Multivariate Approach in Varietal Recommendations

M. Anputhas , S. Samita[1] and S.De.Z. Abeysiriwardena[2]

International Water Management Institute,
P.O.Box 2075, Colombo, Sri Lanka.

ABSTRACT. A crop variety (genotype) good for cultivation should necessarily give a very good economic return to farmers with low cost of production along with good consumer preference. In addition, a good crop variety should be consistent across the locations over different seasons over a period of time. These aspects are usually measured by relative stability, performance and superiority, and are normally been set as goals in all breeding programmes. The yield is often evaluated for the above criteria. However, yield alone will not lead to good economic returns, as other characters such as quality aspects and agronomic aspects need to be considered. Thus univariate statistical methods are not appropriate. This study proposes a methodology for varietal selection based on all the important aspects using Principal Component Analysis (PCA) which can be used efficiently to obtain a solution to the above problem. The eigen values and their vectors from PCA provide the basis to develop a single aggregated index by which the consistency of the crop variety can be investigated. Using this index relative stability, performance and superiority can be computed and accordingly selections can be made. A paddy data set on three months age group of Yala 2002 was used to illustrate the proposed methodology. The traditional univariate analysis (on yield) was also performed for the purpose of comparison.

INTRODUCTION

The economic merit of a crop depends on many individual charters (traits) of a crop. Single trait selection is often utilized to maximize the genetic gain. However, traits negatively correlated to the primary trait can deteriorate the merits of the traits the crop was selected upon. Selection based on multiple traits can be used to correct this deficiency. For instant, in the case of paddy, researchers have identified, in addition to yield, characters such as days taken to flowering (number of days taken to reach fifty percent of heading), days taken to maturity (number of days taken to reach eighty-five percent of the panicle golden brown), brown rice percentage (percentage of rice just after dehusking), total milled rice percentage (percentage of rice after polishing) and head grain percentage (percentage of full grain after milling) need to be considered in varietal selection.

Multivariate statistical techniques pave the way to handle several variables simultaneously. Thus a multivariate approach needs to be adopted in proper varietal selection. Only a few research works has been carried out in Sri Lanka in the direction of establishing selection criteria based on multiple characters (Perera et al 1989; Wigesuriya. et al 1993). Often varieties give different yields under different environmental conditions. Thus in varietal recommendation this phenomenon (which

[1] Department of Crop Science, Faculty of Agriculture, University of Peradeniya, Sri Lanaka.

[2] Rice Research and Development Institute, Batalagoda, Sri Lanka.

is referred to as genotype environmental interaction G X E interaction) is taken into account.

If recommendations are to be based on multi traits then, for each of these traits, the G X E will have to be considered. This will enable farmers to have a better option for cultivation in the long run as well as new directions for future varietal improvement programmes. This study suggests a more efficient methodology for varietal selection based on multi traits.

MATERIALS AND METHODS

Univariate techniques used

A large number of techniques are currently used on univariate basis for varietal selection. Some of the popular technique often used are Yates and Cochran (1938), Finlay and Wilkinson (1963), Eberhart and Russel (1966), Shukla (1972) and Eskridge (1990). Most of these techniques consist of fairly complicated analysis and cannot get performed by standard software available. However, Kamidi (2001) proposed a simple technique for varietal selection. The uniqueness of this method is that the methodology can get implemented easily using standard statistical software. This technique is based on the regression model of the form;

$$y_{ij} = \mu_i + \beta_i x_{ij} + d_{ij} \qquad (1)$$

where

y_{ij} - yield of the i^{th} genotype at the j^{th} environment,

μ_i . i^{th} genotype mean,

β_i . regression coefficient,

d_{ij} - deviation from regression.

x_{ij} . environmental index for the i^{th} genotype at the j^{th} environment given by

$$x_{ij} = \frac{g\bar{y}_{\bullet j} - y_{ij}}{g-1} \qquad (2)$$

where

$\bar{y}_{\bullet j}$ - marginal mean of the j^{th} environment
g – number of genotypes

The three measures (indices) used for varietal recommendation are stability, performance and superiority.

Stability

The stability is defined as the correlation between genotype and environmental index. The correlation coefficient (ρ) significantly different from zero merely signifies the presence of some association between two variables. This association has to be sufficiently strong for a stable genotype. If ρ = 1 then it is regarded that the variety is stable. Thus in testing the stability it is necessary to determine whether the estimate of the correlation (r_{ge}) actually represents $\rho = 1$. The test is then H_o: $\rho = 1$ versus H_A: $\rho < 1$. Depending on the outcome, varieties can be classified. If the ρ is not being significantly different from unity at $\alpha = 0.05$ (i.e. $P > 0.05$) the genotype is regarded as very stable. Similarly, if $0.01 < P < 0.05$, the variety is considered as sufficiently stable, if $0.001 < P < 0.01$ the variety is considered as fairly stable and if $P < 0.001$, the variety is considered as unstable.

Testing ρ with H_o: $\rho = 1$ versus H_A: $\rho < 1$ is fairly complicated. The test used in this circumstance is the test suggested by Gayen (1951). However, one can use the critical values published by Kamidi (2001) and make inference without actually performing the test.

Performance

The relative performance of the i^{th} variety (p_i) is defined as $b_i - 1$, where b_i is the estimated regression coefficient (measure of response across environments) from model (1) i.e., by how much its response lies above or below the average ($b = 1$).

Relative superiority

Superiority (s_i) is measured as a product of relative performance and stability i.e. $(s_i = p_i \times r_{ge})$. This measure is usually taken as the measure for selecting stable high yielding varieties.

Although the method of Kamidi (2001) is easily to be implemented, the limitation is that stability, performance and superiority indices are computed based only on one variable, i.e. yield only. If the above technique need to be used for multivariate situation then y_{ij} should be replaced by z_{ij} (single aggregated index of i^{th} variety in j^{th} environment) which gives a combined value for all the variables.

Multivariate approach

In the past, attempts have been made to select varieties based on multi responses. Most of these methods introduced were simply computing indices (Smith, 1936; Hazel, 1943; Kempthorne and Nordskog, 1959; Johnson *et al.*, 1988; Bernado, 1991 and Dolan *et al.*, 1996) and do not provide a proper statistical basis. Therefore, the conclusions were subjective. This study suggests a methodology that will take into account all the traits with a proper statistical basis using the multivariate statistical technique PCA

A principal component (Q) in general is of the form

$$Q = a_1 y_1 + a_2 y_2 + \dots\dots\dots\dots\dots\dots a_p y_p \qquad (3)$$

where,

y_l to y_p p response variables (traits).

a_l to a_p - eigen vector coefficient corresponding to p response variables respectively.

If only one PC is sufficient to explain the variability of the variables y_l to y_p, Q_{ij} (where Q_{ij} is the principal component score of the ith variety from jth environment) will replace $z_{ij..}$ Otherwise if several (assume m) PCs are necessary to explain the variability; z_{ij} can be computed based on linear combination of PCs as follows.

$$z_{ij} = \frac{\lambda_1}{\sum_{k=1}^{p}\lambda_k} Q_{1ij} + \frac{\lambda_2}{\sum_{k=1}^{p}\lambda_k} Q_{2ij} + \ldots\ldots\ldots + \frac{\lambda_p}{\sum_{k=1}^{p}\lambda_k} Q_{mij} \quad \ldots\ldots \; (4)$$

where,

Q_{1ij} to Q_{mij} – *scores for PC_l to PC_m* for ith variety from jth Environment.

λ_1 to λ_p - eigen values of corresponding PCs.

$\sum_{k=1}^{p}\lambda_k$ - Sum of eigen values from all (p) PCs.

It is to be noted that if PCA is based on standardized variables $\sum_{k=1}^{p}\lambda_k = p$

Data used

The paddy data from six sub stations of Rice Research and Development Institute (RRDI) of Sri Lanaka for Yala 2002 on six (3 months age group) varieties is used to illustrate the suggested method. The six varieties were At-303, Bg-300, Bg-305, Bg-2834, Bg-2845 and Ld-98-3. The sub stations were Ambalantota, Batalagoda, Bowbuwala, Labuduwa, Maha Illuppallama and Vantharumullai. Other than yield the traits considered here were, days taken to flowering, days taken to maturity, brown rice percentage, total milled rice percentage and head grain percentage. The statistical analysis was carried out using SAS Version 6.12 (SAS, 1990). The selection results from illustrated multivariate method were compared against the results from univariate based analysis.

RESULTS AND DISCUSSION

Simple correlations

The correlation coefficients of five traits, days taken for flowering, days taken for maturity, brown rice percentage, total milled grain percentage and head grain percentage to yield were 0.40 ($P = 0.02$), 0.36 ($P = 0.03$), 0.42 ($P = 0.01$), – 0.16 ($P = 0.34$) and – 0.16 ($P = 0.35$) respectively. The correlation coefficients show that certain characters are not related to the yield. This indicates the fact that selection based only on yield can deteriorate the merits of the expected outcome.

Use of PCA

The PCA for the 6 traits gave 6 PCs with eigen values 2.16, 1.57, 1.06, 0.63, 0.34 and 0.26 respectively. The eigen values of the first three principal components are relatively large and these three principal components explained the 79.66% of the variability. Therefore three PCs were used to establish the single aggregated index (z). The eigen vectors of the first three principal components are given in the Table1.

Table 1. Eigen vectors of the first three PCs.

Trait	PC_1	PC2	PC3
Yield (Yd)	0.51	0.03	-0.32
Days taken for flowering (Df)	0.52	0.22	0.45
Days taken for maturity (Dm)	0.55	0.01	0.36
Brown rice % (Br)	0.30	0.32	-0.70
Total milled rice % (Tm)	-0.23	0.61	0.27
Head grain % (Hg) ·	-0.14	0.69	-0.04
Eigen Value (λ_i)	2.16	1.57	1.06
% of variability explained	35.92	26.10	17.63

According to the results the three PCs can be formulated as

$$Q_1 = 0.51Yd + 0.52Df + 0.55Dm + 0.30Br - 0.23Tm - 0.14Hg$$

$$Q_2 = 0.03Yd + 0.22Df + 0.01Dm + 0.32Br + 0.61Tm + 0.69Hg$$

$$Q3 = -0.32Yd + 0.45Df + 0.36Dm - 0.70Br - 0.27Tm - 0.04Hg$$

The PCs were derived using standardized variables.

Yield, days taken to flowering and days taken to maturity contribute largely to the first PC. This implies that these characters are linked, which was revealed from correlation analysis too. The second PC is highly contributed by total milled rice percentage and head grain p:rcentage. The third PC is highly, but negatively contributed by brown rice percentage. Since one PC is inadequate, the Z_{ij} was calculated using the equations (3 and 4) as

$$Z_{ij} = \frac{2.16}{6}Q_{1ij} + \frac{1.57}{6}Q_{2ij} + \frac{1.06}{6}Q_{3ij} \qquad (5)$$

where

Q_{1ij} . First PC score for ith variety from jth environment.

Q_{2ij} . Second PC score for ith variety from jth environment.

Q_{3ij} . Third PC score for ith variety from jth environment.

Using z_{ij} as the response variable, the regression analysis was performed for the model specified in the equation (1). In addition, as a comparison, the regression analysis was performed for the same model taking yield values only as the response variable.

Regression analysis

The estimated regression coefficient (b_i) from the analysis using yield values only is given in Table 2. The relative performance (p_i) computed from b_i is also given in the Table 2. Corresponding output by using z_{ij} as the response variable is given in Table 3. (Note that p values associated with all regression coefficients estimated were < 0.05 and also R^2 for all the models were above 80 %)

Stability analysis

The stability analysis results (r_{ge}) for each variety using yield as the response variable are given in Table 2. Corresponding results using z_{ij} are given in Table 3.

Computation of relative superiority

Relative superiority values computed for each variety taking yield only as the response variable are given in Table 2. Corresponding values using single aggregated index as the response variable are given in Table 3.

According to Table 2, all six varieties considered were found to be adapted to locations at varying levels. Varieties Bg 2845, Bg 300, Ld 98-3 and Bg 305 were very stable ($P > 0.05$) and Bg 2834 and At 303 were sufficiently stable ($P > 0.01$). The variety Bg 305 was very stable across locations with higher yield but had poor relative performance ($p_i < 0$) and thus received lowest superiority ranking. Similarly Bg 2845 ranked one before the last in terms of yield but had relative performance higher than all other varieties and thus received highest superiority.

Table 2. Mean yield and corresponding selection measures based on univariate analysis for six varieties grown in six locations.

Variety	Yield (T/ha)	Stability (r_{ge})	Regression coefficient (b_i)	Relative performance ($p_i = b_i - 1$)	Relative superiority ($s_i = p_i \cdot r_{ge}$)
Bg 2845	4.31	0.98***	1.152	0.152	0.149
Bg 2834	4.52	0.93**	1.042	0.042	0.039
Bg 300	448	0.97***	1.030	0.030	0.029
At 303	4.24	0.93**	0.935	-0.065	-0.061
Ld 98-3	4.52	0.96***	0.895	-0.105	-0.101
Bg 305	4.64	0.99***	0.876	-0.124	-0.123

*, **, *** – r_{ge} not significantly different from one ($P > 0.001$, $P > 0.01$, $P > 0.05$ respectively)

The results are not surprising because in varietal selection stability is considered more important than yield. The variety At 303 has exhibited negative relative superiority. Thus, this variety can not be recommended generally. In fact, this variety has given the lowest average yield. However, this variety has given high yield (6.48 t / ha) at Maha Illuppallama. Thus, although the variety can not be recommended generally, it can be recommended for that location and locations with similar environments. Varieties Bg 2845, Bg 2834 and Bg 300 were the most superior varieties and recommended for all the locations but they were not the best three in terms of yield. This emphasizes the fact that average yield is not a criterion for selection.

Table 3. **Mean single aggregated index and corresponding indices for six varieties grown in six locations based on multivariate analysis.**

Variety	Single aggregated index (z_{ij})	Stability (r_{ge})	Regression coefficient (b_i)	Relative performance ($p_i = b_i - 1$)	Relative superiority ($s_i = p_i \cdot r_{ge}$)
Ld 98-3	0.19	0.95***	1.389	0.379	0.359
Bg 2834	0.21	0.86*	1.215	0.215	0.191
Bg 2845	–0.40	0.97***	1.026	0.026	0.025
At 303	–0.29	0.76[a]	0.850	–0.150	-0.114
Bg 305	0.38	0.80*	0.803	–0.197	-0.158
Bg 300	–0.09	0.95***	0.539	–0.461	-0.436

*, **,***, r_{ge} not significantly different from one ($P > 0.001$, $P > 0.01$, $P > 0.05$ respectively) [a], r_{ge} significantly different from one.

On the basis of z_{ij}, values, varieties Ld 98–3, Bg 2845 and Bg 300 were found to be very stable ($P > 0.05$). while varieties Bg 2834 and Bg 305 can be grouped into the fairly stable category ($P > 0.001$). Variety At 303 is unstable.

The varieties Ld 98–3. Bg 2834 and Bg 2845 occupied the first three ranks by means of relative superiority. So these varieties can be recommended based on all traits. Out of the three varieties recommended based on z_{ij}, the varieties Bg 2834 and Bg 2845 were recommended under univariate analysis too. This emphasizes the fact that even under the multi-trait approach yield plays an important role. However, Ld 98-3 is recommended under multi trait approach where as it is not recommended under univariate approach. Even though the variety Bg 2845 was more stable than Bg 2834, Bg 2834 was superior than Bg 2845 mainly because the latter gave poor head grain percentage (Table 4) and yield.

The variety At 303 is not recommended for generally since the relative performance is low. However, this variety has given highest z_{ij} at Maha Illuppallama (This was revealed by the univariate analysis too). This indicates that this variety is suitable for Maha Illuppallama and locations with similar environments.

The variety Bg 305 recorded the highest average z_{ij} with highest yield and head gain percentage at all the locations except at Labuduwa. In fact this is the only variety that meets RRDI recommendation for head grain percentage. However, superiority is low mainly because of poor relative performance. Late flower initiation and shorter grain filling period (Table 4) could be the reason for poor performance of Bg 305. An interesting point here is that there are indications that late flower initiation and short grain filling period lead to the higher yield with higher head grain percentage.

In calculation of single aggregated index, PC_1 is the most important PC. PC_1 represents yield, days taken to flowering and days taken to maturity more or less equally (Table 1). Accordingly days taken for flowering and days taken for maturity play an equal role as yield in giving higher z. This emphasizes the fact that it is important to improve these characters too. However, these three traits are inter related and when one trait is improved the other two also get improved. The second and third PCs represent other 3 traits that are economically important. Not having these three traits in PC1 implies that they are more or less independent from yield and thus should be considered separately in breeding.

Table 4. Mean values of each trait for different paddy varieties and over all mean.

Variety	Days taken for flowering (Df)	Days taken for maturity (Dm)	Dm-Df	Brown rice %	Total milled rice %	Head grain %
Ld 98-3	66.94	95.79	28.85	78.49	73.18	48.53
Bg 2834	68.28	97.04	28.76	77.66	73.18	43.93
Bg 2845	64.25	94.97	30.72	77.77	72.31	35.37
At 303	65.79	94.42	28.63	77.38	72.33	41.21
Bg 305	68.42	95.33	26.91	78.44	73.64	53.25
Bg 300	65.51	94.33	28.82	78.40	72.99	43.88
Overall Mean	66.53	95.31	28.78	78.02	72.94	44.36

One might argue that farmers are selling their product before processing or, they are getting the price for paddy regardless of the variety. But the consumers actually pay for it. Therefore, it is mandatory to give attention to these traits too. In addition, awareness must be created among the farmers and buyers in this regard.

In this illustration of multivariate approach yield played an important role. However, there can be arguments raised as to 'what assurance can be given that yield will always be taken care of? In general, among the characters considered, yield gives the highest variability. Thus yield will always be in the first PC. Hence, yield will always receive a high weight.

CONCLUSIONS

The recommendation made under the suggested method is more or less in agreement with the univariate methods. The importance of this suggested method is that selection is done taking into account all important yield characters in addition to the yield alone with a proper statistical basis. The methodology suggested here can be implemented by using standard statistical software. Information on the stability, performance and superiority can also be obtained by this method. Thus the additional information obtained in this method is not by sacrificing information obtained in univariate methods. Therefore, this approach is much more superior compared to existing univariate techniques. The recommendations currently being done are based purely on yield (quantity). Therefore it is high time in this country that other characters (quality parameters) are also used in making varietal recommendations so as to avoid long run risk in the breeding programme.

ACKNOWLEDGMENTS

The authors are grateful to the Central Rice Research and Development Institute and its Sub Stations staff for providing the dataset for illustration and their constructive comments. Our sincere thanks are also extended to Ms. Janitha Godamuduna and library staff of International Water Management Institute for their offered help. This piece of work is being dedicated to the memory of late Mrs. Rasamalar Markandu, mother of the first author.

REFERENCES

Bernado, R. (1991). Retrospective index weights used in multiple trait selection in maize breeding programme, Journal of Crop Science, 31 (5), 1174-1179.

Chatfield, C. and Collins, A.J. (1980). Introduction to Multivariate and Analysis. London: Chapman and Hall.

Dolan, D.J., Stuthman, D.D., Kolb, F.L. and Hewings, A.D. (1996). Multiple Trait Selection in a Recurrent Selection Population in Oat (Avena sativa L.), Journal of Crop Science, 36, 1207-1211.

Eberhart, S.A. and Russel, W.A. (1966). Stability Parameters for Comparing Varieties, Crop Science, 6, 36-40.

Eskridge, K.M. (1990). Selection of stable cultivars using a safety first rule, Journal of Crop Science, 30 (2), 369-374.

Finlay, K.W. and Wilkinson. G.N. (1963). The Analysis of Adaptation in a Plant Breeding Programme, Australian Journal of Agricultural Research, 14, 742-754.

Gayen, K.A. (1951). The frequency distribution of the product-moment correlation coefficient in random samples of any size drawn from non normal universes, Biometrika, 38, 219-247

Hazel, L. N. (1943). The genetic basis for constructing selection indexes. Genetics, 28, 476 – 490.

Johnson, B., Gardner, C.O. and Wrede, K.C. (1988). Application of an Optimization model to multi trait selection programmes, Journal of Crop Science, Vol 28 (5), 723-728.

Kamidi, R.E. (2001). Relative stability, Performance, and Superiority of Crop Genotypes Across Environments, Journal of Agricultural, Biological and Environmental Statistics, Volume 6, Number 4, 449-460

Kempthorne, O. and Nordskog, A.W. (1959). Restricted selection indices, Biometrics, 15:10-19.

Kempton, R.A. (1984). The Use of Biplots in Interpreting Variety by Environment Interactions, Journal of Agriculture Science,103, 123-135.

Perera, A L T., Thattil, R.O. and Leuke Bandara. (1989). use of index of selection in chillies , Sri Lankan Journal of Agricultural Science, vol 26, 01, 18-23.

SAS Institute Inc. (1990). SAS / STAT User's Guide, Ver.6 (4th ed), Cary, NC: Author.

Shukla, G.K. (1972). Some Statistical aspects of portioning genotype – environmental components of variance. Heredity 29: 237 – 245.

Smith, H. F. (1936). A discriminate function for plant selection. Annals of Eugenics 7, 240 – 250.

Wigesuriya A., Thattil, R.O. and Perera, A.L.T. (1993). Selection criteria used in colonel evaluation of Sugarcane, Tropical Agriculture Research ,vol 05,109-119.

Yates, F. and Cochran, W.G: (1938). The Analysis of Groups of Experiments, Journal of Agricultural science, 28, 556 – 580.