THE APPLICATION OF THE TECHNIQUE OF ANALYSIS OF COVARIANCE TO FIELD EXPERIMENTS WITH SEVERAL MISSING OR MIXED-UP PLOTS

K. R. Nair

Statistical Laboratory, Calcutta.

Introduction

The techniques of analysis of variance and of covariance developed by R. A. Fisher are well understood in their application to field experiments especially of such orthogonal designs as randomized blocks and Latin squares. Even a research worker with scarcely any pretensions to much mathematical knowledge can easily analyse and interpret the data from such experiments without taking help from a professional statistician. This mechanisation on the computational side is a striking feature of Fisher's contributions to the statistical craft.

The procedure is particularly simple when the various factors involved are orthogonal to one another. Thus in the ordinary randomized block experiment, blocks and treatments are orthogonal to one another on account of the fact that in every block there is one and only one plot for every treatment. For the Latin square, in every row and every column there is one and only one plot for every treatment, so that rows, columns and treatments are mutually orthogonal in their effects.

But accidents sometimes happen which destroy the orthogonality. Such are cases of missing and mixed-up plots. The former is of more common occurrence, and is usually due to extraneous causes not within the control of the experimenter. The latter type of accident occurs only through negligence.

Both types of accidents entail loss of information (in a technical sense) or of efficiency of treatment comparisons. If the additional time required for the analysis of such experiments is also taken into consideration the total loss (in a popular sense) will be still greater. Missing plots may sometimes be unavoidable, but there is little excuse for allowing the plots to get mixed-up. However this actually happened in the case of an agricultural experiment in India; and S. S. Bose and P. C. Mahalanobis solved the problem of reconstructing the individual plot data following a method developed by Yates in the case of missing plots.[2, 3, 6]

The straight-forward method in such cases is that of fitting of constants. Yates evolved a technique by which the labour of fitting of constants was avoided in the case of missing plots, and this technique was used by S. S. Bose and P. C. Mahalanobis in the case of mixed-up plots.

Bartlett's Method for Missing Plots

Bartlett recently suggested an alternative method of tackling the problem of missing plots, by introducing pseudo-characters on the observational side of the experiment[1]; and was thus able to overcome the difficulty due to the non-orthogonality of the design by changing over from analysis of variance to analysis of covariance in the case of a single observed character and by an extension of analysis of covariance if there are more than one observed character. The number of 'observed' pseudo-characters is made equal to the number of missing plots; and Bartlett's technique has the advantage that it again mechanises the process of analysis.

Suppose there are k missing plots. He then introduces k pseudo-variates $x_1, x_2 \ldots x_k$ and in the k missing plots these variates and y, the observed character will be assigned values as shown in Table (1).

TABLE 1. VALUES ASSIGNED TO THE PSEUDO- AND OBSERVED CHARACTERS IN EACH MISSING PLOT.

Missing Plot	Pseudo-character					Observed character
	x_1	x_2	x_3	...	x_k	y
1	1	0	0	...	0	0
2	0	1	0	...	0	0
3	0	0	1	...	0	0
⋮	⋮	⋮	⋮	...	⋮	⋮
k	0	0	0	...	1	0

In the remaining plots, which are unaffected, all the k pseudo-variates will have value o, and y will assume its observed values in those plots. A Multiple Covariance Table can be constructed and if $b_1, b_2, \cdots b_k$ are the partial regression coefficients of y on $x_1, x_2, \cdots x_k$ calculated from the Error line, the value of y in each plot after adjusting for variations in $x_1, x_2 \ldots x_k$ may be taken as y' given by

$$y' = y - b_1x_1 - b_2x_2 - \quad \ldots \quad \ldots \quad - b_kx_k \qquad \ldots (1)$$

In the k missing plots the values of y' will be

$$-b_1, \; -b_2, \ldots\ldots\ldots\ldots \; -b_k$$

and in the other plots there is no adjustment to be made, so that $y'=y$.

The treatment sum of squares of y for test of significance is, as is well known for covariance analysis, obtained by getting the adjusted sum of squares of y from the Treatment+Error' line and from the 'Error' line of the Covariance Table and by finding their difference. The adjusted sum of squares of y got from the Error line of the Covariance Table will be the same as the Error sum of squares of the Variance Table of the reconstructed data, namely, of y'.

The treatment sum of squares of y got from the reconstructed data will be in excess of the one obtained above. It will be convenient in practical work to know what exactly this excess is, so that after getting estimates of missing values, one need not do further calculations from the Covariance Table. But elegant expressions for this excess is available only in the case of a single missing value.

TABLE 2 ANALYSIS OF VARIANCE AND COVARIANCE.

Due to :—	$S(x^2)$	$S(xy)$	$S(y^2)$
Treatments ...	A	C	B
Error ...	A'	C'	B'
Treatments + Error ...	A + A'	C + C'	B + B'

If Table (2) is the Covariance Table for the pseudo-variate x and the observed variate y, the sum of squares of y due to treatments from the reconstructed data is equivalent to adjusting B with the regression coefficient C'/A' giving the result

$$B - \frac{2CC'}{A'} + \frac{C'^2A}{A'^2} \quad \cdots \quad (2)$$

whereas the valid measure of the treatment sum of squares is

$$B - \frac{(C+C')^2}{A+A'} + \frac{C'^2}{A'} \quad \cdots \quad (3)$$

subtracting (3) from (2) we get

$$\frac{A^2}{A+A'}\left(\frac{C}{A} - \frac{C'}{A'}\right)^2 \quad \cdots \quad (4)$$

which is a positive quantity.

COVARIANCE TECHNIQUE FOR MIXED-UP PLOTS

The general problem faced is of k plots getting mixed-up, leaving a known total, say u. Introduce $k-1$ pseudo-variates $x_1, x_2, \ldots \ldots x_{k-1}$ all having value 0 in the unaffected plots and the values shown in Table (3) for the k affected plots.

TABLE 3. VALUES ASSIGNED TO THE PSEUDO- AND OBSERVED CHARACTERS IN EACH MIXED-UP PLOT.

Mixed-up Plot	Pseudo-character					Observed character
	x_1	x_2	x_3	...	x_{k-1}	y
1	1	1	1	...	1	u/k
2	1−k	1	1	...	1	u/k
3	1	1−k	1	...	1	u/k
⋮	⋮	⋮	⋮	...	⋮	⋮
k	1	1	1	...	1−k	u/k

The observed character y is given value u/k in each of the k affected plots and its known values in the unaffected plots. Next, construct a Multiple Covariance Table. If $b_1, b_2 \ldots\ldots b_{k-1}$ are the partial regression coefficients, calculated from the Error line, of y on $x_1, x_2 \ldots\ldots x_{k-1}$ the adjusted value of y say y' is given by

$$y' = y - b_1x_1 - b_2x_2 - \quad \ldots \quad \ldots \quad - b_{k-1}x_{k-1} \quad \cdots \quad (5)$$

19

The values of y' in the k affected plots supply the estimates of the individual yields of those plots and are given in Table (4) in terms of b_1, $b_2 \cdots\cdots b_{k-1}$. In the unaffected plots $y' = y$.

TABLE 4. ESTIMATED VALUE OF THE OBSERVED CHARACTER IN EACH MIXED-UP PLOT.

Mixed-up Plot	Estimated value of y
1	$(u/k) - b_1 - b_2 - \quad \ldots - b_{k-1}$
2	$(u/k) + (k-1)\,b_1 - b_2 - \ldots - b_{k-1}$
3	$(u/k) - b_1 + (k-1)b_2 - \ldots - b_{k-1}$
⋮	
k	$(u/k) - b_1 - b_2 \quad \ldots + (k-1)b_{k-1}$
Total	u

The treatment sum of squares of y for test of significance should be done in the way usually followed for covariance analysis as explained before.

In the case of two mixed-up plots there is only one pseudo-variate x, with values $+1$ and -1 in the two plots; and the correction to be given to the sum of squares of treatments in the Variance Table of the reconstructed data can be easily found, using (4).

STANDARD ERROR OF DIFFERENCES BETWEEN TREATMENT EFFECTS

We have considered the test of significance of all the treatment effects taken together when k plots are missing or get mixed-up. There is still the problem of testing the difference between any two treatment effects. The calculation of the variance of this difference is a complicated process from the point of view of 'fitting of constants' and Yates' method does not open out any simplification. It will be presently seen however that great simplification in this process can be achieved by means of the technique of covariance.

Thus in the case of k missing plots if y_i, x_{1i}, x_{2i}, ... x_{ki} are the means corresponding to the i-th treatment and y_j, x_{1j}, x_{2j}, ... x_{kj} those corresponding to the j-th treatment, the adjusted treatment means $\bar{y}'_i$ and $\bar{y}'_j$ of y are given by

$$\bar{y}'_i = \bar{y}_i - b_1\bar{x}_{1i} - b_2\bar{x}_{2i} - \quad \ldots \quad \ldots \quad - b_k\bar{x}_{ki} \qquad \ldots (6)$$

$$\bar{y}'_j = \bar{y}_j - b_1\bar{x}_{1j} - b_2\bar{x}_{2j} - \quad \ldots \quad \ldots \quad - b_k\bar{x}_{kj} \qquad \ldots (7)$$

so that the difference between the two treatment effects is

$$\bar{y}'_i - \bar{y}'_j = (\bar{y}_i - \bar{y}_j) - b_1(\bar{x}_{1i} - \bar{x}_{1j}) - \quad \ldots \quad - b_k(\bar{x}_{ki} - \bar{x}_{kj}) \qquad \ldots (8)$$

The general expression for the variance of $\bar{y}'_i - \bar{y}'_j$ has been worked out by Wishart.[4]

If none of the plots of the i-th and (or) the j-th treatments is missing, $\bar{x}_{1i}$, $\bar{x}_{2i}$, etc., and (or) $\bar{x}_{1j}$, $\bar{x}_{2j}$, etc., will be zero.

Loss of Information with a Single Missing Plot or with two Mixed-up Plots

'Amount of information' has a technical meaning in Fisherian concepts of statistical theory. The amount of information contained in a field experiment is defined to be proportional to the reciprocal of the mean variance of comparisons between all possible pairs of treatment effects.

With a single missing plot or with two mixed-up plots, we have to introduce only one pseudo-variate x, so that (8) reduces to

$$\bar{y}'_i - \bar{y}'_j = (\bar{y}_i - \bar{y}_j) - b'(\bar{x}_i - \bar{x}_j) \qquad \dots (9)$$

where $b' = C'/A'$ of Table (2).

The variance of $\bar{y}'_i - \bar{y}'_j$ is given by

$$V(\bar{y}'_i - \bar{y}'_j) = V(\bar{y}_i - \bar{y}_j) + (\bar{x}_i - \bar{x}_j)^2 V(b') = s^2 \left\{ \frac{2}{r} + \frac{(\bar{x}_i - \bar{x}_j)^2}{A'} \right\} \qquad \dots (10)$$

where r is the number of replications of each treatment in the original lay-out and s^2 is the error variance calculated from the reconstructed data.

If t be the number of treatments there are $t\ (t-1)/2$ possible comparisons and the mean variance of all these comparisons is

$$s^2 \left\{ \frac{2}{r} + \frac{1}{A'} \cdot \frac{2}{t(t-1)} \sum_{i=1}^{t} \sum_{j=1}^{t} (\bar{x}_i - \bar{x}_j)^2 \right\}_{i \neq j} \qquad \dots (11)$$

Now,

$$\sum_{i=1}^{t} \sum_{j=1}^{t} (\bar{x}_i - \bar{x}_j)^2_{\,i \neq j} = t \sum_{i=1}^{t} (\bar{x}_i - \bar{x})^2 = \frac{tA}{r} \qquad \text{(see Table 2)} \qquad \dots (12)$$

so that (11) reduces to

$$\frac{2s^2}{r} \left\{ 1 + \frac{A}{(t-1)A'} \right\} \qquad \dots (13)$$

If the experiment was free from any accident and remained orthogonal the mean variance would have been $2s'^2/r$, s'^2 being the error variance from its analysis.

Ignoring the possible differences in s and s', the loss of information due to a single missing plot or due to having two mixed-up plots can be expressed in the form

$$\frac{1}{1 + (t-1)A'/A} \qquad \dots (14)$$

The values of A and A' will depend on the lay-out and in the case of mixed-up plots, on the distribution of the plots as well.

(i) *Single Missing Plot*

Thus for a randomized block experiment with a single missing plot

$$\left. \begin{aligned} A &= 1/b - 1/(bt) \\ A' &= 1 - 1/b - 1/t + 1/(bt) \end{aligned} \right\} \qquad \dots (15)$$

so that the loss of information is

$$\frac{1}{1 + (b-1)(t-1)} \qquad \dots (16)$$

For a Latin square experiment having n rows and n columns ($t=n$) and one plot missing,

$$\left.\begin{aligned} A &= 1/n-1/n^2 \\ A' &= 1-3/n+2/n^2 \end{aligned}\right\} \quad \ldots \quad (17)$$

so that the loss of information is

$$\frac{1}{1+(n-1)(n-2)} \quad \ldots \quad (18)$$

(ii) *Two Mixed-up Plots*

In the case of a randomized block experiment there are three possibilities, namely, the two plots may belong to (1) different blocks and different treatments, (2) same block, and (3) same treatment. The values of A/A' in the three cases and the mean value of A/A' are given in Table (5).

TABLE 5. VALUES OF A/A' FOR RANDOMISED BLOCK EXPERIMENT WITH TWO MIXED-UP PLOTS.

	(1)	(2)	(3)	Mean
A/A'	$\frac{t}{bt-b-t}$	$\frac{1}{b-1}$	0	$\frac{2(bt-b-t)+b}{3(b-1)(bt-b-t)}$

As all three possibilities may be taken as equally probable we can substitute for A'/A in (14) the reciprocal of the mean value of A/A'. The average loss of information will therefore be

$$\frac{1}{1+\frac{3(b-1)(t-1)(bt-b-t)}{2(bt-b-t)+b}} \quad \ldots \quad (19)$$

In the case of a Latin square experiment there are four possibilities, namely, the two plots may belong to (1) different rows, different columns and different treatments, (2) same row, (3) same column, and (4) same treatment. The values of A/A' in the four cases and the mean value of A/A' are given in Table (6).

TABLE 6. VALUES OF A/A' FOR LATIN SQUARE EXPERIMENT WITH TWO MIXED-UP PLOTS.

	(1)	(2)	(3)	(4)	Mean
A/A'	$\frac{1}{n-3}$	$\frac{1}{n-2}$	$\frac{1}{n-2}$	0	$\frac{3(n-3)+1}{4(n-2)(n-3)}$

Sustituting for A'/A in (14), the reciprocal of the mean value of A/A' as before, we get for the average loss of information

$$\frac{1}{1+\frac{4(n-1)(n-2)(n-3)}{3(n-3)+1}} \quad \ldots \quad (20)$$

In Table (7) are given the percentage losses of information in a randomized block experiment, for select values of b and t. The upper figures indicate the loss due to a single missing plot and the lower figures the loss due to two mixed-up plots.

Table (8) gives similar values for a $n \times n$ Latin square experiment for select values of n. It is interesting to see that with increase in size of the experiment the loss of information with two mixed-up plots becomes less than the loss with a single missing plot.

TABLE 7. PERCENTAGE LOSS OF INFORMATION IN RANDOMIZED BLOCK EXPERIMENT WITH ONE MISSING PLOT (UPPER FIGURES) AND TWO MIXED-UP PLOTS (LOWER FIGURES).

$b \rightarrow$ / $t \downarrow$	2	3	4	5	6	7	8	9	10
2	50.00	33.33	25.00	20.00	16.67	14.29	12.50	11.11	10.00
	100.00	45.45	30.77	23.40	18.92	15.89	13.70	12.04	10.74
3	33.33	20.00	14.29	11.11	9.09	7.69	6.67	5.88	5.26
	40.00	20.00	13.46	10.16	8.16	6.82	5.86	5.14	4.57
4	25.00	14.29	10.00	7.69	6.25	5.26	4.55	4.00	3.57
	25.00	12.62	8.47	6.38	5.12	4.28	3.67	3.21	2.86
5	20.00	11.11	7.69	5.88	4.76	4.00	3.45	3.03	2.70
	18.18	9.19	6.16	4.64	3.72	3.10	2.66	2.33	2.07
6	16.67	9.09	6.25	4.76	3.85	3.23	2.78	2.44	2.17
	14.29	7.22	4.83	3.63	2.91	2.43	2.08	1.83	1.62
7	14.29	7.69	5.26	4.00	3.23	2.70	2.33	2.04	1.82
	11.76	5.94	3.97	2.99	2.39	2.00	1.71	1.50	1.33
8	12.50	6.67	4.55	3.45	2.78	2.33	2.00	1.75	1.56
	10.00	5.04	3.37	2.54	2.03	1.69	1.45	1.27	1.13
9	11.11	5.88	4.00	3.03	2.44	2.04	1.75	1.54	1.37
	8.70	4.38	2.93	2.20	1.76	1.47	1.26	1.10	0.98
10	10.00	5.26	3.57	2.70	2.17	1.82	1.56	1.37	1.22
	7.69	3.87	2.59	1.05	1.56	1.30	1.11	0.98	0.87

In Table (7), the loss of information due to a missing plot remains symmetrical with respect to b and t, but the loss due to the mixed-up plots decreases more rapidly with increase of t than of b. From the same table it can be seen that given the total number of plots ($b.t$) for a randomized block experiment the losses of information by both the accidents will be a minimum if the number of treatments (t) and number of blocks (b) remain as nearly equal as possible. For a randomized block experiment with $b=t$ the losses of information by both the accidents are smaller than that in the case of a Latin square experiment, with same number of plots, as is evident from a comparison of the diagonal line, corresponding to $b=t$, of Table (7), with the entries of Table (8).

TABLE 8. PERCENTAGE LOSS OF INFORMATION IN LATIN SQUARE EXPERIMENT WITH ONE MISSING PLOT AND TWO MIXED-UP PLOTS.

n	2	3	4	5	6	7	8	9	10
Missing Plot	100.00	33.33	14.29	7.69	4.76	3.23	2.33	1.75	1.37
Mixed-up Plots	100.00	100.00	14.29	6.80	4.00	2.64	1.87	1.30	1.08

If there are more than one plot missing or more than two plots getting mixed-up the loss of information will be greater still. The calculations of Tables (7) and (8) were done simply to give the quantitative measure of the part of the information which it is impossible for the statistician to recover even with small accidents to the experiments as considered there and to focus the attention of the agricultural field experimenter on the fact that the part of the information that can be recovered will go on steadily decreasing with bigger number of plots missing or getting mixed-up. In conclusion, therefore, I have to stress that the experimenter should not lead himself into the belief that he can miss or mix-up his observations on the plots with impunity.

REFERENCES

1. BARTLETT, M. S.: "Some examples of statistical methods of research in agriculture and applied biology", *Suppl. to the J. R. S. S.*, Vol. IV, No. 2, 1937, p. 151.
2. BOSE, S. S. and MAHALANOBIS, P. C.: "On estimating individual yields in the case of mixed-up yields of two or more plots in field experiment," *Sankhyā*, Vol. 4, Pt. 1, 1938, pp. 103-111.
3. BOSE, S. S.: "Appendix. The estimation of mixed-up yields and their standard errors", *Sankhyā*, Vol. 4, Pt. 1, 1938, pp. 112-120.
4. WISHART, J.: "Tests of significance in analysis of covariance". *Suppl. J. R. S. S.* (1936), pp. 79-82.
5. YATES, F.: "The Analysis of replicated experiments when the field results are incomplete." *Emp. Jour. Expt. Agri.*, Vol. 1, Pt. 2, 1933, pp. 129-142.