Efficient Estimation of the Relationship between Plot Size and the Variability of Crop Yields

W. H. Hatheway and E. J. Williams

Institute of Statistics

Mimeo Series No. 174

June, 1957

Efficient Estimation of the Relationship between Plot Size and the Variability of Crop Yields[≠]

W. H. Hatheway and E. J. Williams[*]

1. Introduction

The optimum size of plot in field experimentation depends on the relationship between fixed costs and costs varying with number of units, and on soil variability. Perhaps the most useful measure of soil heterogeneity yet devised is that of Smith (1938), who showed empirically that the logarithm of the variance between plots of a given size was linearly related to the logarithm of the size of the plot. In the present paper we consider only the relationship between size and variability. The objects of the paper are, firstly, to show how efficient estimates of the constants in this relationship may be determined, and secondly, to illustrate a general method of determining efficient linear estimates when the data are, as in the present instance, correlated and of unequal variability.

Koch and Rigney (1951) demonstrated that the regression coefficient of the logarithm of variance on the logarithm of plot size could be estimated from experimental data in which treatment effects are present, as well as from the data of uniformity trials. They noted that Smith had recommended that, in determining the regression coefficient b, the variances of the different sized plots should be weighted by their respective degrees of freedom. In fact, since the variance estimates for different size of plot, both in uniformity trials and experimental data, are built up from common components, they are frequently highly correlated, so that a simple weighting by degrees of freedom is not accurate. Koch and Rigney point out this difficulty for experimental data, but do not seem to have realized that their arguments apply with equal force to uniformity trial data.

≠ Paper No. 57 of the Agricultural Journal Series of the Rockefeller Foundation.

* Rockefeller Foundation, Agricultural Field Staff, and the Institute of Statistics, Raleigh, N. C.

The present paper presents a method of weighting observed variances of different-sized plots which leads to an unbiased estimate $\underline{b}$ with asymptotically minimum variance. It is applicable both to uniformity trial data and to experimental data; in the latter case the analysis of variance is in effect reconstructed to simulate one derived directly from uniformity trial data, in the manner suggested by Koch and Rigney.

2. Estimation from Uniformity Trial Data

Koch and Rigney showed that a uniformity trial subdivided to simulate a split-plot or lattice design could be analyzed in the manner shown below; a randomized block arrangement could similarly be superimposed on the trial, though it would not provide so much information about the relationship of variability to plot size.

Source	Degrees of Freedom	Mean Square	Expectation of Mean Square
Replications	d-1	V_1	$S + aP + abQ + abcR$
Blocks within replications	d(c-1)	V_2	$S + aP + abQ$
Plots within blocks	cd(b-1)	V_3	$S + aP$
Subplots within plots	bcd(a-1)	V_4	S

The variance of plots the size of a complete replication is V_1, the replication mean square as it appears in the analysis of variance. The variance of plots the size of blocks contains, in addition to the variation due to blocks within replications, that removed by the stratification of groups of blocks into replications in the analysis of variance. Thus the total sum of squares for blocks is

$$d(c-1)V_2 + (d-1)V_1 \quad ,$$

and since there are cd blocks, its mean square is

$$V_2' = (d(c-1)V_2 + (d-1)V_1)/(cd-1) \ .$$

Similarly, the variance between plots over the entire area is

$$V_3' = (cd(b-1)V_3 + d(c-1)V_2 + (d-1)V_1)/(bcd-1)$$

and the variance between subplots over the entire area is

$$V_4' = (bcd(a-1)V_4 + cd(b-1)V_3 + d(c-1)V_2 + (d-1)V_1)/(abcd-1).$$

These formulas are formally identical to those given by Koch and Rigney, who expressed their results in terms of components of variance.

Smith's regression coefficient $\underline{b}$ is defined by the formula

$$\log V_x = \log V - b \log x,$$

where x is the number of units per plot, V is the variance among plots one unit in size, and V_x is the variance of mean per unit area for plots of size x units. For purposes of estimating optimum plot size, the coefficient $\underline{b}$ is alone of interest. In the computations suggested by Koch and Rigney, the values of V_x are obtained by dividing each value of V' by the number of units per replication, block, plot or subplot, thus putting them on a unit basis. According to them, b is given as an unweighted regression coefficient:

$$b_1 = \frac{\sum_j y_j(x_j' - \bar{x}')}{\sum_j (x_j' - \bar{x}')^2} \qquad (1)$$

where $y = \log (V'/x)$, $x' = \log x$, and $\bar{x}' = \sum_j x_j'/n$.

As was pointed out to one of the authors by D. D. Mason, Department of Experimental Statistics, N. C. State College, application of this formula for b_1 sometimes gave results less than -1, which are unacceptable on physical grounds. It was realized that the above estimate (1) would often be inaccurate owing to the equal weighting of y-values of differing variability. It was therefore decided to apply to the different terms in the sums of squares and products defining the regression coefficient, weights that would lead to an estimate of minimum variance. The

appropriate weights are the elements of the inverse of the covariance matrix (i.e. the information matrix) of the values of y. If these elements are designated w_{jk}, the estimate is

$$b_2 = \frac{\sum\sum w_{jk} y_j (x_k' - \bar{x}')}{\sum\sum w_{jk} x_j' (x_k' - \bar{x}')} \qquad (2)$$

$= \frac{U}{T}$ (say), where

$$\bar{x}' = \sum\sum w_{jk} x_j' / \sum\sum w_{jk} \quad .$$

The weights w_{jk} will have to be estimated from the data and will be to that extent inaccurate; but apart from this source of error, the effect of which we do not consider, the estimate will be of minimum variance; this variance is in fact

$$\frac{1}{\sum\sum w_{jk} x_j' (x_k' - \bar{x}')} = \frac{1}{T} \quad . \qquad (3)$$

When, as is often the case, there are more than two variance estimates from which to compute the regression, we may also test the significance of departure from regression. The weighted total sum of squares of the y_j is

$$V = \sum\sum w_{jk} y_j (y_k - \bar{y})$$

where

$$\bar{y} = \sum\sum w_{jk} y_j / \sum\sum w_{jk},$$

with n-1 degrees of freedom, n being the number of variance estimates. The sum of squares attributable to regression on x_j' is

$$U^2/T.$$

Hence the sum of squares for departure from regression is

$$V - U^2/T \ ,$$

which is distributed approximately as χ^2 with n-2 degrees of freedom, and may be

tested accordingly.

It now remains to estimate the weights. Since V_1, V_2, V_3, and V_4 are independent, and their variances are $2V_1^2/(d-1)$, $2V_2^2/d(c-1)$, $2V_3^2/cd(b-1)$, and $2V_4^2/bcd(a-1)$ respectively, it is not difficult to determine the variances and covariances of V_1', V_2', V_3', and V_4', which are linear functions of the former set. In fact, not only the variance of V_1', but also its covariances with the other V_i', are proportional to $V_1 2$.

Likewise the variance of V_2' is estimated as

$$2\left[d(c-1)V_2^2 + (d-1)V_1^2\right] / (cd-1)^2$$

and its covariances with V_3' and V_4' are proportional to this. Thus we find the covariance matrix of the V_i' to be as follows:

$$\begin{bmatrix} \frac{D}{(d-1)^2} & \frac{D}{(d-1)(cd-1)} & \frac{D}{(d-1)(bcd-1)} & \frac{D}{(d-1)(abcd-1)} \\ \frac{D}{(d-1)(cd-1)} & \frac{C+D}{(cd-1)^2} & \frac{C+D}{(cd-1)(bcd-1)} & \frac{C+D}{(cd-1)(abcd-1)} \\ \frac{D}{(d-1)(bcd-1)} & \frac{C+D}{(cd-1)(bcd-1)} & \frac{B+C+D}{(bcd-1)^2} & \frac{B+C+D}{(bcd-1)(abcd-1)} \\ \frac{D}{(d-1)(abcd-1)} & \frac{C+D}{(cd-1)(abcd-1)} & \frac{B+C+D}{(bcd-1)(abcd-1)} & \frac{A+B+C+D}{(abcd-1)^2} \end{bmatrix}$$

where

$$D = 2(d-1)V_1^2$$

$$C = 2d(c-1)V_2^2$$

$$B = 2cd(b-1)V_3^2$$

$$A = 2bcd(a-1)V_4^2$$

The inverse matrix is found to be even simpler in form; as may be verified, it is

$$\begin{bmatrix} (d-1)^2(\frac{1}{C}+\frac{1}{D}) & \frac{-(d-1)(cd-1)}{C} & 0 & 0 \\ \frac{-(d-1)(cd-1)}{C} & (cd-1)^2(\frac{1}{B}+\frac{1}{C}) & \frac{-(cd-1)(bcd-1)}{B} & 0 \\ 0 & \frac{-(cd-1)(bcd-1)}{B} & (bcd-1)^2(\frac{1}{A}+\frac{1}{B}) & \frac{-(bcd-1)(abcd-1)}{A} \\ 0 & 0 & \frac{-(bcd-1)(abcd-1)}{A} & \frac{(abcd-1)^2}{A} \end{bmatrix}$$

The weights for y_j (= log V_j') are obtained by multiplying each row and each column of this inverse matrix by the corresponding V_j'. This result follows from the approximate formula

$$\text{Cov}(\log V_j', V_k') \backsim \text{Cov}(V_j', V_k')/V_j'V_k' \quad .$$

If, as is usual in practical computation, logarithms to base 10 rather than natural logarithms are taken, the weights will need to be multiplied by the factor

$$M^{-2} = (\log_{10} e)^{-2}$$

$$= 5.302.$$

We shall deal here with the transformation to natural logarithms, and indicate the adjustments necessary for common logarithms below.

Thus, from the inverse matrix,

$$\text{Wt}(y_j, y_k) = w_{jk} \backsim V_j'V_k' \text{Wt}(V_j', V_k') \quad .$$

The weights may thus be determined from the inverse matrix without too much difficulty.

It will be found that the sum of the elements of the weight matrix is equal to half the total number of degrees of freedom for the sums of squares from which variance estimates are derived. This may be seen in the following way. If the variances are unaffected by size of plot, then all the available sums of squares are estimates of

the same basic variance. The different estimates of the logarithm of the variance, derived from different lines of the analysis of variance, are independent, and have asymptotic variance equal to twice the reciprocal of the corresponding degrees of freedom. Consequently the information from each is half the degrees of freedom, whence the total information is half the total degrees of freedom.

Thus, for data from uniformity trials, the sum of the weights is

$$\tfrac{1}{2}\,(abcd-1) \quad ,$$

while for data from split-plot experiments, the sum of the weights will be

$$\tfrac{1}{2}\,bc(ad-1) \quad .$$

Similar results may be derived for lattices and other types of experimental design. They provide a convenient check on the computation of the weights.

To determine the regression coefficient and to test the departure from regression, the calculation is best carried out in stages, as follows. Let

$$X_k = \sum_j w_{jk} x'_j$$

and

$$Y_k = \sum_j w_{jk} y_j$$

Then the sum of squares of x' is

$$T = \sum_j X_j x'_j - \Big(\sum_j X_j\Big)^2 \Big/ \sum_j \sum_k w_{jk}$$

similarly the sum of products of y with x' is

$$U = \sum_j X_j y_j - \Big(\sum_j X_j\Big)\Big(\sum_j Y_j\Big) \Big/ \sum_j \sum_k w_{jk}$$

$$= \sum_j Y_j x'_j - \Big(\sum_j X_j\Big)\Big(\sum_j Y_j\Big) \Big/ \sum_j \sum_k w_{jk}$$

and the sum of squares of y is

$$V = \sum_j Y_j y_j - \left(\sum_j Y_j\right)^2 / \sum_j \sum_k w_{jk}$$

$$b_2 = U/T$$

The variance of the estimate b_2 is, to the degree of approximation of the analysis,

$$T^{-1}.$$

Hence, approximate confidence limits for the population regression coefficient β are

$$b_2 \pm tT^{-1/2} \quad ,$$

t being the normal deviate at the required level of probability.

Departure from regression is tested, as indicated above, by means of

$$V - U^2/T$$

which is regarded as χ^2 with n-2 degrees of freedom.

When common logarithms are used, the value of b_2 is determined as above, but its variance is now

$$T^{-1}/5.302$$
$$= 0.1886\ T^{-1} \quad ,$$

and the corresponding confidence limits are

$$b_2 \pm 0.4343tT^{-1/2}$$

The sum of squares for departure from regression is also altered, to

$$5.302(V - U^2/T).$$

It should be observed that the key to these computations is the covariance matrix of the variances V_i' of the plots of different sizes. Because these variances are expressed as linear combinations of the original mean squares, which are independent, and not in terms of the variance components, which are correlated with one another, the resulting covariance matrix, and its inverse, take on a relatively

simple form.

3. Estimation from experimental data

When variance components are to be estimated from experimental data, the estimates are calculated in the same way as from uniformity trial data. However, since a number of comparisons are given over to the estimation of treatment effects, the different plot and block variances are estimated with fewer degrees of freedom, and hence less precision, than they could have been in a uniformity trial. Apart from this complication, for which allowance must be made in determining the weights for the various components, the determination of a linear unbiased estimate with asymptotic minimum variance follows the same lines as that given in the previous section. The method is illustrated by the analysis for a split-plot experiment in the form given by Koch and Rigney. It will be noted that, in this model, it is assumed that block-treatment interactions do not exist.

	Degrees of freedom	Mean Square	Expectation of mean square
Replications	d-1	V_1	S + aP + abQ + abcR
Treatments (1)	c-1		S + aP + abQ + treatment effects
Error (1)	(c-1)(d-1)	V_2	S + aP + abQ
Total between whole plots	cd-1		
Treatments (2) and interactions	c(b-1)		S + aP + treatment effects
Error (2)	c(b-1)(d-1)	V_3	S + aP
Split-plots	cd(b-1)		
Sampling error	bcd(a-1)	V_4	S

As for a uniformity trial, the estimated variance of plots the size of a complete replication is V_1. Since it is estimated with the full d-1 degrees of

freedom, its variance is as given in the previous section.

In estimating the mean square for blocks (i.e. whole plots) we must allow for the fact that, of the d(c-1) comparisons between blocks within replications, only (c-1)(d-1) are available for estimating the variance, the other c-1 containing treatment effects. Thus, as before, the estimated variance between blocks is

$$V_2' = (d(c-1)V_2 + (d-1)V_1)/(cd-1) \quad ,$$

but its estimated variance is now increased to

$$2[d^2(c-1)V_2^2/(d-1) + (d-1)V_1^2]/(cd-1)^2 \quad .$$

The variance of V_3 has similarly to be adjusted by a factor $\frac{d}{d-1}$.

The analysis now proceeds as for uniformity trials, and the inverse matrix is as given above, provided we redefine

$$D = 2(d-1)V_1^2$$

$$C = 2d^2(c-1)V_2^2/(d-1)$$

$$B = 2cd^2(b-1)V_3^2/(d-1)$$

$$A = 2bcd(a-1)V_4^2$$

4. Numerical example

The computations required in the proposed method are illustrated in a numerical example, the data for which, set out in Table 1, were kindly furnished by D. D. Mason.

Table 1

Soybean Yield Trial Conducted by C. A. Brim,
U. S. Department of Agriculture, at Willard, North Carolina, 1956.

Source	Degrees of Freedom	Mean Square
Replications	2 = d-1	452 = V_1
Varieties	11 = c-1	30,401
Experimental Error	22 = (d-1)(c-1)	10,589 = V_2
Rows in Plots	36 = cd(b-1)	5,938 = V_3
Subplots in Rows	72 = bcd(a-1)	2,862 = V_4

Here a = 2, b = 2, c = 12, d = 3.

In the determination of the weights we can work with multiples of the V_j' more conveniently than with the V_j' themselves. This device makes the computations simpler as well as more accurate. We have

$$(d-1)\ V_1' = 2\ V_1 = 904$$
$$(cd-1)\ V_2' = 35\ V_2' = 2\ V_1 + 33\ V_2 = 350341$$
$$(bcd-1)\ V_3' = 71\ V_3' = 2\ V_1 + 33\ V_2 + 36\ V_3 = 564109$$
$$(abcd-1)\ V_4' = 143\ V_4' = 2\ V_1 + 33\ V_2 + 36\ V_3 + 72\ V_4 = 770173$$

This gives

$$V_1' = 452$$
$$V_2' = 10010$$
$$V_3' = 7945$$
$$V_4' = 5386$$

The number of units per plot corresponding to the different-sized plots the variances of which are V_1', V_2', V_3', and V_4' are 48, 4, 2, and 1 respectively. Putting the variances V' on a unit basis and taking logarithms, we obtain the values given in Table 2.

Table 2

Logarithms of Relative Plot Sizes (x') and Unit Variances (y)

x'	y
0.0000	3.7313
0.3010	3.5891
0.6021	3.3984
1.6812	0.9739

The unweighted regression coefficient is then

$$b_1 = \left[\sum x'y - \frac{(\sum x')(\sum y)}{n}\right] / \left[\sum (x')^2 - \frac{\sum (x')^2}{n}\right]$$

$$= \frac{4.7638 - 7.5544}{3.2796 - 1.6697}$$

$$= -1.7334$$

As Koch and Rigney point out, b is an index of soil variability; it should vary between zero and minus one. A value of zero indicates perfect correlation (extreme uniformity) among the units making up a plot; a value of minus one indicates no correlation. Clearly the value obtained in the present example can have no unambiguous physical interpretation. Here it is apparent that weighting a low mean square for replications based on only two degrees of freedom, equally with others based on many more degrees of freedom, has led to an unreasonable estimate of soil variability.

Using the method proposed in the present paper, y and x' are as before. The weights are the elements w_{jk} of the information matrix of the y. To obtain these numbers it is convenient first to calculate

$$A = 2bcd(a-1)V_4^2 = 1,179,500,000$$

$$B = 2cd^2(b-1)V_3^2/(d-1) = 3,808,100,000$$

$$C = 2d^2(c-1)V_2^2/(d-1) = 11,100,600,000$$

$$D = 2(d-1)V_1^2 = 800,000$$

then

$$w_{11} = \left[(d-1)V_1'\right]^2 \left(\frac{1}{C}+\frac{1}{D}\right) = 1.00$$

$$w_{12} = w_{21} = -\left[(d-1)V_1'\right]\left[(cd-1)V_2'\right]/C = -0.03$$

$$w_{13} = w_{31} = w_{14} = w_{41} = 0$$

The remaining elements are computed in similar fashion. The completed information matrix is

$$W = \begin{bmatrix} 1.00 & -0.03 & 0.00 & 0.00 \\ -0.03 & 43.29 & -51.90 & 0.00 \\ 0.00 & -51.90 & 353.36 & -368.34 \\ 0.00 & 0.00 & -368.34 & 502.89 \end{bmatrix}$$

the sum of whose elements is $\frac{1}{2}\,bc(ad-1) = 60$, as may be verified.

We now compute the set

$$Y_k = \sum_j w_{jk} y_j \quad .$$

Thus

$$Y_1 = (1.00)(0.9739) - (0.03)(3.3984) = 0.87 \quad ;$$

similarly $Y_2 = -39.19$,

$Y_3 = -282.52$,

$Y_4 = 554.42$.

$\sum_j Y_j = 233.58$

In the same way we compute the X_k :

$X_1 = 1.66$

$X_2 = 10.39$

$X_3 = 75.11$

$X_4 = -110.87$

$\sum_j X_j = -23.71$

Then

$$T = \sum_j X_j x_j' - \left(\sum_j X_j\right)^2 / \sum_j \sum_k w_{jk}$$

$$= 31.65 - (-23.71)^2/60$$

$$= 22.28$$

Similarly $U = -14.87$

and $V = 13.05$

$$b_2 = -14.87/22.28$$

$$= -0.667$$

$$V(b_2) = 0.1886/22.28$$

$$= 0.00846$$

Standard error = 0.092

As a matter of interest, the variance of b_1 was also determined. This variance is given by

$$\frac{\sum_j \sum_k w^{jk}(x_j' - \bar{x}')(x_k' - \bar{x}')}{\left[\sum_j (x_j' - \bar{x}')^2\right]^2}$$

where the w^{jk} are the elements of the inverse of the weight matrix; in other words, they are the elements of the covariance matrix of the y_j's. Here $\bar{x}'$ is the unweighted mean of the x_j'.

The variance of b_1 was found to be 0.1644, giving a standard error of 0.406.

Thus the efficiency of this estimate is

$$\frac{0.00846}{0.1644} = 5 \text{ per cent.}$$

With such a large standard error, any estimate of a quantity lying between 0 and 1 is of little value.

To test departure from regression, we have

$$\chi^2_{(2)} = 5.302\ (V-U^2/T)$$
$$= 5.302 \times 3.125$$
$$= 16.57$$

Since this value exceeds the 1 per cent point of the χ^2 distribution, the data depart significantly from the assumed linear relationship.

5. Allowance for departure from empirical relationship

Departure from linearity (i.e., from the empirical law of Fairfield Smith) may cause concern in some examples. In such cases, provided cost data are available, optimum plot size may be estimated with reasonable accuracy without the assumption that the empirical law holds.

Suppose the cost of r replications is

$$r(K_1+K_2x) \quad ,$$

where K_1 is the cost of a plot (regardless of size), K_2 is the cost per unit of plot and x is the number of units per plot.

We then require to minimize V_x/r, subject to the condition that $r(K_1+K_2x)$ be fixed. This is equivalent to minimizing

$$F(x) = (K_1+K_2x)V_x$$

with respect to x. If F(x) can be determined from experimental data for a few values of x, its minimum may be fairly easily determined graphically.

Example 2

Johnson and Hixon (1952) have reported a 100 per cent cruise of 40 acres of old-growth Douglas fir timber in Oregon. The data consist of timber volume on each of 1600 1/40-acre plots in a 40 x 40 square. The analysis of variance, with stratification to eliminate systematic variation between sets of 8 rows and sets of 8

columns, has been worked out in the manner shown in the table below:

Source	Degrees of Freedom	Mean Square
Among 1.6-acre plots	16	277106 = V_1
Among 0.4-acre plots in 1.6-acre plots	75	93947 = V_2
Among 0.1-acre plots in 0.4-acre plots	300	73012 = V_3
Among 0.025-acre plots in 0.1-acre plots	1200	100744 = V_4

The large mean square among 0.025-acre plots indicates competitive effects. Since the average diameter of trees measured was 45 inches, such a result is hardly surprising.

Here

		Number of 0.025-acre units per plot
V_1'	= 277106	64
V_2'	= 138349	16
V_3'	= 89223	4
V_4'	= 97869	1

When the values of the V' are adjusted to a unit basis and plotted, departures from linearity appear serious.

Johnson and Hixon also estimate the number of plots of different sizes which can be measured in a four-hour cruise of 40 acres. Converting these data to mean minutes per plot for plots of different sizes, we obtain

Kind of Plot	Number of 0.025-acre Units (x)	Mean Minutes per Plot (t)
$\frac{1}{2}$ X 1 chain	2	7.50
$\frac{1}{2}$ X 2 chains	4	11.10
$\frac{1}{2}$ X 4 chains	8	16.14
$\frac{1}{2}$ X 6 chains	12	22.16

Assuming cost per plot (in minutes) to be linearly related to size of plot, we may write

$$T = K_1 + K_2 x ,$$

where T is total cost per plot (measured in minutes per plot)

K_1 is a constant (measured in minutes per plot)

K_2 is a constant (measured in minutes per unit area)

x is the number of 0.025-acre units per plot.

From the data given above we obtain

$$T = 4.9 + 1.43 x$$

Thus we may compute F(x)

x (number of 0.025-acre units)	$K_1 + K_2 x$	$V'/x = V_x$	$F(x) = V_x (K_1 + K_2 x)$
1	6.33	97869	619500
4	10.62	22306	236900
16	27.78	8647	240200
64	96.42	4330	417500

Plotting F(x) as a function of x we find that its minimum occurs between x = 4 and x = 16 units. It is suggested that in this region departures from linearity in the relation

$$\text{Log } V_x = \log V_1 - b \log x$$

will not be serious. Hence b is given approximately by