

ON THE POSSIBILITY OF EXPLOITING INTRA-ROW CORRELATIONS TO
ESTIMATE GENETIC AND ENVIRONMENTAL VARIANCE IN A PLANTATION

by

BU-401-M

D. S. Robson
Cornell University

December, 1971

Abstract

Uniformity analysis of a genetically heterogeneous plantation has been used as a technique for separating total variance into genetic and environmental components. By superimposing a plot structure on a plantation of **randomly** allocated genotypes and calculating an analysis of variance between plots and within plots, Shrikhande (1957) showed that application of H. F. Smith's empirical variance law to the "between plots" mean square results in identifiability of the two variance components. An alternative procedure consists of exploiting the regular relationship between environmental correlation of two plants in the same row and the intervening distance. Various empirical studies have shown that in uniformity trials the correlation ρ_k between the first and the k^{th} plant in a row decreases as $\rho_k = \rho^k$. Such a model for environmental correlation (with no correlation between the first and k^{th} genotype) results in identifiability of genetic and environmental components of variance.

ON THE POSSIBILITY OF EXPLOITING INTRA-ROW CORRELATIONS TO
ESTIMATE GENETIC AND ENVIRONMENTAL VARIANCE IN A PLANTATION

by

BU-401-M

D. S. Robson
Cornell University

December, 1971

Introduction

If genetically heterogeneous plants are grown in a regular plantation with random assignment of genotypes to the given planting sites, then for any plant trait X and any genotype g the following identity obtains

$$X = \mathcal{E}(X|g) + [X - \mathcal{E}(X|g)]$$

where \mathcal{E} denotes the operation of averaging over all possible random assignments to planting sites. Letting $G = \mathcal{E}(X|g)$ and $E = X - G$ then $\mathcal{E}(E|G) = 0$; hence E and G are uncorrelated so that

$$\sigma_X^2 = \sigma_G^2 + \sigma_E^2 .$$

Superficially, the measurement of X for each plant would not appear to provide enough information to identify the two components of σ_X^2 separately. This view, however, neglects the information in the order of the X measurements; in usual practice this collection of measurements from a rectangular plantation is recorded in a corresponding rectangular array. A basic premise of field plot experimentation is that plants which are growing near to one another share similar environments and hence tend to grow alike; thus, variability within a small cluster of entries in the data matrix will tend to be less than the total variance $\sigma_G^2 + \sigma_E^2$.

V. J. Shrikhande (1) ingeniously noted that only the environmental contribution to variance will be reduced by this intra-cluster correlation, and that the amount of reduction can be described by H. Fairfield Smith's empirical law (2) relating variance to plot (cluster) size. Equivalently, the variance among means of clusters of size k is correspondingly greater than $(\sigma_G^2 + \sigma_E^2)/k$, and applying Smith's law Shrikhande concluded that the between cluster mean square must estimate $k(\sigma_G^2/k + \sigma_E^2/k^b) = \sigma_G^2 + k^{1-b}\sigma_E^2$, where $0 < b < 1$ is an unknown constant to be estimated from the data. The procedure for estimating b by calculating mean squares for clusters of several different sizes (k) simultaneously provides estimates of σ_G^2 and σ_E^2 . Other authors (3, 4) have followed Shrikhande's example in successfully applying this procedure to tree plantations. Here we offer the alternative possibility of empirically fitting the intra-row correlation as a function of the number of spacings between two plants in order to arrive at estimates of σ_G^2 and σ_E^2 .

Intra-row correlations

If k consecutive plants in a row are numbered 1, 2, ..., k then the observed variance between the i^{th} and j^{th} plant is an estimate of

$$s\left\{\frac{(X_i - X_j)^2}{2}\right\} = \sigma_G^2 + \sigma_E^2(1 - \rho_{ij})$$

where ρ_{ij} is an intra-row correlation between the i^{th} and j^{th} plant. We assume that this correlation depends only on the distance between the two plants and not on their particular location, thus implying that these correlations can be indexed by the single subscript $|i - j|$ instead of the pair (i, j) . For plants h units apart in the row we may therefore write

$$s\left\{\frac{(X_{i+h} - X_i)^2}{2}\right\} = \sigma_G^2 + \sigma_E^2(1 - \rho_h)$$

which is independent of i .

This variance, as a function of h , should approach $\sigma_G^2 + \sigma_E^2$ as h gets large and should approach σ_G^2 as h approaches zero. Calculating such variance estimates for many values of h and plotting against h should thus produce a graph from which estimates of $\sigma_G^2 + \sigma_E^2$ and σ_G^2 can be empirically extrapolated. Empirical evidence from a variety of crop plants indicates that ρ_h does decrease to zero, and several investigators have independently shown that the particular model $\rho_h = \rho^h$ provides a good fit (5, 6, 7).

Earlier investigations into the correlation function have been motivated primarily by the objective of determining optimal size and shape of plot, and hence each plant was considered to be correlated with neighboring plants at different distances in all directions. Correlations at a fixed distance were found to vary with direction, as might be expected merely from consideration of the fact that plants derive their energy directly from sunlight, and in some directions one plant may fall in the shadow of the other. While this complication does require careful consideration in determining plot size and shape for field experiments, and also in designing plantations, there is little reason to include such considerations in the present context where the primary objective is to obtain estimates of σ_G^2 and σ_E^2 . For this purpose we need consider only correlations within rows; if the geometric design of the plantation does define rows in several directions then we have the option of analyzing each direction separately and combining the resulting estimates of σ_G^2 and σ_E^2 obtained from these separate analyses.

As Shrikhande has pointed out in his application of Smith's variance law, both the within clusters and between clusters mean squares are affected by the intra-cluster correlation, and both mean squares could be employed in the estimation of σ_G^2 and σ_E^2 . Analogous results obtain with the present approach in that

these two variances are expressible as functions of σ_G^2 , σ_E^2 , and

$$\bar{R}_k = \frac{1}{\binom{k}{2}} \sum_{i=1}^{k-1} i \rho_{k-i};$$

thus, the variance among k consecutive plants in a row is

$$\sigma_k^2 = \sigma_G^2 + \sigma_E^2(1 - \bar{R}_k)$$

and the variance among means of k consecutive plants is

$$\begin{aligned} \sigma_{\bar{X}_k}^2 &= \frac{1}{k} \left[\sigma_G^2 + \sigma_E^2 \{1 + (k-1)\bar{R}_k\} \right] \\ &= \frac{\sigma_k^2}{k} + \sigma_E^2 \bar{R}_k. \end{aligned}$$

The difference, or the "between cluster" component of variance

$$\sigma_{\bar{X}_k}^2 - \frac{\sigma_k^2}{k} = \sigma_E^2 \bar{R}_k$$

is thus independent of σ_G^2 , but like σ_k^2 and $\sigma_{\bar{X}_k}^2$ this difference depends on a linear function of the $k-1$ correlations $\rho_1, \rho_2, \dots, \rho_{k-1}$. To eliminate this awkward feature we introduce an operator ψ defined on a sequence f_1, f_2, f_3, \dots

$$\psi(f_k) = \binom{k+1}{2} f_{k+1} - 2 \binom{k}{2} f_k + \binom{k-1}{2} f_{k-1}$$

which gives

$$\psi(\sigma_k^2) = \sigma_G^2 + \sigma_E^2(1 - \rho_k)$$

and

$$\psi(\sigma_E^2 \bar{R}_k) = \sigma_E^2 \rho_k.$$

If the earlier empirical evidence is generally valid then a regression analysis of

$\log \psi(\hat{\sigma}_{\bar{X}_k}^2 - \frac{1}{k} \hat{\sigma}_k^2)$ should provide an estimate of the equation

$$\log \psi(\sigma_E^2 \bar{R}_k) = \log \sigma_E^2 + k \log \rho$$

and then

$$\hat{\sigma}_G^2 = \hat{\sigma}_X^2 - \hat{\sigma}_E^2.$$

In any event, empirical extrapolation to $k = 0$ should provide an estimate of σ_E^2 and hence of $\sigma_G^2 = \sigma_X^2 - \sigma_E^2$.

Remarks

Consideration of these methods of exploiting empirical laws concerning intra-cluster (or row) correlations in order to separate genetic and environmental variances raises several questions concerning the earlier work where these empirical laws were derived. Thus, Smith's law

$$\sigma_{X_k}^2 = \sigma_X^2 / k^b$$

becomes

$$\sigma_{X_k}^2 = \frac{\sigma_G^2}{k} + \frac{\sigma_E^2}{k^b} \neq (\sigma_G^2 + \sigma_E^2) / k^b$$

when genetic variability is present, and the question then arises whether the many experimental studies testing Smith's law all involved genetically homogeneous plants. It is true that the genetic component can be statistically eliminated by a variance component analysis:

<u>Source</u>	<u>d.f.</u>	<u>M.S. Expectation</u>
Between clusters	$c-1$	$\sigma_W^2 + k\sigma_B^2 = \sigma_W^2 + k \frac{k^{1-b} - 1}{k - 1} \sigma_E^2$
Within clusters	$c(k-1)$	$\sigma_W^2 = \sigma_G^2 + \frac{k}{k - 1} (1 - k^{-b}) \sigma_E^2$

since the "between clusters" variance component

$$\sigma_B^2 = \frac{k^{1-b} - 1}{k - 1} \sigma_E^2$$

does not depend on σ_G^2 . Earlier investigators, not mindful of the contaminating effect of genetic variance, however, probably did not bother estimating this "between clusters" variance component and attempting to fit this more complicated function of k,

$$\sigma_B^2 = \frac{k^{1-b} - 1}{k - 1} \sigma_E^2 ,$$

but instead treated the "between clusters" mean square as an estimate of σ_X^2/k^b .

Analogous questions arise concerning the earlier work testing the empirical relation $\rho_k = \rho^k$. If genetic variability is present then an analysis of variance produces:

<u>Source</u>	<u>d.f.</u>	<u>Mean Square Expectation</u>
Between clusters	c-1	$\sigma_W^2 + k\sigma_B^2 = \sigma_W^2 + k(\sigma_E^2 \bar{R}_k)$
Within clusters	c(k-1)	$\sigma_W^2 = \sigma_G^2 + \sigma_E^2(1 - \bar{R}_k)$

and, again, earlier investigators would have had to base their analysis of correlation on

$$\sigma_B^2 = \sigma_E^2 \bar{R}_k$$

if genetic variability was present in their "uniformity trial".

As in all genetic variance component analysis, scaling to eliminate heterogeneity of environmental variance is an important consideration in the present analysis. If each genotype g generates a different environmental variance $\sigma_{E \cdot G}^2$ then correlations between environmental effects must also be expected to depend on genotypes and ρ_k is then, at best, a weighted average of such correlations.

If for each pair of genotypes the environmental correlation obeys some regular law such as $\rho_k = \rho^k$, a weighted average of correlations will not follow a recognizable law. The same restriction holds with respect to H. F. Smith's variance law.

These empirical methods provide only estimates of total genetic variance and are thus unaffected by linkage, epistasis, or other genetic factors which complicate many statistical genetic procedures. Heritability defined as $\sigma_G^2 / (\sigma_G^2 + \sigma_E^2)$ is of very limited usefulness, however, in the context of predicting gains due to selection.

References

1. Shrikhande, V. J. (1957). "Some considerations in designing experiments on coconut trees." J. Ind. Soc. Agri. Stat. 9:82-98.
2. Fairfield Smith, H. (1938). "An empirical law describing heterogeneity in the yield of agricultural crops." J. Agr. Sci. 28:1-23.
3. Kan-Ichi Sakai and Suekichi Hatakeyama (1963). "Estimation of genetic parameters in forest trees without raising progeny." FAO Conference on Forest Genetics and Tree Improvement, Stockholm. Limited dist.
4. Kedharnoth, S. et.al (1969). "Estimation of genetic parameters in teak without raising progeny." Ind. For. 95:238-245.
5. Osborne, J. G. (1942). "Sampling errors of systematic and random surveys of cover-type areas." J. Amer. Stat. Assoc. 37:256-264.
6. Matern, B. (1947). "Methods of estimating the accuracy of line and sample plot surveys." In Swedish, English resume, pp. 118-136. Meddelanden fran Statens Skogsforskningsinstitut, Band 36:1.
7. Taylor, H. L. (1948). "An examination of the effect of plot shape on experimental error." M.S. Thesis, Iowa State Univ.