Listy Biometryczne - Biometrical Letters
Vol. 33 (1996), No. 2, 79-88

A method of yield component analysis

Janusz Gołaszewski

Department of Plant Breeding and Seed Production,
Olsztyn University of Agriculture and Technology
Plac Łódzki 13, 10-724 Olsztyn

SUMMARY

The paper presents a method of yield component analysis, developed by Eaton, called two-dimensional partitioning method (TDP). The method combines multiple regression and ANOVA and enables concise tabular presentation and simple interpretation of the distribution of traits in one direction and the sources of variance, according to the ANOVA model, in the second direction. Computational procedure and conclusions from TDP table on the basis of the data from breeding field trial with 10 lines of faba bean cultivated in normal conditions of fertilization have been shown.

KEY WORDS: yield component analysis, two-dimensional partitioning, regression analysis, ANOVA, faba bean.

1. Introduction

The main research activities in crop breeding and cultural practices are focused on studying the changes in yield and its components affected by various experimental treatments. From the beginning of concisely planned and conducted agricultural trials, different statistical approaches to solve the problem of variation and interrelationship between traits inducing crop performance have been improved. Among the most commonly used statistical methods: analysis of variance, simple correlation and linear regression, partial correlation, multiple correlation and regression, path-coefficient analysis, multiple stepwise regression, factor and principal component analysis and cluster analysis have been used. Little or no application in yield component analysis has had multivariate procedures like canonical correlation, discriminant function analysis, multivariate analysis of variance, etc. Abundance of papers related to these topics makes listing of all attempts of yield component analyses impossible.

Some new interesting approaches were based on a modification of computational procedures of above-mentioned methods (Jolliffe, 1982; Eaton, 1986; Idźkowska *et al.* 1993; Sparnaaij and Bos, 1993) or giving more readable graphical form of the presented results (Gołaszewski *et al.* 1993; Piech and Stankowski, 1990).

Fraser and Eaton (1983) reviewed different methods of yield component analysis (without the computational procedures), their advantages and disadvantages, and gave examples of their application. Individually applied statistical methods in yield component analysis give incomplete information about entire set of components affecting the yield. For example, ANOVA alone gives no information on the nature of interdependence, and multiple regression alone does not take into account the problems of multicolinearity and decomposition of the total variation according to sources of variation (e.g. blocks, varieties, rates of a fertilizer). These authors concluded that many statistical techniques related to yield component analysis are based upon the multiple regression and are in a sense unified by its theoretical bases in the general linear model. This makes good prognosis for further theoretical developments of different approaches to the problem of yield component analysis.

The objective of this paper was to present an approach to yield component analysis called two-dimensional partitioning (TDP) in which total variation in yield is partitioned by the regression procedure into increments of variation of successive components (first direction), and due to the sources of variation like treatments, blocks, experimental error, etc. (second direction) by ANOVA procedure. The data from a breeding experiment with faba bean were analysed using this method.

2. Method

Two-dimensional partitioning method (TDP) was developed by Eaton in 1986, but practical use of the method has been limited mainly to the Eaton and co-workers' studies. Recently, two Dutch breeders, Bos and Sparnaaij (1993), have revealed another advantage of the main assumption of the method. The authors have used Eaton's idea to predict values for traits in a sequential order and finally for yield in a breeding program with heterosis.

The main computational steps in TDP may be summarized as follows:

Step 1. Select and measure primary variables on a common basis, such as a per plant basis. The variables should be recorded in a sequential order of their appearance at the consecutive stages of plant growth (but to get orthogonalized variables this restriction is not obligatory). The construction of ratios of those variates in chronological order (for instance: number of nods/height of plant, number of pods/node, number of seeds/pod, etc.) and then transformation of the ratios to logarithms, proposed by Eaton, can be omitted.

Step 2. Make the transformation yielding uncorrelated variables, using e.g. Gram-Schmidt orthogonalization process (Winer, 1971).

Let us consider a set of $p = 4$ independent variables X_1, X_2, X_3, X_4 and a dependent variable Y. The extension to arbitrary p is straightforward. Let us calculate the following prediction equations in which the weights are defined by the least-squares criterion:

$$\begin{aligned} \hat{X}_2 &= b_{20} + b_{21}X_1 \\ \hat{X}_3 &= b_{30} + b_{31}X_1 + b_{32}X_2 \\ \hat{X}_4 &= b_{40} + b_{41}X_1 + b_{42}X_2 + b_{43}X_3 \end{aligned} \tag{1}$$

A new set of variables is obtained:

$$\begin{aligned} X_1 &= X_1 \\ X_{2.1} &= X_2 - \hat{X}_2 \\ X_{3.12} &= X_3 - \hat{X}_3 \\ X_{4.123} &= X_4 - \hat{X}_4 \end{aligned} \tag{2}$$

The new set of uncorrelated variables in (2) is equivalent to the set X_1, X_2, X_3, X_4, in the sense that:

- variable $X_{2.1}$ represents that part of X_2 from which X_1 has been partialled out, thus X_2 is completely predictable from X_1 and $X_{2.1}$,

- variable $X_{3.12}$ represents that part of X_3 from which X_1 and X_2 have been partialled out, thus X_3 is completely predictable from X_1, $X_{2.1}$ and $X_{3.12}$,

- variable $X_{4.123}$ represents the part of X_4 from which X_1, X_2 and X_3 have been partialled out, thus the X_4 is completely predictable from X_1, $X_{2.1}$,$X_{3.12}$ and $X_{4.123}$.

The variables in the set (2) are uncorrelated. For example, $r_{X_1X_{2.1}} = 0$, because $X_{2.1}$ has all information that is linear function of X_1 partialled out and similarly, $r_{X_1X_{3.12}} = 0$, because the linear information on X_1 (and also X_2) has been partialled out of $X_{3.12}$.

The transformation defined by (2), known as the Gram-Schmidt orthogonalization process, expressed in terms of matrix operations is as follows (Winer 1971). Let $\mathbf{M}$ will be the $(p \times p)$ covariance matrix for original variables $X_1, ..., X_p$; $\mathbf{X}$ will be the $(n \times p)$ matrix of observations on original variables scaled so that $\frac{1}{n-1}\mathbf{X}'\mathbf{X} = \mathbf{M}$; $\mathbf{Z}$ will be the $(n \times p)$ matrix of observation in terms of the transformed variables $Z_1, ..., Z_p$ (where $Z_1 = X_1$, $Z_2 = X_{2.1}$, $Z_3 = X_{3.12}$, and so on). According to the Dwyer algorithm, the matrices $\mathbf{T}$ and $\mathbf{U}$ are defined as:

$$\underset{p,p}{\mathbf{T}} = \text{lower triangular matrix such that } \mathbf{M} = \mathbf{T}\mathbf{T}';$$

$$\underset{p,p}{\mathbf{U}} = \text{lower triangular matrix such that } \mathbf{U} = \mathbf{T}^{-1}.$$

Hence

$$\mathbf{UMU'} = \mathbf{U(TT')U'} = \mathbf{I}.$$

Under the Gram-Schmidt orthogonalization process,

$$\underset{n,p}{\mathbf{Z}} = \underset{n,p}{\mathbf{X}}\,\underset{p,p}{\mathbf{U'}} \tag{3}$$

defines the transformed set of variables $Z_1, ..., Z_p$. The matrix $\mathbf{U'}$ is called the matrix of the transformation. The covariance matrix for the transformed set is

$$\frac{1}{n-1}\mathbf{Z'Z} = \frac{1}{n}\mathbf{UX'XU'} = \mathbf{UMU'} = \mathbf{I}.$$

Except of the scaling factor the transformation defined by (3) is equivalent to the transformation defined by (2). The transformation forms the orthogonal set of variables.

Step 3. Measure incremental contribution of successive orthogonal variates.

Let us consider a prediction system in which Y is to be predicted from variables in the set (2). For the standardized variables the linear prediction system has the form:

$$\hat{Y}^* = b_1^* X_1^* + b_{2.1}^* X_{2.1}^* + b_{3.12}^* X_{3.12}^* + b_{4.123}^* X_{4.123}^*. \tag{4}$$

The vector of weights $\mathbf{b}^*$ is obtained from the solution to the normal equations:

$$\mathbf{Rb}^* = \mathbf{r}$$

or

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_1^* \\ b_{2.1}^* \\ b_{3.12}^* \\ b_{4.123}^* \end{bmatrix} = \begin{bmatrix} r_{Y(1)} \\ r_{Y(2.1)} \\ r_{Y(3.12)} \\ r_{Y(4.123)} \end{bmatrix},$$

hence

$$\begin{bmatrix} b_1^* \\ b_{2.1}^* \\ b_{3.12}^* \\ b_{4.123}^* \end{bmatrix} = \begin{bmatrix} r_{Y(1)} \\ r_{Y(2.1)} \\ r_{Y(3.12)} \\ r_{Y(4.123)} \end{bmatrix}.$$

The correlation between Y and $X_{2.1}$, denoted $r_{Y(2.1)}$, is a semipartial correlation and measures the correlation between Y and the part of X_2 from which X_1 has been partialled out; $r_{Y(3.12)}$ is a semipartial correlation between Y and $X_{3.12}$ from which X_1, X_2 have been partialled out and $r_{Y(4.123)}$ is a semipartial correlation between Y and the part of $X_{4.123}$ from which X_1, X_2 and X_3 have been partialled out.

The square of the multiple correlation associated with the prediction system given in (4) is:

$$r^2_{Y_{1234}} = r^2_{Y_1} + r^2_{Y_{2.1}} + r^2_{Y_{3.12}} + r^2_{Y_{4.123}} \tag{5}$$

or

$$r^2_{Y(1234)} = 1 - r^2_{Y(Y_{.1234})},$$

where $r^2_{Y(Y_{.1234})}$ is the part of total variation which is not predictable and $Y_{.1234}$ may be calculated in a similar way as $X_{2.1}$, $X_{3.12}$ and $X_{4.123}$ e.i. $Y_{.1234} = Y - \hat{Y}$ where

$$\hat{Y} = b_{50} + b_{51}X_1 + b_{52}X_2 + b_{53}X_3 + b_{54}X_4.$$

The coefficients of determination calculated as a sequential contribution into Y can be computed from:

$$\begin{aligned} r^2_{Y(1)} &= r^2_{Y(1)}, \\ r^2_{Y(12)} &= r^2_{y(1)} + r^2_{Y(2.1)}, \\ r^2_{Y(123)} &= r^2_{Y(1)} + r^2_{Y(2.1)} + r^2_{Y(3.12)}, \\ r^2_{Y(1234)} &= r^2_{Y(1)} + r^2_{Y(2.1)} + r^2_{Y(3.12)} + r^2_{Y(4.123)} \end{aligned} \tag{6}$$

Step 4. Calculate the ANOVA for the standardized variables, residual $Y_{.1234}$ and dependent variable Y according to the experimental design. Scaling to units of Y before ANOVA should be made to balance the sum of squares for components and Y. It may be accomplished by multiplying each orthogonal variate by its regression coefficient.

Step 5. Partition the sums of squares and cross-products of the appropriate sources of variation for each variable as a percentage of total variation of Y or of real values of sums of squares (after scaling the variates). The cross-products are all possible interactions between treatments and component pairs which can affect Y in different ways, so that the value of cross-product for any source of variation may be zero, negative or positive. For example, a treatment may increase one component and decrease another and although both components may have positive effect their interaction may be negative, thus cross-product will be also negative. The sum of cross-products for all analysed sources of variation is equal to zero.

3. Example

In the field breeding trial conducted in 1994, eight inbred lines (Polish and foreign) of faba bean and two standard cultivars (t=10) cultivated in the conditions of natural fertilization were tested. The trial was designed as randomized complete block design with r=3 replications. Height of plant (X_1), number of nodes with pods (X_2),

number of pods (X_3), number of seeds (X_4) and yield (Y) on a per plant basis have been measured. The mean values of 20 characterised individuals from each plot were recorded. The size of the plot was three rows, 1.5 m long, 0.3 m apart. The results of the computational procedure are presented below.

The original and transformed data are presented in Table 1. The column ($Y_{.1234}$) contains the residuals from the regression of Y on X_1, X_2, X_3 and X_4.

Table 1. Original and transformed data of faba bean traits

		Original data					Transformed data				
t	r	X_1	X_2	X_3	X_4	Y	X_1	$X_{2.1}$	$X_{3.12}$	$X_{4.123}$	$Y_{.1234}$
1	1	94.4	5.0	11.5	23.2	9.05	94.4	0.58	1.19	-0.15	-0.34
1	2	78.5	3.6	7.0	17.7	6.72	78.5	-0.10	-0.42	1.38	-0.31
1	3	93.4	4.3	10.2	24.1	9.74	93.4	-0.07	1.09	1.87	0.56
2	1	83.3	5.8	11.5	18.5	8.65	83.3	1.88	0.25	-2.25	-0.21
2	2	83.3	5.2	10.5	20.8	9.67	83.3	1.28	0.25	0.73	0.47
2	3	67.0	4.2	9.4	20.4	9.75	67.0	1.02	1.39	2.19	1.31
3	1	69.2	2.9	6.9	16.8	6.26	69.2	-0.38	0.97	0.39	0.18
3	2	83.9	3.6	8.6	20.5	7.81	83.9	-0.34	0.99	0.82	0.23
3	3	84.9	3.4	7.8	17.2	6.54	84.9	-0.59	0.48	-1.55	0.25
4	1	78.0	3.9	7.8	20.1	6.80	78.0	0.22	-0.10	2.97	-1.30
4	2	81.5	3.7	8.1	19.7	6.60	81.5	-0.13	0.41	1.32	-0.97
4	3	86.9	3.8	8.7	20.9	7.67	86.9	-0.28	0.65	1.08	-0.21
5	1	66.0	3.3	5.5	13.6	6.01	66.0	0.16	-0.98	0.69	0.25
5	2	78.6	3.8	6.9	15.4	7.71	78.6	0.10	-0.86	-0.40	1.26
5	3	72.1	3.2	5.3	12.2	5.37	72.1	-0.21	-1.23	-1.19	0.32
6	1	80.7	3.4	9.2	17.6	6.34	80.7	-0.40	2.03	-3.14	0.12
6	2	83.4	3.1	7.7	21.8	8.53	83.4	-0.82	0.94	2.85	0.86
6	3	84.2	3.0	7.7	19.2	7.46	84.2	-0.96	1.07	-0.02	0.88
7	1	75.3	3.4	7.0	13.3	4.94	75.3	-0.16	0.02	-3.04	-0.31
7	2	78.7	3.3	6.8	16.6	5.91	78.7	-0.41	-0.13	0.07	-0.39
7	3	70.8	2.8	5.2	14.1	4.45	70.8	-0.55	-0.62	0.31	-0.86
8	1	82.7	3.8	5.5	10.0	4.89	82.7	-0.09	-2.40	-3.80	0.19
8	2	80.2	3.6	6.0	13.5	6.24	80.2	-0.18	-1.48	-1.26	0.56
8	3	89.2	4.3	6.7	14.4	6.98	89.2	0.12	-2.26	-1.29	0.48
9	1	79.1	3.5	8.4	16.2	4.86	79.1	-0.23	1.12	-2.80	-1.18
9	2	72.2	3.8	8.4	19.3	6.90	72.2	0.38	0.87	1.57	-0.77
9	3	82.3	4.0	9.5	16.4	6.19	82.3	0.13	1.28	-3.97	-0.21
10	1	92.0	4.4	7.9	20.1	7.82	92.0	0.09	-1.33	2.20	-0.56
10	2	81.9	3.7	5.6	15.8	6.34	81.9	-0.15	-2.11	1.74	-0.35
10	3	83.8	4.0	7.2	19.2	7.94	83.8	0.06	-1.07	2.68	0.04
Mean		80.4	3.8	7.8	17.6	7.00	80.6	0	0	0	0

Prediction equations are as follows:

$$\hat{X}_2 = 0.161 + 0.04507X_1$$
$$\hat{X}_3 = -1.339 + 0.0354X_1 + 1.6626X_2$$
$$\hat{X}_4 = 2.202 + 0.10629X_1 - 1.7831X_2 + 1.7420X_3$$

The entries of transformed data in the column $X_{2.1}$ are calculated as residuals from the prediction equation: $X_{2.1} = X_2 - \hat{X}_2$; for example, the first entry of $X_{2.1}$ is $5.0 - (0.161 + 0.04507 \cdot 94.4) = 0.58$. Identical computational procedure is used for the columns $X_{3.12}$ and $X_{4.123}$.

The results of the multiple correlation analysis of original and transformed data are presented in Table 2.

Table 2. Coefficients of simple and semipartial correlation.

Variable	X_1	X_2	X_3	X_4	Y
original					
X_1	1				0.432
X_2	0.484	1			0.642
X_3	0.457	0.713	1		0.708
X_4	0.466	0.395	0.749	1	0.809
transformed					
X_1	1	0.484	0.457	0.466	0.432
$X_{2.1}$	0	0.875	0.562	0.194	0.494
$X_{3.12}$	0	0	0.689	0.619	0.338
$X_{4.123}$	0	0	0	0.602	0.503
$X_{.1234}$	0	0	0	0	0.449

Thus the summarized contribution of the sequentially recorded traits into total variation of faba been yield, according to (6), gives:

$$r^2_{Y(1)} = 0.187$$
$$r^2_{Y(12)} = 0.431$$
$$r^2_{Y(123)} = 0.545$$
$$r^2_{Y(1234)} = 0.798$$

Unpredictable part of the total variation is $r^2_{Y(Y.1234)} = 1 - r^2_{Y(1234)} = 0.202$.

Sums of squares from regression analysis and ANOVA of all the standardized studied traits and residuals ($Y_{.1234}$) are shown in Table 3. The results after proportional partitioning are presented in two-dimensional Table 4.

Table 3. Sums of squares from ANOVA of the transformed data (not scaled)

Sources	df	X_1	$X_{2.1}$	$X_{3.12}$	$X_{4.123}$	$X_{.1234}$	Y
from ANOVA							
Lines	9	696.82	8.514**	36.341**	52.72	8.413**	42.42**
Blocks	2	11.64	0.478*	0.361	15.72	1.673*	2.95
Error	18	790.44	0.981	4.757	50.47	2.551	17.23
Total	29	1498.90	9.974	41.459	118.90	12.637	62.60
SS Regres.[1)]	1	11.69*	15.300**	7.143	15.83**	12.637**	
% of total SS of Y		18.7*	24.4**	11.4	25.3	20.2	100

[1)] SS Regres. - sums of squares for regression from simple regression analyses of the transformed variables with yield (in the case of scaling of transformed variables to units of Y the total SS from ANOVA and SS Regres. will be the same)

*, ** - significant at $P = 0.05$ or $P = 0.01$, respectively

Table 4. Two-dimensional partitioning of yield variation as a percentage of the total sum of squares for yield.

Sources	Height of plants	No of nodes with pods	No of pods/ plant	No of seeds/ plant	Residual	Cross-products	Seed yield/ plant
Lines	8.7	20.9**	10.0**	11.2	13.4**	3.6	67.8**
Blocks	0.1	1.2	0.1	3.3	2.7*	-2.7	4.7
Error	9.8	2.4	1.3	10.7	4.1	-0.8	27.5
Total	18.7*	24.4**	11.4	25.3**	20.2*		100.0

*, ** - significant at $P = 0.05$ or $P = 0.01$, respectively

4. Interpretation of the results

TDP analysis (Table 4) indicated that variation in seed yield of faba bean cultivated in natural conditions of fertilization in 1994 was mainly accounted by variation of lines (67.8%). The main contribution into lines' yield variation had the number of nodes with pods (20.9%) while the rest of traits had similar influence – about 10%. Taking into consideration the total sums of squares for the traits it was shown that total yield variation was primarily derived from number of nodes with pods (24.4%) and number of seeds per plant (25.3%). A high contribution of unexplained variation for sums of squares for lines (13.4%) and total (20.2%) might point to an incomplete set of predicted characters to yield component analysis.

Discrepancies in estimated values of correlation coefficients for original and transformed variables point to different conclusions according to which data were analysed

(last column in Table 2). The "pure" relation of number of pods per plant with yield was not too strong. Low value of semipartial correlation with yield (0.338), relatively low percent (11.4%) for total contribution into yield variation and significant lines' variation for number of pods per plant has proved high variability of this character in faba been lines and weak relation with yield although the simple correlation coefficient was high and significant.

Acknowledgments

I would like to acknowledge Prof. G.W. Eaton from the Department of Plant Sciences of British Columbia University (Canada) and Dr Jadwiga Milewska from the Department of Plant Breeding and Seed Production of Olsztyn University of Agriculture and Technology for valuable discussions during preparing this paper.

References

Bos I., Sparnaaij L.D. (1993). Component analysis of complex characters in plant breeding: II. The pursuit of heterosis. *Euphytica* **70**, 237-245.

Eaton G.W., Bowen P.A., Jolliffe P.A. (1986). Two-dimensional partitioning of yield variation. *HortScience* **21**, 1052-1053.

Fraser J., Eaton G.W. (1983). Application of yield component analysis to crop research. *Field Crop Abstracts* **36**, 787-797.

Idźkowska M., Koczowska I, Gołaszewski J., Grabowski S. (1993). Analiza współczynników ścieżek u żyta ozimego (Secale cereale L.). Cz.I i II. *Acta Acad. Agricult. Tech. Olst., Agricult.* **56**, 17-30.

Jolliffe P.A., Eaton G.W., Lovett Doust J. (1982). Sequential analysis of plant growth. *The New Phytologist* **92**, 287-296.

Gołaszewski J., Koczowska I., Korona A., Idźkowska M. (1993). Path coefficients method in estimation of relationship between some characters of winter rye and spring triticale. *Zesz. Nauk. AR Wrocław, Rolnictwo LVIII* **223**, 158-163.

Piech M., Stankowski S. (1990). Wpływ terminu siewu i poziomu nawożenia azotem na plon i jakość ziarna dwóch odmian pszenżyta na glebie lekkiej. Cz.I. Plon ziarna i komponenty plonu. *Biul. Inst. Hod. i Aklim. Rośl.*, **159**, 15-25.

Sparnaaij L.D., Bos I. (1993). Component analysis of complex characters in plant breeding. I. Proposed method for quantifying the relative contribution of individual components to variation of the complex character. *Euphytica* **70**, 225-235.

Winer B.J. (1971). *Statistical principles in experimental design.* 2nd Ed. McGraw-Hill, New York, 126-133.

Received 15 September 1995; revised 19 September 1996

Jedna z metod analizy komponentów plonu

STRESZCZENIE

W pracy przedstawiono pewną procedurę, zwaną TDP (two-dimensional partitioning), umożliwiającą ocenę komponentów plonu na podstawie dwukierunkowego podziału sum kwadratów: wg. komponentów plonu (jeden kierunek) oraz źródeł zmienności wyszczególnionych w analizie wariancji (drugi kierunek). Procedura jest kompilacją regresji wielokrotnej i ANOVA. Końcowe, tabelaryczne zestawienie danych pozwala na ocenę niezależnego udziału w ostatecznym plonie kolejnych zmiennych niezależnych (komponentów) sukcesywnie wprowadzanych do modelu oraz równoczesną ocenę efektów związanych ze źródłami zmienności uwzględnianymi w analizie wariancji danego układu doświadczalnego. Tok obliczeniowy oraz sposób interpretacji wyników zaprezentowano na przykładzie polowego doświadczenia hodowlanego z 10 liniami bobiku, prowadzonego w warunkach naturalnego zapylenia.

SŁOWA KLUCZOWE: analiza komponentów plonu, podział dwukierunkowy, analiza regresji, ANOVA