▼ Journal of Research (Science), Bahauddin Zakariya University, Multan, Pakistan.
Vol.14, No.1, June 2003, pp. 61-72 ISSN 1021-1012

ON QUANTIFICATION OF THE QUALITY AND ECONOMIC RETURN OF COTTON LINT: PRINCIPAL COMPONENT ANALYSIS AND RANK-SUM ANALYSIS

Muhammad Mutahir Iqbal[1] and G. M. Clarke[2]
[1]Department of Statistics, Bahauddin Zakariya University, Multan, Pakistan, [2]Institute of Mathematics and Statistics, Cornwallis Building, University of Kent at Canterbury, Canterbury, Kent CT2 7NF, UK.
email: [1]drmmi@brain.net.pk, [2]g_clarke@rss.org.uk

Abstract: Exploring the best laboratory or field conditions for maximizing the yield is usually the focus of studies. We think that improving economic value of the crop should be one of the most useful aspects of an experiment relating to an important cash-generating crop like cotton.

The economic value of cotton, like any commodity, is based upon quality. The quality of the crop is assessed subjectively by the experts of the field, which we believe needs quantification. Several responses are recorded for the crop, none of which serves the purpose itself. However there are responses, which define properties of the crop and could jointly be studied for the purpose.

Farmers will certainly be interested in improving the quality of the crop to get a higher price but will be more interested to maximize their net profit. So recommendations leading to higher production costs with less net profit will not attract farmers' and growers' attention. So economic return is given its due weight.

We have used four main classification criteria of cotton fiber-namely *length*, *strength, uniformity* and *fineness* - to define the ultimate quality and economic value of the crop. The criteria are bundled together into different indices using Principal Component Analysis and related techniques.

Keywords: Applied Statistics, economic indices, principal component analysis, rank-sum.

THE EXPERIMENT

The Central Cotton Research Institute (CCRI), Multan, Pakistan has been running a long-term rotation experiment on cotton and wheat crops since the cotton season of 1981. The effect of three dose levels of each of Nitrogen (N), Phosphorus (P_2O_5) and Potassium (K_2O) is of main interest. Crop records were not taken in 1981, and this can be regarded as a *settling-down* stage in the experiment. If all possible combinations of dose-level of three factors had been utilized it would be a 3^3 factorial experiment but only two different sets of 10 or less distinct treatment combinations were chosen from 27 and named as *phase-A* and *phase-B* each for cotton and wheat crops. We used data from phase-A of the experiment as it contains maximum, 10, distinct treatment combinations. The treatment combinations are presented in Table 1. Several responses related to N, P, K concentration, and uptake, dry-matter weight, growth, earliness, shedding, yield and fiber were observed. Four responses related to fiber namely *length*, *strength*, *uniformity* and *fineness* are used here.

Table 1: Treatment Combinations.

Treatment Combination No.	N	P_2O_5	K_2O
1	0	0	0
2	0	50	50
3	50	50	50
4	100	50	50
5	50	0	50
6	50	100	50
7	50	50	0
8	50	50	100
9	50	0	0
10	100	100	100

Unit of fertilizer application was kg ha^{-1}.

ECONOMIC VALUE OF THE CROP

Finding the best combination of treatments and their levels and climatic conditions for improving the economic value of the crop should be one of the most useful aspects of an experiment relating to an important cash-generating crop like cotton. Unfortunately none of the individual responses recorded for the experiment serves the purpose itself. However there are responses, which define properties of the crop, which could jointly be studied to define the economic value of the crop.

Following are the four main classification criteria of cotton fiber, which define the ultimate economic value of the crop.

1. ***LENGTH*:** The average length of cotton fiber after the ginning process, measured in inches.
2. ***STRENGTH*:** Power of the fiber to sustain the application of force (as applied in spinning) without breaking, measured in 000 lbs in^{-2}.
3. ***UNIFORMITY*:** The degree to which the fibers in a sample are uniform based on the ratio of mean length to the upper half mean length, given as a percentage.
4. ***FINENESS*:** Thickness of the fiber measured in different ways - the size of an individual cotton fiber taken in cross-section, or micro gram (μgm) per inch length.

NEED OF A SINGLE CRITERION

If we look at the four responses mentioned above we find at least two reasons that none of them can work as the sole classification criterion for cotton lint, the ultimate product of the cotton crop.

1. **EMPIRICAL CHARACTERISTICS**
   An observed response should be variable enough to be potentially explored for the underlying structure of variability. If it is not so then there is a very slim chance to find a model, which could help explain the situation. Referring to Fig. 1 we do not see much variability in any of the four classification criteria mentioned above.

It would be good if we could find a way to induce some extra variability by combining some or all of them in a meaningful way.

2. **UNIQUENESS**
   Potentially a different classification can be obtained if any of the criteria is used on its own, which leads us to search for a single criterion based on all the four properties.

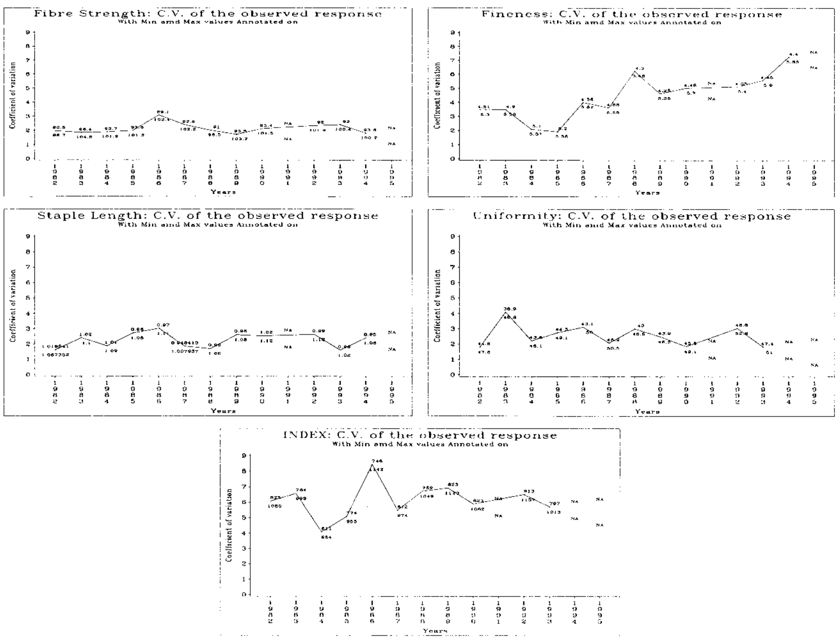

Fig. 1: Variability of the classification criterion: year 1982.

PRINCIPAL COMPONENTS OF THE ACTUAL DATA

The first thing in this regard that crosses one's mind is the scores of the first principal component. The first principal component for the four responses calculated using the covariance matrix explained 62.67 to 86.18% of the total variation for different years. It thus qualifies for further investigation to be used as a possible *index* of economic value of the crop. Initial results showed the following problems:

- The smallest score of the first principal component for various years varies from -7.877 to -3.733, and the corresponding largest score ranges from 2.795 to 5.74. Negative values in the score restrict the use of log-linear response function (LLRF) (Ali [1982], Iqbal [1999]),
- If the logarithmic transformation is omitted for the sake of fitting the model, the fitted surface is very flat for most of the N, P, K ranges with a negative fitted response which cannot be the case in a biometric experiment (Fig. 2),

That is why the score of the first principal component cannot be considered to use as an *economic index* any more.

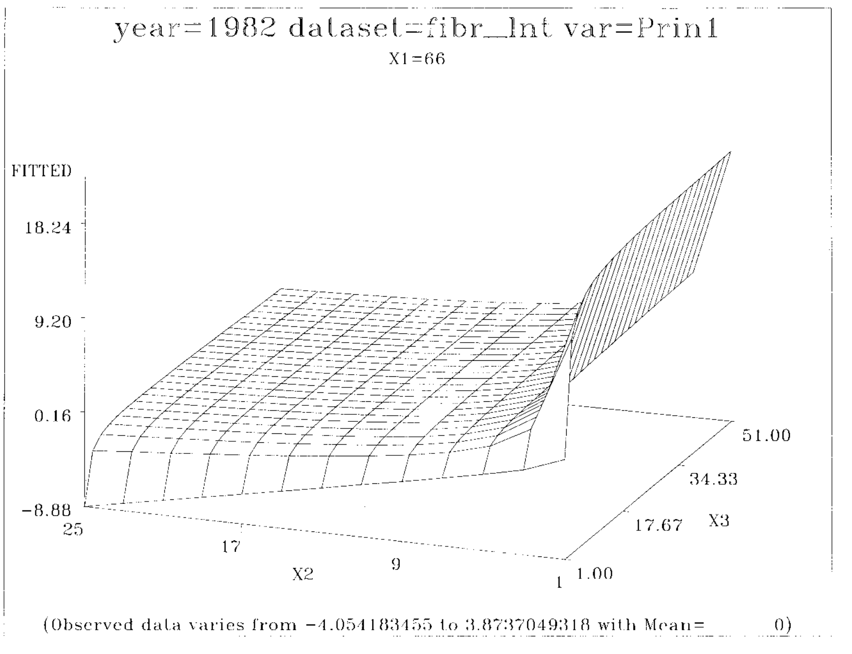

Fig. 2: Response surface for scores of First Principal Component.

ECONOMIC INDEX

After the *failure* of the principal component, our search of relevant printed and electronic literature, like Cotlook [1997] and Cotton-Council [1998], and consultation with some experts in the relevant fields could not provide us with a ready-made *index*. This was unfortunate, because we feel that there must be some quantitative rules in the mind of the people who set the market price for the crop. However, we must now devise an *index* of our own.

Referring to the four properties of cotton fiber, we noted that the first three properties should be as high as possible while the fourth should be as small as possible. The proposed *index* should also reflect this. We thus propose the following *index* to define economic value of cotton fiber.

$$index = \frac{length \times strength \times uniformity}{fineness}$$

To compare the variability pattern of the proposed *index* with that of individual responses we prepare a summarizing graphical layout (Fig. 1) of the coefficients of variation of all five criteria over the entire period of time. This tells us, for example about the year 1982 that:

- among observed responses:
  - *length* varies from 1.02 to 1.09 with a CV of 1.62
  - *strength* varies from 90.5 to 98.7 with a CV of 1.98
  - *uniformity* varies from 44.6 to 47.6 with a CV of 1.81
  - *fineness* varies from 4.51 to 5.3 with a CV of 3.47

whereas

- value of the proposed *index* varies from 825.34 to 1080.30 with a CV of 6.07, i.e. moderately variable and necessarily positive

moreover,

- logarithmic transformation is possible thus LLRF can be tried

PRINCIPAL COMPONENT ANALYSIS: REVISITED

Our urge to use the scores of first principal component, as a natural choice for the *economic index*, leads us to revisit it with transformed data instead. If log (*index*) is used, as in LLRF, then

$$\log(index) = \log\left(\frac{length \times strength \times uniformity}{fineness}\right)$$

or

$$\log(index) = \log(length) + \log(strength) + \log(uniformity) - \log(fineness)$$

That is the proposed *index* is actually a weighted average of the four characteristics over the logarithmic scale with weight 1, 1, 1 and -1 for *length*, *strength*, *uniformity* and *fineness*, respectively. In an attempt to make use of the two techniques simultaneously we can perform the principal component analysis on logarithmically transformed responses. Then if the first principal component explains sufficient variation the score of the first principal component could be used as an *economic index*. The score of the first principal component will then be:

$$Score\ 1 = k_1 \log(length) + k_2 \log(strength) + k_3 \log(uniformity) + k_4 \log(fineness)$$

where k_i*s* are the elements of the eigenvector corresponding to the first principal component. As this score will already be in a logarithmic scale the LLRF can be used no matter if some of the values turn up negative. In this sense the corresponding score of the first principal component will be nothing but another version of the proposed *index* with different weights. For this experiment, the proportion explained by the first principal component for different years varies from 0.303834 to 0.52660, which is not very exciting. So we decided not to take it any further for this analysis. This version of the proposed *index* provides an alternative to cover the situations where scores of the first principal component, based on the actual data, involve negative values.

ANALYSIS OF PROPOSED *INDEX*: THE RESPONSE SURFACE

In response surface methodology introduced by Box and Wilson [1951], the researcher is interested in finding a suitable approximating function for the purpose of predicting the future response. In biological and industrial factorial experiments interest lies in investigating to what extent a response is sensitive to a certain treatment/level combination. We used Ali's [1982] Log Linear Response Function (LLRF), which is an improved form of variance stabilizing log transformation of the response. LLRFs are sensitive to the origin as they cannot be fitted with (0, 0, 0) origin if any factor has a level at zero, which was the case in this experiment. A procedure for model and origin selection is discussed by Iqbal [1999], which was used for the proposed *index*. After having chosen the model and origin, we proceeded to explore the fitted response surface. For this we fitted the surface for a variety of feasible N, P, K combinations and different parts of the surfaces were locally checked to locate a stationary point in the region of the mean of the *index* computed from observed data. The stationary points defined by the combination of the lowest possible values of N,P,K were not very close to the mean values although studied independently for three different years. Figs. 3, 4 and 5 show that stationary points could not be located, but instead there are ridges.

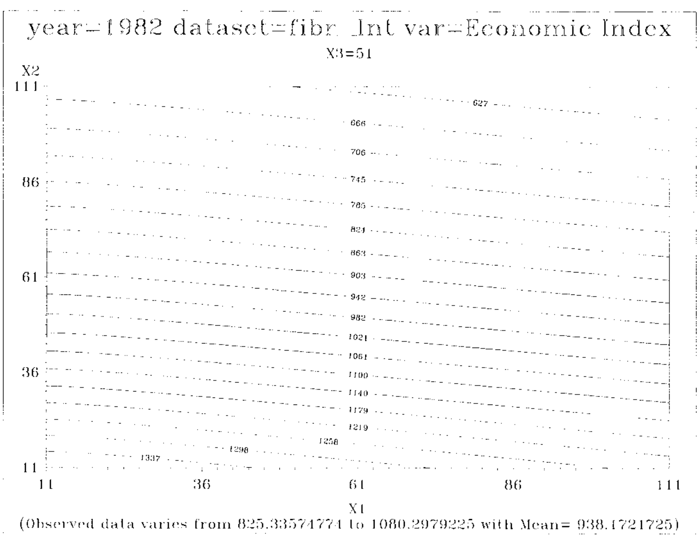

Fig. 3: Response surface for *economic Index*: year 1982.

ANALYSIS OF PROPOSED *INDEX*: WAYS TO GROUP TREATMENT COMBINATIONS

Having failed to locate a stationary point with expected response close to the mean value, we tried to group treatment combinations in two ways:

- ranking *economic index*,
- profile analysis.

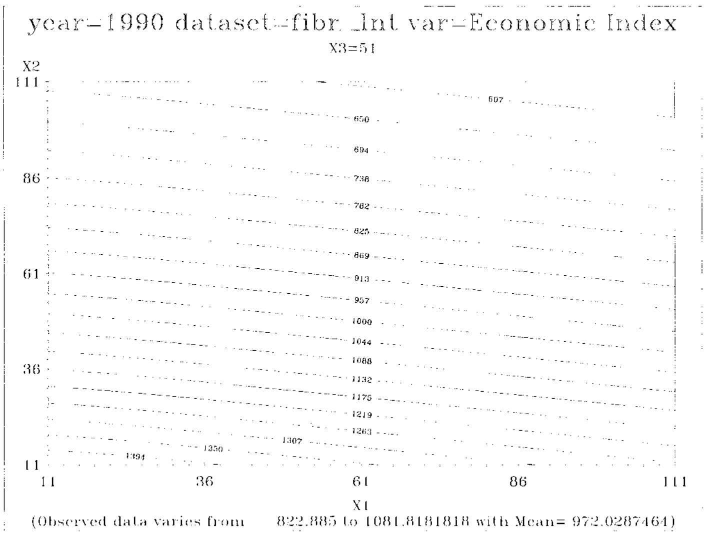

Fig. 4: Response surface for *economic index*: year 1990.

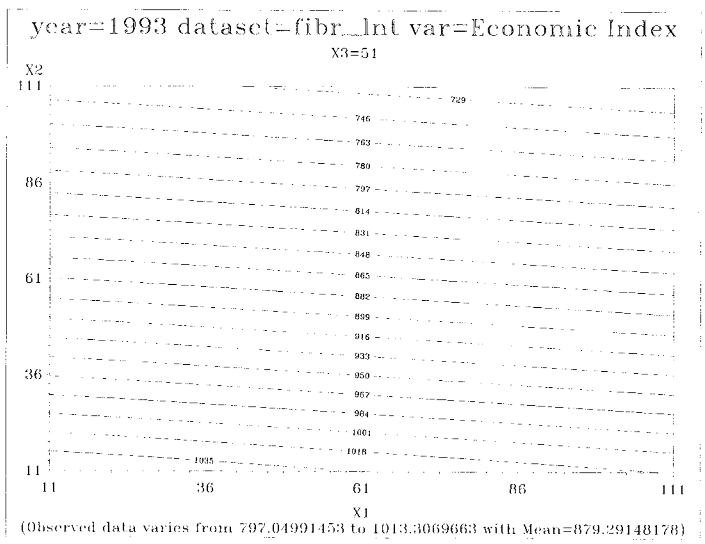

Fig. 5: Response surface for *economic index*: year 1993.

We need a method based on ranking the treatment combinations, in the hope of locating those, which generally gave higher values of the *index*. Table 2 shows the sums of the ranks for each of the ten treatments for the eleven years when there were complete data available. The treatment having the lowest value of the *index* was given rank 1 and the highest 30 (each treatment was replicated three times in each set of data; so, for example, if the same treatment in the three blocks gave the three lowest values of the *index* the sum of its ranks would be 1 + 2 + 3 = 6, while if it gave the three highest values in that set of data the sum would be 28 + 29 + 30 = 87). The last column is the sum of the entries for each treatment over the 11 years.

It is obvious from Table 2 that the rankings show considerable variation and much less tendency towards any stability. It was hoped that the 11-year total in the final column of Table 2 might provide information. A quick, rough and ready, method of looking at this is to carry out a completely randomized design type analysis, taking years as replicates, for the 110 rank sums in the body of the table. An F-test of the null hypothesis that there is no difference between the treatments showed a result that did not approach significance, but this is not really surprising when we look carefully at the behavior of individual treatments. Treatment 8, for example, goes from 74 to 16 in 1992 and 1993.

Table 2: Summary of the ranks of economic index.

N	P	K	Treatment Combination No.	1982	1983	1984	1985	1986	1987	1988	1989	1990	1992	1993	All
0	0	0	1	33	32	20	41	36	54	64	20	36	16	59	411
0	50	50	2	49	44	60	50	63	47	30	40	49	37	60	529
50	50	50	3	69	33	22	57	39	33	78	53	56	57	64	561
100	50	50	4	42	60	35	68	76	53	33	65	27	29	53	541
50	0	50	5	45	37	39	24	46	61	43	29	66	65	28	483
50	100	50	6	38	45	58	36	36	40	50	65	36	57	58	519
50	50	0	7	34	45	60	22	40	36	54	46	34	55	45	471
50	50	100	8	39	64	71	53	21	39	29	41	34	74	16	481
50	0	0	9	54	52	33	48	29	52	59	76	56	48	36	543
100	100	100	10	62	53	67	66	79	50	25	30	71	27	46	576

We can look at the last column of Table 2 to see if it makes any sense, and in particular we might consider possible hypotheses suggested by the leaf N, P and K levels. In studying leaf N see Iqbal (2003), for example, we found that treatments 1 and 2, with N at level 0, were consistently lowest, while 4 and 10, with N at level 100, were consistently highest; the N 50 treatments were in between. The two high-N treatments are reasonably consistent over most years although there are three occasions (1984, 1989, 1990) when they did not agree at all. The two low-N treatments show less good agreement and medium-N treatment combinations indicate no consistency at all. Fig. 6 uses three different line styles to denote three N-levels. From this, and from Table 2, we can reasonably say that treatment 1 (N0, P0, K0) gives poor results. Over the whole experiment the N0 treatments never did best, although treatment 2 was not too bad overall.

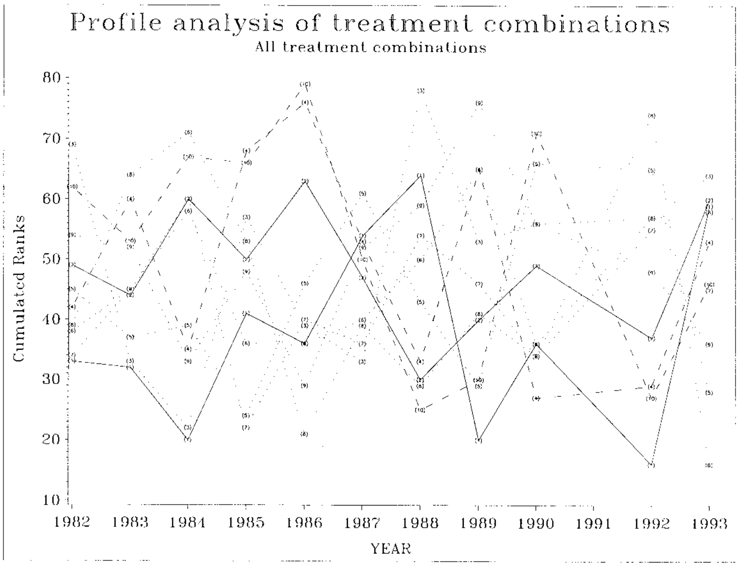

Fig. 6: Profile analysis of cumulated ranks of *economic index*: Levels of N.

PROFIT (RETURN) *INDEX*

The *economic index* discussed in the previous section mainly takes *quality* of the crop into consideration. It may be used to assess market value of the crop. Farmers will certainly be interested in improving the quality of the crop to get a higher price but will be more interested to maximize their net profit. So recommendations leading to higher production costs with less net profit will not attract farmers' and growers' attention. It is therefore necessary to take into account all the costs involved, to modify the *economic index*, which will help in recommending treatment combinations leading to higher profits. The economic value of the crop depends upon the four classifying characteristics mentioned earlier. It depends also upon the total crop yield for a particular treatment and upon the cost involved in applying that treatment. These costs are of two types: *variable*, depending on the costs of the fertilizers applied, and *fixed*, such as the costs of irrigation, pesticides, routine cultivation, labor and administrative costs. (There may also be small contributions to variable costs if some treatments affect the plants or soil in different ways from others, e.g. require extra weeding, or come to harvest at slightly different times from most of the others, or produce substantially more crop which requires extra labor or transport.)

If we sum up all those costs in one variable *cost* then it could easily be linked with the *economic index* where cost can be defined as a weighted average of various costs involved.

$$\cos t = \sum w_i c_i \quad where \sum w_i = 1.$$

The cash return will be inversely proportional to the cost involved so the *profit index* can be defined as

$$profit\ index = \frac{economic\,index}{\cos t}$$

Actual costs of purchasing the amount of fertilizers required for each experimental plot could be calculated from the price information, but we do not have information on all the other costs, especially the fixed ones. We know the amount of fertilizers applied but not their exact prices. The three fertilizers are not manufactured and sold in their generic form but in different compositions. Some price information for several compositions were used, and suggested an approximate price weighting of 1:1:1 for the same amount of the three fertilizers. Thus the weights come to $w_i = 1/3$ for i = 1, 2, 3. If the amount of fertilizer applied is used as c_i for i =1, 2,3 then calculation of profit index for first treatment combination (where N = P = K = 0) will not be possible. Moreover in treatment like N = P = K = 100 its units will highly influence the *index*. To overcome such extreme problems we suggest to use c_i =1 for no fertilizer application and c_i =2 for 50 kg ha^{-1} and c_i =3 for 100 kg ha^{-1} applications. This will help keep the cost within the reasonable range of 1 and 3. Even then the values and variability structure of the cost variable are very different from others, which is why cost will be the most dominant factor in the *profit index*. Standardizing all the variables will rectify the problem but we have to choose a mean other than zero as variables standardized with zero mean and unit variance are bound to get negative values which will not be acceptable when we take logarithms to try principal component analysis. We should thus standardize them aiming for a mean of 5 as Finney [1971] which provides necessarily positive values for all the variables. With this specific definition, we will analyze it in the same way as we analyzed *economic index* earlier.

PRINCIPAL COMPONENT ANALYSIS: COST INVOLVED

If we forget the *profit index* for the time being and consider cost as the fifth classifying characteristic, the principal component analysis could be attempted on original as well logarithmic scale for completeness.

The results of principal components, because of variability structure calculated from correlation matrix, were again not exciting. For the original scale, first principal component was able to explain 28.5% to 45.3% of the total variation whereas on logarithmic scale this proportion varied from 28.2% to 42.0%.

ANALYSIS OF *PROFIT INDEX*: THE RESPONSE SURFACE

The response surface of *profit index* was analyzed. Stationary points could not get located again. This once again leads us to do the alternative analysis.

ANALYSIS OF *PROFIT INDEX*: WAYS TO GROUP TREATMENT COMBINATIONS

Ranking of the *profit index* and profile analysis was done as explained earlier. Summary of the ranks is given in Table 3.

Table 3: Summary of the ranks of profit index

N	P	K	Treatment Combination No.	1982	1983	1984	1985	1986	1987	1988	1989	1990	1992	1993	All
0	0	0	1	68	54	59	77	65	77	79	58	79	46	67	729
0	50	50	2	54	45	70	61	68	49	41	48	52	45	63	596
50	50	50	3	60	33	18	48	26	39	73	47	42	56	67	509
100	50	50	4	32	56	16	61	63	41	28	57	17	21	38	430
50	0	50	5	51	44	46	29	49	59	59	36	69	64	54	560
50	100	50	6	26	37	42	26	32	33	33	52	21	34	35	371
50	50	0	7	43	51	66	34	56	43	55	48	48	60	49	553
50	50	100	8	26	51	53	41	10	30	20	30	28	71	10	370
50	0	0	9	69	64	69	61	46	64	66	79	74	57	65	714
100	100	100	10	36	30	26	27	50	30	11	10	35	11	17	283

Contrary to the results for *economic index*, the quick method of complete randomized design type analysis showed a highly significant result for treatment differences. The *profit index* for all the years shows a pattern in the cumulated ranks. It reflects an inverse relation with the cost factor (numerator of the w_i used in cost). This is not an unexpected relation but it should not be that obvious. Profile analysis (not shown) was done for the profit analysis but no clear patterns could be located for any treatment over time. Complete information of fertilizers' price structure and other costs involved will be essential for correct definition of w_i*s* and will essentially improve the situation and a clearer pattern should be obtained.

CONCLUSIONS

Three indices to quantify the quality of cotton lint were introduced and analyzed in this paper.

The scores of the first principal component of *length*, *strength*, *uniformity* and *fineness* were used as a first *index*. Initially, some of the negative values of the scores restricted the use of LLRF to estimate the response surface. The response surface became too flat if we compromise on logarithmic transformation of the scores.

An alternative *index* based on higher values for *length*, *strength* and *uniformity* and smaller values of *fineness* was proposed as $index = \dfrac{length \times strength \times uniformity}{fineness}$. We then used the scores of first principal component of log(*index*), as in LLRF. These scores were already in a logarithmic scale so the LLRF could be used no matter if some of the values turn up negative. Both the approaches were similar but used different weights i.e.

$$Score\ 1 = k_1 \log(length) + k_2 \log(strength) + k_3 \log(uniformity) + k_4 \log(fineness)$$

where $k_i s$ were the elements of the eigenvector corresponding to the first principal component. Again no stationary point could be located close to the mean values in the surface.

The *economic index* mainly took *quality* of the crop into consideration, which helps assessing market value of the crop. Farmers and growers are naturally more interested to maximize their net profit. So the cash return, which is inversely proportional to the cost involved, was introduced. The *profit index* was defined as $profit\ \ index = \dfrac{economic\, index}{\cos t}$. Analytical results were similar to those of the first two.

We then used an alternate analytical method of grouping treatment combinations in two ways: ranking indices and profile analysis. For the first, we needed a method based on ranking the treatment combinations to locate those, which generally gave higher values of the *index*. To look at this we carried out a completely randomized design type analysis, taking years as replicates. We could find some significant patterns this way. For the second, we found similar patterns using profile graphs.

References

Ali, A. (**1982**) “Interpretation of Multivariate Data: Comparison of Several Methods of Interpreting Multivariate Data from a Series of Fruit Nutrition Experiments”, PhD Thesis, School of Mathematical and Physical Sciences, University of Sussex, UK.

Box, G.E.P. and Wilson, K.B. (**1951**) “On the experimental attainment of optimum conditions”, *Journal of Royal Statistical Society, Series-B* XIII, 1-45.

Cotlook (**2003**) “Cotton Outlook: Gateway to the World of Cotton; Glossary of Terms”, http://www.cotlook.com/information/glossary_of_terms.php

Cotton_Council (**2003**) “Cotton Council International: Buyers’ Guide, Official Cotton Standards for 2001”. http://www.ams.usda.gov/standards/cvrstd4.pdf

Finney, D. (**1971**) “Probit Analysis”, 3rd ed., Cambridge University Press.

Iqbal, M.M. (**1999**) “Analysis of Long-Term Experiment on Cotton using a Blend of Theoretical and New Graphical Methods to Study Treatment Effects Over Time”, PhD Thesis, Institute of Mathematics and Statistics, University of Kent, UK.

Iqbal, M.M. (**2003**) “Profile Analysis in Agriculture” *Journal of Pure and Applied Sciences*, Islamia University, Bahawalpure, (accepted paper).