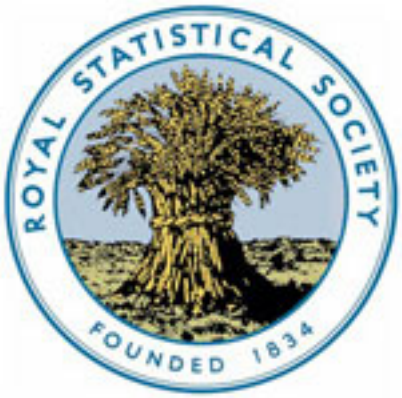

Principal Component Analysis of Designed Experiment

J. N. R. JEFFERS

1. Introduction

The increasing availability of electronic digital computers for statistical analysis has focused a great deal of attention upon techniques of multivariate analysis. The books of Rao, 1952, and Kendall, 1957, have, in particular, made clear the interrelationships between the various types of multivariate analysis, and have underlined the basic theory upon which these techniques depend. Although few of the standard texts deal with the application of multivariate techniques to data arising from designed experiments, there have been a number of papers dealing with this subject. Quenouille, 1950, reviewed and extended some methods of multivariate experimentation, distinguishing between dependent and independent variables. Steel, 1955, suggested a multivariate test of significance for the effects of varieties or treatments, and the use of orthogonal transformations in the analysis of experiments, with particular emphasis upon the use of canonical variates. Rao, 1958, suggested two methods for the special problem of repeated measurements of perennial crops, and Danford *et al.*, 1960, suggested multivariate procedures for situations where a valid univariate analysis is not justified. Fairfield-Smith, 1958, discussed a special case of the related problem of the multivariate analysis of covariance.

The small number of papers listed above is by no means exhaustive, and it is clear that a number of statisticians have been thinking about the multivariate analysis of data from designed experiments. Despite this interest, remarkably little use of multivariate methods is evident in the day-to-day analysis of experimental results, and general procedures for such analysis do not appear to be in common use.

This paper gives an example of the application of one form of multivariate analysis, that of principal component analysis, to experimental data. The methods described have been applied to many experiments, and have been found to combine a reasonably simple method of analysis, if an electronic digital computer is available, with an approach sufficiently general to be usefully applied to the widest possible range of experimental procedures.

2. Methods of Multivariate Analysis

It is desirable, first to consider some of the many forms of multivariate analysis that might be adopted in the interpretation of designed experiments. In considering these, the simplest form which the data from a typical experiment might take is that of a two-way table giving the arithmetic mean of each variable assessed for each of the experimental treatments. Such a data-set most nearly conforms to the type of data usually assumed in texts on multivariate analysis, and may be looked at in two ways. First, greatest interest may be expressed in the degrees of similarity between pairs of treatments, calculated over the full range of the variables assessed. Techniques with this emphasis, sometimes known as Q-techniques, have come to form the basis of many methods of numerical classification, and lead to the setting up of a "taxonomy" of the treatments. (Sneath, P. H. A., 1957, and Michener, C. D. and Sokal, R. R., 1957). For certain types of experiment, they may prove to be of particular value, as, for example, in experiments seeking to evaluate closely related groups of plants or animals. This approach will not, however, be adopted in this paper.

The alternative emphasis in examining the data directs attention to the variates themselves and the ways in which they are correlated. By seeking for "dimensions" of variability which are more general than the individual variates, the experimenter attempts to gain a better understanding of the response to his "treatments" and a sounder knowledge of the variates which are important in future experiments. If these more general "dimensions" are discovered, the treatments may be compared in terms of these new variables and most of the information provided by the experiment summarized by only a small number of comparisons. The techniques with this emphasis are sometimes grouped together under the name of *factor analysis*, but this terminology is unfortunate, since the term is also applied to a particular type of analysis with this emphasis. In fact, a considerable number of closely related techniques of this broad class can be generated by different choices of procedure at various stages in the analysis. In this paper, only one of the possible methods will be examined, that of *principal component analysis*.

The data-set referred to above is the most general that can be derived from designed experiments, and enables data derived from different stages of the experiment to be included in the same analysis, even if the actual designs of the stages of the experiment are not the same. Thus, in an experiment to compare the progeny from different trees, assessments of important variables may be made on the seed, on the seedlings raised from the seed, on transplants taken from the seedlings, on the trees planted in several sites, and on the timber cut from the mature trees, and all of the stages of this experiment may

require different experimental designs. The ability to analyse the data together helps to correlate a great number of separate variates.

At a lesser level of summary, there may be a number of assessments within any single stage of the experiment, so that, for each plot of the experimental design, there are corresponding values of a number of variates. It is therefore possible, for this more limited number of variates, to express the correlations between the variates for several components isolated by the experimental design. Thus, it is possible to calculate the correlations for the total range of variability of the experiment, for the more limited range expressed as "treatments-plus-error," and for the experimental error alone. Factor analysis of the correlations shown by the various components may therefore reveal the precise effects of the treatments, and, sometimes more important, the structure of the unexplained error (Pearce and Holland, 1960). An alternative approach by Steel, 1955, made use of canonical correlations to maximize the ratio of treatment to treatment-plus-error correlations, in other words, to identify those "dimensions" which were most affected by the treatments. Yet another approach has been adopted in this paper, that of Rao, 1958, in which a principal component analysis is based upon the total correlation matrix, followed by the calculation of the new variables defined in this analysis, and an analysis of variance of these variables.

3. Methods of Calculation

As indicated above, only one of the many possible forms of multivariate analysis is described in this paper. The reasons for this choice are not intended to be the subject of this paper, but may be summarized briefly—

1. The method of principal component analysis described below is objective and free from the dubious practices of estimation of communalities and rotation of axes present in some other methods of factor analysis.

2. Principal component analysis is relatively easy to programme on electric digital computers, and can be applied to a wide variety of situations.

3. Although obviously a mathematical artefact, experience in its application frequently suggests meaningful and valuable interpretations.

4. The method directs attention to the wider problem of the "dimensions" which assessed variates seek to express, and gives guidance to the choice of variates in future experimentation.

The actual method of computation can be described very simply. Given a matrix of variates.

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdot & \cdot & \cdot & x_{1n} \\ x_{21} & x_{22} & x_{23} & \cdot & \cdot & \cdot & x_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{t1} & x_{t2} & x_{t3} & \cdot & \cdot & \cdot & x_{tn} \end{bmatrix}$$

where n is the number of variates, and t is the number of treatments or plots, the coefficients of correlation between every pair of columns are calculated to form the correlation matrix.

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} & \cdot & \cdot & \cdot & r_{1n} \\ r_{21} & r_{22} & r_{23} & \cdot & \cdot & \cdot & r_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ r_{n1} & r_{n2} & r_{n3} & \cdot & \cdot & \cdot & r_{nn} \end{bmatrix}$$

where the principal diagonal, composed of the elements $r_{11}, r_{12}, \ldots, r_{nn}$ consists of 1's. The latent roots and vectors of this symmetric correlation matrix define an orthogonal set of linear combinations of the original variates

$$\begin{aligned} z_1 &= a_{11}\ x_1 + a_{12}\ x_2 + \ldots + a_{1n}\ x_n \\ z_2 &= a_{21}\ x_1 + a_{22}\ x_2 + \ldots + a_{2n}\ x_n \\ &\cdot \quad \cdot \quad \cdot \quad \cdot \qquad\qquad \cdot \quad \cdot \\ &\cdot \quad \cdot \quad \cdot \quad \cdot \qquad\qquad \cdot \quad \cdot \\ z_n &= a_{n1}\ x_1 + a_{n2}\ x_2 + \ldots + a_{nn}\ x_n \end{aligned}$$

such that the first linear combination (component) has, subject to the restraint of statistical independence, maximum variance. Furthermore, the second component is uncorrelated with the first and has as large a variance as possible, and so on. The new linear combinations therefore describe the original variance in as small a number of uncorrelated dimensions as possible, and, if the components so obtained have any physical interpretation, may lead to a better understanding of the measured variates.

The calculations described above may be readily programmed for any electronic digital computer which has an efficient sub-routine for the calculation of latent roots and vectors of symmetric matrices. The computations for the example of this paper were carried out on a Ferranti Pegasus computer, using a special programme for principal component analysis written by the British Iron and Steel Research Association (Head, 1961).

4. An Example

As an example of the application of the methods described above, the data from an actual experiment will be analysed in this section of the paper. The investigation from which the data are taken was designed to compare the growth and characteristics of thirteen *provenances* of a tree species (*Thuya plicata*) introduced into this country from North America. *Provenance*, in this sense, is defined as a group of trees grown from seed collected at a given geographical source of place of origin. The experiment was laid out in a randomized block design, with four replications, the seed from the thirteen separate origins being sown in square yard plots. Before sowing, the percentage of seeds germinating on standard germination tanks, and the weight of 1,000 pure seed, were determined for samples of the seed from each origin. After sowing. counts of the numbers of seedlings on each plot where made when there were germinating seedlings on all plots of any one provenance, and at four and eight weeks after this. At the end of the first year, the number of seedlings remaining on each plot were counted, and the heights of a random sample of seedlings from each plot were measured. In addition, seedlings were selected at random from each plot for determination of the root length, shoot length, and root collar diameter. Thus, data illustrating the two stages of summary discussed above are available for analysis. Table 1 summarizes the variables assessed.

TABLE 1

Variables Assessed in Experiment.

Key	Variable	Sampling units
A	Percentage of seed germinating in laboratory	3 samples from each provenance
B	Weight of 1,000 pure seed	3 samples from each provenance
C	First seedling count	Experimental plots
D	Second seedling count	,, ,,
E	Third seedling count	,, ,,
F	Total number of seedlings	,, ,,
G	Mean height in inches	,, ,,
H	Root length in millimetres	,, ,,
I	Shoot length in millimetres	,, ,,
J	Root collar diameter in millimetres	,, ,,

4.1. Analysis of Treatment Means

Table 2 gives the half-matrix of the coefficients of correlation between every pair of variables, calculated from the mean values of the variables for each of the thirteen provenances. The principal diagonal and the upper half of the matrix has been omitted. A diagrammatic form of the correlations is also given in Figure 1.

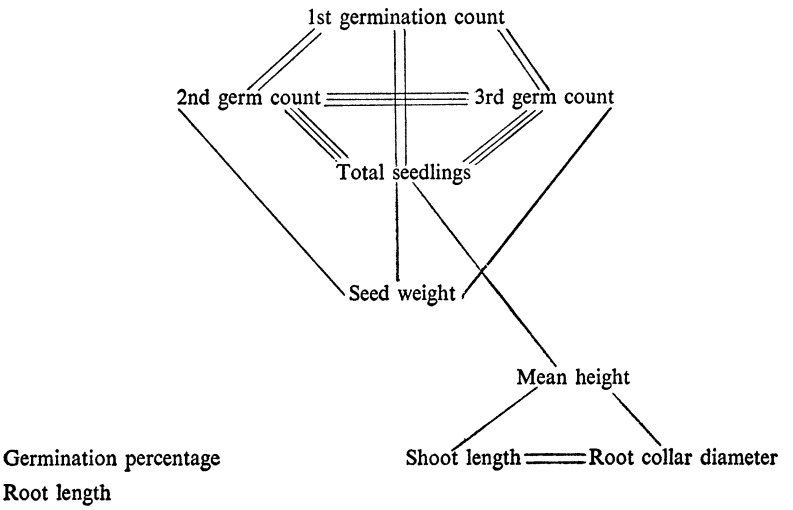

FIG. 1

Diagrammatic representation of correlations between variables

Thus, there is a large, closely correlated group of variables representing the numbers of seedlings at the various counts, and the total number of seedlings at the end of the growing season, and all except the numbers of seedlings at the first count are linked to seed weight. A second group of variables is composed of the mean height of the seedlings, the shoot length and the root collar diameter, and this second group is linked to the first by the correlation between the total number of seedlings and the mean height. Percentage germination in the laboratory, and root length, are not significantly correlated with either of these groups, or with each other.

The first five latent roots of the correlation matrix are given in table 3, together with the percentage of the total variability accounted for by each component, and the cumulative percentages.

TABLE 2

Analysis of Treatment Means: Correlation Coefficients

A								
0·163	B							
−0·212	0·444	C						
−0·192	0·614*	0·786**	D					
−0·185	0·571*	0·758**	0·992**	E				
−0·113	0·607*	0·789**	0·971**	0·951**	F			
0·253	0·346	0·314	0·504	0·520	0·567*	G		
0·387	−0·337	−0·314	−0·413	−0·357	−0·350	−0·061	H	
−0·137	−0·034	0·483	0·398	0·402	0·457	0·643*	−0·012	I
−0·051	−0·001	0·479	0·286	0·265	0·325	0·579*	−0·232	0·878** J

* indicates significance at 0·05 level of probability.

** „ „ „ 0·01 „ „ „

TABLE 3

Analysis of Treatment Means: Latent Roots of the First Five Components

Component	Latent root	Percentage of variability	
		Component	Cumulative
M1	5·05	50·5	50·5
M2	1·80	18·0	68·5
M3	1·42	14·2	82·7
M4	0·76	7·6	90·3
M5	0·46	4·6	94·9

The first five components therefore account for 94·9 per cent of the variability contained by the ten variables assessed so far, and a worthwhile condensation of the data has been achieved. Furthermore, the first two components account for 68·5 per cent of the variability, and the first alone for 50·5 per cent. The identification of the components may be achieved by examination of the latent vectors for these five components, given in table 4.

TABLE 4

Analysis of Treatment Means: Latent Vectors for the First Five Components

	Variable	Vectors for component—				
		M1	M2	M3	M4	M5
A	Germination per cent of seed	0·072	−0·189	−0·728	0·221	0·334
B	Seed weight	−0·266	0·316	−0·376	0·311	0·176
C	First seedling count	−0·373	0·054	0·112	−0·234	0·644
D	Second seedling count	−0·419	0·196	−0·043	−0·160	−0·116
E	Third seedling count	−0·411	0·178	−0·057	−0·223	−0·193
F	Total number of seedlings	−0·421	0·129	−0·097	−0·168	−0·087
G	Mean height	−0·294	−0·358	−0·294	0·199	−0·550
H	Root length	0·193	−0·296	−0·368	−0·755	0·032
I	Shoot length	−0·280	−0·528	0·168	−0·086	−0·007
J	Root collar diameter	−0·249	−0·529	0·226	0·281	0·284

The first component clearly represents the yield of seedlings given by the provenances, while the second represents the vigour of these seedlings. These two components correspond, therefore, to the main groups of Figure 1. The third gives greatest weight to the laboratory germination percentage of the seed, and the fourth and fifth components to root length and to speed of germination respectively. From the vectors of table 4, it is possible to calculate the value of each component for each of the thirteen provenances, and these values are given in table 5.

TABLE 5

Analysis of Treatment Means: Values of the First Five Components for each Provenance

Provenance	Value of component number—				
	M1	M2	M3	M4	M5
Terrace	−1·35	−0·48	−2·74	−0·71	0·52
Massett	0·63	0·93	−0·71	0·46	1·34
Queen Charlotte Islands	3·00	0·74	0·31	−1·68	−0·03
Shuswap Lake	−1·27	0·40	2·11	−0·09	0·53
Courtenay	−3·90	−2·87	1·11	0·14	0·55
Alberni	1·94	−1·19	−0·98	0·88	−0·60
Ladysmith	−0·42	−1·49	0·09	−0·94	−1·11
Sooke	−0·35	1·27	−0·12	1·79	−0·33
Joyce	−2·46	2·08	0·00	−0·51	0·26
Sequim	−2·92	1·73	0·02	−0·04	−1·23
Tenino	2·22	−0·23	−0·11	−0·53	0·16
Ashford	1·40	−0·83	−0·74	0·85	−0·11
Vernonia	3·47	−0·05	1·76	0·39	0·05

These values of the five components provide the experimenter with five independent and mutually orthogonal comparisons of the thirteen provenances, while the weighting given to the original variables in each component provide him with further information upon which to base his comparison. Thus, the relatively high weighting given to seed weight in the first two components would lead the experimenter to discount the value of early seedling yields and vigour as a useful measure for the selection of provenances, since it suggests that differences in seedling yield and vigour may be largely dependent upon climatological conditions in the summer preceding the formation and collection of the seed. Greater weight might therefore be given to comparisons based upon the other components.

The principal component analysis of the treatment means has therefore given a worthwhile condensation of the data, and has given some information on the ways in which the several variables assessed in the experiment interact. Definable factors have emerged which may be sensibly interpreted by the experimenter, the more so in that he is guaranteed their mutual orthogonality.

4.2. Analysis of Plot Means

For the more limited set of variables (C – J) determined for each plot of the same experimental design, the analysis of the preceding section may be repeated using the 13 × 4 values available for each variable. The latent roots and vectors derived in this way are given in tables 6 and 7.

TABLE 6

Analysis of Plot Means: Latent Roots of the First Five Components

Component	Latent root	Percentage of variability	Identification
P1	3·83	47·9	Seedling yield
P2	2·08	26·0	Seedling vigour
P3	0·96	12·0	Root length
P4	0·70	8·7	Seedling height
P5	0·24	3·0	Speed of germination

TABLE 7

Analysis of Plot Means: Latent Vectors of the First Five Components

		Vectors for component—				
	Variable	P1	P2	P3	P4	P5
C	First seedling count	0·434	0·028	−0·216	−0·332	−0·800
D	Second seedling count	0·499	0·041	−0·036	−0·043	0·176
E	Third seedling count	0·496	0·049	−0·024	−0·019	0·317
F	Total number of seedlings	0·484	0·015	−0·035	−0·074	0·378
G	Mean height	0·267	−0·281	0·160	0·865	−0·262
H	Root length	−0·103	−0·195	−0·947	0·171	0·117
I	Shoot length	0·019	−0·673	0·031	−0·156	0·055
J	Root collar diameter	−0·005	−0·651	0·167	−0·284	0·031

The component P1 clearly corresponds to that of M1, the reversal of the signs having no practical meaning in these artefacts. Similarly P2, P3 and P5 correspond to M2, M4 and M5 respectively. The component P4, giving greatest weight to the variable of mean height accounting for nearly 9 per cent of the total variability, is given greater emphasis in this analysis than in the analysis of treatment means.

From these vectors, values of the components may be calculated for each plot in the randomized block design, and the new values so calculated subjected to the usual analysis of variance for randomized blocks, in order to test the significance of differences between provenances. The result of such an analysis is given in table 8.

TABLE 8

Analysis of Plot Means: Treatment Means for the Components P1 *to* P5

Provenance	Mean value of component				
	P1	P2	P3	P4	P5
Terrace	1·06	−0·27	−0·75	0·41	0·17
Massett	−0·49	0·43	−0·10	−0·62	−0·07
Queen Charlotte Islands	−1·40	0·84	−0·88	−0·26	0·11
Shuswep Lake	0·79	−0·01	0·46	−0·87	−0·12
Courtenay	1·69	−1·88	0·35	−0·23	−0·27
Alberni	−1·20	−0·12	0·19	0·75	−0·08
Ladysmith	0·33	−0·63	−0·30	0·57	0·29
Sooke	0·27	0·51	1·04	0·04	0·15
Joyce	1·87	0·57	0·01	−0·24	−0·08
Sequim	1·91	0·18	0·37	0·13	0·84
Tenino	−1·18	0·28	−0·46	0·00	−0·10
Ashford	−1·28	−0·26	−0·01	0·60	−0·45
Venonia	−2·36	0·36	0·12	−0·28	−0·39
Standard error	±0·665**	±0·641	±0·470	±0·822	±0·206*

Thus, only for the components P1 and P5 were there any significant differences between the provenances, corresponding to differences in seedling yield and speed of germination. Since the principal component analysis has already demonstrated that these factors are linked to seed weight, it is clear that the experiment has not yet given any information which would lead to a worthwhile selection from the provenances and that suggested differences between them are a reflection of the energy stored in the seed.

4.3. Conclusions and Discussion

Finney, 1956, has suggested that analyses of the type illustrated above are not desirable if they remove from the experimenter the necessity of formulating meaningful hypotheses about the variables that he has chosen to assess, particularly if these variables represent successive measurements of the same character, as is frequently the case in perennial crop experiments. It must be admitted that this is fair criticism. There can be no substitute for proper formulation of hypotheses in the analysis of designed experiments. Nevertheless, in many fields of research, knowledge of essential "dimensions" of variability is still incomplete, and variables for assessment are frequently chosen for their convenience and ease of measurement. In forestry, for example, the general concept of tree vigour is important, but little work has so far been done on the relative value of such measurements as height, diameter, taper, etc., as measures of "vigour." It is, therefore, of value to know whether, in any particular experiment, these variables can be regarded as measuring some common factor, or whether they are measures of distinct factors.

In the experiment used as an example above, normal statistical analysis of the separate variables gave no significant differences between the provenances. The application of principal component analysis to the plot means has revealed significant differences between the provenances in two of the factors of which the individual variables may be regarded as expressions, and principal component analysis of the treatment means has further revealed the nature and possible origin of these differences, and suggested that they do not provide useful criteria for the selection of provenances in the future. The analysis has in fact pointed the way to useful hypotheses for future research, and to other variables which should be assessed, e.g. the climatological conditions in the years preceding seed collection.

Perhaps even more important in a world in which important subjects for research jostle for the attention of experimenters, the analysis above has revealed a certain wastefulness in the number of assessments made. In future experiments of this kind, assessments could well be reduced to five variables without any important loss of information, and the vector loadings of tables 4 and 7 give clear indications as to which these should be.

As more and more evidence becomes available on such an experiment, for example, by later measurements of the growth of the seedlings, of the growth of the surviving trees when they are planted in the forest, of the properties of the timber of the resulting trees, etc., principal component analysis of experimental results would inform the experimenter when he had obtained information which genuinely pointed to the existence of new dimensions, as opposed to merely strengthening the evidence for the dimensions he had already

found. Such a facility is not to be despised in perennial experiments, where the greatest difficulty may be experienced in knowing where to terminate the experiment and when to expect the onset of the treatment effects. Moreover, most experimenters would readily agree that they themselves think about their problems in a multivariate context, but that they have conditioned themselves to expressing their requirements in a univariate form, largely because of the insistence of statisticians that this is the only form that can readily be handled.

The analysis described above is relatively simple to perform and interpret, though tedious if attempted on desk-calculating machines. In the simple account given, much of the complication which can be introduced has been omitted, as, for example, in the testing of the "significance" of individual components as proposed by Bartlett, 1950. The theory underlying such an analysis has been understood for some time, what is now needed is its application to a wide field of practical research in order that the usefulness of the technique may be tested.

REFERENCES

BARTLETT, M. S. (1950). "The tests of significance in factor analysis." *Brit. J. Psych.* (Stat. Sect.), **3**, 77.

DANFORD, M. B. *et al.* (1960). "On the analysis of repeated measurements experiments." *Biometrics*, **16** (4), 547–565.

FAIRFIELD-SMITH, H. (1958). "A multivariate analysis of variance." *Biometrics*, **14** (1), 107–127.

FINNEY, D. J. (1956). "Multivariate analysis and agricultural experiments." *Biometrics*, **12** (1), 67–71.

HEAD, A. E. (1961). *Handbook on the use of Compan—a Component Analysis. Programme.* Unpublished.

KENDALL, M. G. (1957). *A Course in Multivariate Analysis.* Griffin.

MICHENER, C. D. and SOKAL, R. R. (1957). "A quantitative approach to a problem of classification." *Evolution*, **11**, 130–162.

PEARCE, S. C. and HOLLAND, D. A. (1960). "Some applications of multivariate methods in botany." *Applied Statistics*, **9** (1), 1–62.

QUENOUILLE, M. H. (1950). "Multivariate experimentation." *Biometrics*, **6** (3), 303–316.

RAO, C. R. (1952). *Advanced statistical methods in biometric research.* Wiley.

RAO, C. R. (1958). "Some statistical methods for comparison of growth curves." *Biometrics*, **14** (1), 1–17.

SNEATH, P. H. A. (1957). "The application of computers to taxonomy." *J. gen. Microbiol.*, **17**, 201–226.

STEEL, R. G. D. (1955). "An analysis of perennial crop data." *Biometrics*, **11** (2), 201–212.