Sur la recherche de critères de sélection : la régression géno-phénotypique

Patrick Vincourt, André Gallais

► To cite this version:

Patrick Vincourt, André Gallais. Sur la recherche de critères de sélection : la régression géno-phénotypique. *Agronomie*, 1983, 3 (9), pp.827-830. hal-00884577

HAL Id: hal-00884577

<https://hal.science/hal-00884577>

Submitted on 1 Jan 1983

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sur la recherche de critères de sélection : la régression géno-phénotypique

Patrick VINCOURT & André GALLAIS

I.N.R.A. - Station d'Amélioration des Plantes fourragères, F 86600 Lusignan

RÉSUMÉ

La régression géno-phénotypique, dont les bases théoriques proviennent des travaux d'HENDERSON (1975), est présentée avec deux exemples. Cet outil, qui intègre héritabilité et corrélations génétiques, semble présenter quelques avantages par rapport à la régression classique, pour juger de l'efficacité de critères de sélection. Un programme FORTRAN, développé sur Mini-6 CII-HB, cohérent avec la bibliothèque statistique AMANCE (BACHACOU *et al.*, 1981) est disponible à la demande.

Mots clés additionnels : *Prédiction, héritabilité, corrélations génétiques, régression multiple, meilleur prédicteur linéaire sans biais.*

SUMMARY

The search for breeding criteria : the genotype-on-phenotype regression.

The genotype-on-phenotype regression, whose theoretical bases were given in HENDERSON'S paper (1975), is presented through two examples. This tool, which integrates the heritability concept and genetic correlations, has shown a number of advantages over classical regression for the choice of breeding criteria. A HENDER program, written in FORTRAN for the CII-HB Mini-6 computer and compatible with the statistical software AMANCE (BACHACOU *et al.*, 1981), can be provided on request.

Additional key words : *Prediction, heritability, genetic correlation, multiple regression, best linear unbiased predictor.*

I. INTRODUCTION

Dans le calcul d'index de sélection, la première étape consiste à estimer les valeurs prises par des variables aléatoires (valeurs génétiques additives ou totales), à partir de variables aléatoires mesurées (caractères phénotypiques) sur les individus candidats à la sélection ou leurs apparentés — dans un plan d'expérience conçu à cet effet. HENDERSON (1975) a résolu ces problèmes d'estimation dans le cadre d'un modèle linéaire mixte général et a établi que le meilleur prédicteur (BLUP) était de la forme :

$$(1) \hat{G} = \Sigma_{GP} \Sigma_{PP}^{-1} (Y - E(Y))$$

où $\hat{G}$ est l'estimation des valeurs génétiques

Σ_{GP} la matrice de covariance entre valeurs génétiques et valeurs phénotypiques, qui est généralement proportionnelle à la matrice de covariance des valeurs génétiques.

Σ_{PP} la matrice de covariance des variables phénotypiques mesurées.

Y le vecteur des variables phénotypiques mesurées est analysé selon le modèle :

$$Y = E(Y) + XG + Z$$

La matrice X de ce modèle dépend notamment du dispositif expérimental et des relations de parenté entre les génotypes testés.

Z est le vecteur aléatoire des erreurs résiduelles.

$E(Y)$ représente l'ensemble des effets fixes : effets dus au milieu, moyennes de populations, etc...

L'objet de cet article est de souligner, à partir de quelques exemples, l'intérêt de cette formulation pour la recherche de prédicteurs, c'est-à-dire souvent de critères de sélection, en amélioration des plantes. La question de la précision des estimations des matrices Σ_{GP} et Σ_{PP} ne sera pas abordée.

II. MATÉRIEL ET MÉTHODES

A. La régression géno-phénotypique

La formule (1) est en fait une équation de régression dont les coefficients sont les termes de la matrice $\Sigma_{GP} \Sigma_{PP}^{-1}$ et dont les variables prédictrices sont les $(Y - E(Y))$. Nous l'appelons « régression géno-phénotypique » pour la distinguer de la régression multiple classique :

— parce qu'elle établit un hiatus entre les valeurs génoty-

piques G , inconnues et que l'on veut estimer, et les valeurs phénotypiques mesurées Y ;

— parce qu'elle permet d'intégrer en une seule formulation deux réalités souvent disjointes dans la pratique de la sélection : l'héritabilité (au sens strict ou au sens large) et la prise en compte des corrélations entre caractères mesurés.

En effet, lorsqu'une seule variable phénotypique $Y^{(k)}$ est décrite par le modèle :

$$(2) \quad Y_{i,l}^{(k)} = m^{(k)} + G_i^{(k)} + Z_{i,l}^{(k)}$$

$1 < i < I$ (indice génotype) ; $1 < l < L$ (indice répétition)
k indice caractère

$$\text{Var } G_i = \sigma_G^2 ; \quad \text{var } Z_{i,l} = \sigma_E^2$$

A la variable prédictrice $Y_{i,l}$, qui résume l'information phénotypique sur le génotype i , est affecté le coefficient

$$\sum_{GP} \sum_{PP}^{-1} = h_U^2 = \frac{\sigma^2 G}{\sigma^2 G + \frac{\sigma_E^2}{L}}$$

qui est l'héritabilité (au sens large) dans le dispositif considéré. Le coefficient h_U est l'héritabilité univariable du caractère k lorsque seul le caractère k est mesuré. C'est aussi le carré du coefficient de corrélation (ou coefficient de prédiction) entre la vraie valeur $G_i^{(k)}$ et son estimation $\hat{G}_i^{(k)}$ à partir du phénotype $Y^{(k)}$.

$$h_U^2 = r^2(G, \hat{G})$$

Lorsque plusieurs variables phénotypiques sont mesurées, soit parce qu'elles sont toutes objets de sélection, soit parce qu'elles contribuent à augmenter la précision du choix effectué sur une seule variable (VINCOURT & GALLAIS, 1981 ; GALLAIS *et al.*, 1983), il est également possible de calculer une héritabilité multivariable (h_m^2) pour chaque caractère et qui mesurera l'adéquation $r^2(G, \hat{G})$ entre vraie valeur et estimation des variables génétiques. La matrice $\sum_G$ de covariance entre estimateurs des valeurs génétiques s'écrit $\sum_{GP} \sum_{PP}^{-1} \sum_{PG} = \sum_G$ et pourra avec profit être comparée à la matrice $\sum_G$ de covariances entre les « vraies » valeurs génétiques. Pour chaque caractère, l'héritabilité multivariable est le rapport des termes diagonaux des matrices $\sum_G$ et $\hat{\sum}_G$.

Lorsque l'on recherche des critères de sélection, l'attitude classique consiste à raisonner uniquement à partir de la matrice $\sum_{PP}$, en faisant des régressions multiples entre phénotypes des caractères à améliorer et phénotypes des critères testés.

On perd alors une information précieuse sur les précisions relatives du caractère agronomique à améliorer (variable à prédire) et du test utilisé en sélection (variable prédictrice).

Le choix des critères de sélection peut être fait :

- en fonction de la précision, mesurée par $r^2(G, \hat{G})$
- dans le cas de sélection multicaractère, en fonction de la distorsion existant entre les matrices $\sum_G$ et $\hat{\sum}_G$, ce qui revient à raisonner sur les gains attendus en sélection.

B. Données expérimentales

1. Prédiction du rendement en matière sèche plante entière et de la teneur en matière sèche du maïs ensilage

35 plantes S_0 de la population NP 5 ont été étudiées pour leur valeur en croisement avec le testeur hybride simple A 654 × W 629 A. NP 5 est une sélection effectuée par EAGLES dans le pool HEYD (Pool 5) du CIMMYT qui possède des caractéristiques d'adaptation aux basses tempé-

ratures (MIEDEMA, 1979 ; EAGLES & HARDACRE, 1979). Les descendances ont été mises en essai en 1982 dans un lattage 7×6 à 3 répétitions. Le protocole expérimental précis est décrit par POTIER (1982).

Les matrices $\sum_{GP}$ et $\sum_{PP}$ sont estimées à partir de l'analyse de variance selon la méthode d'HENDERSON n° III (1953) et la régression géno-phénotypique a été utilisée notamment pour étudier la prédiction de caractère du stade adulte à partir de caractères phénotypiques du stade jeune (tabl. 1) (POTIER, 1982).

TABEAU 1

Variable à prévoir (G) et variables prédictrices (P) mesurées sur les descendances NP 5* (A 654* W 629 A).

Genetic values to be estimated : (G) and predictors (P) measured in trial on progenies NP 5* (A 654* W 629 A).

P	B2 VIS	Vitesse d'apparition des feuilles visibles (dont la pointe apparaît dans le cornet foliaire), estimée par régression linéaire sur la somme de température en base 6 °C (sol, - 10 cm)
	NBF TOT	Nombre total de feuilles
G	MÂLE	Date de floraison mâle
	FEM	Date de floraison femelle
	RDT	Rendement en matière sèche plante entière
	% MS	Teneur en matière sèche

2. Analyse d'essais multilocaux

40 génotypes de maïs ont été testés pour leur rendement en grain dans 8 lieux (tabl. 2). Ces données nous ont été fournies par M. DERIEUX (Station de Mons-en-Chaussée).

TABEAU 2

Lieux de tests pour 40 génotypes de maïs grain (essais de la Station de Mons en Chaussée, 1978).

Locations for grain yield tests — 40 genotypes of maize.

Lieux	Rendement moyen (q/ha)	Coefficient de variation (intra-lieu-)
Compiègne	66,8	7,4
Rennes	78,5	9,0
La Minière	70,2	10,8
Clermont-Ferrand	61,7	7,3
Chateaudun	76,4	8,9
Dieppe	53,8	9,3
Boigneville	74,2	7,9
Verneuil-L'Etang	60,6	7,8

Les différents lieux sont considérés comme des caractères, approche souvent utilisée en amélioration forestière (KREMER, 1979, Comm. pers. ; BURDON, 1979). La covariance entre lieux se réduit alors à la covariance génétique. L'adoption d'un modèle à effets aléatoires (3) pour traiter ce type de données peut sembler très critique. Il est cependant certain que tout sélectionneur serait tenté d'examiner les corrélations entre lieux et que ce modèle nous semble le plus simple pour exploiter rationnellement cette information. Une discussion plus générale sur cette question a été exposée par LEFORT (1980).

$$(3) \quad Y_{i,l}^{(k)} = m^{(k)} + G_i^{(k)} + b_l^{(k)} + Z_{i,l}^{(k)}$$

k indice lieu, i indice variété, l indice bloc (hiérarchisé à lieu)

$$\text{Cov}(Y_{i,l}^{(k)}, Y_{i,l}^{(k')}) = \text{cov}(G_i^{(k)}, G_i^{(k')}) = {}_G\rho_{kk'} \cdot {}_k\sigma_G \cdot {}_k\sigma_G =$$

$$= \text{covariance génétique entre lieux}$$

$$\text{Var } G_i^{(k)} = \sigma_G^{2(k)}$$

$$\text{Var } Z_{i,l}^{(k)} = \sigma_e^{2(k)} = \text{variance résiduelle intra-lieu}$$

III. RÉSULTATS

A. Prédiction stade adulte/stade jeune

Uniquement à partir du nombre total de feuilles (NBFTOT) et de leur vitesse d'apparition dans le cornet foliaire (B2 VIS), il est possible de prédire les dates de floraison, la teneur en matière sèche (p. 100 MS) à la récolte et le rendement (RDT) avec des coefficients de détermination génétique r^2 (G, $\hat{G}$) relativement importants (tabl. 3) : de 0,344 à 0,677. Ces valeurs sont assez proches de celles qui

TABLEAU 3

Coefficients de détermination (sur la diagonale), corrélations génétiques (sous la diagonale), corrélations entre prédicteurs (au-dessus de la diagonale) pour les variables prédites à partir de B2 VIS et NBFTOT.

Coefficients of determination (on the diagonal), genetic correlations (under the diagonal), correlations between predictors (above the diagonal) for variables predicted from B2 VIS and NBFTOT traits.

	Mâle (jours)	Femelle (jours)	% MS points	RDT t/ha
Mâle	0,677	0,883	-0,797	0,189
Femelle	0,824	0,527	-0,988	-0,298
% MS	-0,489	-0,543	0,447	0,444
RDT	0,259	-0,202	0,132	0,344

auraient été obtenues par régression multiple classique (tabl. 4). On peut cependant remarquer que la prédiction génétique est d'autant meilleure, par rapport à la prédiction phénotypique, que le caractère possède une héritabilité — au sens univariable — plus faible. Ce résultat est d'ailleurs prévisible par le calcul, car les matrices Σ_G et Σ_P sont d'autant plus semblables que les caractères sont plus héréditaires : peu sensibles au milieu ou mesurés sur un grand nombre de répétitions.

TABLEAU 4

Effet de l'héritabilité (au sens univariable) sur l'écart entre régression géno-phénotypique et régression classique.

(1) : héritabilité (sens large) au sens univariable ;

(2) : rapport du coefficient de détermination génétique au coefficient de détermination phénotypique.

Effects of univariate heritability on the divergence between the geno-phenotypic and standard regressions.

(1) univariate heritability (in the broad sense) ;

(2) ratio of the coefficient of genetic determination to the coefficient of phenotypic determination.

	Mâle	Femelle	% MS	RDT
(1) h^2	0,781	0,872	0,765	0,695
(2) $R^2(\text{Phen})/R(G, \hat{G})$	0,929	1,036	0,847	0,735

La matrice Σ'_G de covariance des estimateurs des valeurs génétiques est sensiblement différente de Σ_G . En particulier, regardées au travers des variables B2 VIS et NBF TOT, les précocités de floraison et d'ensilage apparaissent regroupées en un seul facteur. La corrélation (RDT, p. 100 MS) est très renforcée (0,44), ce qui implique qu'une sélection fondée sur les critères B2 VIS et NBF TOT constituerait en soi un index de sélection permettant l'amélioration simultanée de la population pour RDT et p. 100 MS.

B. Interactions génotype $\times$ milieu

A partir de la matrice de covariance inter-lieux et des erreurs commises dans chacun des lieux, on estime la valeur génétique du génotype i dans le lieu k $\hat{G}_i^{(k)}$ par une combinaison linéaire des valeurs phénotypiques observées sur chacun des lieux. La contribution d'un lieu k' à la définition de la valeur génétique attendue dans un autre lieu k est d'autant plus forte que les lieux k et k' sont proches l'un de l'autre — au sens du carré de la corrélation $\rho_{kk'}$. Lors de la sélection, le choix des génotypes est fait sur la base d'un index $\hat{I} = \sum_k b_k \hat{G}_i^{(k)}$. Les pondérations b_k à affecter aux différents lieux peuvent être déterminées selon différents critères : représentativité d'un lieu au sein d'une zone agricole, méthode des gains désirés, etc...

La qualité d'un lieu de test peut être définie de la façon suivante : la réponse, dans l'ensemble des lieux, à la sélection pratiquée dans ce lieu est forte en moyenne et assez peu variable. Ces qualités sont symétriques de celles qui sont classiquement utilisées pour caractériser les génotypes : valeur moyenne et stabilité (EBERHART & RUSSELL, 1966).

Le tableau 5 contient les corrélations génétiques entre

TABLEAU 5

Corrélations génétiques entre lieux (sous la diagonale) et coefficients de prédiction (sur la diagonale) dans un essai multilocal de maïs grain.

Genetic correlations between environments (under the diagonal) and coefficients of prediction (on the diagonal) in a multilocation trial of corn.

Lieux	1	2	3	4	5	6	7	8
1	0,879							
2	-0,049	0,858						
3	0,640	0,619	0,871					
4	0,375	0,489	0,730	0,818				
5	-0,042	0,720	0,106	0,061	0,724			
6	0,644	0,575	0,355	0,684	0,306	0,852		
7	0,767	0,489	0,761	0,436	0,128	0,732	0,827	
8	-0,020	0,437	0,428	0,688	0,473	0,287	0,189	0,762

TABLEAU 6

Comparaison des différents lieux de test (détails dans le texte).
Comparison of locations in a multilocation trial (details in text).

Lieux	1	2	3	4	5	6	7	8
$r^2(G, \hat{G})$	0,88	0,86	0,87	0,82	0,72	0,85	0,83	0,76
$r'^2(G, \hat{G})$	0,83	0,69	0,65	0,72	0,58	0,42	0,69	0,68
ΔG_k	4,2	4,8	6,1	5,6	3,1	5,5	5,8	4,3
(qx/ha)								
CV_k	1,0	0,7	0,5	0,5	0,9	0,4	0,6	0,7

lieux. Les valeurs de coefficients de détermination $r_{G, \hat{G}}^{2(k)}$ (tabl. 6) sont à rapprocher de celles qui auraient été obtenues ($r'^{2(k)}$) si l'on n'avait pas pris en compte les corrélations entre lieux. L'amélioration est surtout sensible pour les milieux assez bien corrélés avec chacun des autres milieux (par ex., CHATEAUDUN, lieu 6).

Si $\Delta G_k'$ est le gain attendu dans le lieu k' , lorsqu'on sélectionne dans le lieu k , on peut calculer $\Delta G_k = \sum_{k'} c_k$

$\Delta G_k'$, efficacité moyenne du lieu k lorsqu'on a affecté à chaque lieu k' un poids économique c_k . Pour simplifier, prenons tous les c_k égaux à 1. On peut en outre caractériser chaque lieu par le coefficient de variation CV_k des $\Delta G_k'$. Ces différents critères figurent dans le tableau 6 pour une intensité de sélection de 2. Les lieux 3, 4, 6 semblent avoir une valeur de prédiction plus intéressante que les autres.

IV. DISCUSSION-CONCLUSION

La recherche de critères de sélection repose sur la mesure de corrélation entre le caractère à améliorer et un caractère plus simple ou mesurable à moindre coût. Les sélectionneurs sont confrontés dans cette démarche à plusieurs difficultés :

— éviter la confusion entre corrélation génétique et corrélation environnementale,

— trouver un compromis satisfaisant entre la précision du critère et sa justesse évaluée, dans leur discipline, comme l'aptitude à « bien classer »,

— apprécier l'ampleur des réponses corrélatives dont l'origine peut tenir autant à l'imprécision du critère de test qu'aux relations génétiques vraies entre les caractères.

La régression géno-phénotypique présente l'avantage de guider le sélectionneur à l'aide de quelques critères quantitatifs.

Le coefficient $r^2(G, \hat{G})$ de détermination génétique — pour reprendre une terminologie en usage en sélection animale — doit être distingué du « coefficient de prédiction génétique » de BARADAT (1976). Ce dernier critère donne des rôles symétriques aux variables prédictives et à prédire, ce qui peut présenter un inconvénient dans un problème de prédiction.

GALLAIS (1973) a proposé de comparer les méthodes d'appréciation de la valeur génétique — ou de sélection multicaractère — d'après l'« héritabilité généralisée » qui désigne la matrice $\sum_{GP} \sum_{PP}^{-1}$. Il est cependant délicat de comparer 2 régressions d'après les valeurs des coefficients de régression. Les matrices $\sum_G = \sum_{GP} \sum_{PP}^{-1} \sum_{PG}$ et $\sum_G$ constituent vraisemblablement un résumé nécessaire et suffisant à une évaluation plus juste des méthodes de prédiction. Le coefficient de détermination $r^2(G, \hat{G})$ et les réponses à la sélection en sont des fonctions. Les rapports entre ces matrices ont d'ailleurs été étudiés, notamment par ROUVIER (1969) pour orienter le choix entre plusieurs index.

Reçu le 15 décembre 1982.

Accepté le 27 avril 1983.

RÉFÉRENCES BIBLIOGRAPHIQUES

- Bachacou J., Masson J. P., Millier C., 1981. *Manuel de la programmation statistique Amance*, 81, 516 p., I.N.R.A. Ed.
- Baradat Ph., 1976. Use of juvenile-mature relationships and information from relatives in combined multitrait selection. *Proc. IUFRO joint Meet. genet. Work. Parties on adv. Generations Breed.*, Bordeaux, 121-138.
- Burdon R. D., 1979. Generalization of multitrait selection indices using information from several sites. *N. Z. J. For. Sci.* (9), 145-152.
- Eagles H. A., Hardacre A. K., 1979. Genetic variation in maize for germination and emergence at 10 °C. *Euphytica*, 28, 287-295.
- Eberhart S. A., Russell W. A., 1966. Stability parameters for comparing varieties. *Crop Sci.*, 6, 36-40.
- Gallais A., 1973. Sélection pour plusieurs caractères : synthèse critique et généralisation. *Ann. Amél. Plantes*, 23 (3), 183-208.
- Gallais A., Vincourt P., Bertholleau J. C., 1983. Etude de critères de sélection chez le maïs fourrage. Héritabilités et corrélations génétiques. *Agronomie*, 3 (8), (sous presse).
- Henderson C. R., 1953. Estimation of variance and covariance components. *Biometrics*, 9, 226-252.
- Henderson C. R., 1975. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, (2), 423-447.
- Lefort G., 1980. Le modèle de base de la sélection : justification et limites, in « *Biométrie et génétique* », J. M. Légaré Ed. p. 1-14.
- Miedema P., 1979. Potential use of gene pool material for improving low temperature adaptation in *Zea mays*. *Euphytica*, 28, 661-664.
- Potier F., 1982. *Etude de la variabilité génétique à l'intérieur d'une population de maïs adaptée aux basses températures*. D.A.A., E.N.S.A. de Rennes, Rapport ronéotypé, 60 p.
- Rouvier R., 1969. *Contribution à l'étude des index de sélection sur plusieurs caractères*. Thèse 3ème cycle Fac. Sci. Paris, 92 p.
- Vincourt P., Gallais A., 1981. Application of index selection in maize breeding for silage. *Proc. 4th meet. section « Biometrics in plant breeding »*. Eucarpia, Poitiers.