots of Σ .

Then, the Mahalanobis' squared generalized distance can be estimated by

$$\begin{aligned} \hat{D}^2_{ij} &= (\bar{y}_i - \bar{y}_j)' D^{-1} (\bar{y}_i - \bar{y}_j) \\ &= \sum_{k=1}^p \left[(y_{ik} - y_{jk})^2 / d_k \right] \end{aligned}$$

where $\bar{y}_i = (y_{i1}, y_{i2}, \dots, y_{ip})$ is the estimated vector of principal component means for genotype i , and D is a diagonal matrix whose elements, d_k , are the characteristic roots of the sample covariance matrix S .

Estimation of Standard Error for D^2

The squared Mahalanobis distance D^2 between genotype 1 and genotype 2 was estimated by

$$D^2 = (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2)$$

where

$\bar{x}_i$ = vector of means for genotype $i, i=1,2$; and

$S = \{(n_1 - 1)S_1 + (n_2 - 1)S_2\}/N$ where $N = n_1 + n_2 - 2$ is the pooled variance covariance

matrix of genotype 1 and 2.

Assuming that the samples sizes are large enough that $S^{-1} = \Sigma^{-1}$ and $\bar{x}_i$, is $N_p(\mu_i, \Sigma)$, $i=1,2$; then an approximate variance for the squared D^2 distance can be obtained by using Searle's(1971) result.

$$\begin{aligned} \text{Var}(D^2) &= \text{Var}((\bar{x}_1 - \bar{x}_2)' \Sigma^{-1} (\bar{x}_1 - \bar{x}_2)) \\ &= 2\text{tr}(\Sigma^{-1}(\frac{\Sigma}{n_1} - \frac{\Sigma}{n_2}))^2 + 4(\mu_1 - \mu_2)' \Sigma^{-1}(\frac{\Sigma}{n_1} + \frac{\Sigma}{n_2}) \Sigma^{-1}(\mu_1 - \mu_2) \\ &= 2\text{tr}((\frac{1}{n_1} - \frac{1}{n_2})^2 \Sigma^{-1}(\Sigma)^2) + 4(\mu_1 - \mu_2)' \Sigma^{-1}(\frac{1}{n_1} + \frac{1}{n_2}(\Sigma)) \Sigma^{-1}(\mu_1 - \mu_2) \\ &= 2\text{tr}((\frac{1}{n_1} - \frac{1}{n_2})^2 (I_p)^2) + 4(\mu_1 - \mu_2)' \Sigma^{-1}(\frac{1}{n_1} + \frac{1}{n_2})(\mu_1 - \mu_2) \\ &= 2(\frac{1}{n_1} + \frac{1}{n_2})^2 \text{tr}(I_p)^2 + 4(\frac{1}{n_1} + \frac{1}{n_2})(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2) \\ &= 2(\frac{1}{n_1} + \frac{1}{n_2})^2 (1 + 1 + \dots + 1) + 4(\frac{1}{n_1} + \frac{1}{n_2})(\mu_1 - \mu_2)' \Sigma^{-1}(\mu_1 - \mu_2) \\ &= 2p(\frac{1}{n_1} + \frac{1}{n_2})^2 + 4(\frac{1}{n_1} + \frac{1}{n_2})D^2 \end{aligned}$$

where D^2 is the population Mahalanobis' distance. This approximate variance agrees with the exact variance of the studentized- D^2 statistic obtained by Bose and Roy. The studentized- D^2 statistic is a linear transformation of the squared Mahalanobis' distance D^2 .

From above, an estimator for the variance of D^2 is

$$\widehat{\text{Var}}(D^2) = 2p(\frac{1}{n_1} + \frac{1}{n_2})^2 + 4(\frac{1}{n_1} + \frac{1}{n_2})D^2$$

3.2.2.2 Test of significance of D^2 values

Each D^2 value will be compared to the calculated chi-squared values at p degree of freedom (i.e. χ_p^2). If D^2 values are greater than the calculated chi-squared (χ_p^2) values they are all significant (i.e. significantly different from zero).

3.2.2.3 Contribution of individual environments for divergence

Literature abounds with reports on using the rank method of d^2 to get the total rank, where lowest rank-total means highest contribution and vice-versa for all combinations $n(n-1)/2$ of each environment ranked from 1 to p on the basis of $d_i = (y_i^j - y_i^k)$ values, where the super scripts j and k to y represent a pair of any two genotypes, eg., 1-2, 1-3, 1-4 ... 1- n ; 2-3, 2-4 ... 2- n ; 3-4, 3-5 ...; ...-1)- n , (where n is number of populations) and d_i is the differences between any two genotypes for each of the uncorrelated variables Y_s . Rank 1 refers to the highest mean difference while rank p (number of environments) to the lowest mean difference. Percent contribution of an environment towards divergence is calculated by counting number of times an environment appear first in ranking. However, this method has been criticized for lack of any weighted assigned to different ranks and also for the fact that ranking is done on uncorrelated variables (Y_s), which might not be applicable to the original correlated means (E_s). But, it can be easily determined by coefficients of variation for each E_s at inter - cluster level. The relative contribution of each environment to divergence is of great importance in identifying those environments having the highest contribution, and also in excluding those with the least (Missio *et al.*, 2007).

3.2.2.4 Cluster Analysis (Group Constellation)

Cluster analysis is a numerical classification technique that defines groups or clusters of individuals. Cluster analysis methods have been known also as Q-analysis, typology, grouping, clumping, classification, numerical taxonomy and unsupervised pattern recognition methods. Basically, this technique for analyzing multivariate data deals with the grouping of a sample of n

objects or individuals with p variates into k classes (Everitt, 1974, 1979). Cluster analysis is the most widely used technique for classifying environments or genotypes into homogeneous groups i.e. it has been used to study the adaptation of genotypes to environments by simplifying the pattern of responses and by subdividing genotypes and environments into more homogeneous categories (Lin et al. 1986; Crossa 1990; Piepho 1998). It operates on a matrix of dissimilarity (or similarity) indexes for all possible pairs of genotypes or pairs of environments, depending on which is being clustered (Ghaderri *et al.*, 1980). The performance of the most stable genotypes in subsets of clusters can also be used to determine and compare the stability of other genotypes. Cluster analysis is a potentially very useful tool for the exploration of complex multivariate data, and is especially valuable for grouping together locations, environments or genotypes that show similar response patterns. Classification of genotypes, environments or locations into homogeneous groups is of major importance for improving the efficiency of breeding programmes (Abou-El-Fittouh et al. 1969; Jensen 1988).

Two types of classification can be distinguished. The first is non-hierarchical classification, which assigns each item into a class. The second type is hierarchical classification, which groups individuals into clusters and arranges these into a hierarchy for the purpose of studying relationships in the data (Crossa, 1990) i.e. has used to process the values of a similarity matrix resulting in a diagram known as a tree or dendrogram that shows the hierarchy of similarities among all pairs of objects (Lefkovitch 1980; Romesburg 1984; Afifi & Clark 1996). The most important of the choice is the hierarchical classification techniques. A hierarchical strategy always optimizes a 'route' between the entire population and the set of individuals of which it is composed. The route may be defined by progressive fusions, beginning with individuals and ending with the complete population; or by progressive divisions, beginning with the population and decomposing into individuals or sub-groups.

In the hierarchical method, it is important to evaluate the adjustment between the original distance coefficient matrix (phenetic matrix ph) and the matrix that results from clustering process (cophenetic matrix C). Thus, the cophenetic correlation coefficient (CCC) proposed by Sokal & Rohlf (1962) was used, which is estimated based on the following formula:

$$CCC = r_{\text{coph}} = \frac{\sum_{j=1}^{n-1} \sum_{j'=j+1}^n (c_{jj'} - \bar{c})(ph_{jj'} - \overline{ph})}{\sqrt{\sum_{j=1}^{n-1} \sum_{j'=j+1}^n (c_{jj'} - \bar{c})^2} \sqrt{\sum_{j=1}^{n-1} \sum_{j'=j+1}^n (ph_{jj'} - \overline{ph})^2}}$$

where:

$$\bar{c} = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{j'=j+1}^n c_{jj'} \quad \overline{ph} = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{j'=j+1}^n ph_{jj'}$$

n = number of subjects.

The higher the CCC value, the lower is the distortion caused by the grouping. In practice, it should be considered that dendrogram with CCC higher than 0.7 indicate that the clustering method is adequate to resolve the information from a dataset (Rohlf, 1970).

A dendrogram linking together clusters of samples that is more genetically similar to each other than to samples in other clusters is produced from the resulting distance matrix, mostly using the Unweighted Pair Group Method with Arithmetic Averages (UPGMA) or the Neighbour Joining Method which is one of the most popular hierarchical cluster analysis approaches from a variety of alternative agglomerative procedures among numerical taxonomists (and paleontologists).

- ❖ Unweighted - all pairwise distances contribute equally.
- ❖ Pair-Group - groups are combined in pairs (dichotomies only).
- ❖ Arithmetic mean - pairwise distances to each group (clade) are mean distances to all members of that group.

In this method, the distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters. In most instances UPGMA maximizes the cophenetic correlation coefficient of a cluster analysis and so produces results with minimum levels of distortion (Farris 1969; Sokal and Rohlf 1970). The UPGMA approach does this by attempting to use a greater proportion of the information present in the similarity matrix.

A process of division, in which all individuals start by being alone in groups of one and then closer groups are merged gradually until finally all individuals form a single group, is used to form clusters. Individual values of genotypes or environments, the subgroups are made which further are continued to be successively grouped until all variation is covered in one large group. The whole

structure from the individual values to final groups is represented like branches of a tree in a dendrogram that depicts composition of groups and degree of dissimilarity among groups both for genotypes as well as environments. The genotypes or environments in each group or cluster are expected to have a similar contribution towards as compared to the constituents of the other group at each level of clustering (Ramagosa and Fox, 1993).

3.2.2.4.1 Aims of clustering analysis

A good clustering will explain a large portion of the sums of squares that exist in a data set in the most parsimonious manner. Cluster analysis technique have been employed by researchers to meet wide-ranging objectives. These include:

- Clustering genotypes when there were many to compare across a smaller of environments, so that the average or best genotype from each cluster could be compared(Mungomery et al., 1974).
- Clustering genotypes to ensure they are compared to the most similar check cultivars (Lin and Binns,1985).
- Clustering environments when there were a smaller number of genotypes to compare, so that genotypes could be compared to the average environments from each cluster (Ivory et al, 1991).
- Researchers can establish which environments are so similar that they give the same results, and therefore do not all need testing in future thus leading to resource efficiency(Baril et al., 1994).
- Clustering both genotypes and environments to reduce the number of comparisons necessary (Baril et al, 1994; Cooper and DeLacy, 1994).

3.2.2.4.2 General guide-lines for cluster analysis

- ❖ The limit of divergence within groups should be based on the range of all $\frac{n(n-1)}{2} D^2$ - values. However, as group - formation proceeds from one to the next, this limit gradually increases.
- ❖ Divergence between groups (inter-cluster average D^2 - values) should not be less than that within groups (intra-cluster average D^2 - values).
- ❖ Ideally, the divergence within a cluster must not decrease on considering some additional characters. Conversely, increase in divergence must be relatively small so that group constellation based on a chosen set of characters are not distorted. This, in fact, is merely a practical necessity.
- ❖ A recurrent exercise of grouping of genotypes leads to perfection in the technique of cluster-formation.

3.2.2.4.3 Criterion for Clustering

Tocher's optimization method: All D^2 values are arranged in matrix form, based on the degree of divergence (D^2 values) between any two genotypes/populations. All the $\frac{n(n-1)}{2} D^2$ values were considered for determining the population constellation. This was realized by using Tocher's (1952) criterion used in clustering by this is that any two varieties belonging to the same cluster, should at least, in average, show a smaller D^2 value than those belonging to different clusters. As per the device it was to start with two closely associated population and find a third population, which had the smallest average D^2 from these two. Similarly, the fourth was chosen to have a smallest average D^2 from the first three and so on. The permissible increase in D^2 values for clustering into the same group was fixed approximately nearer the maximum D^2 value shown by a population to the nearest population. This procedure was continued till D^2 values of all the pairs of genotypes were exhausted. After the formation of the clusters inter and intra group distance was calculated. The square root of the average D^2 values obtained from the above represents the distance (D) between and within clusters.

3.2.2.5 Intra and inter cluster divergence

Once, the genotypes are grouped into clusters, the meaningful quantitative measures may be constructed to indicate the average distance between any two genotypes included in the same cluster, this being determined for each cluster formed. Likewise it would be useful to know the average distance between any two different clusters, this being determined for all pairs of clusters.

3.2.2.6 Intra cluster divergence

If there are g_k genotypes in a cluster and D_{ij}^2 is the Mahalanobis D^2 value between i^{th} and j^{th} ($j = 1, 2, \dots, k$) with $D_{ii} = 0$ and $D_{jj} = 0$, there will be;

$\frac{g_k(g_k-1)}{2}$ i.e., $\binom{g_k}{2}$ pairs of genotypes and hence that many D^2 values.

Then the (average distance)² is defined as the average of all such D^2 values i.e.,

$$\text{Intra-cluster distance } (D^2) = \frac{\sum_i^{g_k} \sum_j^{g_k} D_{ij}^2}{\frac{g_k(g_k-1)}{2}}, i \neq j.$$

Finally, the intra cluster distance is obtained as the square root of this average D^2 .

That is average intra cluster distance.

$$D = \left(\frac{2 \sum_{i=1}^{g_i} \sum_{j=1}^{g_i} D_{ij}^2}{g_i(g_i-1)} \right)^{\frac{1}{2}}$$

Or in short,

Intra-cluster is calculated by the formula $\sum D_i^2 / n$, where $\sum D_i^2$ is the sum of distances between all possible pairs of the genotypes/populations included in a cluster, n is the number of pairs in a cluster.

3.2.2.7 Inter cluster divergence

If there are g_1 genotypes in cluster I and g_2 genotypes in cluster II, there will be $g_1.g_2$ pairs in all. The average D^2 of all such $g_1.g_2$ pairs will be the (average distance)² between cluster I and cluster II.

$$(\text{Inter Cluster Distance})^2 = \frac{\sum_{i=1}^{g_2} \sum_{j=1}^{g_1} D_{ij}^2}{g_1 g_2}$$

The square root of the above quantity represents the average distance between cluster I and cluster II.

$$\text{Average distance between cluster I and cluster II} = \left(\frac{\sum_{i=1}^{g_2} \sum_{j=1}^{g_1} D_{ij}^2}{g_1 g_2} \right)^{\frac{1}{2}}$$

These two average distances among clusters will then reflect the extent to which genotypes within a cluster resemble among themselves and how they are dissimilar when grouped in different clusters.

Or in short,

Inter-cluster distance is calculated by averaging all possible D^2 values among genotypes/populations belonging to different clusters concerned, divided by number of pairs involved. The formula is: $\frac{\sum D_{ij}^2}{n_i n_j}$, where, $\sum D_{ij}^2$ is the sum of D^2 values of genotype pairs of clusters i and j ; and n_i and n_j are the number of pairs of cluster i and j , respectively.

3.2.2.8 Cluster diagram

Diagrammatic representation of cluster divergence showing different genotypes/populations can be pictorially presented using $\sqrt{D^2}$ values between and within the clusters.

3.2.2.9 Cluster mean analysis

Mean performance of all genotypes within a cluster will be obtained as the mean of over - all values of individual correlated variables i.e. environments included in that particular cluster.

3.2.3 Canonical Variate Analysis

Canonical variate analysis (CVA) is an alternative multivariate method that can be used to classify individual genotypes or environments of pre-known classes into two or more alternative categories on the basis of a set of measurements (Tabachnick & Fidell 1989; Afifi & Clark 1996). The pre-known classes of individuals could be breeding lines, tissue culture derived materials, standard checks, or environmental factors, such as localities, years, soil types,

climates, management practice, etc. If some of these dummy or indicator variables are not known, then clustering analysis should be used since it is a method for constructing an unknown classification (Afifi & Clark 1996). CVA is an extension of multiple regression and applies where there is more than one moderately interconnected dependent variable (Tabachnick & Fidell 1989; Afifi & Clark 1996). It is often applied when there is more interest in differences between groups than between individuals (Van Lill et al. 1995).

During CVA, the variability in a large number of variates is first reduced to a smaller set of variates that accounts for most of the variability in the data set. Then the new variates (canonical variates) are linear combinations of the original measurements. This helps to maximize variability between groups of individuals. A set of genotypes or localities can, thus, be grouped into various subgroups in accordance with their defined observations. CVA is also used to show which variables discriminated more among the genotypes and localities by using their standardized model coefficients (Van Lill et al. 1995; Afifi & Clark 1996). In CVA, it is easier to interpret the first canonical variate than the subsequent ones. Moreover, a graphical analysis with just two values is valuable for illustrating the maximum possible separation among the groups. The advantage of the canonical procedure is that it helps to interpret the results when there are two or more groups.

This technique is also known as discriminant function analysis, multiple discriminant, multiple-group discriminant, step-wise discriminant and canonical analysis. It is a generalization of the two group discriminant or discriminant function analysis. Green (1978) summarized the general objectives of canonical variate analysis (CVA) as follows:

- Find linear transformations of variables with the property of maximizing among groups to within groups variability, subject to each linear transformation being uncorrelated with previously obtained transformations. The transformations are computed so that each successive transformation accounts for a smaller proportion of the total variation.
- Test if group centroids are different, and if so, the number of discriminant axes (i.e., the dimensionality of the discriminant space) for which this is the case.
- Find which variables contribute most to discriminating among groups.

- Assign new objects to groups on the basis of the objects' measurements.

Whitehouse (1971) considered that the most important result of CVA is an estimate of the observer's ability to discriminate between populations. Blackith and Reyment (1971) indicated that the main use of CVA in morphometrics is for disclosing relationships among groups.

Statistically, this approach does not assume any particular parametric form for the distributions of the populations, but merely looks for a "sensible" rule to discriminate between them (Mardia *et al.* , 1979) . To get the linear transformation $\underline{b}'\underline{X}$, which maximizes the ratio of the between groups sum of squares to the within groups sum of squares, consider that:

$$y_i = \underline{b}' \underline{x} = b_1x_1 + b_2x_2 + \dots + b_px_p ,$$

where $\underline{b}$ is a $p \times 1$ vector of constants and $\underline{b}'\underline{X}$ is a scalar with mean $\underline{b}'\underline{U}$ and variance $\underline{b}'\underline{\Sigma}\underline{b}$. The total sum of squares of the new transformed variable y is defined as:

$$SST_y = \sum_{j=1}^k \sum_{i=1}^{n_j} (\underline{y}_{ij} - \underline{\bar{y}}) (\underline{y}_{ij} - \underline{\bar{y}})' = \underline{b}'\underline{T}\underline{b} ,$$

which can be partitioned as a sum of the within-groups sum of squares

$$SSW_y = \sum_{j=1}^k \sum_{i=1}^{n_j} (\underline{y}_{ij} - \underline{\bar{y}}_j) (\underline{y}_{ij} - \underline{\bar{y}}_j)' = \underline{b}'\underline{E}\underline{b}$$

plus the between-groups sum of squares

$$SSB_y = \sum_{j=1}^k n_j (\underline{\bar{y}}_j - \underline{\bar{y}}) (\underline{\bar{y}}_j - \underline{\bar{y}})' = \underline{b}'\underline{B}\underline{b}$$

To see if the between-group sum of squares for y is large relative to the within-group sum of squares, the ratio

$$\frac{SSB_y}{SSW_y} = \frac{\underline{b}'\underline{B}\underline{b}}{\underline{b}'\underline{E}\underline{b}} = F ,$$

where F is a scalar, is used.

Mathematically, the problem in CVA is to maximize F with respect to $\underline{b}$. Goodman (1973) has reported that this involves finding the partial derivative of F with respect to $\underline{b}$ and setting it equal to zero as follows:

Let

$$\log F = \log \underline{b}'\underline{B}\underline{b} - \log \underline{b}'\underline{E}\underline{b}$$

$$\frac{\partial \log F}{\partial \underline{b}} = \frac{2B\underline{b}}{\underline{b}'E\underline{b}} - \frac{2E\underline{b}}{\underline{b}'E\underline{b}} = 0,$$

thus $\underline{b}$ is the solution of the equation

$$(\underline{b}'E\underline{b}B - \underline{b}'B\underline{b}E)\underline{b} = 0.$$

For this equation to have a nontrivial solution, the determinant

$$|\underline{b}'E\underline{b}B - \underline{b}'B\underline{b}E| \text{ must be equal to zero or } |B - \frac{\underline{b}'B\underline{b}E}{\underline{b}'E\underline{b}}| = 0. \text{ By pre-multiplying by } E^{-1}, \text{ then}$$

$$|E^{-1}B - \lambda I| = 0.$$

if we let

$$\lambda = \frac{\underline{b}'B\underline{b}}{\underline{b}'E\underline{b}}.$$

The canonical variates are found from the latent roots' (λ_i) and vectors of this determinantal equation, and the largest root λ_1 will be the maximum ratio of variances.

Thus, the solution of $|B - \lambda_i E| \underline{b}_i = 0$ will yield the desired transformation, $y = \underline{b}_i \underline{x}_i$ apart from an arbitrary constant multiple of b_i usually chosen as $1/\sqrt{\underline{b}_i' E \underline{b}_i}$ (i.e., $\underline{b}_i^* = \underline{b}_i / \sqrt{\underline{b}_i' E \underline{b}_i}$ so that $\underline{b}_i^* E$

$\underline{b}_i^* = 1$) in which case the variance of y_i among populations is $\underline{b}_i^* A \underline{b}_i^* = \lambda_i$. The remaining eigen values, $\lambda_2, \lambda_3, \dots, \lambda_p$, and their corresponding eigenvectors $\underline{b}_2, \underline{b}_3, \dots, \underline{b}_p$ represent transformations of the data resulting in successively lower F values as the corresponding values of λ_p decrease.

To test the statistical significance of individual discriminant axes, Green (1978) presents a modification of the Bartlett's V statistic which is defined as

$$V = - \left[(n-1) - \frac{(p+k)}{2} \right] \ln \Lambda$$

where $V \sim \chi^2$ with $p(k-1)$ df, where

$$\frac{1}{\Lambda} = |E^{-1}T| = |E^{-1}(E + B)| = |I + E^{-1}B|.$$

Since the determinant of a matrix is equal to the product of its eigenvalues λ_i / Λ

$(1 + \lambda_1)(1 + \lambda_2) \dots (1 + \lambda_p)$, where λ_p are the p non-zero eigen values of $I + E^{-1}B$. Since $-\ln(a) = \ln(1/a)$, the Bartlett V can be modified to

$$V = - \left[(n-1) - \frac{(p+k)}{2} \right] \sum_{j=1}^p \ln(1 + \lambda_j).$$

This test allows removal of one or more of the smaller $(1 + \lambda_j)$ terms and the carrying out of further tests. The first null hypothesis to test is that the group centroids are equal in the full discriminant space. If this hypothesis is rejected, separate axes of the space are tested starting with the first (highest variance accounted for) axis.

The rejection of the overall null hypothesis implies that at least the first axis is statistically significant. However, significance testing in morphometries is considered definitely of subordinate value by Blackith and Reyment (1971).

Rempe and Weber (1972) and Albrecht (1980) presented a geometric description of the CVA process as follows: CVA starts with a rigid rotation of the original coordinate axis so as to be parallel with the major and minor axis of the within group dispersion ellipses, The transformed variates will be uncorrelated and all morphological relationships and morphometric relationships are preserved. The next step is to rescale the transformed variates so as to standardize the within group dispersions along all the axes (so that the within-group dispersion ellipses become circular). The deformation of transformed data space affects both the within-group and the between-group differences. If the within-group dispersion in the original data space represents uncorrelated variables, the first step is not necessary. This is equivalent to PCA based on the within-group dispersion.

The final step is to rigidly rotate the rescaled, standardized axes so as to be parallel with the major and minor axes of the between group dispersion. The canonical variates are constructed to be mutually orthogonal and statistically independent. This step is equivalent to a PCA based on the between-group dispersion matrix calculated from the rescaled standardized variates.

The method of mahalanobis distance has also well known application for assessing multivariate normality and outlier described as follows.

3.2.4 Graphical Method For Assessing Multivariate Normality using Mahalanobis Distance

In order to check multivariate normality, one can also use a graphical method.

Similar to the Quantile-Quantile plot for assessing univariate normality, there is a chi-squared plot based on Mahalanobis Distance (MD) (Sharma, 1996). According to Johnson (2007), when the population is normal and sample size is at least 25, MD behave like random variable with a

chi-squared distribution. In order to obtain the chisquared plots the following steps should be done:

i) Calculate MD by using the following formula;

$$MD = \sqrt{\left((x_i - \bar{x})' \hat{\Sigma}^{-1} (x_i - \bar{x}) \right)}, i=1,2,\dots,n.$$

ii) Arrange squared Mahalanobis distances (MD^2) in ascending order;

iii) For each MD^2 , compute $\left(\frac{j-0.5}{n} \right)$ percentile, where j is the ordered observation number;

iv) Get the corresponding χ^2 values for the percentile obtained in step (iii) from the χ^2 distribution with p df, where p is the groups / environments;

v) Create a scatter plot with (MD^2, χ^2) .

Under the null hypothesis of multivariate normality, the plot should be linear and any deviation indicates non-normality of the data.

3.2.6 Multivariate Outliers

One of the core assumptions of most parametric multivariate techniques is multivariate normality; which implies the absence of multivariate outliers. It is important to realize, cases which are multivariate outliers may not necessarily be univariate outliers. Stated another way; being an outlier on one of the variables under consideration is not a necessary condition of being

. O k f ,

(Mahalanobis, 1936).

3.2.6.1 Mahalanobis Distance Method

Mahalanobis distance is often used to determine multivariate outliers due to its several strengths:

- It features numerical and graphical thresholds to determine outliers.
- It is exible and permits the use of robust and independent choices for the centroid and covariance matrix values.
- It is amenable to techniques for mitigating the inuence of outliers during the search for the outliers.
- It can detect unusual patterns within a multivariate observation.

- It provides an alternative to regression techniques when there is no obvious value to be predicted.

Outlier detection belongs to an essential task in the data analysis. Usually the outliers in a dataset describe the unusual behavior of data. Many methods have been proposed to detect outliers in a univariate case. They are based on the estimation of location and shift parameters. The basis for multivariate outlier detection is the Mahalanobis distance. Rousseeuw (1990) proposed the method to detect multivariate outliers based on robust estimation of the parameters in the Mahalanobis distance.

The shape of multivariate data is quantified by the covariance matrix, and a well known distance measure which takes into account the covariance matrix is the Mahalanobis distance;

$$MD_i = \sqrt{(x_i - \bar{x})' \hat{\Sigma}^{-1} (x_i - \bar{x})},$$

where $\bar{x} = 1 \ 2 \ \dots \ p$ the estimated covariance matrix. For multivariate normally distributed data MD_i^2 are approximately chi-square distributed with p degrees of freedom. Multivariate outliers can now be defined as observations having a large (squared) Mahalanobis distance. Single extreme observation or group of observations, departing from the main data structure can have a severe influence to this distance measure, because both the location and the covariance matrix are usually estimated in a non-robust manner. Many robust estimators for the location and the covariance matrix have been introduced in some literatures. Using robust estimators one can get so-called robust distances (RD). Rousseeuw (1990) used these RDs for multivariate outlier detection.

CHAPTER FOUR

RESULT AND DISCUSSION

4.1 VARIATION (SIGNIFICANCE OF DIFFERENCES) AMONG GENOTYPES

4.1.1 UNIVARIATE ANALYSIS OF VARIANCE(ANOVA)

The statistical analyses presented in this thesis were done using SAS , R , STATA and Microsoft Excel statistical (software) packages.

The usual diagnostic plots-including a normal probability plot of residuals (i.e. for separate ANOVA for each environment Fig A.1 of Appendix), plot of residuals versus fitted values, plot of residuals versus level of regressor variable and formal statistical procedures to assess model assumptions for the individual ANOVAs (i.e. for yield data at each environment) were performed. Examination of the results do not reveal any serious violations of the assumptions that errors are normally and independently distributed with mean zero and constant variance. In the Shapiro-Wilk (W) test for normality, the p-value is based on the assumption that the distribution is normal. In this study, the p-value for each of the separate ANOVA model is very large (>0.05), indicating that we cannot reject the hypothesis that the residual is normally distributed (see the output of the proc univariate in Table A.4 of Appendix). For homogeneity of

f NOV B ‘

considering locations as a group. Based on this, the p value for each of the separate ANOVA is much greater than 0.05 (see Table A.3 in Appendix) indicating the hypothesis that the residual variances in each of the separate ANOVA are homogeneous cannot be rejected. Thus, the residuals have constant variance. Plots of residual versus the fitted value for each of the separate ANOVA are given in Fig A.1 of Appendix. There should be no relationship between the size of the residual and the fitted values. These plots reveal nothing of unusual interest. Thus, the usual interpretation for each of the separate ANOVA is valid since all assumptions concerning the data are met.

Separate analyses of variance (ANOVA) were first done with the classical RCBD model with replications (block) at each environment. Summary results of the separate analysis are given in Table A.2 of appendix.

Following the procedures of ANOVA laid down out for RBD on the replicated data (genotype=20, replication=4, environment=8), the following result for the individual environment were obtained (Table 4.1):

Table 4.1 Mean Sum of Square for 20 genotypes in bread wheat

Source of Variation	d.f.	MSS							
		E1	E2	E3	E4	E5	E6	E7	E8
Replication	3	303817.30	624283.79	49100.445	986869.22	1439094.46	742661.087	311891.86	258916.68
Genotype	19	1234073.43**	573069.07**	270052.860**	1083925.93**	1045726.57**	404526.804**	1068196.78**	694304.84**
Error	57	485943.88	217170.16	133636.19	574896.92	123295.46	197121.67	502225.07	366847.15

** = Significant at 1% probability

Significance of difference among genotypes for all environments would justify further calculation of D^2 . However, pooled differences can be determined by Wilk's criterion as below:

4.1.2 WILK'S CRITERION OF AGGREGATE VARIATION

Wilk's statistic which follows V - statistic representing χ^2 - distribution (at pq d.f., p being the number of environments and q the d.f. for genotypes) is

$$\Lambda = \frac{|E|}{|T|} = \frac{|E|}{|E+G|}$$

the $|E|$ and $|E + G|$ values are found out from the inverted/reduced matrices of variance and covariance of error (Table 4.3) and error + genotype (Table A.1 at the appendix), respectively by pivotal condensation method.

Now, the value of $|E|$ was obtained as the product of all bold face elements of (Table 4.3) at LHS.

$$\text{i.e. } |E| = 485943.88 \times 216133.99 \times 120671.23 \times 527901.15 \times 117215.78 \times 94972.99 \times 448090.79 \times 349053.78 = 1.164957525 \times 10^{43}$$

similarly the value of $|E + G|$ was obtained as the product of all bold face elements of (Table A.1 at the appendix) as:

$$|E + G| = 1720017.31 \times 756982.72 \times 56943.35 \times 3313793.78 \times 1677868.13 \times 254491.22 \times 1148128.13 \times 869815.84 = 1.047696694 \times 10^{47}$$

Therefore,

$$\Lambda = \frac{|E|}{|T|} = \frac{|E|}{|E+G|} = \frac{1.164957525 \times 10^{43}}{1.047696694 \times 10^{47}} = 0.0000111$$

Now,

$$V = -\ln \Lambda = -\log_e \Lambda \text{ (at pq d.f.)}$$

$$\text{Where, } m = t - \frac{(p+q+1)}{2} = 76 - \frac{(8+19+1)}{2} = 62$$

t = total degree of freedom (degree of freedoms for error plus genotypes) = $19 + 57 = 76$.

p = number of environments = 8

q = d.f. for genotypes = $20 - 1 = 19$

Hence, $V = -\ln \Lambda = -\log_e \Lambda = -62 \ln(0.0000111) = -62 \log_e (0.0000111) = 707.33$ (at $pq = 152$ d.f.)

$$V(\text{statistic}) = \chi^2 \text{ at } 152 \text{ d.f.} = 707.33$$

The probability of χ^2 at 152 d.f. falls within the range of 0.99 - 1.0, hence the difference is highly significant. Therefore, pooled differences (aggregate of all environments) among genotypes are significant justifying thereby the need to calculate D^2 values.

4.2 D^2 ANALYSIS

Genetic divergence in a population, especially with respect to the environments in which improvement is sought, is an indispensable prerequisite for successful crop improvement program. The D^2 statistic has found favor as a tool for estimating genetic divergence, which is the basis in choosing parents for hybridization in a breeding program. Progenies derived from diverse crosses are expected to show a broad spectrum of genetic variability providing greater scope for isolating high yielding segregants in the succeeding generations. In the present study, the D^2 was analyzed using 20 bread wheat genotypes on eight environments.

Test of significance: The analysis of variance for each individual character showed highly significant differences among the genotypes for all the 8 environments studied (Table 4.1). So that all the 20 genotypes of bread wheat were analyzed by the D^2 technique. The square of the distance (D^2 value) between any two genotypes calculated as sum of squares of difference between the mean values of all the 8(E1-E8) transformed variables were obtained for further analysis. Each genotype produced 19 combinations and totally, 190 D^2 values were obtained for all the genotypes considered. The D^2 values were tested for their significance using $\chi^2 = 15.51$ for 8 d.f. at 5% level of significance, hence they are all significant(i.e. significantly different from zero). Hence the analysis of genetic divergence among genotypes used in the study was considered relevant.

Formula for computation of D^2 values, which requires inversion of higher order determinant of the matrix, becomes complicated especially when the number of variables under consideration are large. Therefore, the original correlated unstandardised variables (E_i) was transformed to standardized uncorrelated variables (Y_j), so that the computation of D^2 values reduce to simple summation of squares of the difference between values of transformed variables of the two populations i , e D_i^2 . The result of the transformation done by using the pivotal condensation method were obtained (Table 4.3). This transformation requires the error matrices obtained by the ANOVA and ANCOVA procedures presented as in table 4.2.

Table 4.2 variance and covariance matrix at different levels*

Env't	Error Matrix(E)							
	E1	E2	E3	E4	E5	E6	E7	E8
E1	485943.88							
E2	22525.47	217170.16						
E3	-70121.97	21708.27	133636.19					
E4	-168458.48	-33551.44	10713.97	574896.92				
E5	-12187.06	-24500.54	-9500.28	42439.45	123295.46			
E6	29493.91	-3773.83	11644.20	207672.82	-30998.05	197121.67		
E7	16063.72	8320.90	14191.12	87695.39	12564.98	92177.83	502225.07	
E8	-76381.76	-29199.46	19590.14	14719.90	-4919.66	-14072.33	-953.78	366847.15

Genotype Matrix(G)								
Env't	E1	E2	E3	E4	E5	E6	E7	E8
E1	1234073.43							
E2	215021.00	573069.07						
E3	815542.47	220623.68	270052.86					
E4	181416.98	166879.83	550975.67	1083925.93				
E5	301397.68	210184.78	473289.84	134567.96	1045726.57			
E6	535141.00	202875.26	589453.86	465089.74	81467.89	404526.80		
E7	-79606.94	127781.17	24687.29	268376.08	42522.00	344483.53	1068196.78	
E8	261866.74	104914.08	281334.38	135816.88	209980.67	199901.44	-50875.42	694304.84

(E + G) Matrix								
Env't	E1	E2	E3	E4	E5	E6	E7	E8
E1	1720017.31							
E2	237546.47	790239.23						
E3	745420.5	242331.95	403689.05					
E4	12958.5	133328.39	561689.64	1658822.85				
E5	289210.62	185684.24	463789.56	177007.41	1169022.03			
E6	564634.91	199101.43	601098.06	672762.56	50469.84	601648.47		
E7	-63543.22	136102.07	38878.41	356071.47	55086.98	436661.36	1570421.85	
E8	185484.98	75714.62	300924.52	150536.78	205061.01	185829.11	-51829.2	1061151.99

*variances and covariances (diagonal and off-diagonal values, in order) were calculated as procedures of ANOVA and ANCOVA

Applying the pivotal condensation method of transformation to the error matrix by following its computational procedures the results are presented in table 4.3.

Table 4.3 Reduction of common dispersion (error) matrix by pivotal condensation method

LHS(Error Matrix)									RHS(Unit matrix)								Sum check
Env't	E1	E2	E3	E4	E5	E6	E7	E8	1	2	3	4	5	6	7	8	
E1	485943.88	22525.47	-70121.97	-168458.48	-12187.06	29493.91	16063.72	-76381.76	1	0	0	0	0	0	0	0	22687 8.71
E2	22525.47	217170.16	21708.27	-33551.44	-24500.54	-3773.83	8320.90	-29199.46	0	1	0	0	0	0	0	0	17870 0.53
E3	-70121.97	21708.27	133636.19	10713.97	-9500.28	11644.20	14191.12	19590.14	0	0	1	0	0	0	0	0	13186 2.64
E4	-168458.48	-33551.44	10713.97	574896.92	42439.45	207672.82	87695.39	14719.90	0	0	0	1	0	0	0	0	73612 9.53
E5	-12187.06	-24500.54	-9500.28	42439.45	123295.46	-30998.05	12564.98	-4919.66	0	0	0	0	1	0	0	0	96195. 3
E6	29493.91	-3773.83	11644.20	207672.82	-30998.05	197121.67	92177.83	-14072.33	0	0	0	0	0	1	0	0	48926 7.22
E7	16063.72	8320.90	14191.12	87695.39	12564.98	92177.83	502225.07	-953.78	0	0	0	0	0	0	1	0	73228 6.23
E8	-76381.76	-29199.46	19590.14	14719.90	-4919.66	-14072.33	-953.78	366847.15	0	0	0	0	0	0	0	1	27563 1.2
01	1	0.046	-0.144	-0.347	-0.025	0.061	0.033	-0.157	0.0000021	0	0	0	0	0	0	0	0.4670 021
02	0	216133.99	24951.94	-25735.10	-23936.86	-5147.88	7577.56	-25662.96	-0.05	1	0	0	0	0	0	0	16818 1.64
03	0	24933.88	123538.63	-13618.35	-7747.23	15921.64	16505.15	8580.99	0.15	0	1	0	0	0	0	0	16811 5.86
04	0	-25802.35	13544.05	516441.83	38227.99	217948.79	93254.52	-11728.08	0.35	0	0	1	0	0	0	0	84188 8.1
05	0	-23939.94	-11255.22	38210.54	122990.78	-30254.64	12967.15	-6833.03	0.03	0	0	0	1	0	0	0	10188 6.67
06	0	-5130.55	15891.32	217907.21	-30260.70	195322.54	91204.53	-9441.79	-0.06	0	0	0	0	1	0	0	47549 3.5
07	0	7581.97	16504.30	93269.50	12966.57	91197.94	501694.97	1568.22	-0.03	0	0	0	0	0	1	0	72478 4.44
08	0	-25685.90	8591.17	-11784.57	-6829.20	-9413.04	1566.82	354855.21	0.16	0	0	0	0	0	0	1	31130 1.65

09	1	0.115	-0.119	-0.111	-0.024	0.035	-0.119	-0.0000002	0.0000046	0	0	0	0	0	0	0	0.7770 044
10	0	120671.23	-10651.22	-4979.57	16520.05	15632.46	11548.12	0.15	-0.11	1	0	0	0	0	0	0	14874 2.11
11	0	165111.32	513371.35	35363.93	217329.53	94157.60	-14798.56	0.34	0.12	0	1	0	0	0	0	0	10105 36.63
12	0	-8502.13	35361.69	120333.45	-30829.20	13805.05	-9681.88	0.03	0.11	0	0	1	0	0	0	0	12048 8.12
13	0	16481.33	217296.67	-30830.19	195199.41	91384.10	-10052.33	-0.06	0.02	0	0	0	1	0	0	0	47947 9.95
14	0	15632.37	94171.75	13808.17	91379.91	501429.60	2470.47	-0.03	-0.03	0	0	0	0	1	0	0	71889 3.21
15	0	11545.05	-14841.19	-9680.33	-10029.50	2465.83	351798.59	0.15	0.12	0	0	0	0	0	0	1	33125 9.72
	16	1	-0.088	-0.041	0.137	.130	0.096	0.0000012	-0.0000009	0.0000083	0	0	0	0	0	0	1.2340 086
	17	0	527901.15	42133.49	194709.28	72693.13	-30649.25	0.14	0.27	-1.37	1	0	0	0	0	0	80678 7.84
	18	0	34613.50	119984.86	-29664.41	14910.33	-8865.68	0.04	0.10	0.07	0	1	0	0	0	0	13097 9.81
	19	0	218747.03	-30154.46	192941.47	89241.53	-11634.54	-0.08	0.03	-0.14	0	0	1	0	0	0	45914 1.84
	20	0	95547.40	14449.10	89238.28	499397.39	969.76	-0.05	-0.02	-0.13	0	0	0	0	1	0	69960 2.73
	21	0	-13825.23	-9206.98	-11611.17	964.97	350690.27	0.14	0.13	-0.10	0	0	0	0	0	1	31701 3.03
	22	1	0.080	0.369	0.138	0.138	-0.058	0.0000003	0.0000005	-0.0000026	0.0000019	0	0	0	0	0	1.5290 001
	23	0	117215.78	-42436.79	10133.67	10133.67	-6858.10	0.03	0.08	0.16	-0.07	1	0	0	0	0	78055. 76
	24	0	-47654.22	112223.82	59054.44	59054.44	1052.79	-0.15	-0.08	0.43	-0.42	0	1	0	0	0	12467 7.61
	25	0	6805.31	53981.29	486211.85	486211.85	6511.51	-0.08	-0.07	0.12	-0.18	0	0	0	1	0	55351 0.75
	26	0	-8100.96	-6509.66	2872.85	2872.85	349888.41	0.14	0.14	-0.14	0.03	0	0	0	0	1	33815 1.81

27	1	-0.362	0.086	-0.059	0.0000003	0.0000007	0.0000014	-0.0000006	0.0000085	0	0	0	0.6650103
	0	94972.99	63152.70	-1758.81	-0.14	-0.05	0.50	-0.45	0.41	1	0	0	156368.15
	0	56444.81	485626.59	6913.02	-0.08	-0.07	0.11	-0.18	-0.06	0	1	0	548985.14
	0	-9442.21	3569.53	349410.45	0.14	0.15	-0.13	0.03	0.07	0	0	1	343539.03
31		1	0.665	-0.019	-0.0000015	-0.0000005	0.0000053	-0.0000047	0.00000043	0.0000105	0	0	1.64600953
32		0	448090.79	7985.47	0.00	-0.04	-0.19	0.09	-0.30	-0.59	1	0	456076.23
33		0	9848.60	349231.05	0.13	0.15	-0.08	-0.01	0.11	0.10	0	1	359081.05
34			1	0.018	0	-0.0000001	-0.0000004	0.0000002	-0.0000007	-0.0000013	0.0000022	0	1.017999
35			0	349053.78	0.13	0.15	-0.08	-0.01	0.12	0.11	-0.02	1	349055.18
36				1	0.0000004	-0.0000004	0.0000002	0.0000000	0.0000003	0.0000003	-0.0000001	0.0000029	1.0000036

From the result of the table 4.3:

- i. The rows numbers 01, 09, 16, 22, 27, 31, 34, and 36 are the pivotal rows at each stage of the reduction of the matrix.
- ii. The last column (sum check) is the check of all calculations in the row concerned such that, row total = sum check.
- iii. Elements of the bold face on LHS are determinants of $|E|$.

From the inverted error (common dispersion) matrix (Table 4.3), we can locate the coefficients of transformation (elements of the bold face on LHS). From these coefficients, we develop the following factors:

$$A = \sqrt{485943.88} = 697.0968 \quad E = \sqrt{117215.78} = 342.3679$$

$$B = \sqrt{216133.99} = 464.9021 \quad F = \sqrt{94972.99} = 308.1769$$

$$C = \sqrt{120671.23} = 347.3776 \quad G = \sqrt{448090.79} = 669.3958$$

$$D = \sqrt{527901.15} = 726.5681 \quad H = \sqrt{349053.78} = 590.8077$$

Then, the Y (Y1, ..., Y8) uncorrelated linear functions (equations) are constructed using the above factors (A,B,C,D,E,F,G and H) , the 8 variables (environments E1, ..., E8) and the bold numbered faces on the RHS of the inverted or common dispersion matrix.

$$Y1 = \frac{1}{697.0968}E1 = 0.0014E1$$

$$Y2 = \frac{-0.05}{464.9021}E1 + \frac{1}{464.9021}E2 = -0.0001E1 + 0.0022E2$$

$$Y3 = \frac{0.15}{347.3776}E1 + \frac{-0.11}{347.3776}E2 + \frac{1}{347.3776}E3 = 0.0004E1 - 0.0003E2 + 0.0029E3$$

$$Y4 = \frac{0.14}{726.5681}E1 + \frac{0.27}{726.5681}E2 - \frac{1.37}{726.5681}E3 + \frac{1}{726.5681}E4 = 0.0002E1 + 0.0004E2 - 0.0019E3 + 0.0014E4$$

$$Y5 = \frac{0.03}{342.3679}E1 + \frac{0.08}{342.3679}E2 + \frac{0.16}{342.3679}E3 - \frac{0.07}{342.3679}E4 + \frac{1}{342.3679}E5 = 0.0001E1 + 0.0002E2 + 0.0005E3 - 0.0002E4 + 0.0029E5$$

$$Y6 = \frac{-0.14}{308.1769}E1 - \frac{0.05}{308.1769}E2 + \frac{0.50}{308.1769}E3 - \frac{0.45}{308.1769}E4 + \frac{0.41}{308.1769}E5 + \frac{1}{308.1769}E6$$

$$= -0.0005E1 - 0.0002E2 + 0.0016E3 - 0.0015E4 + 0.0013E5 + 0.0032E6$$

$$Y7 = \frac{0.00}{669.3958}E1 - \frac{0.04}{669.3958}E2 - \frac{0.19}{669.3958}E3 + \frac{0.09}{669.3958}E4 - \frac{0.30}{669.3958}E5 - \frac{0.59}{669.3958}E6 + \frac{1}{669.3958}E7$$

$$= -0.0001E2 - 0.0003E3 + 0.0001E4 - 0.0004E5 - 0.0009E6 + 0.0015E7$$

$$Y8 = \frac{0.13}{590.8077}E1 + \frac{0.15}{590.8077}E2 - \frac{0.08}{590.8077}E3 - \frac{0.01}{590.8077}E4 + \frac{0.12}{590.8077}E5 + \frac{0.11}{590.8077}E6 - \frac{0.02}{590.8077}E7 + \frac{1}{590.8077}E8$$

$$= 0.0002E1 + 0.0003E2 - 0.0001E3 + 0.0002E5 + 0.0002E6 + 0.0017E8$$

By substituting the corresponding mean values of E1 - E8 in the above Y- equations, we can get the uncorrelated Y_s matching to individual E_s for each genotype .

4.2.1 COMPUTATION OF D² VALUES

The D² between any two genotypes is the sum of squares of difference (d_i^2) of their corresponding Y values:

$$D^2 = \sum_i^y (Y_i^1 - Y_i^2)^2 = \sum_i^y d_i^2 = d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 + d_7^2 + d_8^2$$

where, i = 1 to y, and the superscripts 1 and 2 over Y are the genotype 1 and 2, respectively. Thus, the values of D² between all $\binom{20}{2} = \frac{20 \times 19}{2} = 190$ pairs, 1-2, 1-3, ..., 1-20; 2-3 ... (n-1)-n, are calculated and presented in matrix form as in Table 4.4. The average D² value for each genotype, $\bar{D}^2$ (last row) is calculated by averaging all D² for ith genotype occurring in n-1 = 19 pairs.

The closest and distant relatives: T²_{ij}) of bread wheat genotypes, considering eight quantitative environments ranging from 13.85 - 83.35 indicates the degree of divergence among 20 bread wheat genotypes to be high were shown on Table 4.4. Arranging the D² values in ascending/descending order, enabled to detect the closest and distant relatives. The highest genetic distances (the D² value of the most distantly related combination) (83.35) were observed between genotype G20 and G1 followed by combining genotype G20 with genotype G3 with D² value of 74.95 (Table 4.4). The highest D² distance

values were observed in the combination of line G20 with G1 and of G20 with G3. These combinations deserve special attention in bread wheat breeding programs, since they are the most suitable genotypes for hybrid combinations, due to the high genetic diversity and higher probability of finding favorable gene combinations for the selection of superior genotypes. The wide range of D^2 and the high estimated values for pairs of lines show the great genetic variability in this group of genotypes, which makes the identification of parents for the formation of a population with a broad genetic base possible, thereby increasing the probability of obtaining superior genotypes in segregating generations. Machado et al. (2002), in order to obtain the best combinations, choose parents which have the greatest genetic divergence. However, not only the genetic divergence might be used to choose parents for crossing (breeding program), but also their *per se* performance. As reported by Abreu *et al.* (1999) the best hybrids to be tested in a breeding program might involve both divergent and highly productive parents. The least D^2 value (13.85) were obtained by combining genotype G11 with genotype G12, indicating high similarity of these two genotypes for the environments studied (Table 4.4).

As per the average ($\bar{D}^2$) values, genotypes 20, 16, 1 and 3 (with $\bar{D}^2 = 50.66, 31.26, 30.83,$ and 30.28 , respectively) are, on an average, maximum divergent from all genotypes; while genotypes 6, 7, 17, and 4 (with $\bar{D}^2 = 22.12, 22.29, 22.66,$ and 23.14 , in order) are the least divergent.

Table 4.4 Divergence as measured by D²* statistic among 20 genotypes of bead wheat

Gen.	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G16	G17	G18	G19	G20
G1	0.00																			
G2	28.09	0.00																		
G3	19.71	32.87	0.00																	
G4	20.42	19.71	20.57	0.00																
G5	33.38	27.42	25.60	21.40	0.00															
G6	30.47	25.38	27.08	17.20	15.87	0.00														
G7	25.33	23.14	22.00	18.62	14.66	17.05	0.00													
G8	27.70	39.57	26.69	25.65	22.50	21.17	20.63	0.00												
G9	26.47	20.86	22.46	15.95	21.52	18.94	21.13	32.08	0.00											
G10	32.16	29.43	31.23	21.86	17.36	20.78	17.28	22.99	25.78	0.00										
G11	42.59	39.45	38.08	27.15	20.37	17.64	23.91	24.62	20.41	16.51	0.00									
G12	44.76	39.12	41.34	27.89	21.59	17.70	25.76	25.39	30.06	18.20	13.85	0.00								
G13	23.43	16.47	39.58	19.20	26.50	23.86	22.01	33.43	18.73	26.54	36.46	35.11	0.00							
G14	32.12	31.42	29.20	22.20	15.73	14.99	16.85	16.90	25.10	16.44	17.15	17.62	27.78	0.00						
G15	18.06	22.11	25.15	18.95	26.29	23.56	21.57	28.01	20.75	24.79	33.93	34.67	16.58	25.23	0.00					
G16	20.54	31.67	21.67	26.27	39.59	32.43	25.06	34.19	25.88	30.31	40.59	43.88	24.55	32.63	18.75	0.00				
G17	21.58	23.91	20.23	18.94	26.70	18.64	14.64	20.05	21.28	19.54	25.37	27.44	21.55	17.67	18.53	20.26	0.00			
G18	37.09	33.86	34.35	23.96	18.27	15.99	20.94	23.83	24.03	16.08	16.12	16.06	27.85	15.84	25.55	33.00	22.12	0.00		
G19	18.52	29.87	22.54	23.72	28.43	27.95	24.02	26.19	24.75	31.62	39.87	41.24	22.78	27.16	16.03	18.46	18.68	30.60	0.00	
G20	83.35	54.23	74.95	49.94	38.40	33.55	48.94	59.79	43.39	39.20	31.95	28.95	53.54	38.23	58.20	74.15	53.42	29.43	69.01	0.00
$\bar{D}^2$	30.83	29.93	30.28	23.14	24.29	22.12	22.29	27.97	24.19	24.11	27.69	28.98	27.16	23.17	25.09	31.26	22.66	24.47	28.50	50.66

*Excepting the bold faces, all are significant(P < 0.05).

Contribution of different environments towards divergence: The analysis of the contribution of each environment towards the expression of genetic divergence (Table 4.5) indicated that E1 contributed maximum (18.58%) followed by E7 (16.61%) were the potential contributor to the total genetic divergence in the genotypes of bread wheat studied. These two environments followed by E8 (14.10%), E5 (13.61%) , E2 (12.41%) and E6(9.23%) totally accounted for more than 84% of total genetic divergence. E4 (6.94%) and E3 (8.52%) were the least contributor to the divergence.

Table 4.5. Contribution of each environment to the total divergence

Environment	E1	E2	E3	E4	E5	E6	E7	E8
Percent contribution	18.58	12.41	8.52	6.94	13.61	9.23	16.61	14.10

E1(environment 1) = Adet 2007

E5(environment 5) = Adet 2008

E2(environment 2) = Holeta 2007

E6(environment 6) = Holeta 2008

E3(environment 3) = Kulumsa 2007

E7(environment 7) = Kulumsa 2008

E4 (environment 4) = Sinana 2007

E8(environment 8) = Sinana 2008

4.3 CLUSTER ANALYSIS

T', (Rao, 1952), x
(D^2), utilizes the D^2 value. When D^2 values were arranged in ascending order (data not shown) for each genotype, the D^2 value of close relatives varied from 13.85 to 83.35. Hence the maximum level of cluster formation was fixed as 83.35. Starting with two nearest entries, new entries were added in the cluster until the average D^2 values of added genotype remain below the maximum level fixed for cluster formation otherwise, eliminated and used in other cluster formation. The above criterion holds true for the remaining cluster constellation too. Based on the D^2 values between the 20 genotypes, all the studied genotypes could be separated into 6 different clusters (Table 4.6). There is a wide distribution of genotypes in different groups, indicating a wide diversity among the genotypes. Among these, cluster I was

the largest and consists of 7 genotypes, followed by cluster III and IV with 4 genotypes each. Clusters II, VI and VII contained 3, 1 and 1 genotypes, respectively.

Table 4.6. Cluster number with their respective genotypes

Clusters	No of Genotypes	Genotype name	Clusters	No of Genotypes	Genotype name
I	7	G5	II	3	G7
		G6			G8
		G10			G17
		G11	IV	4	G1
		G12			
		G14			
		G18			
III	4	G2			
		G4			
		G9			
		G13			
V	1	G3	VI	1	G20

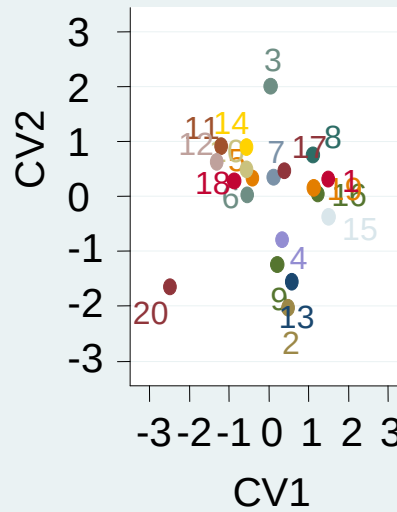


Fig 4.1. Scatter plot of the six cluster groupings in relation to the two representative axes of the first two canonical variates (CV1 and CV2) in 8 environments evaluated.

According to Figure 4.1, the results demonstrated a scatter plot related to the first two canonical variate, in a bi-dimensional space. The first two canonical variate explained more of the variation among accession variance. Each canonical variate is the linear combination of the independently measured variables that has the greatest possible variance. This graphic representation also

f T , f to clusters.

4.3.1 Intra-and inter-cluster distances

The average intra-and inter-cluster D^2 values and their square root ($\sqrt{D^2}$ value) are depicted in (Table 4.7). The magnitude of the intra - cluster distances indicates the extent of genetic diversity

among genotypes of the same cluster. The intra-cluster distances ranged from 0 (in mono genotypic cluster) to 18.49 in cluster III. Cluster I had the smallest intra-cluster distance relative to clusters II and VI (Table 4.7). The minimum inter-cluster distance were observed between the clusters I and II (21.49) and VI and V (22.27) (Table 4.7) ,indicating that the intercross of the genotypes that make up these respective pair groups may not produce superior genotypes in segregating generations. However, Nick et al. (2008) emphasized the importance of knowledge of similarity between pairs of plants or groups; this information can be used in programs involving backcrosses, where the use of similar parentals, which differ basically in terms of the allele one wants to transfer, facilitates the recovery of the recurrent parental. Both these clusters together accounted for 15 genotypes of all the genotypes of diverse origin. The maximum distance (74.95) was recorded between clusters V and VI followed by clusters IV and VI (71.18) (Table 4.7), which together covered only 7 genotypes of all the materials, indicating the importance of group VI as a source of parentals. Rahman et al. (2002) claimed that the identification of superior genotypes based on genetic divergence is the most appropriate strategy to start a breeding program. It is important to emphasize that crosses between highly divergent genotypes with a good yield potential are most effective. All other inter-cluster D^2 values were lying between these two values. Cluster VI exhibited a distant relationship to most of the clusters.

4.3.2 The nearest and distant clusters

As indicated in Table 4.8, cluster I showed a close relationship with cluster II followed by cluster III . Cluster II was nearest to cluster I, but farthest to cluster VI. Cluster II and IV exhibited an intimate relationship with cluster V. A very close proximity was observed between clusters IV and V followed by clusters I, II and III. A distant relationship was observed between cluster VI and Clusters II,IV, and V.

Table 4.7. Intra- (main diagonal) and inter-cluster D^2 values along with their $\sqrt{D^2} = D$ value in parenthesis

Clusters	I	II	III	IV	V	VI
I	17.15 (4.14)	21.49 (4.64)	27.06 (5.20)	33.07 (5.75)	32.41 (5.69)	34.24 (5.85)
II		18.44 (4.29)	25.11 (5.01)	24.26 (4.93)	22.97 (4.79)	54.05 (7.35)
III			18.49 (4.30)	24.14 (4.91)	28.87 (5.37)	50.28 (7.09)
IV				18.39 (4.29)	22.27 (4.72)	71.18 (8.44)
V					0.00 (0.00)	74.95 (8.66)
VI						0.00 (0.00)

Table 4.8. The nearest and distant clusters from each cluster based on D value

Clusters	No of genotypes grouped	Nearest clusters	Farthest clusters
I	7	II, III	IV, VI
II	3	I, III	IV, VI
III	4	I, II, V	IV, VI
IV	4	III, V	I, VI
V	1	III, IV	VI
VI	1	IV, V	From all clusters

The dendrogram (Figure 4.2) was a result of intra- and intergroup cluster analysis of the 20 bread wheat genotypes, based on the method of Tocher optimization. The generalized distance (D^2),

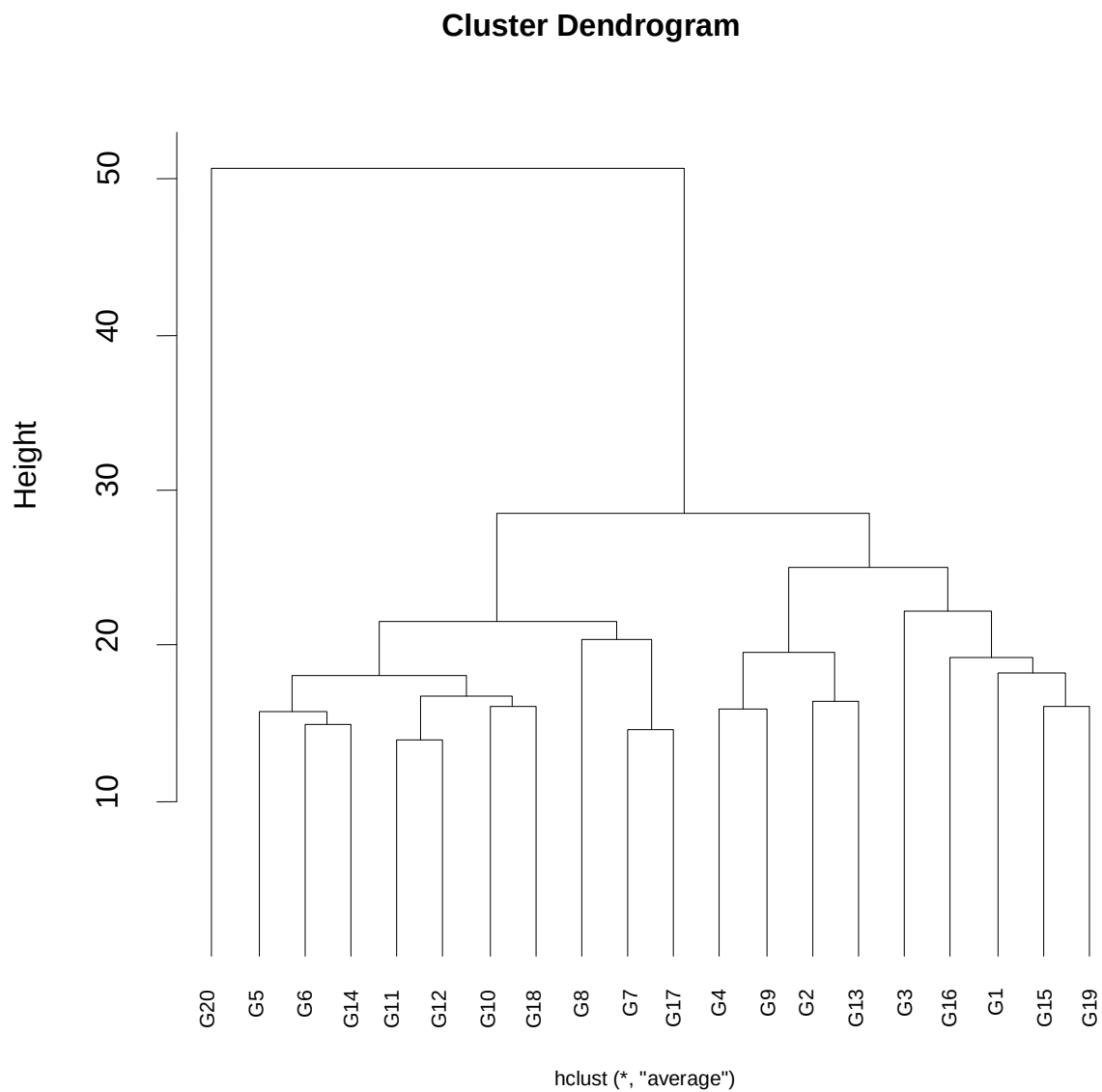


Fig. 4.2. Dendrogram depicting the hierarchical clustering of 20 genotypes of bread wheat tested in 8 environments ,o

(D²) as

dissimilarity measure

Hierarchical clustering produces a tree (dendrogram) which can be cut at any level. In fact, there are one less of joints than there were observations. Calling them classification may be somewhat misleading, because they do not give one classification, but several alternative classifications, with extreme cases of all observations in one cluster or each observation in its private cluster. Therefore the user commonly wants to cut the classification at a certain level, and pretend that this was the classification produced by the method. The result below shows the classification of genotypes to six desired number of groups based on the method of Tocher optimization . Fig 4.3 displays the grouping by drawing boxes in the tree (dendrogram) separating clusters at six number groups (clusters).

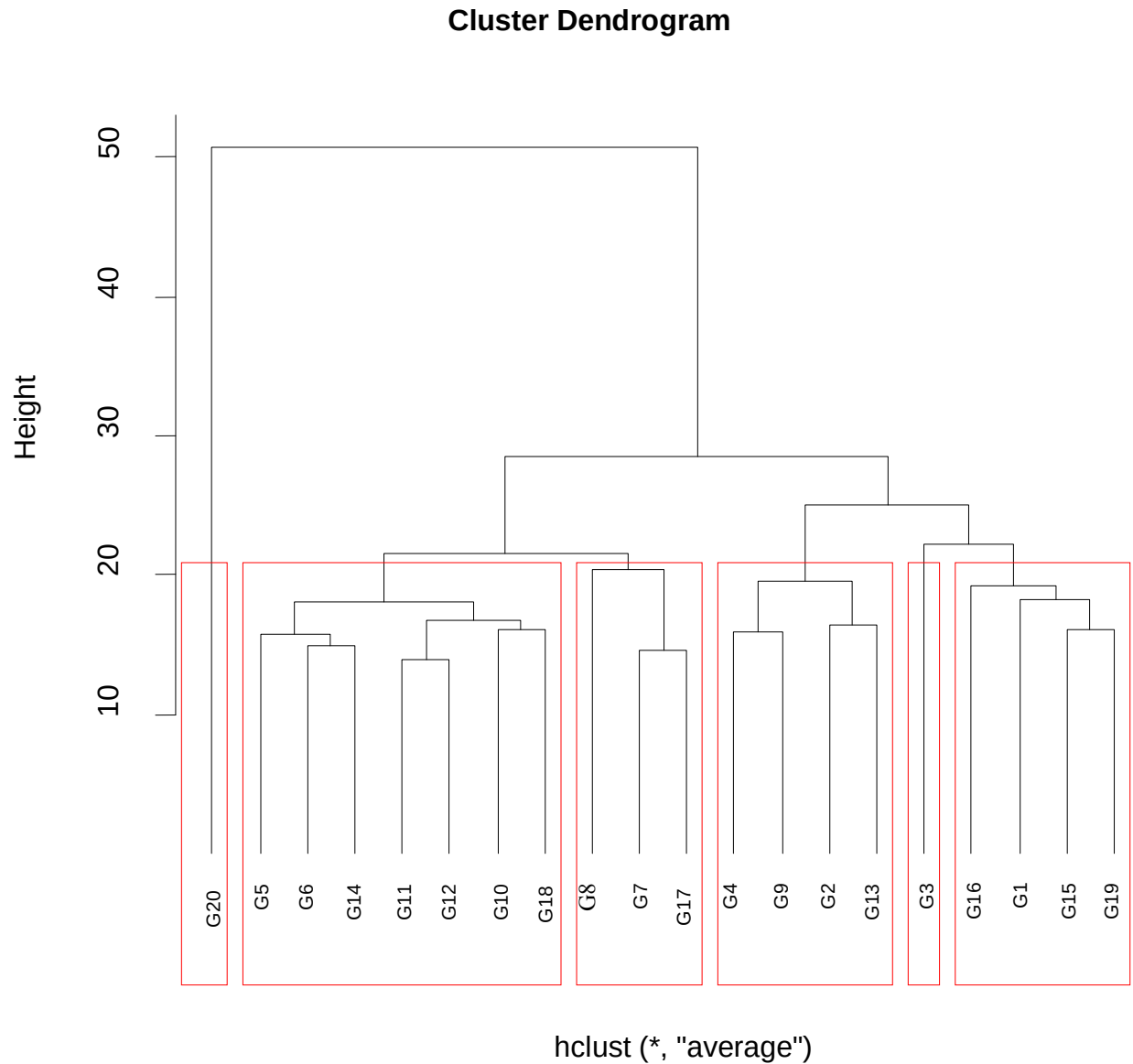


Fig 4.3 Dendrogram depicting grouping by drawing boxes into six of the hierarchical clustering of 20 genotypes of bread wheat tested in 8 environments ,o

(D²) as dissimilarity measure

The analysis of the dendrogram that represents the genetic divergence between lines obtained using the UPGMA method (Figure 4.3) shows six different clusters. Taking into consideration the changes in the degree of dissimilarity to produce the clusters, the first one comprises lines G5,G6,G10,G11,G12 and G14, the second comprises lines G7,G8 and G17, the third comprises

lines G2,G4,G9 and G13, the fourth comprises G1,G15,G16 and G19, the fifth comprises only G3 and the sixth comprises only G20. The above cluster dendrogram shows that genotype G20 and G3 were confirmed as the most divergent of all, considering that it is in a group separated clearly from other groups. The 20 genotypes were separated into 6 groups as:

G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G16	G17	G18	G19	G20
4	3	5	3	1	1	2	2	3	1	1	1	3	1	4	4	2	4	4	6

, when a cut was made at an average distance in the hierarchical grouping of genotypes based on the measured response variables. This cutoff value allows the visualization of 6 groups formed by the Tocher method and separation of the genotype of the genotypes in different groups and to join in a same group, confirming these genotypes as the most similar among the evaluated.

The produced dendrogram can largely simplify the original information and some distortions on the dissimilarity patterns between the groups may result. Thus, the last step in the joining process is to evaluate if the results are adequate (Cruz & Crneiro, 2003).

A criterion used to evaluate whether the results are adequate is the cophenetic correlation coefficient (CCC), which should preferably be higher than 0.70 (Rohlf, 1970). In the present study , CCC was 0.77, indicating that the clustering method was adequate to resolve the information from the dataset. The estimated cophenetic distance between two points is the level at which they are fused in the dendrogram, or the height of the root. A good clustering method correctly reproduces the actual dissimilarities. In unweighted pair-group method of average linkage clustering, the cophenetic distance is the average of observed distances (Fig.4.4).

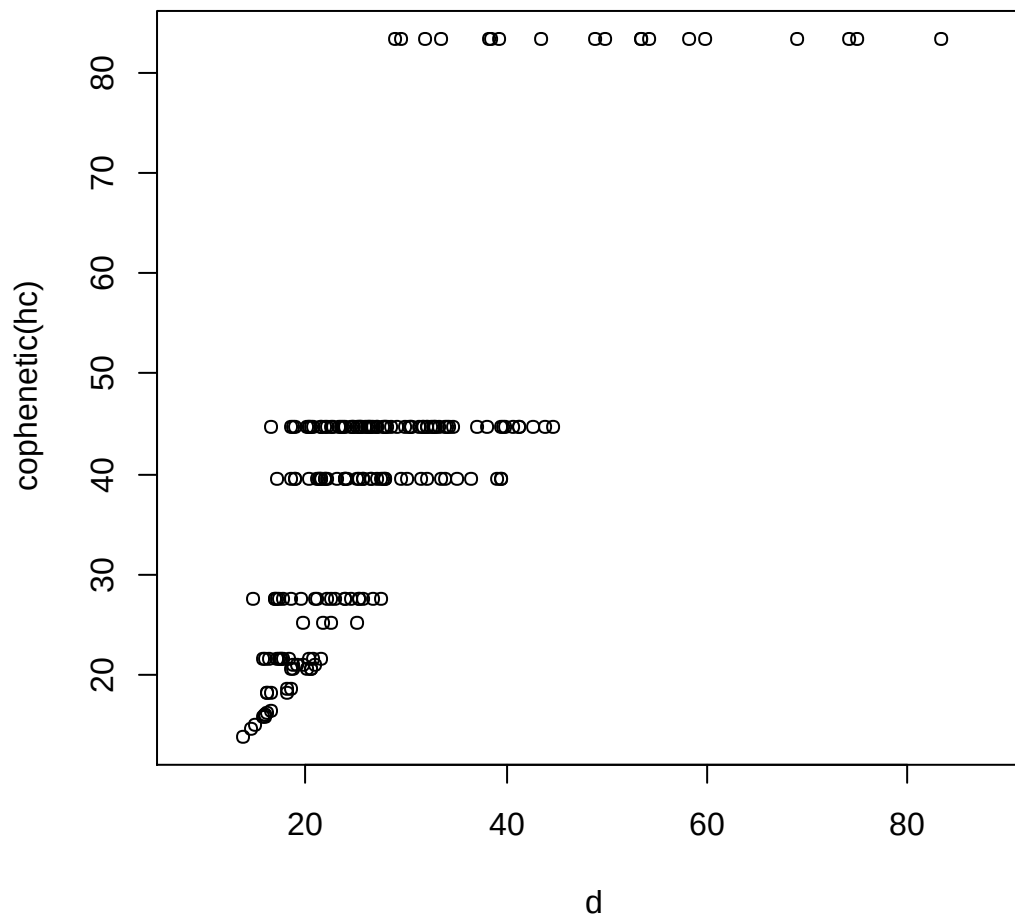


Fig. 4.4 Cophenetic distance defined by UPGMA

The correlation between observed and cophenetic distances (the cophenetic correlation:)

[1] 0.7710455

To validate clusters, that is, to verify the ability of the dendrogram to reproduce the dissimilarity matrix, the cophenetic correlation coefficient (CCC) was calculated. The CCC is the Pearson correlation coefficient between the distance matrix (D^2) and the cophenetic matrix (C).

The cophenetic correlation coefficient (CCC) between the generalized Mahalanobis distance matrix (D^2) and the cophenetic distance matrix (C), given by the dendrogram for UPGMA, were of high magnitude and significant, equal to 0.77, demonstrating consistency in clustering. Cruz

and Carneiro (2003) report that cophenetic correlation is a good criterion to evaluate the consistency of a graphical representation, with values close to one indicating better performance.

4.3.3 Diagrammatic representation of cluster divergence

Using $\sqrt{D^2} = D$ values (Table 4.7), the groups - distances are pictorially presented as in fig.4.5:

The diagram is an arbitrary depiction, since many distances cannot be adjusted in two dimensional representation as per scale.

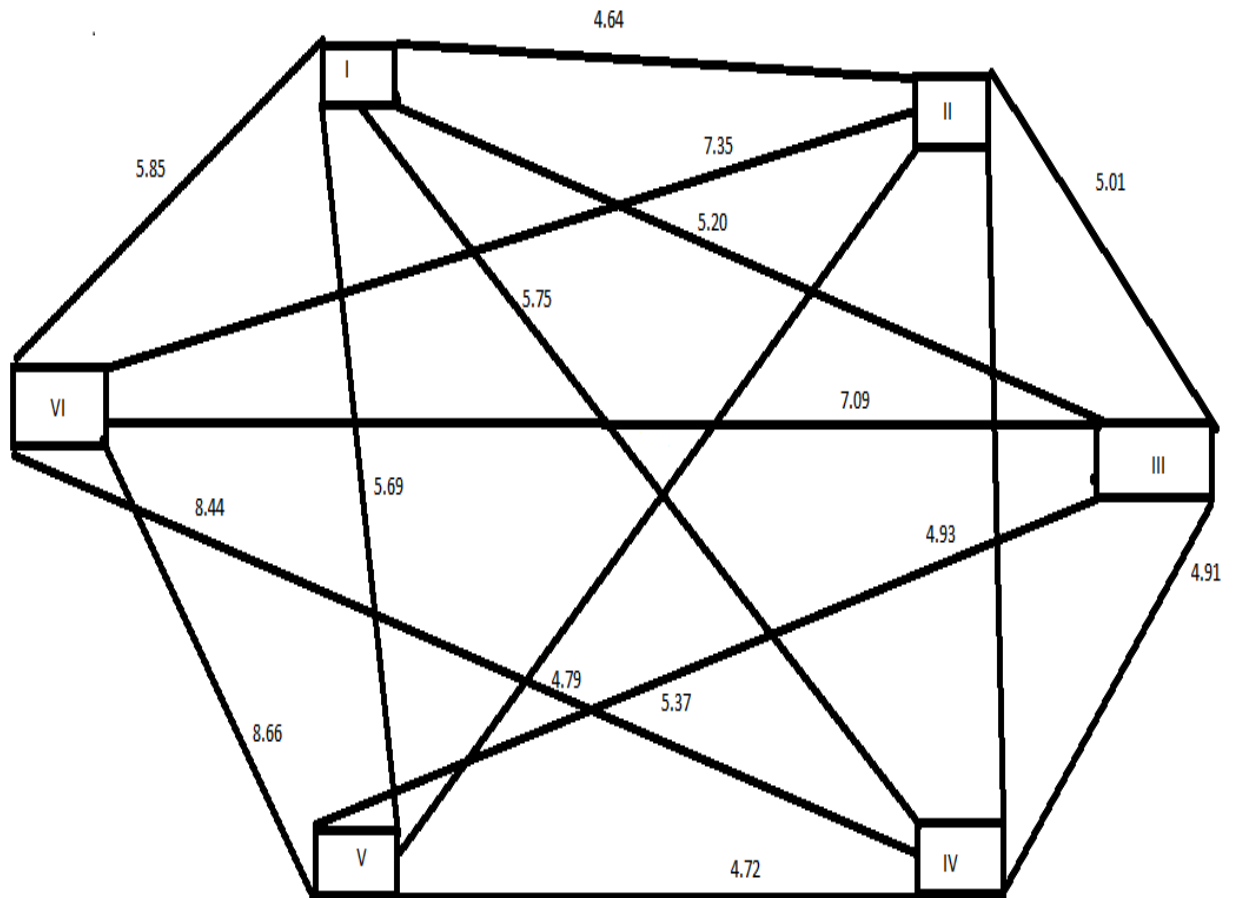


Fig. 4.5 Generalized distance - inter-cluster divergence ($\sqrt{D^2}$) among six clusters of 20 bread wheat genotypes

4.3.4 Cluster Mean Analysis

As indicated in Table 4.9, the genotypes in clusters II, III, IV, and V were the most productive (i.e. produce more yield) in E4 and E5, whereas those in clusters I and VI were of the lowest. Cluster V means for E1 was the highest, while those in cluster I showed the lowest in E2 and E6, and the genotypes were the lowest productive. The highest mean yield E4 and E5 were recorded by genotypes in cluster V. The highest mean cluster value for E1 and E7 was recorded by cluster IV, while the lowest by cluster VI. Genotypes in cluster VI exhibited the lowest in producing yield across all environments, whereas the highest in V. With respect to E8, genotypes grouped in clusters IV, and V recorded large mean values, while those grouped in cluster II recorded the lowest mean value. Average yield value of E2 varied from 1867.18 for cluster I to 2648.76 for cluster IV. The mean of E3 ranged from 2271.48 to 2819.85 for genotypes of clusters V and IV, respectively.

Table 4.9. Cluster Means for 8 quantitative environments in 20 bread wheat genotypes

Cluster	E1	E2	E3	E4	E5	E6	E7	E8
I	3902.82	1867.18	2302.50	4841.49	5097.02	1870.00	2955.11	2250.68
II	4451.52	2045.91	2330.95	5510.96	5648.38	2117.68	3134.91	2206.45
III	3970.03	2253.48	2689.49	5122.37	5541.70	2437.56	3776.33	2358.32
IV	4675.39	2648.76	2819.85	5322.13	5825.80	2004.04	3118.83	2843.89
V	4089.01	2478.46	2271.48	5317.35	6155.72	1963.92	4018.89	3092.53
VI	2479.81	1905.68	2338.78	4404.01	3932.61	2340.13	2448.44	2229.01

E1(environment 1) = Adet 2007

E2(environment 2) = Holeta 2007

E3(environment 3) = Kulumsa 2007

E4 (environment 4) = Sinana 2007

E5(environment 5) = Adet 2008

E6(environment 6) = Holeta 2008

E7(environment 7) = Kulumsa 2008

E8(environment 8) = Sinana 2008

Cluster mean for E6 varied from 3932.61 for genotype in cluster VI to 6155.72 for genotypes in cluster V. Cluster V and VI, with a single genotype exhibited the lowest mean yield in E6 and E2, while they showed the highest mean cluster value for E5 and E4 respectively. The highest

yield was recorded by genotypes of cluster V in E5 , while the lowest in genotypes of cluster I in E2.

Choices of divergent parent(s) can be made for hybridization purposes on the basis of D^2 - values between two genotypes or two clusters. Members from most divergent clusters can be chosen taking into account their desirable or complementary environments.

A cluster analysis of the four test localities resulted in two classes (Fig. 4.6). The first cluster consisted of the two localities of Holeta and Kulumsa. The remaining two sites (Adet and Sinana) grouped in the second cluster. Similarly, the 8 test environments were grouped into four clusters (Fig. 4.7). Cluster I consisted of four environments (Kulumsa 2007, Sinana 2008, Holeta 2008 and Holeta 2007), while the second comprised two environments (Adet 2008 and Sinana 2007). This again indicates the close similarity of the Kulumsa, Sinana and Holeta environments.

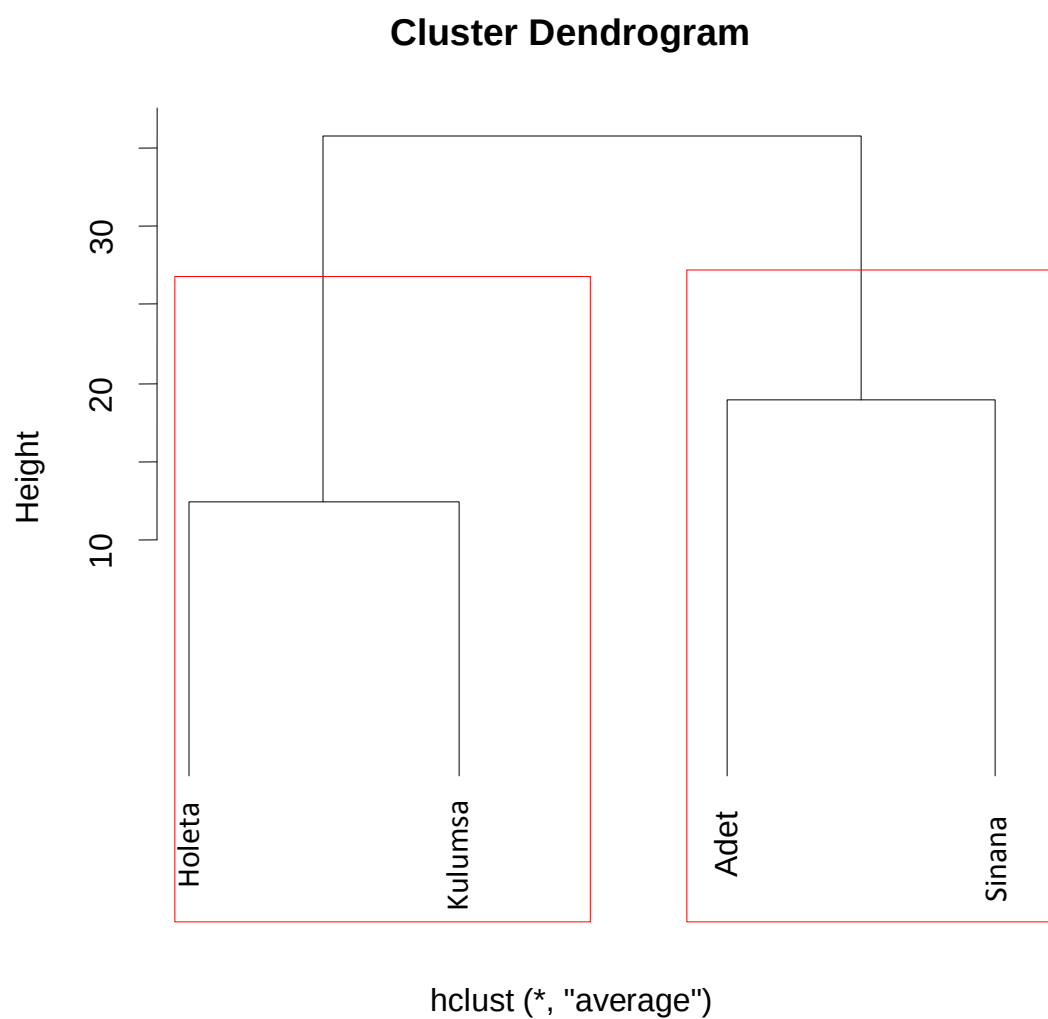


Fig. 4.6. Dendrogram depicting the hierarchical clustering of four localities used in bread wheat multi- environment trials ;Cophenetic correlation=0.89,o

(D²) as dissimilarity measure.

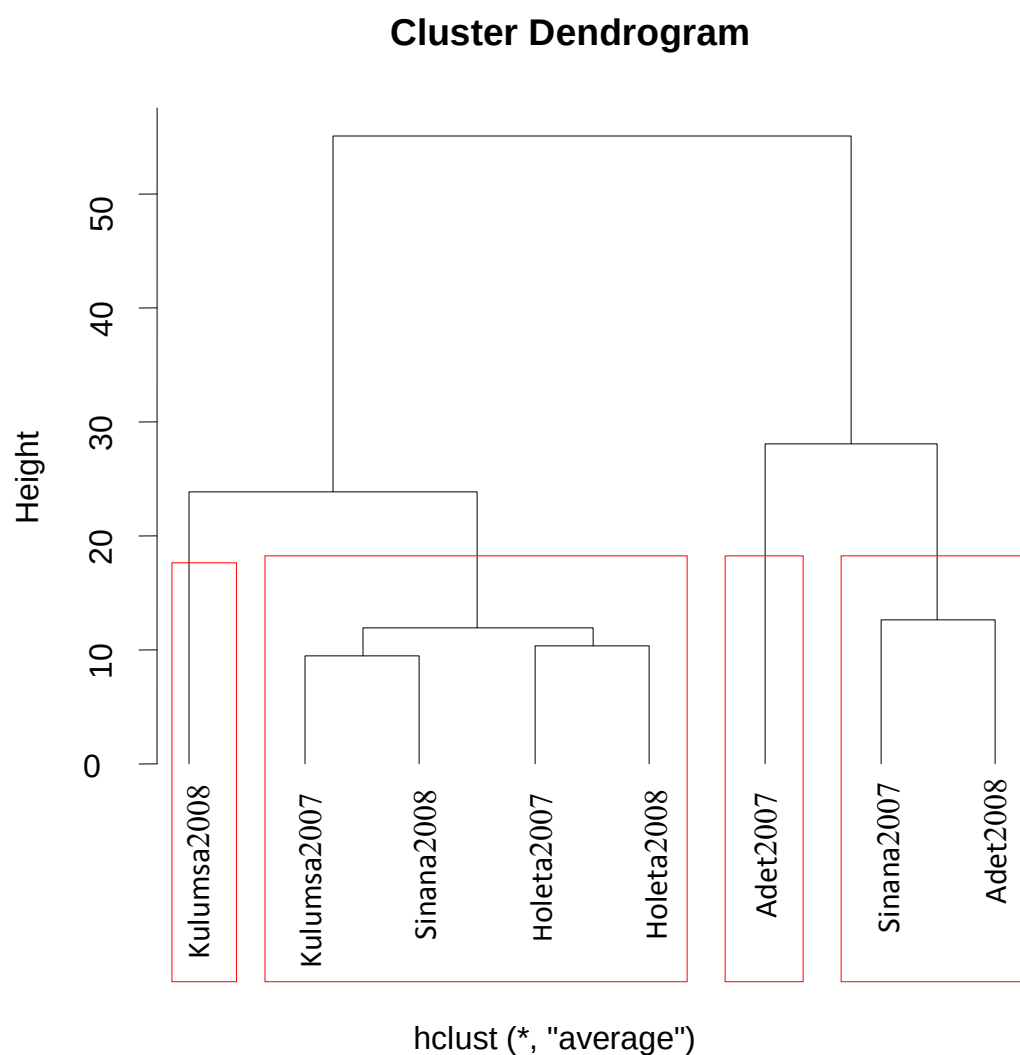


Fig. 4.7. Dendrogram depicting the hierarchical clustering of 8 environments (location $\times$ year) used in testing 20 genotypes of bread wheat at four locations, 2007–08. Cophenetic correlation=0.78, obtained by the Unweighted Pair-Group Method distance (D^2) as dissimilarity measure

The remaining two clusters were the individual environments of Kulumsa 2008 and Adet 2007. These two environments were too variable to be clustered with any of the other environments.

4.4 Mahalanobis distance for assessing multivariate normality and outlier

As one of the core assumptions of most parametric multivariate techniques is multivariate normality; which implies the absence of multivariate outliers the data used for this paper does not show the evidence of deviation from multivariate normality as shown on graph presented in the Figure 4.8. The plot based on Mahalanobis distance (Figure 4.9) , shows that the data value for all observation considered did not contained an outlying observations.

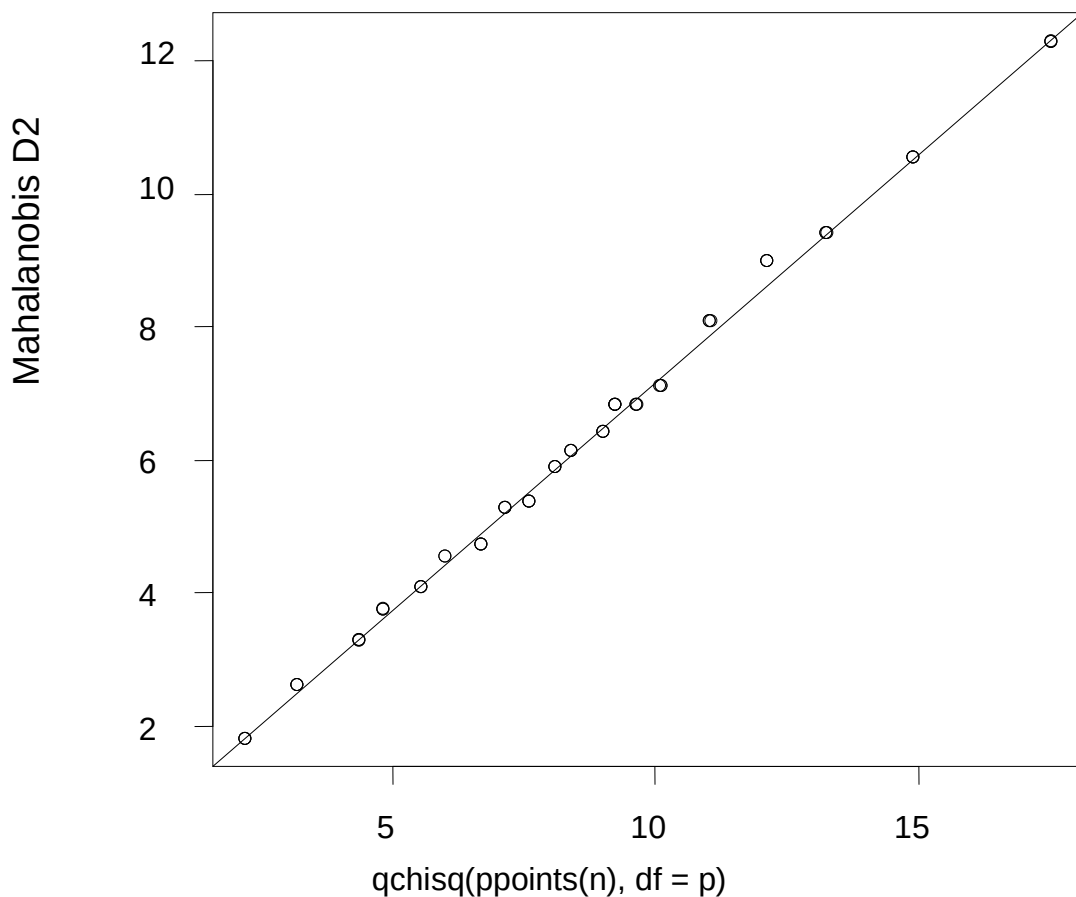


Figure 4.8. Mahalanobis distance plotted versus chi-square quantiles for checking multivariate normality.

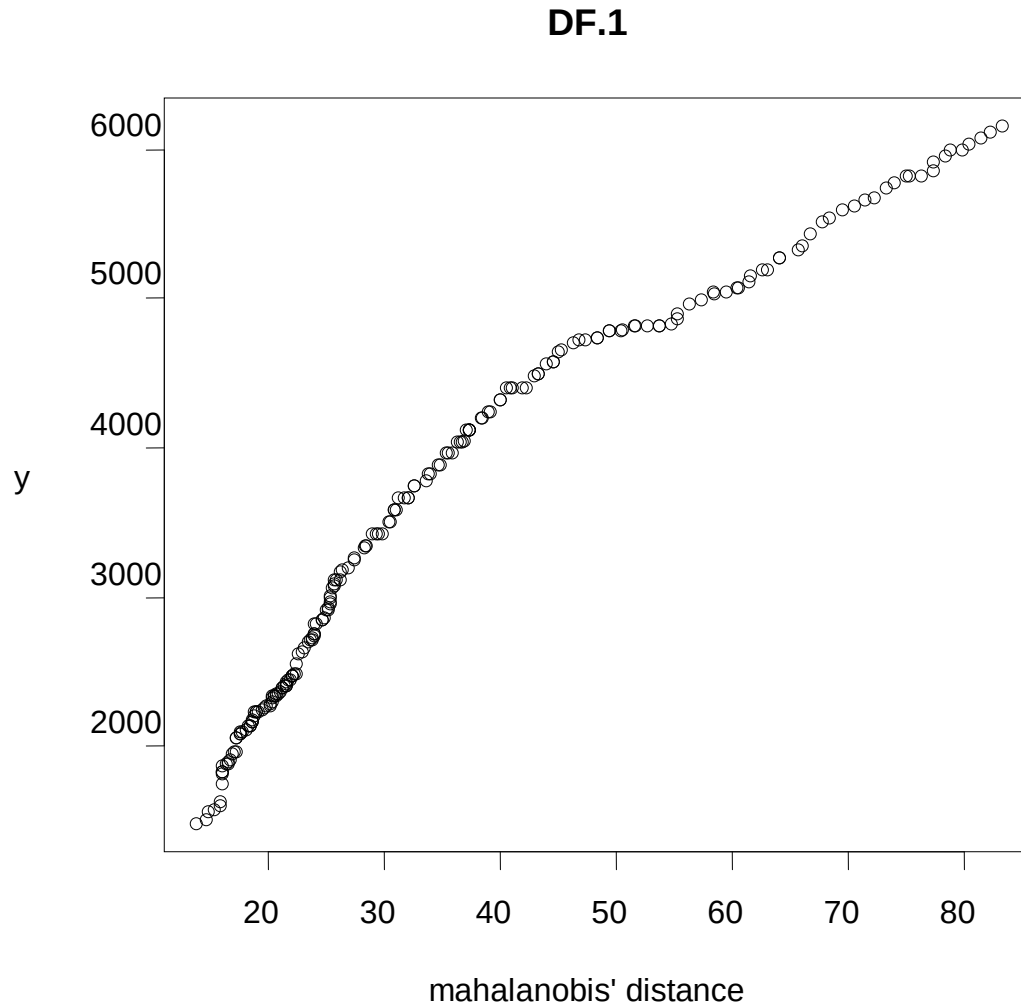


Figure 4.9. Plot of Mahalanobis distance for checking multivariate outlier.

In the plot above (Fig. 4.9); cases with seeming extreme Mahalanobis distances (upper right) are not likely to be true multivariate outliers.

4.5 Discussion

Crop breeders ascribe genetic improvement in any crop plant to the amount of variability present in the genotypes. This is judged based on the knowledge of the extent of variability and the genetic diversity available in the breeding material at hand. In this study, 20 genotypes were evaluated to study the genetic potential they possess with respect to 8 environments. The results

obtained on some genetic parameters and genetic divergence study through multivariate technique was discussed as follows.

The univariate analysis of variance and pooled differences for 20 genotypes indicated that the genotypes varied among themselves with respect to the 8 environments studied. The development of new varieties is mainly governed by the magnitude of genetic variability in the base material and extent of variability for the desired environments. This finding is consistent with other studies (Malhotra *et al.* (1973)) showing that univariate ANOVA revealed high significant differences among genotypes for all studied characters and also pooled differences for the characters were highly significant when tested by Wilk's criterion.

Environment-wise ranks have shown that no single environment had a lion share to total divergence. However, E1 contributed the largest share to the divergence, followed by E7 , E8 , E5 and E2 respectively. This could imply that the genotypes included in the study were divergent for most of the environments studied. It is clear that these environments are the basic attributes of plant architecture, which need greater attention, as they accounted for more of total divergence. Malhotra *et al.* (1973) and Singh *et al.* (1997) observed no relationship between geographic diversity and genetic distance.

Based on the D^2 value, all the 20-genotypes of bread wheat were grouped into 6 clusters. It is essential to note that out of 20 genotypes introduced 7 of them were grouped in the same cluster, cluster I. The trend was repeated in almost all the clusters except the monogenotypic clusters, V and VI. Earlier workers have discussed genetic drift; selection pressure and environment could cause greater diversity than geographic distance (Bhatt,1970; Murthy and Arunchalm, 1966). The results reported in this paper are also in agreement to these findings. The results presented by Tochers method demonstrated that the most divergent, G1 and G20 were placed in clusters IV and VI, respectively, and the most similar G11 and G12 in the same cluster I. The clustering process also formed a monogenotypic cluster. This may be associated with intensive natural and human selection for diverse adaptive gene complexes. The inter-cluster distances were greater than intra-cluster distances, revealing considerable amount of genetic diversity among the genotypes studied . Inter-cluster distance is the main criterion for selection of genotypes using D^2

analysis. Genotypes belonging to the clusters with maximum inter-cluster distances are genetically more divergent and hybridization between genotypes of divergent clusters is likely to produce wide variability with desirable segregant. This finding is consistent with other studies (Sarma and Roy, 1994). However, it is a general notion that the larger is the divergence between the genotypes, the higher will be the heterosis. In this context, it is important to consider the practical significance of grouping the genotypes into different clusters and estimating the genetic distance between them, which represents an index of genetic diversity among clusters (Bhatt, 1970). It may be useful to produce crosses between genotypes belonging to the clusters separated by large estimated distances (Bhatt, 1970). Success might therefore be expected through making crosses between the genotypes from cluster V and cluster VI, followed by the one between clusters IV and cluster VI. Genotypes from these clusters can be selected for hybridization program that can evolve highly heterotic crosses, which might prove potential in isolating superior segregants. Arunchalam (1981) and Ghaderi et al. (1984) indicated that too high a divergence might not produce the highest frequency of heterotic crosses. Parents chosen to be genetically divergent through D^2 analysis can fail at times, to show higher gene frequency difference. According to Falconer and Mckay (1996) heterosis is a direct function of the square of gene frequency difference between parent population and their dominance effect.

Cluster means revealed appreciable variation for various environments. A close look into the cluster means revealed that clusters differ in respect of different environments. Singh *et al.* (1997) reported the importance of cluster mean of a character within the different clusters and its significance for improvement, which is in agreement with the present result. Intercrossing between genotypes of these diverse clusters would generate a broad spectrum of variability for effective selection in the segregating generations for the development of high yielding cultivars.

Here, it is important to note that in calculating cluster mean, the superiority of a particular genotype with respect to a given environment can be diluted by other genotypes that are related and grouped in the same cluster which are inferior or intermediary for that environment in question. Hence, apart selection of a line from the cluster, which has high inter-cluster distance for hybridization, one can also think of selecting parents based on the extent of divergence with respect to an environment of interest.

The four test localities were clustered into two groups, while 8 environments (sites $\times$ years) were categorized into four classes. Some of the environments, such as Kulumsa 2008 and Adet 2007 were very different from the others. Brown et al. (1983) discussed how to improve the efficiency of variety testing programmes through grouping similar test environments and by identifying optimum environments within the clusters. Moreover, Ghaderi et al. (1980) used cluster analysis to classify locations into similar groups, homogeneous clusters in which the genotype $\times$ locations interactions were non-significant.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

In this study twenty genotypes of bread wheat were evaluated at four locations Adet, Holeta, Kulumsa, and Sinana under rain fed locations on 2007 and 2008 in a univariate analysis of variance using a randomized complete block design with four replications to elucidate information on the nature and magnitude of genetic variability and the degree of genetic divergence. Significant genetic variability among tested genotypes was observed that indicates the presence of excellent opportunities to bring about improvement through wide hybridization by crossing genotypes with high genetic distance. The D^2 distance matrix reflected highest diversity between genotypes G1 and G20 while the minimum between genotypes G11 and G12. The most influential environments associated with the variation on the genotypes in this study are E1, E7, E8 and E5. Additionally, the use of cluster analysis is successful in differentiating the genotypes into similar groups on the basis of environment. The characterization of the variation of these environments among the genotypes that will permit plant breeders to focus on particular environment that are linked to genotype variation and to discern environments that are not close indicators of genotype variability.

Cluster analysis and the canonical variate analysis have analogous efficacy to determine the most suitable combinations for a breeding program. Both multivariate methods showed similar results i.e. resulted in the formation six genetically similar clusters of genotypes. The inter-cluster distances were greater than intra-cluster distances, revealing considerable amount of genetic diversity among the genotypes studied. Inter-cluster distance is the main criterion for selection of genotypes using D^2 analysis. The future breeding program utilizing the studied accession should be based on the genetic analysis of various parts and hybridization carried out between clusters rather than within clusters. The most suitable crosses might involve between the genotypes from cluster V and cluster VI, followed by the one between clusters IV and cluster VI.

RECOMMENDATION

From this study the analysis based on multivariate methods using D^2 statistic procedure is recommended to identify divergent genotypes according to their adaptation environment. The categorization of genotypes according to their response to the different environments has an agronomic advantage for breeders and farmers. With D^2 statistic it is also possible to target the different production environments by developing genotypes for specific adaptation.

REFERENCE

- A.A. Jewel, A. K. Chowdhury, A. K. M. M. Alam, M. A. Latif, and M. M. Hassan 2010. Multivariate Analysis In Lentil (*Lens culinaris* L.). Bangladesh J. Pl. Breed. Genet., 23(1): 09 - 12.
- Abebe, D. & Giorgis, H.M. 1991. Wheat genetic resources in Ethiopia. In: Plant genetic resources of Ethiopia. Engels, J.M.M., Hawkes, J.G.
- Abou-El-Fittouh, H. A., Rawlings, J.O. & Miller, P.A. (1969). Classification of environments to control genotype by environment interactions with an application to cotton. Crop Science 9, 135–140.
- Abreu, A. De F. B.; Ramalho, M. A. P.; Ferreira, D. F. Selection potential for seed yield from intra end inter-racial populations in common bean. Euphytica, v.108, n. 2, p. 121-127. 1999.
- Afifi, A.A. & Clakr, V. (1996). Computer–Aided Multivariate Analysis. London: Chapman and Hall.
- Albrecht, G. H. 1980. Multivariate analysis and the study of form with special reference to canonical variate analysis. Amer. Zool. 20:679-693.
- Arega G, Hussein M, Harjit S (2007). Genetic divergence in selected durum wheat genotypes of Ethiopian plasm. Afr. Crop Sci. J., 15(2): 67 - 72.
- Ariyo OJ (1987a). Multivariate analysis and the choice of parents for hybridization in okra (*Abelmoschus esculentus* (L.) Moench). Theor. Appl. Genet. 74: 361–363.
- Arunachalam, V. 1981. Genetic distance in plant breeding. *Indian Jour. Genetics* 41: 226-236.
- Atchley, W. R., 1980 M-statistics and morphometric divergence. Science 208: 1059-1060.
- Ayed Sourour and Slim-Amara Hajer (2009). Distribution And Phenotypic Variability Aspects of Some Quantitative traits Among Durum Wheat Accessions. African Crop Science Journal, Vol. 16, No. 4, pp. 219 - 224.
- Baril, C.P., J.B. Denis, and P.Brabant (1994) "Selection of environments using simultaneous clustering based on genotype $\times$ environment interaction." Canadian Journal of Plant Science 74, 311 - 317.
- Barnett, V., & Lewis, T. (1994). Outliers in statistical data (3rd ed.). New York: Wiley.

- Bechere, E., Belay, G., Mitiku, D. & Merker, A. 1996. Phenotypic diversity of tetraploid wheat landraces from the northern and north-central regions of Ethiopia. *Hereditas* 124, 165-172.
- Belay, G. & Merker, A. 1999. C-band polymorphism and chromosomal Rearrangements in tetraploid wheat (*Triticum turgidum* L.) landraces from Ethiopia. *Wheat Infor Serv* 88, 6-14.
- Belay, G., Tesemma, T., Becker H.C. & Merker, A. 1993. Variation and interrelationships of agronomic traits in Ethiopian tetraploid wheat landraces. *Euphytica* 71, 181-188.
- Bhatt, G.M. 1970. Multivariate analysis approach for selection of parents for hybridization aiming at yield improvement in self-pollinated crops. *Aust. Jour. Agric. Res.* 21:1-7.
- Blackith, R. E. and R. A. Reyment. 1971. *Multivariate Morphometrics*. Academic Press, New York.
- Briggle, L.W. & Curtis, B.C. 1987. Wheat worldwide. In: *Wheat and wheat improvement*, Heyne EG(ed), ASA, CSSA, SSSA, Madison, Wisconsin, USA, pp 1-32.
- Chandrasekhariah, S.R., B.R. Murty, and V. Arunachalam. 1969. Multivariate analysis of genetic divergence in Eu – Sorghum. *Proc. Nat. Inst. Sci. (India)* 35: 172–195.
- Cooper, M., and I.H. Delacy (1994) " Relationships among analytical methods used to study genotypic variation and genotype-by-environment interaction in plant breeding multi-environment experiments." *Theoretical and Applied Genetics* 88, 561 - 572.
- Crossa, J., 1990. Statistical analysis of multilocation trials. *Advances in Agronomy* 44:55-85.
- Cruz CD, Carneiro PCS. *Modelos Biometricos aplicados ao melhoramento genetico*. Volume 2.1 ed. vicoso(Mg) : UFV, Imprensa universitaria;2004.
- CSA. 2010. Federal Democratic Republic of Ethiopia Central Statistics Agency Crop Production Forecast Sample Survey, 2010/11 (2003 E.C.) In: Report on Area and Crop Production forecast for Major Grain Crops (For Private Peasant Holding, Meher Season) Statistical Bulletin.
- Curtis, B. 2002. Wheat in the world. In: *Bread wheat: Improvement and production*. Plant production and protection series, Curtis, B.C., Rajaram, S., Gomez & Macpherson H. (eds). No. 30, pp. 1–17.
- Das PK and Das Gupta T (1984). Multivariate analysis in black grain (*Vigna mungo* (L.) Hepper. *Indian J. Genet.* 44(2): 243-247.

- Denton, F. M. and R. del Moral. 1976. Comparison of multivariate analysis using taxonomic data of *Oxalis*. *Can. J. Bot.* 54:1637-1646.
- Dixon, J., Nally, L., Aquino, P., Kosina, P., La Rovere, R. & Hellin, J. 2006. Adoption and economic impact of improved wheat varieties in developing countries. *J Agric Sci (Cambridge)* 144, 489–502.
- Eticha, F., Bekele, E., Belay, G. & Börner, A. 2005. Phenotypic diversity in Tetraploid wheats collected from Bale and Wello regions of Ethiopia. *Plant Gen Res: Charac Util* 3, 35-43.
- Faris H., Arnulf M., Harjit S., Getachew B. and Johansson, E., 2006, Multivariate Analysis of Diversity of Tetraploid Wheat Germplasm from Ethiopia. *Gen. Resour. and Crop Evol.*, 53(6): 1089-1098.
- Farris, J. S. 1969. On the cophenetic correlation coefficient. *Systematic Zoology*, 16, 174–175.
- Franco, J., J. Crossa, J. Villaseñor, S. Taba, and S.A Eberhart. 1997. Classifying Mexican maize accessions using hierarchical and density search methods. *Crop Sci.* 37:972–980.
- Gebre-Mariam, H. 1991. Wheat production & research in Ethiopia. In: Wheat Research in Ethiopia: A Historical Perspective. Gebre-Mariam, H., Tanner, D.G. & Huluka, M. (eds), IAR/CIMMYT, pp. 1–15. Addis Ababa, Ethiopia.
- Ghaderi, A, Adams, M.W. and Nassib, A.M. 1984. Relationship between genetic distance and heterosis for yield and morphological traits in dry edible bean and faba bean. *Crop Sci.* 24:37-42.
- Ghaderi, A., M. Shishegar, A. Rezai and B. Ehdaie. 1979. Multivariate analysis of genetic diversity for yield and its components in mung bean. *J. Amer. Soc. Hort. Sci.* 104(6):728-731.
- GMR. 2011. Grain Market Report of International Grain Council Number 415 Sept 22/2011 (<http://www.igc.int/downloads/gmrsummary/gmrsumme.pdf>).
- Goodman, M. M. 1967. The races of maize: I. The use of Mahalanobis' generalized distances to measure morphological similarity. *Fitotecnica Latinoamericana* 4(1):1-22.
- Goodman, M. M. 1968. The races of maize: II. Use of multivariate analysis of variance to measure morphological similarity. *Crop Science* 8;693-698.

- Goodman, M. M. 1973. Genetic distances: Measuring dissimilarity among populations. *Yearbook of Physician Anthropology* 17:1-38.
- Goodman, M. M. and R. M. Bird. 1977. The races of maize. IV. Tentative grouping of 219 Latin American races. *Econ. Botany* 31:204-221.
- Gorfu, A., Girma, B., & Girma, K. 2001. Present status and future direction of wheat research and development in Ethiopia. In: *Wheat and Weeds: Food and Feed Proceedings of two stakeholder workshops*. Wall, P. C. (ed.), Addis Ababa, Ethiopia: Ethiopian Agricultural Research Organization, pp. 167–206.
- Green, P. E. 1978. *Analyzing Multivariate Data*. Dryden Press, Hinsdale, Illinois.
- Hair, J.R., Anderson, R.E., Tatham, R.L. and Black, W.C. 1995. *Multivariate Data Analysis with Readings*. 4th Ed., Prentice-Hall, Englewood Cliffs, NJ.
- Harlan, J.R. 1975. Our vanishing genetic resources. *Science* 188, 618–621.
- Ivory, D.A., S. Kaewmeechai, I.H. DeLacy, and K.E. Basford (1991) "Analysis of the environmental component of genotype $\times$ environment interaction in crop adaptation evaluation." *Field Crops Research* 28, 71- 84.
- Jagadev, P. N and K. M. Samal. 1991. Multivariate anlysis in niger (*Guizotic abyssinica* cass.). *Indian J. Genet.* 51(4): 392-394.
- Jain, S.K., Qualset, C.O., Bhatt, G.M. & Wu, K.K. 1975. Geographic pattern of phenotypic diversity in a world collection of durum wheat. *Crop Sci* 15, 700-704.
- Jaylal, Mahto. 1994. Genetic divergence in soybean for physiological and yield attributes under rainfed condition. *Indian Jour. Genetics* 54: 418-424.
- Jensen, N. F. (1988). *Plant Breeding Methodology*. New York: Wiley-Interscience.
- Johnson, R.A. (2007). *Applied Multivariate Statistical Analysis*. 6th edition, USA, Pearson Education, Inc.
- Johnson, R.A., & Wichern, D.W. (Eds.). (2007). *Applied multivariate statistical analysis* (2nd ed.). Upper Saddle River, New Jersey: Pearson Prentice Hall.
- Kundu B.C., M.M. Hossain, M.A. Khaleque Mian And I.H. Mian 2012. Genetic Divergence in Bitter Gourd (*Momordica Charntia* L.). *J. Asiat. Soc. Bangladesh, Sci.* 38(2): 125-134.
- Lefkovitch, L. P. (1980). Conditional clustering. *Biometrics* 36, 43–58.
- Lin, C. S., Binns, M.R. & Lefkovitch, L. P. (1986). Stability analysis: where do we stand? *Crop Science* 26, 894–900.

- Lin, C.S., and M.R. Binns (1985) " Procedural approach for assessing cultivar- location data: Pairwise genotype - environment interactions of test cultivars with checks." *Canadian Journal of Plant Science* 65, 1065 - 1071.
- Machado, C. F.; Nunes, G. H. S.; Ferreira, D. F. *et al.* Divergência genética entre genótipos de feijoeiro a partir de técnicas multivariadas. *Ciência Rural*, v. 32, n. 2, p. 251-258. 2002.
- Mahalanobis, P.C. 1930. On test and measure of group divergence. *Jour. Asiatic Soc. Bengal* 26: 541-588.
- Mahalanobis, P.C. 1936. On the generalized distance in statistics. *Proc. Nat. Acad. Sci. (India)* 2: 49-55.
- Malhotra, V.V, Singh, S. and Singh, K. B. 1973. Relationship between geographic diversity and genetic divergence and the relative role of each character towards maximizing divergence in green gram. *Indian Jour. Agric. Sci.* **44**: 811-815.
- Mardia, K. V., J. T. Kent and J. M. Bibby. 1979. *Multivariate Analysis*. Academic Press, London.
- Messele, T. 2001. Multidisciplinary approach in estimating genetic diversity of Ethiopian tetraploid wheat (*Triticum turgidum* L.) landraces. PhD dissertation, Wageningen University, Wageningen, 106pp.
- Missio, R. F.; Moraes, M. L. T.; Dias, L. A. S. Efeito do desbaste seletivo sobre a divergência genética em progênies de *Pinus caribaea* Morelet var. *bahamensis*. *Scientia Forestalis*, v. 73, n. 1, p. 27-36, 2007.
- Mohammadi, S. A. and Prasanna, B. M. 2003. Analysis of genetic diversity in crop plants - Salient Statistical Tools and Considerations. *Crop Sci.* 43:1235–1248.
- Mungomery, V.E., R.Shorter, and D.E. Byth (1974) " Genotype × Environment interactions and environmental adaptation. I Pattern analysis - Application to soya bean populations." *Australian Journal of Agricultural Research* 25, 59 - 72.
- Murthy, B.R. and Arunachalam, V. 1966. The nature of divergence in relation to breeding system in some crop plants. *Indian Jour. Genetics* 26: 188-198.
- Murthy, B.R. and M.I. Quadri. 1966, Analysis of divergence in some self compatible forms of *Brassica campestris* L. Var. Brown Sarson. *Indian J. Genet.* 20: 45-48.
- Murty, B.R., V. Arunachalam, and M.B.L. Saxena. 1967. Catalogue of world collection of *Sorghum* and *Pennisetum*. *Indian Journal of Genetics and Plant Breeding* 27(Suppl.): 1–394.

- Nair, K.R. and Mukherjee, H.K., 1960, Classification of natural and plantation teak (*Tectona grandis*) grown at different localities of India and Burma with respect to its mechanical and physiological properties. *Sankhya*, 22: 1-20.
- Nair, N.V., R. Balakrishnan and T.V. Sreenivasan, 1998. Variability for quantitative traits in exotic hybridgermplasm of sugarcane. *Gen. Res. Crop Evo.* 45(5): 459-463.
- Negassa, M. 1986. Estimates of phenotypic diversity and breeding potential of Ethiopian wheat. *Hereditas* 104, 41–48.
- Ortiz, R., Braun, H.J., Crossa, J., Crouch, J.H., Davenport, G., Dixon, J., Dreisigacker, S., Duveiller, E., He, Z., Huerta, J., Joshi, A.K., Kishii, M., Kosina, P., Manes, Y., Mezzalama, M., Morgounov, A., Murakami, J., Nicol, J., Ferrara, G.O. Ortiz-Monasterio, J.I., Payne, T.S., Peña, R.J., Reynolds, M.P., Sayre, K.D., Sharma, R.C., Singh, R.P., Wang, J., Warburton, M., Wu, H. & Iwanaga, H.M. 2008. Wheat genetic resources enhancement by the International Maize and Wheat Improvement Center (CIMMYT). *Gen Res Crop Evol* 55, 1095–1140.
- Pecetti, L. & Damania, A.B. 1996. Geographic variation in tetraploid wheat (*Triticum turgidum* ssp. *turgidum* convar. *durum*) landraces from two provinces in Ethiopia. *Gen Res Crop Evol* 43, 395- 407.
- Perrino, P. & Porceddu E. 1991. Wheat genetic resources in Ethiopia and the Mediterranean. In : *Wheat Genetic Resources: Meeting diverse needs*. Srivastava, J.P. (ed) . John Wiley & Sons. pp . 161-186.
- Piassi M, Silva MA, Regazzi AJ, Torres RA, Soares PR, Torres Junior AA. Estudo da divergencia genetica entre oito grupos de aves de postura, por meio de techicas de analise multivariada. *Revista da Sociedade Brasileira de Zootecnia* 1995; 24(5): 715-727.
- Piassi MA. Avaliacao do desempenho de linhagens de postura mantidas na Universidade Federal de vicoso, em competicao com marcas comerciais. [Dissrtacao]. Vicoso (MG): Universidade Federal de vicoso;1994.
- Porceddu, E., Ceoloni, C., Lafiandro, D., Tanzerella, O.A. & Mugnozza, G. T.S. 1988. Genetic resources and plant breeding: problems and prospects. In: *Proceedings Symposium of Seventh International Wheat Genetics*. Miller, T.E. & Koebner R.B.D. (eds). Cambridge, pp. 7-21.

- Porceddu, E., Perrino, P. & Olita, G. 1973. Preliminary information on an Ethiopian wheat germplasm collection mission. In: Proceedings Symposium Genetics & Breeding of Durum Wheat. Mugnozza, G.T. S. (ed), Bari, Italy, pp. 181–200.
- Rabindra Babu, V., Kisore Sadhya, N. Shobha Rani, and Ravichanderan. 2006. Genetic divergence analysis using quality traits in rice genotypes. *Oryza* 43: 260–263.
- Ramagosa, I. and Fox, P.N., 1993. Genotype x Environment interaction and adaptation. In: Hayward M.D., Bosemark N.O. and Ramagosa I. (eds.). Plant breeding: Principles and prospects. Chapman and Hall, London. pp. 373-390.
- Ramanujam, S., A.S. Tiwari, and R.B. Mehra. 1974. Genetic divergence and hybrid performance in Mungbean. *Theor. Appl. Genet* 45: 211–214.
- Rao, C.R. 1952. *Advanced Statistical in Biometric Research*, John Wiley and Sons, Inc., New York.
- Rao, C.R. 1960. Multivariate analysis: An indispensable tool in statistical and in applied research. *Sankhya* 22: 317-338.
- Rao, C.R. 1973a. *Linear Statistical Inference and its Applications*, Second Edition, Wiley, New York.
- Rhodes, A. M., S. E. Malo, C. W. Campbell and S. G. Carmer. 1971. A numerical taxonomic study of the avocado (*Persea americana* Mill.). *Amer. Hor. Sci.* 96(3):391-395.
- Rohlf FJ. Adaptive hierarchical clustering schemes. *Systematic Zoology* 1970; 19(1): 58-82
- Romesburg, H. C. (1984). *Cluster Analysis for Researchers*. Belmont, CA: Lifetime Learning Publications.
- Rousseeuw, P.J. (1990). Unmasking multivariate outliers and leverage points, *J. Amer. Stats. ass.*, 85(411), 633-651.
- Royo, C., Elias, E.M. & Manthey, F.A. 2009. Durum wheat breeding. In: Cereals. Carena, M.J. (ed.), pp. 199-226. DOI: 10.1007/978/-0-387- 72297-9 http://www.springerlink.com/content/tuwq50k67742357_q/fulltext.pdf
- Sakaguti ES, Silva MA, Regazzi AJ, Torres RA, Martins LN. Analise de divergencia geneica entre nove grupos geneticos de coelhos. *Revista da Sociedade Brasileira de Zootecnia* 1996; 25(3):647-660.
- Sarma, R.N and Roy, A.1994. Genetic divergence in early maturing pigeonpea. *Indian Jour. Genet.* 54:84-87.

- Searle, S. R. 1971. Linear Models. John Wiley & Sons, Inc., N. Y.
- Shamsuddin, A. K. M., 1985, Genetic diversity in relation to heterosis and combining ability in spring wheat. *Theor. Appl. Genet.*, 17(3):306-308.
- Sharma, S. (1996). *Applied Multivariate Techniques*. John Wiley & Sons, Inc.
- Singh, G., Singh, S.P. and Chawdhary, B.S. 1997. Genetic divergence analysis in rice bean *Indian Jour. Genetics* 58:101-105.
- Skovmand, B., Warburton, M.L., Sullivan, S.N. & Lage, J. 2005. Managing and collecting genetic resources. In: Durum Wheat Breeding: Current Approaches and Future Strategies. Royo, C., Nachit, M., Di Fonzo, N., Araus, J.L., Pfeiffer, W.H. & Slafer G.A. (eds.). Food Products Press, New York, pp. 143- 163.
- Sohoo, M. S. and K. S. Gill. 1978. Multivariate analysis in cluster bean (*Cyanopsis tetragonoZoba* (L.) Taub.). *Genet. Agr.* 32:327333
- Sokal RA, Rohlf FJ. The comparison dendrogram by objective methods. *Taxonomy*; 1962:1-33-40.
- Sokal, R. R. and F. J. Rohlf. 1970. The intelligent ignoramus, an experiment in numerical taxonomy. *Taxon*, 19, 305–319.
- Somayajulu, P. L. M., A. B. Joshi and B. R. Murty. 1970. Genetic divergence in wheat. *Indian J. of Genetics and Plant Breed.* 30:47-58.
- Spagnoletti-Zeuli, P.L. & Qualset, C.O.1987. Geographical diversity of quantitative spike characters in a world collection of durum wheat. *Crop Sci* 27, 235-241.
- Tabachnick, B.G. & Fidell, L. S. (1989). *Using Multivariate Statistics*, Second Edition. Cambridge: Harper and Row.
- Tai, G. C. C. and H. De Jong. 1980. Multivariate analysis of potato hybrids I. Discrimination between tetraploid-diploid hybrid families and their relationship to cultivars. *Can. J. Genet. Cytol.* 22:227-235.
- Tk Y. & H K. 2006. F , f landraces in Ethiopia. *Gen Res Crop Evol* 43, 395-407
- Tesemma, T. & Belay, G. 1991. Aspects of Ethiopian tetraploid wheat with emphasis on durum wheat genetics and breeding. In: Wheat research in Ethiopia. Tanner, D.G., Gebre-Mariam, H. & Huluka, M. (eds), Addis Ababa, Ethiopia, pp 95-103.

- Tesemma, T., Belay, G. & Worede, M. 1991. Morphological diversity in tetraploid wheat landrace populations from the central highlands of Ethiopia. *Hereditas* 114, 171-176.
- Tsegaye, B. & Berg, T. 2007a. Utilization of durum wheat landraces in East Shewa, central Ethiopia: Are home uses an incentive for on-farm conservation? *Agriculture and Human Values*. 24, 219- 230.
- Tsegaye, S., Becker, C.H., & Tessema, T.1994. Isozyme variation in Ethiopian tetraploid wheat (*Triticum turgidum*) landrace agrotypes of different seed color groups. *Euphytica* 75, 143-147.
- USDA. 2011. United States Department of Agriculture: Foreign Agricultural Service. World Agricultural Production Circular Series WAP 09-11 September 2011. <http://www.fas.usda.gov/wap/circular/2011/11-09/productionfull09-11.pdf>.
- Van Lill, D., Purchase, J. L., Smith, M. F., Agenbag, G. A. & DE Villiers, O. T. (1995). Multivariate assessment of environmental effects on hard red winter wheat. II. Canonical correlation and canonical variate analysis of yield, biochemical and bread-making characteristics. *South African Journal of Plant Soil* 12, 164–169.
- Whitehouse, R. N. H. 1971. Canonical analysis as an aid in plant breeding. In *Barley Genetics II. Proc. Second Int. Barley Genet. Symp.* Washington State Univ. Press, Pullman, Washington.
- Wicklin, R. (2012). *What is Mahalanobis' distance?* The DO loop (SAS programming blog). Retrieved on 2013-07-11 from: <http://blogs.sas.com/content/iml/2012/02/15/what-is-mahalanobis-distance/>
- Wikipedia. (2013). *Mahanobis' distance*. Retrieved on 2013-07-11 from: http://en.wikipedia.org/wiki/Mahalanobis_distance
- Williams, W. T. 1971. Principles of clustering. *Annual Review of Ecology and Systematics* 2:303-326.

APPENDIXES

Table A.1 Reduction of common dispersion (error + genotype) matrix by pivotal condensation method

Env't	E1	E2	E3	E4	E5	E6	E7	E8
E1	1720017.31	237546.47	745420.5	12958.5	289210.62	564634.91	-63543.22	185484.98
E2	237546.47	790239.23	242331.95	133328.39	185684.24	199101.43	136102.07	75714.62
E3	745420.5	242331.95	403689.05	561689.64	463789.56	601098.06	38878.41	300924.52
E4	12958.5	133328.39	561689.64	1658822.85	177007.41	672762.56	356071.47	150536.78
E5	289210.62	185684.24	463789.56	177007.41	1169022.03	50469.84	55086.98	205061.01
E6	564634.91	199101.43	601098.06	672762.56	50469.84	601648.47	436661.36	185829.11
E7	-63543.22	136102.07	38878.41	356071.47	55086.98	436661.36	1570421.85	-51829.2
E8	185484.98	75714.62	300924.52	150536.78	205061.01	185829.11	-51829.2	1061151.99
01	1	0.14	0.43	0.01	0.17	0.33	-0.04	0.11
02	0	756982.72	140186.97	130952.93	145301.34	120711.09	145603.93	49584.51
03	0	137973.08	83158.24	554235.44	337068.08	355109.30	68695.23	218928.27
04	0	131514.20	556117.49	1658693.27	174804.47	668486.26	356589.81	149111.35
05	0	145194.75	339428.99	174115.30	1119856.22	-44969.66	66655.40	173247.84
06	0	120052.54	358305.05	667116.21	-45518.09	415318.95	459246.76	123719.27
07	0	144998.12	163425.65	356706.90	65889.33	457630.62	1567880.12	-44839.45
08	0	49746.72	221165.98	148681.93	173528.56	124619.97	-44409.80	1040748.64
09	1	0.19	0.17	0.19	0.16	0.19	0.19	0.07
10	0	56943.35	530780.02	310853.19	333033.61	42480.34	209270.15	
11	0	531129.79	1636335.86	149816.77	647443.99	331602.11	139905.36	
12	0	311841.99	149432.19	1092269.22	-68200.82	39068.40	163084.21	
13	0	335495.07	646707.28	-68328.07	369110.54	436436.78	115315.59	
14	0	135877.91	332057.22	38339.69	434430.92	1540332.38	-54989.32	
15	0	211714.10	140224.99	164076.68	116659.59	-53861.68	1037266.37	
16	1	0.32	0.46	0.85	0.75	0.68		
17	0	3313793.78	-2750151.88	-2459665.28	-66745.23	-1814652.27		
18	0	-2756935.16	-610388.05	-1892476.46	-194813.09	-984494.31		
19	0	-2480106.77	-1900131.15	-1566535.62	184815.48	-1119306.27		
20	0	-934324.90	-703553.70	-360454.85	1438423.95	-555020.03		
21	0	-1832950.42	-991882.31	-1121867.90	-212647.26	258158.48		
22	1	0.83	0.74	0.02	0.55			
23	0	1677868.13	147655.56	-139674.39	531820.03			
24	0	158357.47	268743.39	234417.62	244752.45			
25	0	71935.97	330945.58	1457110.45	-41141.34			
26	0	529466.54	234515.41	-175988.25	1266281.21			
27	1	0.09	-0.08	0.32				
28	0	254491.22	247086.22	194078.06				
29	0	324471.34	1462865.33	-64160.85				
30	0	186863.42	-133630.93	1096851.92				
31	1	0.97	0.76					
32	0	1148128.13	-310759.07					
33	0	-314888.45	954835.72					
34	1	-0.27						
35	0	869815.84						
36	1							

* The rows numbers 01, 09, 16, 22, 27, 31, 34, and 36 are the pivotal rows at each stage of the reduction of the matrix.

* Elements of the bold face are determinants of $|E + G|$.

Table A.2 Individual (separate) analyses of variance (RCB design) for a trial with 20 genotypes and four replications, by environment.

Environment	Sources of variation	Degree of freedom	Sum of squares	Mean squares
1.	Replication	3	911451.90	303817.30
	Genotype	19	23447395.23	1234073.43
	Error	57	27698801.25	485943.88
2.	Replication	3	1872851.38	624283.79
	Genotype	19	10888312.31	573069.07
	Error	57	12378698.93	217170.16
3.	Replication	3	147301.334	49100.445
	Genotype	19	5131004.336	270052.860
	Error	57	7617262.98	133636.19
4.	Replication	3	2960607.67	986869.22
	Genotype	19	20594592.71	1083925.93
	Error	57	32769124.21	574896.92
5.	Replication	3	4317283.39	1439094.46
	Genotype	19	19868804.89	1045726.57
	Error	57	7027841.01	123295.46
6.	Replication	3	2227983.262	742661.087
	Genotype	19	7686009.276	404526.804
	Error	57	11235935.45	197121.67
7.	Replication	3	935675.59	311891.86
	Genotype	19	20295738.81	1068196.78
	Error	57	28626828.93	502225.07
8.	Replication	3	776750.05	258916.68
	Genotype	19	13191792.03	694304.84
	Error	57	20910287.38	366847.15

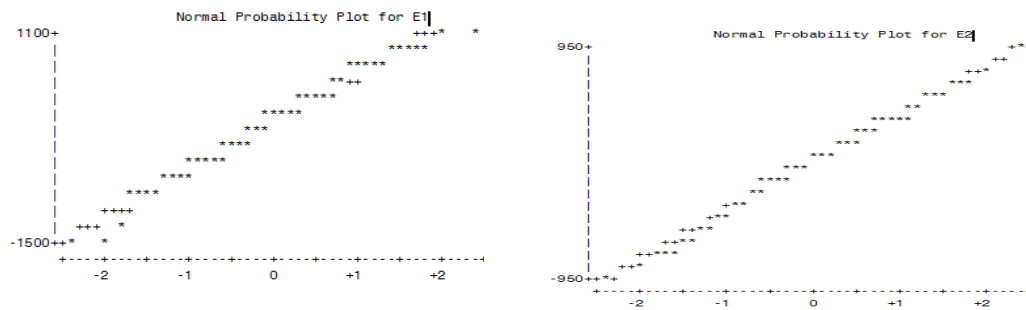
Table A.3: B ' f f
of the separate ANOVA.

Environment	Source	DF	Chi-Square	Pr > Chi-Square
1.	Genotype	19	14.1906	0.7725
2.	Genotype	19	12.7759	0.8498
3.	Genotype	19	20.7092	0.3531
4.	Genotype	19	25.1756	0.1548
5.	Genotype	19	17.8538	0.5322
6.	Genotype	19	17.3993	0.5628
7.	Genotype	19	18.5856	0.4837
8.	Genotype	19	23.3652	0.2216

Table A.4: Test for Normality of residuals in each of the separate ANOVA model using the Shapiro-Wilk (W) statistic.

Environment	Value of W	P-value=Pr<W
E1	0.984037	0.4207
E2	0.988484	0.6968
E3	0.98672	0.5804
E4	0.980998	0.2803
E5	0.987432	0.6269
E6	0.983928	0.4148
E7	0.993185	0.9520
E8	0.987091	0.6045

Figure A.1: Normal probability plot of residuals for each environment from the separate ANOVA.



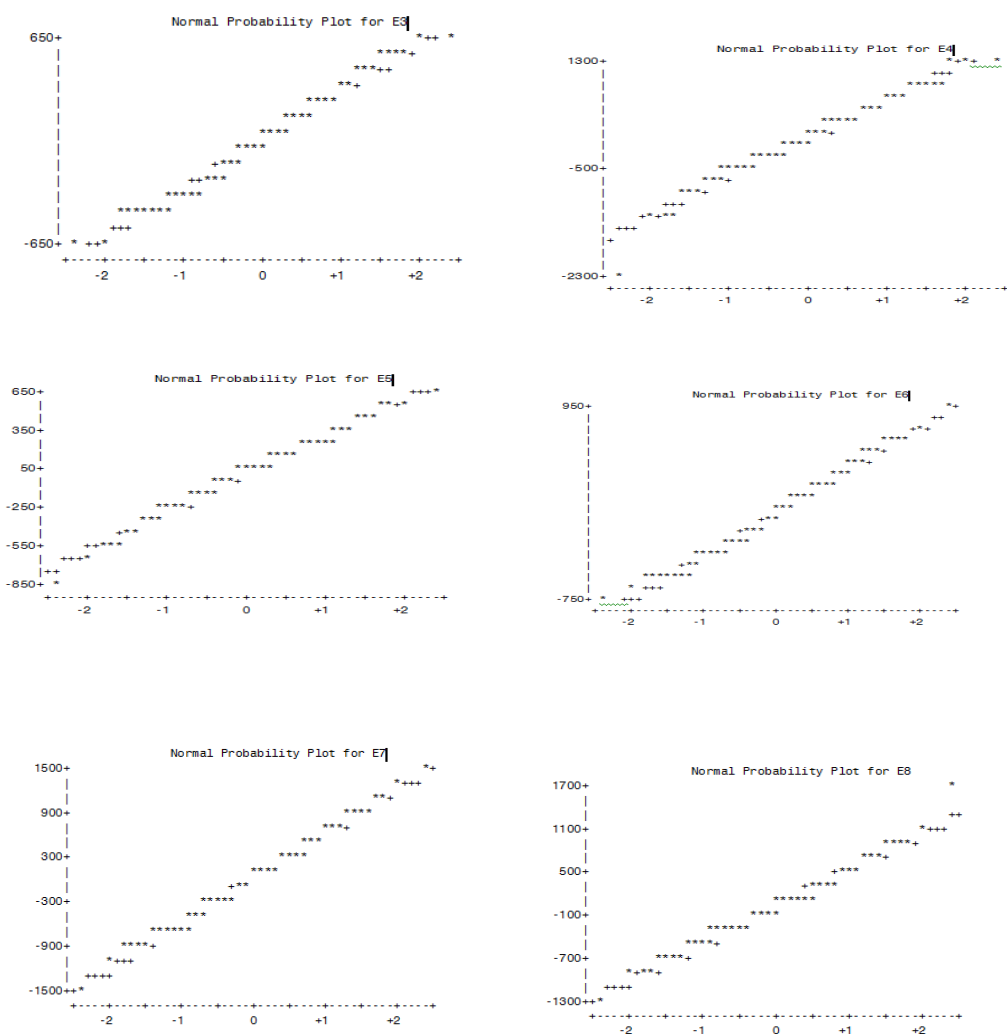
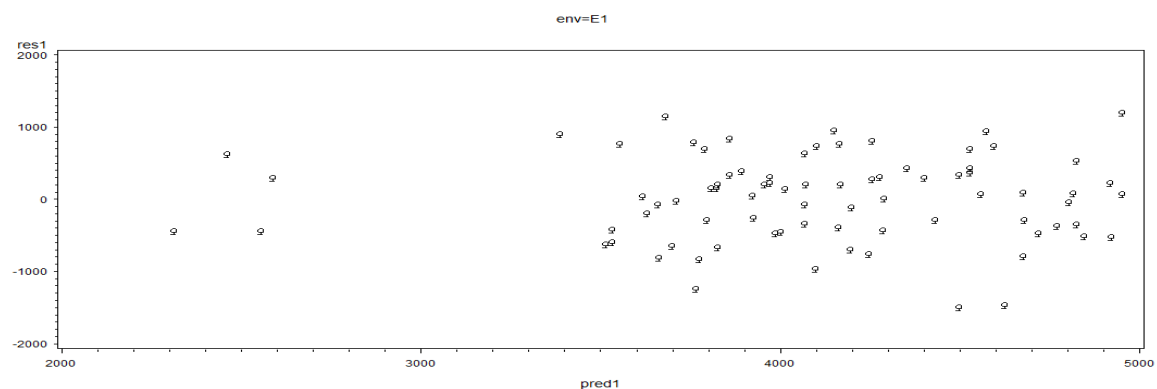
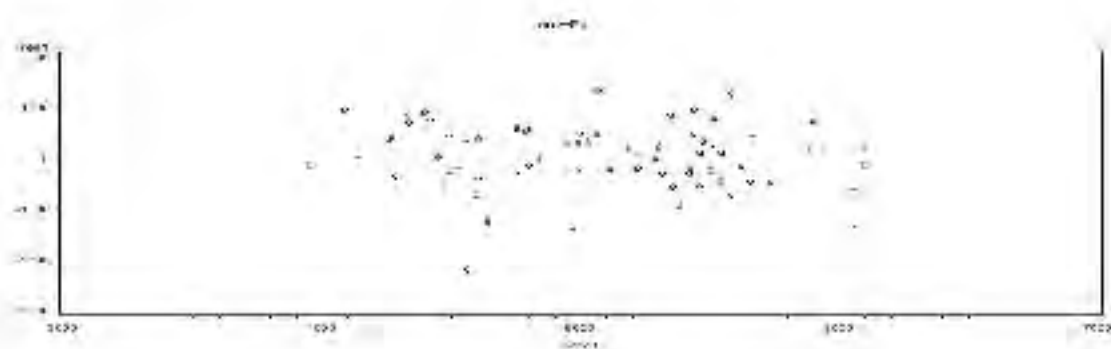
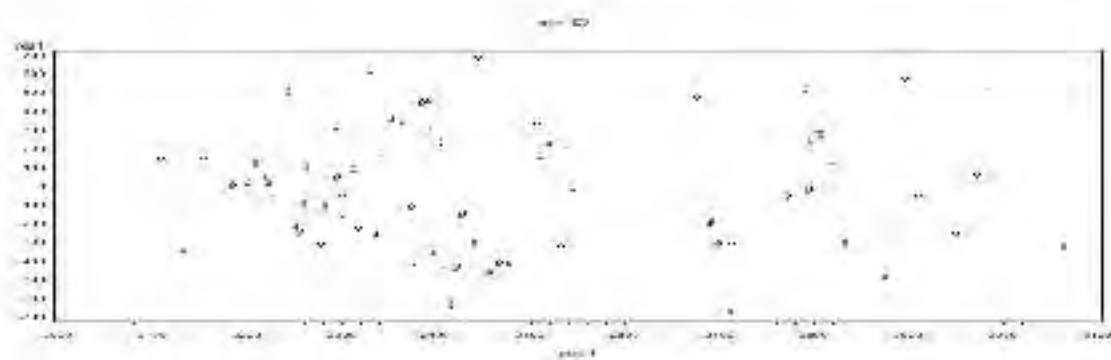
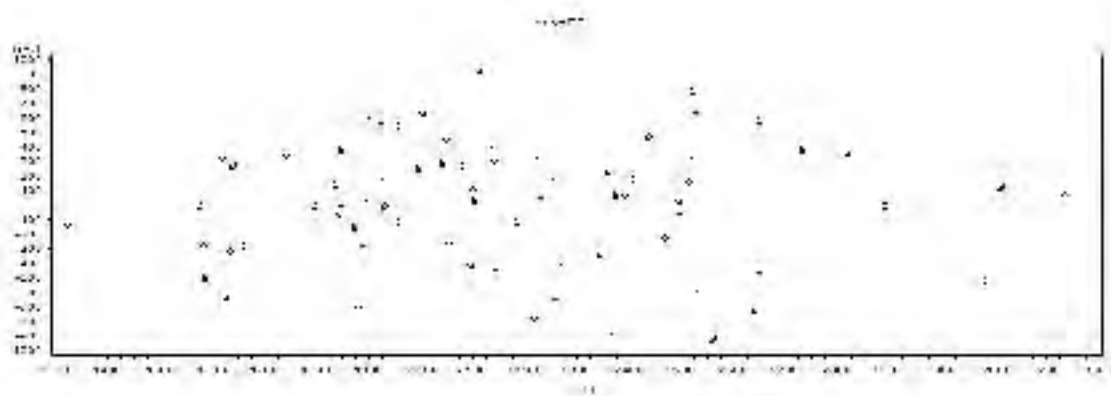
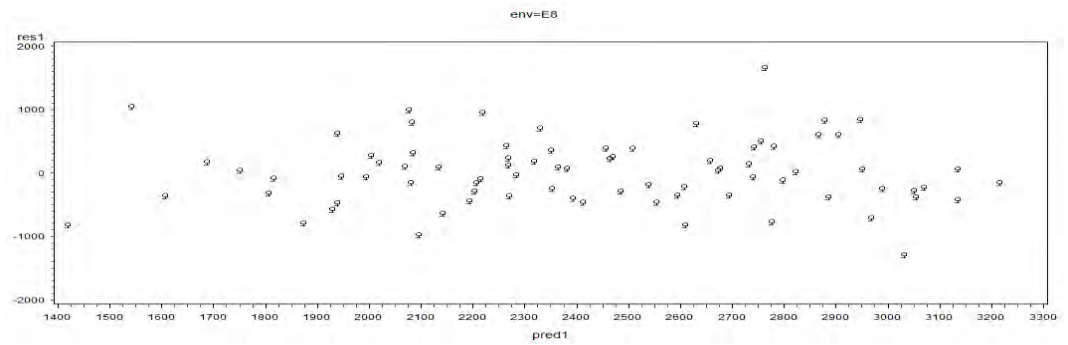
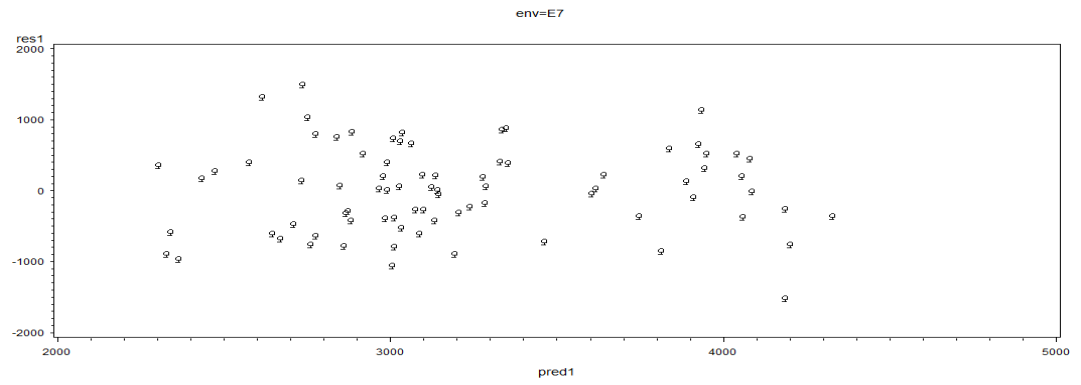
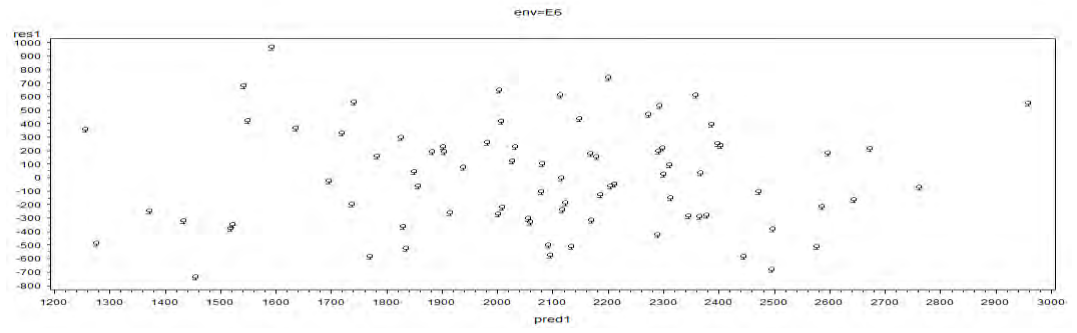


Figure A.2: Plot of residuals against the predicted value in each of the separate ANOVA model.







Declaration

I, the undersigned, declare that the thesis is my original work, has not been presented for degree in any other university and that all sources of material used for the thesis have been duly acknowledged.

Declare by

Name-----

Signature-----

Date-----

This thesis has been submitted for examination with my approval as a University advisor.

Name-----

Signature-----

Date-----