7907550

SEMUGURUKA, GEOFFREY HABIMANA
CLUSTER ANALYSIS APPLICATION TO CLASSIFYING ENVIRONMENTS IN SOYBEAN TESTS.

THE UNIVERSITY OF NEBRASKA - LINCOLN, PH.D., 1978

University Microfilms International 300 N. ZEEB ROAD, ANN ARBOR, MI 48106

CLUSTER ANALYSIS APPLICATION TO CLASSIFYING ENVIRONMENTS IN SOYBEAN TESTS

by

Geoffrey Habimana Semuguruka

A DISSERTATION

Presented to the Faculty of

The Graduate College in the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Department of Agronomy

Under the Supervision of Professor Wilfred M. Schutz and
Professor James H. Williams

Lincoln, Nebraska

December, 1978

TITLE

Cluster Analysis Application to Classifying Environments

In Soybean Tests

BY

Geoffrey Habimana Semuguruka

APPROVED	DATE
Dr. W.M. Schutz, Co-Chmn	11/10/78
Dr. J. H. Williams, Co-Chmn	11/10/78
Dr. R. F. Mumm	11/10/78
Dr. W. A. Compton	11/10/78
Dr. J. W. Schmidt	11/10/78

SUPERVISORY COMMITTEE

GRADUATE COLLEGE UNIVERSITY OF NEBRASKA

GRAD 82002 800 5-66(T)

ACKNOWLEDGEMENTS

The author expresses his sincere gratitude to his major advisers, Dr. Wilfred M. Schutz and Dr. James H. Williams, for their guidance, counsel and advice throughout the period of graduate study and also for their suggestions and constructive criticism in the preparation of this manuscript.

Sincere appreciation is also expressed to Dr. R. F. Mumm, Dr. W. A. Compton and Dr. J. W. Schmidt for serving on the supervisory committee and all members of the Biometrics Center for their assistance in the analysis of the data.

My special thanks goes to my brother, Dr. William D. Semuguruka, without whose encouragement these graduate studies would not have been undertaken.

The author wishes to thank the Ministry of Agriculture of the Government of Tanzania for granting a leave of study and the International Institute of Tropical Agriculture for providing the financial support which made his graduate studies possible.

Appreciation is extended to the Agronomy Department, Biometrics and Information Systems Center and the University of Nebraska for providing facilities and materials.

TABLE OF CONTENTS

LIST OF TABLES

TABLE		PAGE

LIST OF FIGURES

FIGURE		PAGE

FIGURE PAGE

INTRODUCTION

Evaluating differences among improved crop varieties over several environments has long been practiced by plant breeders as part of a breeding program. Different locations have different environmental effects and, hence, the expression of phenotypes depends upon its variety and the environment, and these two are not always independent. The interplay in the genetic and non-genetic effects on the development is called variety x environment (V x E) interaction. V x E interaction acts to confound comparisons among lines in most field situations.

A knowledge of the nature and relative magnitudes of the various types of variety x environment interactions is important in making decisions concerning breeding methods, selection programs, and testing procedures in crops. The presence of large variety x year interaction requires that selection programs include more than one year of testing. Similarly, a large variety x location interaction requires testing at several locations. Variety x location x year interaction can be considered in a selection program by either testing for several years or by testing at several locations. The relative magnitude of the plot-to-plot error component has a bearing on the number of replications that should be included in a test.

Disregarding V x E interaction effects will lead to erroneous conclusions. Thus, several statistical techniques have been used to assist in analyzing variation between and within varieties and environments. V x E interaction could be minimized by either subsetting a

set of environments into homogeneous groups or by grouping lines of varieties that reveal similar responses across environments. Grouping of homogeneous locations can be used to control variety x location interaction while grouping varieties can control variety x year and variety x year x location interaction effects. Nevertheless, the effectiveness of subsetting any set of locations depends upon the magnitude of the variety x year, variety x year x location, and the experimental error component of variance. Minimizing the variety x environment interaction will increase precision or information for comparing varieties.

Since V x E interaction is important, many states and countries have regional testing programs to provide information for evaluation of genetic materials, to determine adaptation of varieties to certain regions or zones, and to select varieties for release to growers and farmers. In defining these regions, the basic criterion has been to evaluate mean differences in certain environmental factors with respect to the level of performance per se, without regard for the relationship of varieties to environments. As far as varietal testing is concerned, the interaction effects of environmental factors with the variety may be more important than their main effects.

The purpose of this study is to classify environments for soybean varieties based on the similarity of locations such that locations with the same response to the same group of varieties are zoned as one group. Once such homegeneous subsets are established, it should then be possible to define proper test zones for variety adaptation trials.

The technique of cluster analysis was used to classify environments, utilizing two methods of measuring similarity of locations: (1) product moment correlation coefficient, and (2) distance coefficient. Both measure the resemblance between pairs of locations on the basis of an estimated V x E interaction effect for yield. The measures of similarity were computed for all possible pairs of locations. Then locations having similar patterns of interactions were clustered into groups at different successive stages. The zones of adaptation should minimize the magnitude of the variety x location interaction effects mean square.

Before the clustering process is performed, it is imperative that the existence of V x E interaction be established. Thus, in this study, several analyses of variance were carried out and pertinent components of variance and stability parameters estimated. An unstable variety is one that interacts with the environments.

LITERATURE REVIEW

Variety x Environment Interaction and Stability Parameters

Variety x environment (V x E) interaction plays a major role in the study of plant adaptation. It is defined by Comstock and Moll (1963) as the interplay of the genetic and non-genetic effects on the development of plants. Since the effects of varieties and environments are not independent, the consequences of variation in varieties depends also on the environment.

The study of V x E interaction has been approached in two ways. One is the estimation of components of variance, and the other is the estimation of stability parameters. Sprague and Federer (1951) used estimation of components of variance technique and proposed the separation of variance due to the effects of varieties, environments, and V x E interactions. This was done by equating the observed mean squares to their expectations and solving the resulting equations. The components of variance approach has been used in several ways to study V x E interactions.

Schutz and Bernard (1967) pointed out that the genotype x environment interaction consists of two parts. One part is due to differences in genetic correlation from one environment to another, and the other is due to differences in genotypic variances. Thus, consideration of the make-up of the genotype x year and genotype x location interaction and genotype x year x location interaction is quite useful in assessing their effect on selection. They also stated that

since, in a soybean breeder's program, the turnover rate is largely determined by the number of lines that can be evaluated each year in regional tests, it is essential that neither poor estimates nor consistent over-estimates of genotype x environment interactions are used in eliminating undesirable lines. It is especially true for the genotype x year interaction, since an over-estimate leads to unnecessary testing over long periods of time.

Miller et al. (1962) found that the three-factor interaction of varieties with sites and years to be particularly important. While studying lint yield of cotton, it was noted that the variety x year x location interaction was relatively large and statistically significant. The first-order interaction of varieties x location was considerably smaller than that for second-order interaction, although still statistically significant. The variety x year source of variation was very small and non-significant. The larger second-order interaction indicated that the varieties showed differential responses when grown in different environments. The presence of a variety x location interaction indicated further that there was at least some consistent location effects on these differential responses. The very small and non-significant variety x year interaction indicated that, averaged over locations, the varieties ranked essentially the same in yield performance in each of the years of testing.

In the study to characterize the nature of G x E interaction in cotton varieties, Abou-El-Fittouh et al. (1969) estimated components of variances of several traits using data from four national standard

varieties grown over a wide range of environments. Results of estimates of pertinent components of variance obtained for yield indicated that the error component and the interaction components, except genotype x year, to be from 1¼ to almost 3 times as large as the component due to genotypic differences. The genotype x location component was the largest of the interaction components. For the remaining traits, the error and interaction components were small and mostly unimportant when compared with the genotypic component. For all traits other than yield, the three-factor interaction was the predominant interaction component. Except for seed index and lint percentage, the genotype x year component was least important. Traits considered in this study included lint yield, boll size, lint percentage, seed index, fiber length, fiber fineness, fiber strength, fiber elongation, and yarn strength.

Kaltsikes (1970) reported that significant cultivar x location interactions obtained indicated that plant breeding in Western Canada should be limited to producing cultivars that perform well on a more restricted environmental range. When the number of years and replications were kept constant, the expected standard error of a cultivar mean decreased as the number of locations increased. An increase in the number of years of testing was more effective where there were only a few locations. Overall, the estimates of cultivars x year interaction variance was relatively small, although it accounted for 40% of the variance of a cultivar mean when only three years of testing were considered.

Based on the knowledge of the relative magnitudes of genotye x environment interaction component of variance, varietal testing regions have been developed (Horner and Frey, 1957; Liang et al., 1966; Abou-El-Fittouh et al., 1969). Liang et al. (1966) estimated variety x environmental interaction components of variation in yield tests conducted in three years involving 10 wheat varieties at 13 locations, 4 barley varieties at 10 locations, and 5 oat varieties at 5 locations. They emphasized that the significant variety x location interactions for wheat and barley indicated that Kansas should be divided into sub-areas for varietal testing. By grouping appropriate locations together, the magnitude of the variety x location components of variation was reduced.

Horner and Frey (1957) suggested that a reduction in the genotype x environment interaction, and thus more efficient use of locations, might best result from subdivision of a region into 2, 3, 4, and 5 zones. Upon doing this the genotype x environment interaction resulted in a decrease of 11, 21, 30, and 40 percent, respectively, within the sub-region, compared to estimates involving the whole state.

Miller et al. (1959) conducted an experiment involving 15 cotton varieties tested at 9 locations in North Carolina over a 3-year period. The variety x year and variety x location interactions were small and non-significant for yield. As a result of the small variety x location interaction, they suggested that it is unnecessary to divide the state into sub-areas for cotton variety testing. However, the size of the variety x year x location interaction indicated that it was

essential to test cotton varieties over a number of different environments within the state. If too many tests were carried out in any one year, to shorten the time required for evaluation, one will not obtain as much information as would be gathered by using one location over several years. When the number of test locations is increased in an area, these locations will be closer together and thus more likely to have similar weather patterns.

Abou-El-Fittouh et al. (1969) reduced the within-group genotype x location interaction in regional cotton variety performance trials by classifying locations into homogeneous groups. They then suggested some modifications in the currently recognized zones of cotton adaptation.

It was suggested by Miller et al. (1962) that, in general, if a large variety x location interaction is observed in a variety performance test over a range of environments, it would indicate that the region concerned should be subdivided into smaller breeding areas. In their cotton variety testing trials, the first-order interaction over 11 locations was relatively small, being about one-fourth the magnitude of the variety source of variation. Elimination of 3 Texas locations from the analysis decreased the variety x location interaction to about one-thirtieth of the variety component.

Hanson (1964) has noted that biases are introduced when data are analyzed on the assumption that a sample of environments is random. In quantitative genetics research, physical restrictions normally limit the extent to which an investigator can sample years and sites. In

practice, he suggested that investigators have a tendency to select sites that represent the range of conditions found within an area rather than a sample of the area. Hanson, then, collected data in y years at s sites in the context of random samples of environments, and expectation for mean squares involved a correction resulting from certain covariances associated with the fact that ys environments were not a random sample of ys independent environments. The corrections involve covariance terms whose expectations are not zero and thus contribute to biased estimates of parameter if ignored. So he concluded that bias in estimation is introduced by assuming that the data collected at s sites for y years form ys random environments unless the two intraclass correlations ρ_y and ρ_s are zero. These intraclass correlations are estimated as:

$$\rho_y = \sigma^2_{gy} / [\sigma^2_{gys} + \sigma^2_{gy} + \sigma^2_{gs}]$$

$$\rho_s = \sigma^2_{gs} / [\sigma^2_{gys} \quad \sigma^2_{gy} \quad \sigma^2_{qs}]$$

where ρ_y = the intraclass correlation involving the variability of the genotype x environment interaction effects among and within the genotype x year classes,

ρ_s = the intraclass correlation involving the variability of the genotype x environment interaction effects among and within the genotype x site classes.

He represented the genotype x environment effect $(GE)_{ijk}$ as a linear model of the interaction effects of genotype and year $(gy)_{ik}$, of genotype and site $(gs)_{ij}$ and of genotype, year, and site $(gys)_{ijk}$; that is:

$$(GE)_{ijk} = (gy)_{ik} + (gs)_{ij} + (gys)_{ijk}$$

In terms of variance components estimated from the analysis of variance, the model becomes:

$$\sigma^2_{ge} = \sigma^2_{gy} + \sigma^2_{gs} + \sigma^2_{gys}$$

where σ^2_{ge} = the genotype x environment variance component,

σ^2_{gy} = the genotype x year variance component,

σ^2_{gs} = the genotype x site variance component, and

σ^2_{gys} = the genotype x year x site variance component.

If σ^2_{ge} is small relative to σ^2_{g} (genotype variance component), a bias in the estimate for σ^2_{ge} is of little practical significance. The bias in the estimate for σ^2_{g} is negligible when σ^2_{ge} is small. He noted that the ratio observed for grain yield involving corn or soybeans normally falls in the range of 0.5 to 1.0. The ratio tended to be smaller for other commonly measured characteristics of these crops.

The estimation of stability parameters provides a statistical tool that aids in the measurement and interpretation of genotype x environment interactions. The regression technique that is used in the estimation of stability parmeters measures that portion of the interaction between genotypes and environments, which is a linear function of the environment. Thus, the stability analysis method is based on the idea that genotype x environment interaction can be explained wholly or in part by linear regression of the performance

of the various genotypes on a measure of the environment. Several measures of the environmental quality have been suggested. Yates and Cochran (1938) first proposed the technique of regression by regressing yield of a genotype on the mean yield of all genotypes at each location to measure the stability of each barley variety. This regression technique was not used again until Finlay and Wilkinson (1963) used it in their study of adaptation of 277 barley varieties tested over three years in three locations. For each variety, a linear regression of yield on the mean yield of all varieties in each environment was used to measure variety stability. In order to induce a high degree of linearity, the yields were transformed to a logarithmic scale. Later, Jowitt (1972) showed this logarithmic transformation scale to be more useful than non-transformed data when the technique was applied to sorghum yields.

Finlay and Wilkinson (1963) concluded from their study that regression coefficients of approximately 1.0 indicated average stability. Thus, when this was associated with high mean yield, varieties have general adaptability and, when this was associated with low mean yield, varieties are poorly adapted to all environments. Regression coefficients above or less than 1.0 described varieties with less than or greater than average stability, respectively. Varieties with below-average stability ($b > 1$) had greater adaptability to high-yielding environments, and varieties with above-average stability ($b < 1$) had greater adaptability to low-yielding environments. If the regression coefficient was equal to zero ($b = 0$), then the variety was, for all

practical purposes, stable because the variety would yield the same in all kinds of environments.

The use of regression technique has further been elaborated by Eberhart and Russel (1966). Their study included a second stability parameter, the variance due to deviations from regression. The following linear additive model was proposed:

$$Y_{ij} = \mu_i + \beta_i I_j + \delta_{ij}$$

where Y_{ij} = the variety mean yield of the i^{th} variety in the j^{th} environment,

μ_i = the mean of the i^{th} variety over all environments,

β_i = the regression coefficient that measures the response of the i^{th} variety to varying environments,

δ_i = the deviation from regression of the i^{th} variety in the j^{th} environment, and

I_j = the environmental index measured as the mean of all varieties in the j^{th} environment minus the grand mean of all environments.

With this model, the sums of squares due to environments and variety x environments are partitioned into environments (linear), variety x environments (linear), and deviations from the regression model. At the same time, the model provides a means of partitioning the genotype x environment interaction of each variety into two parts: (1) the variation due to the response of the variety to varying environmental indices, which is the sum of squares due to regression and (2) the

deviations from regression on the environmental index.

Eberhart and Russel (1966) and Eberhart (1969) defined a stable variety as one with a regression coefficient equal to 1 (b = 1), and non-significant deviations from regression ($S_d^2 = 0$). Marquez-Sanchez (1973) added that, under the conditions where the regression coefficient was equal to 1, and deviations from regression are equal to zero, a stable variety was one that does not interact with the environment.

Joppa et al. (1971) made several observations on the application of the Eberhart and Russel model (1966). First, he computed the regression coeffiicent (b) and mean square for deviations from regression S_d^2. Then he defined that statistic b as a measure of the average increase in yield of a cultivar per unit increase in the environmental index and the statistic S_d^2 as a measure of how well the predicted response agrees with the observed response and the genotype x environment interaction. He suggested that a cultivar may deviate from the slope of the average regression line (b = 1) for at least two reasons: (1) A cultivar may yield relatively more than other cultivars in low-yielding environments and consequently b < 1 for these cultivars or the converse is true. (2) The yield of most of the cultivars could be reduced by some pathogen or factors common to a large number of environments to which the cultivar in question is resistant. Conversely, a cultivar may be susceptible to a factor or pathogen to which most cultivars are resistant. So the regression coefficient is a measure of the relative, general stability or adaptation of a cultivar.