Genetics and Molecular Biology, 21, 4, 551-555 (1998)

METHODOLOGY

A new method to estimate genetic gain in annual crops*

Flávio Breseghello, Orlando Peixoto de Morais and Paulo Hideo Nakano Rangel

ABSTRACT

The genetic gain obtained by breeding programs to improve quantitative traits may be estimated by using data from regional trials. A new statistical method for this estimate is proposed and includes four steps: a) joint analysis of regional trial data using a generalized linear model to obtain adjusted genotype means and covariance matrix of these means for the whole studied period; b) calculation of the arithmetic mean of the adjusted genotype means, exclusively for the group of genotypes evaluated each year; c) direct year comparison of the arithmetic means calculated, and d) estimation of mean genetic gain by regression. Using the generalized least squares method, a weighted estimate of mean genetic gain during the period is calculated. This method permits a better cancellation of genotype x year and genotype x trial/year interactions, thus resulting in more precise estimates. This method can be applied to unbalanced data, allowing the estimation of genetic gain in series of multilocational trials.

INTRODUCTION

Genetic gain estimates in breeding programs are important to critically analyze efficiency and to plan new actions and strategies. Institutions working with annual crop breeding routinely conduct a series of trials to compare elite lines and to release new cultivars. Each year lines are replaced in the trials with the expectation that the new ones will be superior. The change in yield mean as a consequence of these substitutions may be considered an estimate of genetic gain. This approach provides information based on data already available from different years and locations without additional costs.

To evaluate maize breeding efficiency in Brazil, Vencovsky *et al.* (1988) analyzed 20 years of trials. Gain was estimated for each pair of consecutive years as being the variation of annual means minus the variation of the means for the lines common to the two years. Later, Toledo *et al.* (1990) used the method of Vencovsky *et al.* (1988) to calculate genetic gain obtained by soybean (*Glycine max*, Merr.) breeding in Paraná State. They calculated mean genetic gain by the weighted least squares method to avoid cancellation of information obtained in intermediate years. Rodrigues (1990) estimated the variances of annual gains and the covariances of consecutive gains from the number of treatments shared by each pair of years, to obtain a covariance matrix that was used in the generalized least squares method to obtain mean and variance estimates.

Soares (1992) estimated genetic gain for rice (*Oryza sativa* L.) breeding in the State of Minas Gerais using, in addition to the method of Vencovsky *et al.* (1988), a method based on the behavior of standard checks, i.e., those present every year of the series. In this method, the deviation between the mean for the lines and the mean for the standard checks is calculated for each year. The estimate of genetic progress is the linear regression coefficient of these deviations in relation to years.

All of these studies were developed to calculate genetic gain per location by analysis of series of balanced data. However, when a series of many years of trials conducted over a large geographic region is analyzed, the data may be unbalanced due to differences between the sets of treatments, number of replications, missing plots, changes in test locations, etc.

The objective of the present study was to propose

*Part of a thesis presented by F.B. to the Universidade Federal de Goiás, Goiânia, GO, in partial fulfillment of the requirements for the Master's degree.

Embrapa Arroz e Feijão,Caixa Postal 179, 75375-000 Santo Antônio de Goiás, GO, Brasil. Send correspondence to F.B.

a new, generalized methodology to estimate genetic gain of quantitative traits, grain yield in particular, using series of balanced or unbalanced data originating from networks of multiple location trials.

MATERIAL AND METHODS

General case

Let us consider a series of data from a certain number of genotypes, each one tested during one or several years of a given period of time, with variable number of locations/year and replications/genotype/location/year. These data can be described by the following simplified mathematical model:

$$Y_{ijkr} = \mu + A_k + L/A_{jk} + R/L/A_{jkr} + G_i + \varepsilon_{ijkr},$$

where Y_{ijkr}: r^{th} observation ($r = 1, ..., s_{ijk}$) of genotype i at location j during year k; μ: constant associated with observation Y_{ijkr}; A_k: effect of year k ($k = 1, ..., a$); L/A_{jk}: effect of location j within year k ($j = 1, ..., m_k$); $R/L/A_{jkr}$: effect of replication r within location j for year k ($r = 1, ..., s_{jk}$); G_i: effect of genotype i ($i = 1, ..., n$), and ε_{ijkr}: error associated with observation Y_{ijkr}, considered to be independent and normally distributed, with a null mean and a common variance σ_e^2.

The interaction effects related to genotype x year and genotype x location/year should be excluded from the model, i.e., they should be considered as components of experimental error. Thus, the marginal means for each genotype, adjusted for effect of year, location/year and replication/location/year, will represent estimable functions (Searle, 1971). The analysis of variance scheme is presented in Table I.

Analysis of variance should be performed using a procedure compatible with the structure of unbalanced data, such as GLM of the SAS statistical package (SAS Institute Inc., 1985) that provides the least square solutions of the vector of parameters θ° and the generalized inverse $(X'X)^G$ of X'X, where X is the matrix of the coefficients of parameters in the model.

Table I - Scheme used for joint analysis of variance of Y_{ijkr} observations.

Sources of variation	d.f.	MS	E (MS)
Year	$a - 1$	Q_1	–
Location/year	$\sum_{k=1}^{a} (m_k - 1)$	Q_2	–
Replication/location/year	$\sum_{k=1}^{a} \sum_{j=1}^{m_k} (s_{jk} - 1)$	Q_3	–
Genotype (adjusted)	$n - 1$	Q_4	$\sigma_e^2 + c_1 \phi_G$
Residue	$\sum_{k=1}^{a} \sum_{j=1}^{m_k} \sum_{i=1}^{n_{ijk}} (s_{ijk} - s_{jk}) - n + 1$	Q_5	σ_e^2

The vector Y, of the adjusted genotype means ($\overline{Y}_i...$), is obtained as follows:

$$\hat{Y} = C\,\theta^\circ,$$

where C: coefficient matrix, $n \times (1 + a + \sum_{k=1}^{a} m_k + \sum_{k=1}^{a} \sum_{j=1}^{m_k} s_{jk} + n)$, which can be represented by:

μ	— years —			— loc/year —			— repl/loc/year —			— genotype —			
1	1/a	...	1/a	$1/am_1$	...	$1/am_a$	$1/am_1s_{11}$	...	$1/am_as_{m_aa}$	1	0	0	... 0
1	1/a	...	1/a	$1/am_1$	...	$1/am_a$	$1/am_1s_{11}$	...	$1/am_as_{m_aa}$	0	1	0	... 0
1	1/a	...	1/a	$1/am_1$	...	$1/am_a$	$1/am_1s_{11}$	...	$1/am_as_{m_aa}$	0	0	1	... 0
.	.		.	.		.	.		.	.	.	.	.
.	.		.	.		.	.		.	.	.	.	.
.	.		.	.		.	.		.	.	.	.	.
1	1/a	...	1/a	$1/am_1$	...	$1/am_a$	$1/am_1s_{11}$	...	$1/am_as_{m_aa}$	0	0	0	... 1

In the previous expression, $\hat{Y} = C\,\theta^\circ$, and in others to follow, the sign "^" indicates that it is an estimator. The covariance matrix of the adjusted means, V(Y), is given by the expression:

$$\hat{V}(\hat{Y}) = C\,(X'X)^G\,C'\,RMS,$$

where RMS = residual mean square estimator.

Special cases

In some cases the data for each available trial are balanced, i.e., all treatments have the same number of replications for each experiment, with no missing plots. Two different situations are possible:

Experiments with the same number of replications

If $r_{jk} = r_{j'k'}$ for any j and k value, it is possible to use the mean for the lines at the trial level for analysis, i.e., $\overline{Y}_{ijk\bullet}$. Thus, it is possible to use trial information where there is only mean information. Furthermore, there is an additional advantage of a substantial reduction in computational resource requirements given by elimination of the source of variation due to replication/location/year, with a consequent reduction in the X'X matrix.

However, a disadvantage of this alternative is a reduction in estimate precision. The residual sum of squares loses its component due to the interaction between genotypes and replication/location/year, usually the one of lowest magnitude and of highest weight (larger number of degrees of freedom). However, when the residual mean squares of each assay (MSR_{jk}) and their respective degrees of freedom are available, the residual of joint analy-

sis can be easily corrected, and the precision of the original analysis can be recovered.

Experiments with different numbers of replications ($r_{jk} \neq r_{j'k'}$ for at least one j or j' value and any k or k' values)

In this case, it is also convenient to use the $\overline{Y}_{ijk\bullet}$ means, especially when the advantages mentioned in the previous case are relevant. Considering that in this case the means are not formed by an equal number of replications, it is necessary to repeat each mean as many times as the number of replications that produced it, i.e., s_{ijk} times in order to recover the real adjusted treatment means. In this case the replications should not be included in the model because the delineation of each trial is considered to be fully randomized.

By repeating each $\overline{Y}_{ijk\bullet}$ mean s_{ijk} times, the same number of degrees of freedom is recovered for the residual source of variation that would be obtained if the original Y_{ijkr} information was used, but replication of the means does not contribute to the residual sum of squares since $Y_{ijkr} = Y_{ijkr'} = \overline{Y}_{ijk\bullet}$. If the sums of squares of residuals for each trial are available, the residual sum of squares of joint analysis can be corrected. If this information is not available, one should use only the number of degrees of freedom corresponding to genotype x year and genotype x location/year interactions. This, however, reduces the power of the method in identifying significant estimates.

Genetic gain estimate

Let $\overline{Y}_{i}...$ be the mean for genotype i, adjusted for year and trials/year, obtained from the combined analysis of the entire period studied. Let Y_k^* be the arithmetic mean of the adjusted $\overline{Y}_{i}...$ means exclusively for the genotypes tested during year k. The Y_k^* mean is an estimate of the mean grain yield of the set of lines tested during year k, which recovers the information obtained in other years k', with $k' \neq k$.

To calculate the Y_k^* means and their variances and covariances, it is helpful to construct an auxiliary matrix S with a x n dimensions, in which each row refers to one year k and each column to one genotype i. If genotype i was evaluated during year k, the *ki* cell is filled with $1/n_k$, where n_k is the number of genotypes evaluated during year k; if the genotype was not in the trials conducted during year k, the value of the cell is zero.

The Y^* column vector of the Y_k^* means is obtained by the following equation:

$$\hat{Y}^* = S\hat{Y}.$$

Genetic gain over two years, consecutive or not, can be estimated by the difference in the respective Y_k^* means. To estimate the weighted mean genetic gain it is necessary to obtain a matrix of covariances of the Y_k^* means, with a x a dimensions, given by:

$$\hat{V}(\hat{Y}^*) = S\,\hat{V}(\hat{Y})\,S'.$$

Mean annual genetic gain is estimated by the linear regression coefficient b_1 of Y_k^* as a function of year k, which is obtained by the generalized least squares method (Hoffmann and Vieira, 1987).

$$\hat{\beta} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \end{bmatrix} = (x'\,D^{-1}x)^{-1}(x'\,D^{-1}\hat{Y}^*)$$

where $\hat{b}_0$: intercept; $\hat{b}_1$: linear regression coefficient for Y_k^* means as a function of year, weighted estimate of mean annual genetic gain; x: known constant matrix with a x 2 dimensions consisting of a column vector of 1's, relative to b_0 and a column vector of 1, 2, ..., a, relative to b_1, and $D = \hat{V}(\hat{Y}^*)$ estimator of the covariance matrix of Y_k^* means.

The b_1 variance is the value of the cell in the second row, second column of the V(ß) matrix, which is estimated by:

$$\hat{V}(\hat{\beta}) = (x'\,D^{-1}x)^{-1}$$

If the estimate of genetic gain is significant, it is interesting to calculate the percent mean annual gain, using as a base the mean referring to the genotypes tested during the first year of the considered period, Y_1^*.

To illustrate the application of the proposed method, genetic gain was calculated from regional trial data of irrigated rice. The data was the result of nine yield trials of lines and varieties of irrigated rice in the states of Piauí and Maranhão from 1986 to 1990. Table II presents the means for the genotypes in each trial. Only part of the trial results was used. Data were selected only to provide a simple example. Thus, the results obtained here should be considered as hypothetical. This example follows the method proposed for general cases. The example has no missing plots, but there is a trial with a smaller number of replications. Therefore, this example could also be applied to the second special case cited simply by using the data presented in Table II.

RESULTS AND DISCUSSION

The example's combined analysis of variance results are presented in Table III. This combined analysis, which was conducted according to the GLM procedure of the SAS program, provided the generalized inverse $[(X'X)^G]$, the least squares solution vector (θ°), and the residual mean square (RMS).

In this example case, the coefficient matrix C had dimensions 29 x 79 and can be partially represented as follows:

Table II - Mean genotype production (kg/ha) in each trial and adjusted mean for the entire period.

Genotypes	Year 1	Year 2		Year 3		Year 4	Year 5			Adjusted mean
	Trial 1	Trial 1	Trial 2	Trial 1	Trial 2	Trial 1	Trial 1	Trial 2	Trial 3	
CNA 4899				7174	7338		7206	7001	5186	6600
CNA 3459				7916	6762		7479	8009	5781	7008
CNA 5191				8005	6338	7336	7775	7729	4210	6705
CNA 3762						6432				6175
CNA 4893				6710	5985	6299	7284	6469	4319	5984
CNA 3887		7151	6231	6746	5858	7120	8130	7705	5587	6766
CNA 3739	6494	5805	5483	6929	5289		6915	7860	4833	6219
CNA 5394				7560	4879	6395	7103	7085	4527	6064
CNA 5383							5742	6672	5909	6003
CNA 5247							7303	7120	5644	6585
CNA 5244							6786	7599	5733	6601
CNA 5719							7928	7797	4417	6609
CNA 4223		5525	2568	7228	7337				4377	5867
CNA 3888							7960	8515	4767	6976
CICA8	6411	6625	6171	7364	6252	6584	7727	7874	5161	6666
METICA 1	6284	7055	6319	7266	7024	6617	6797	6445	5429	6555
CNA 3815		6444	5336	7803	6548	7105			5282	6679
CNA 3891		6750	5904	7055	5094	6620	7785	8125		6502
CNA 6083							6844	8660		6878
CNA 5743							7705	8379		7167
CNA 3760	5198									5428
CNA 3852	5369	6152	5436	7139	5573					6017
CNA 3889					7132	7231			5394	7070
CNA 3924	6124			7136						6312
CNA 3949	6174	6366	5447							6360
CNA 3950	6041									6271
CNA 5058			5708		5027					5857
CICA 4	5838									6068
CICA 9	6216	6646		7479						6557
Mean	6015	6452	5460	7301	6162	6774	7322	7591	5091	
C.V. (%)	14.43	12.17	15.14	11.93	14.87	8.54	13.33	8.24	13.10	
N° of treat.	10	10	10	15	15	10	17	17	17	
N° of repl.	4	4	3	4	4	4	4	4	4	

$$C = \begin{array}{c} \mu \quad \text{year} \quad \text{trial/year} \quad \text{rep/trial/year} \quad \text{genot.} \\ \begin{bmatrix} 1 & 1/5 \ldots 1/5 & 1/5\ 1/10\ 1/10\ 1/10\ 1/10\ 1/5\ 1/15\ 1/15\ 1/15 & 1/20 \ldots 1/60 & 1\ 0\ 0 \ldots 0 \\ 1 & 1/5 \ldots 1/5 & 1/5\ 1/10\ 1/10\ 1/10\ 1/10\ 1/5\ 1/15\ 1/15\ 1/15 & 1/20 \ldots 1/60 & 0\ 1\ 0 \ldots 0 \\ 1 & 1/5 \ldots 1/5 & 1/5\ 1/10\ 1/10\ 1/10\ 1/10\ 1/5\ 1/15\ 1/15\ 1/15 & 1/20 \ldots 1/60 & 0\ 0\ 1 \ldots 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1/5 \ldots 1/5 & 1/5\ 1/10\ 1/10\ 1/10\ 1/10\ 1/5\ 1/15\ 1/15\ 1/15 & 1/20 \ldots 1/60 & 0\ 0\ 0 \ldots 1 \end{bmatrix} \end{array}$$

$$S = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1/10 & 0 & 0 & 0 & \ldots & 1/10 \\ 0 & 0 & 0 & 0 & 0 & 1/11 & 1/11 & 0 & 0 & 0 & \ldots & 1/11 \\ 1/17 & 1/17 & 1/17 & 0 & 1/17 & 1/17 & 1/17 & 1/17 & 0 & 0 & \ldots & 1/17 \\ 0 & 0 & 1/10 & 1/10 & 1/10 & 1/10 & 0 & 1/10 & 0 & 0 & \ldots & 0 \\ 1/20 & 1/20 & 1/20 & 0 & 1/20 & 1/20 & 1/20 & 1/20 & 1/20 & 1/20 & \ldots & 0 \end{bmatrix}$$

Table III - Results of joint analysis of variance of the example.

Sources of variation	d.f.	SS	MS
Year	4	39373596	9843399
Location/year	4	311668664	77917166
Replication/location/year	26	32879990	1264615
Genotype (adjusted)	28	54842148	1958648
Residue	411	332968675	810143
Total	473	771733073	

The adjusted mean vector, Y_i, corresponds to the last column of Table II. The auxiliary matrix S used to obtain mean vector Y^*, and the $V(Y^*)$ matrix, had dimensions 5 x 29 and can be partially represented as follows:

Thus, the Y^* vector of arithmetic means of the adjusted means for genotypes is:

$$\hat{Y}^* = \begin{bmatrix} 6245 \\ 6368 \\ 6437 \\ 6517 \\ 6575 \end{bmatrix}$$

and the respective covariance matrix is:

$$D = \begin{bmatrix} 13738.53 & 2187.01 & 812.61 & -1930.96 & -1545.98 \\ 2187.01 & 5147.77 & 3051.30 & 1463.31 & 1611.35 \\ 812.61 & 3051.30 & 3779.57 & 2715.91 & 2886.26 \\ -1930.96 & 1463.31 & 2715.91 & 6810.48 & 3349.74 \\ -1545.98 & 1611.35 & 2886.26 & 3349.74 & 5749.27 \end{bmatrix}$$

By applying the generalized least squares method, the ß vector was estimated as follows:

$$\hat{\beta} = \begin{bmatrix} 6213.6 \\ 71.4 \end{bmatrix};$$

and the covariance matrix:

$$\hat{V}(\hat{\beta}) = \begin{bmatrix} 10149.7 & -2428.4 \\ -2428.4 & 777.0 \end{bmatrix}$$

The variance of b_1 is 777.0; therefore, the mean genetic gain for the period represented by the data was 71.4 ± 27.9 kg/ha/year, significant at 1% level of probability by the *t*-test. Considering that the mean for the first year in the period was 6245 kg/ha, the relative genetic gain was 1.14% per year.

The proposed method allows the estimation of genetic gain obtained in breeding programs using regional trial data from any geographic and temporal amplitude. It is possible, for example, to calculate grain yield growth rate for a region, a state, or the entire area of interest in a breeding program. The gain for subregions within the same program may show whether the new lines are adapted to specific environmental conditions.

The versatility and strength of this method permit its use even when the data available are unbalanced. This allows maximum use of existing files, and opens possibility to study more remote periods. The use of $\overline{Y}_{ijk\bullet}$ means permits analysis of experimental series for which the original data are not available, but only the genotype means, replication number, and experimental precision parameters. The use of all available information for each line, resulting in adjusted means that are used in place of point observations, reduces the importance of genotype x year and genotype x trial/year interactions, which are the major sources of error in genetic gain studies. The reliability of the estimated gain depends only on the quality of the experimental results available.

The most cumbersome phase in the application of the proposed method is the organization of the files for combined analysis. It is necessary to standardize the identification of each genotype in all the files within the period. However, a subproduct of this procedure is a perfect organization of the files and a retrospective view of the experiments performed. Construction of auxiliary matrix C is also a laborious phase, especially when the number of trials is high and the analysis is done with replications. It is necessary to observe nesting so that the sum of the cells referring to each effect in the model will be equal to one. Automation of this phase via a specific program would greatly simplify application of the method.

Genetic gain estimation by the method proposed in the present study is efficient, precise, and provides highly useful results to critically evaluate genetic plant breeding programs.

ACKNOWLEDGMENTS

We are indebted to Embrapa Meio Norte and to Empresa Maranhense de Pesquisa Agropecuária (Emapa) for providing the experimental data. F.B. is the recipient of a CNPq fellowship.

RESUMO

Os ganhos genéticos obtidos pelo melhoramento de caracteres quantitativos podem ser estimados utilizando resultados de ensaios regionais de avaliação de linhagens e cultivares. Um novo método estatístico para esta estimativa é proposto, o qual consiste em quatro passos: a) análise conjunta da série de dados dos ensaios regionais através de um modelo linear generalizado de forma a obter as médias ajustadas dos genótipos e a matriz de covariâncias destas médias; b) para o grupo de genótipos avaliados em cada ano, cálculo da média aritmética das médias ajustadas obtidas na análise conjunta; c) comparação direta dos anos, conforme as médias aritméticas obtidas, e d) estimativa de um ganho genético médio, por regressão. Aplicando-se o método de quadrados mínimos generalizado, é calculada uma estimativa ponderada do ganho genético médio no período. Este método permite um melhor cancelamento das interações genótipo x ano e genótipo x ensaio/ano, resultando assim em estimativas mais precisas. Este método pode ser aplicado a dados desbalanceados, o que possibilita a estimativa dos ganhos genéticos em séries de ensaios multilocais de qualquer amplitude e duração.

REFERENCES

Hoffmann, R. and **Vieira, S.** (1987). *Análise de Regressão: Uma Introdução à Econometria*. 2nd edn. Hucitec, São Paulo.

Rodrigues, J.A.S. (1990). Progresso genético e potencial de risco da cultura do sorgo granífero (*Sorghum bicolor* (L.) Moench) no Brasil. Doctoral thesis, ESALQ, Piracicaba.

SAS Institute Inc. (1985). *SAS User's Guide: Statistics*. Version 5. SAS Institute Inc., Cary, NC.

Searle, S.R. (1971). *Linear Models*. John Wiley & Sons, New York.

Soares, A.A. (1992). Desempenho do melhoramento genético do arroz de sequeiro e irrigado da década de oitenta em Minas Gerais. Doctoral thesis, ESAL, Lavras.

Toledo, J.F.F. de, Almeida, L.A. de, Kiihl, R.A. de S. and **Menosso, O.G.** (1990). Ganho genético em soja no Estado do Paraná, via melhoramento. *Pesq. Agropec. Bras. 25*: 89-94.

Vencovsky, R., Morais, A.R., Garcia, J.C. and **Teixeira, N.M.** (1988). Progresso genético em vinte anos de melhoramento de milho no Brasil. In: *Congresso Nacional de Milho e Sorgo, 1986, Sete Lagoas. Anais*. Embrapa-CNPMS, Sete Lagoas, pp. 300-306.

(Received June 17, 1997)