Ann. Génét. Sél. anim., 1982, **14** (3), 309-326

Ecriture des équations du BLUP multicaractères

J.L. FOULLEY, S. CALOMITI et D. GIANOLA [1]
I.N.R.A., Station de Génétique quantitative et appliquée,
Centre de Recherches zootechniques, F 78350 Jouy-en-Josas

Résumé

Cet article formalise l'écriture des équations du BLUP multicaractères dans la situation où tous les caractères sont affectés par les mêmes facteurs de variation. Deux cas sont distingués selon que l'information sur les caractères est complète ou non.

En présence d'un seul facteur aléatoire et quand tous les caractères sont contrôlés, le problème se simplifie beaucoup grâce à une transformation canonique des données (ou directement des seconds-membres) et revient à calculer le BLUP unicaractère de chacune des variables canoniques.

Dans le cas d'information incomplète, on est conduit à subdiviser l'échantillon en ensembles d'individus homogènes entre eux quant aux caractères contrôlés et tels que les résiduelles du modèle relatives aux variables mesurées sur ces individus soient non corrélées entre elles d'un ensemble à l'autre. Le système obtenu présente un nombre d'équations multiple du nombre de caractères ce qui pose des problèmes numériques importants et limite beaucoup la taille des applications actuelles. Cependant dans tous les cas, il est possible d'obtenir très simplement les BLUP d'individus apparentés et de caractères corrélés n'apparaissant pas dans la décomposition des données.

Un exemple très simple est présenté en détail pour illustrer le raisonnement.

I. Introduction

Depuis son introduction en 1949 par HENDERSON, la méthode d'évaluation des reproducteurs dite du BLUP a connu un très grand développement en sélection animale tant du point de vue de la théorie que de celui de ses applications. Du fait des problèmes numériques posés par la mise en œuvre de ces techniques sur de gros fichiers, celles-ci concernent jusqu'à maintenant une évaluation sur des caractères considérés isolément. Il est néanmoins possible, du moins sur des fichiers de taille réduite, d'appliquer une évaluation multidimensionnelle prenant en compte l'ensemble de l'information recueillie sur plusieurs caractères chez différents individus (HENDERSON & QUAAS, 1976). La réalisation de cette évaluation pluricaractères pose de nombreuses difficultés. L'une à laquelle on se heurte généralement dès le départ est l'écriture des équations à

(1) Adresse permanente : Department of Animal Science, University of Illinois, Urbana, Illinois 61801, U.S.A.

résoudre. L'objet de cette note est de donner une certaine présentation matricielle de ces équations, notamment dans le cas d'information inégale sur les caractères présents chez les individus contrôlés.

II. Ecriture des équations

Pour un caractère C (C = 1, 2, ..., t), le modèle linéaire de décomposition des données Y en vue de l'évaluation génétique de q individus s'écrit :

$$Y_C = X_C b_C + Z_C u_C + e_C \qquad (1)$$

où :

- b_C est le vecteur colonne des p effets fixés,
- u_C est le vecteur colonne des q effets génétiques aléatoires à prédire,
- e_C est le vecteur colonne des résidus supposés non corrélés à u_C,
- X_C, Z_C sont des matrices de coefficients traduisant l'incidence des effets des facteurs b_C et u_C respectivement sur le vecteur colonne Y_C des données.

Deux situations doivent être *a priori* distinguées selon que tous les individus possèdent ou non une information complète sur l'ensemble des caractères.

1. *Cas d'information complète*

Ce cas a été déjà abordé par LEE (1979) et ARNASON (1981). Une présentation synthétique en sera donnée ici. En se restreignant à la situation où les mêmes facteurs (y compris des covariables) affectent chacun des t caractères mesurés sur n individus, on peut écrire :

$$X_C = X \text{ matrice } (n, p)$$
$$Z_C = Z \text{ matrice } (n, q)$$

Si on classe les données par individus (i) et caractère (C), les paramètres par facteur (f), niveau (j) intra-facteur et caractère intra-niveau soit :

$$\mathcal{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{iC} \\ Y_{it} \end{bmatrix} \quad b_{f,j} = \begin{bmatrix} b_{f,j,1} \\ b_{f,j,2} \\ b_{f,j,C} \\ b_{f,j,t} \end{bmatrix} \quad u_k = \begin{bmatrix} u_{k1} \\ u_{k2} \\ u_{kC} \\ u_{kt} \end{bmatrix}$$

Le modèle s'écrit alors :

$$\mathcal{Y} = (X \otimes I_t)\, b + (Z \otimes I_t)\, u + e \qquad (2)$$

où :

- $\mathcal{Y}$, b, u et e sont des vecteurs colonne de dimension nt, pt, qt et nt respectivement,
- I_t est la matrice identité d'ordre t,
- $\otimes$ est le symbole du produit direct ou produit de KRONECKER.

La matrice $\mathscr{R}$ de variances et covariances de e s'écrit, sous l'hypothèse d'indépendance des résidus relatifs à des individus mesurés différents :

$$\mathscr{R} = I_n \otimes R$$

Sous un modèle à un seul effet aléatoire, la matrice $\mathscr{G}$ de variances et covariances de u s'explicite en :

$$\mathscr{G} = A \otimes G$$

avec :

$A = 2\,\Phi$ où Φ est la matrice des coefficients de parenté entre les q individus, selon MALECOT,

R et G les matrices de variances et covariances entre les t caractères relatives aux résidus et aux valeurs génétiques respectivement.

Compte tenu des propriétés du produit direct appliquées à :

$$\mathscr{X} = X \otimes I_t,\ \mathscr{Z} = Z \otimes I_t,\ \mathscr{R},\ \mathscr{G} \text{ et } \mathscr{Y},$$

les équations du BLUP correspondant au modèle (2) s'écrivent :

$$\left[\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix} \otimes R^{-1} + \left(\begin{array}{c|c} 0\,(pt, pt) & 0\,(pt, qt) \\ \hline 0\,(qt, pt) & \mathscr{G}^{-1} \end{array}\right)\right] \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \tag{3}$$

$$\text{vec } [(SM_1,\ SM_2,\ ...,\ SM_C,\ ...,\ SM_t)\ R^{-1}]'$$

où :

$SM_C = \begin{pmatrix} X'Y_C \\ Z'Y_C \end{pmatrix}$ est le vecteur (p + q, 1) des seconds membres du modèle mixte (1) appliqué au caractère C,

vec est l'opérateur qui réduit une matrice en un vecteur colonne selon les colonnes de la matrice.

Dans le cas d'un seul effet aléatoire ($\mathscr{G} = A \otimes G$), diverses simplifications peuvent être adoptées, en vue de la résolution.

En multipliant les 2 membres de (3) à gauche par $I_{p+q} \otimes R$, on aboutit au système simplifié suivant :

$$\left[\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z \end{pmatrix} \otimes I_t + \begin{pmatrix} 0 & 0 \\ 0 & A^{-1} \otimes RG^{-1} \end{pmatrix}\right] \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \tag{4}$$

$$\text{vec } (SM_1,\ SM_2,\ ...,\ SM_C,\ ...,\ SM_t)'$$

En rangeant les données et les paramètres par caractère en majeur, on obtiendrait alors les équations suivantes équivalentes :

$$\begin{bmatrix} I_t \otimes X'X & I_t \otimes X'Z \\ I_t \otimes Z'X & (I_t \otimes Z'Z) + RG^{-1} \otimes A^{-1} \end{bmatrix} \begin{bmatrix} \hat{b}_o \\ \hat{u}_o \end{bmatrix} = \begin{bmatrix} X'Y_1 \\ \vdots \\ X'Y_t \\ \hline Z'Y_1 \\ \vdots \\ Z'Y_t \end{bmatrix} \tag{5}$$

Une autre simplification importante consiste à faire un changement de variable permettant de diagonaliser RG^{-1}. Les valeurs propres λ de GR^{-1} étant égales aux racines caractéristiques de l'équation $|G - \lambda R| = 0$, on peut avoir recours à une transformation canonique des données ainsi que l'a suggéré LEE (1979) et effectué ARNASON (1981). Si P est la matrice de passage définissant le changement de base, soit $\mathcal{Y} = (I \otimes P)\,\mathcal{Y}^*$ ou encore $Y_i = PY_i^*$, les colonnes de P sont les vecteurs propres R^{-1} normés de la matrice GR^{-1} (ou réciproquement les lignes de P^{-1} sont les vecteurs propres gauches R normés de cette matrice). De façon équivalente, cela revient à écrire :

— $(P^{-1})' \; RP^{-1} = I_t$
— $(P^{-1})' \; GP^{-1} = \Delta$ (matrice diagonale des valeurs propres).

Les calculs se ramènent alors à l'obtention d'un BLUP unicaractère sur chacune des t variables canoniques, soit pour la $C^{ième}$:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}.\lambda_C^{-1} \end{bmatrix} \begin{bmatrix} \hat{b}_C^* \\ \hat{u}_C^* \end{bmatrix} = \begin{bmatrix} X'Y_C^* \\ Z'Y_C^* \end{bmatrix} \tag{6}$$

La transformation canonique peut être effectuée sur chacune des données, ou de façon plus rapide directement sur les seconds-membres.

Soit :

$$SM = (SM_1, SM_2, ..., SM_C, ..., SM_t)$$

On calculera $SM^* = SM\,(P^{-1})'$

et :

$$\begin{pmatrix} X'Y_C^* \\ Z'Y_C^* \end{pmatrix} = SM_C^* = (C^{ième} \text{ colonne de } SM^*)$$

Les solutions originales en b et u s'obtiennent ensuite par la transformation inverse.

Soit :

$$\hat{\gamma}_C = \begin{pmatrix} \hat{b}_C \\ \hat{u}_C \end{pmatrix} \qquad \hat{\Gamma} = (\hat{\gamma}_1, \hat{\gamma}_2, ..., \hat{\gamma}_C, \hat{\gamma}_t)$$

On a :

$$\hat{\Gamma} = \hat{\Gamma}^* P' \tag{7}$$

Il reste maintenant à exprimer la précision des prédictions lorsqu'on a recours à cette transformation canonique.
Par définition :

$$u = (I_q \otimes P)\, u^*$$

$$\hat{u} = (I_q \otimes P)\, \hat{u}^*$$

D'où :

$$\text{var}\,(\hat{u} - u) = (I_q \otimes P)\ \text{var}\,(\hat{u}^* - u^*)\,(I_q \otimes P')$$

avec :

$$u^{*\prime} = u_1^{*\prime}, u_2^{*\prime}, ... u_i^{*\prime}, u_q^{*\prime}$$

Le vecteur colonne (qt, 1) des éléments diagonaux de var (û — u) qui sera noté Diag var (û — u) s'écrit alors :

$$\text{Diag var } (\hat{u} - u) = \begin{Bmatrix} \text{Diag } PV_1^*P' \\ \vdots \\ \text{Diag } PV_i^*P' \\ \vdots \\ \text{Diag } PV_q^*P' \end{Bmatrix}$$

où :

$$V_i^* = \text{var } (\hat{u}_i^* - u_i^*)$$

V_i^* de dimension (t, t) est une matrice diagonale puisque les t variables canoniques sont non corrélées entre elles, ce qui conduit à une expression simple de PV_i^*P'. Soit VEP* la matrice (t, q) dont la $C^{ième}$ ligne représente les q variances d'erreurs de prédiction de la $C^{ième}$ variable canonique, c'est-à-dire (Diag var $(\hat{u}_C^* - u_C^*))'$ il vient :

$$\text{Diag } PV_i^*P' = (P \textcircled{H} P)\, VEP_i^*$$

où $P \textcircled{H} P$ est la matrice (t, t) obtenue en élevant chaque terme de P au carré (cas particulier du produit d'HADAMARD).

VEP_i^* est la $i^{ème}$ colonne de VEP* relative à l'individu i.

Soit, en définitive :

$$\text{Diag var } (\hat{u} - u) = \text{vec } \left[(P \textcircled{H} P)\, VEP^*\right] \tag{8}$$

2. *Cas d'information incomplète sur les caractères*

Le problème devient nettement plus complexe. Avec 3 caractères par exemple, on peut rencontrer sur un même fichier les 7 situations suivantes :

- Les 3 caractères sont présents simultanément : (1) (2) (3).
- 2 caractères sont présents, le troisième absent (—)
  soit (1) (2) — ; (1) — (3) ; — (2) (3).
- 1 seul caractère est présent, les 2 autres étant absents soit (1) — — ; — (2) — ; — — (3).

Par ailleurs, par analogie avec la théorie des indices, on peut considérer la possibilité de prédire les valeurs réalisées de niveaux de la variable aléatoire qui n'ont pas d'incidence sur les données décrites mais dont la structure de variances et covariances est connue. Il peut s'agir par exemple d'individus apparentés, de caractères corrélés.

De façon générale avec t caractères décrits dans les données, le problème se décompose en $2^t - 1$ situations dont chacune ($s = 1, 2, \ldots, 2^t - 1$) est caractérisée par le modèle suivant :

$$\mathcal{Y}_s = \mathcal{X}_s\, b + \mathcal{Z}_s\, u + e_s \tag{9}$$

où :

$\mathcal{Y}_s$ est le vecteur $(n_s t_s, 1)$ des n_s individus mesurés sur les t_s caractères présents dans la situation s,

$\mathcal{X}_s$ et $\mathcal{Z}_s$ les matrices d'incidence de dimension $(n_s t_s, pt)$, $(n_s t_s, q't')$ relatives aux pt et q't' niveaux de b et u respectivement,

e_s le vecteur $(n_s t_s, 1)$ des résidus tel que :
— $\mathcal{R}_s = E(e_s e'_s) = I_{n_s} \otimes R_s$
(R_s matrice des variances et covariances résiduelles des t_s caractères présents)
— $E(e_s e'_s) = 0$ si $s \neq s'$

Du fait de la propriété d'absence de corrélation entre résidus relatifs à des ensembles (s) différents, les solutions générales du modèle mixte s'obtiennent simplement à partir du système suivant :

$$\left[\sum_s \begin{pmatrix} \mathcal{X}'_s \mathcal{R}_s^{-1} \mathcal{X}_s & \mathcal{X}'_s \mathcal{R}_s^{-1} \mathcal{Z}_s \\ \mathcal{Z}'_s \mathcal{R}_s^{-1} \mathcal{X}_s & \mathcal{Z}'_s \mathcal{R}_s^{-1} \mathcal{Z}_s \end{pmatrix} + \begin{pmatrix} \underline{0} & \underline{0} \\ \underline{0} & \mathcal{G}^{-1} \end{pmatrix} \right] \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} \qquad (10)$$

$$= \sum_s \begin{pmatrix} \mathcal{X}'_s \mathcal{R}_s^{-1} \mathcal{Y}_s \\ \mathcal{Z}'_s \mathcal{R}_s^{-1} \mathcal{Y}_s \end{pmatrix}$$

Ici les matrices d'incidence s'écrivent :

$$\mathcal{X}_s = X_s \otimes \mathcal{J}_{ts,t}$$
$$\mathcal{Z}_s = Z_s \otimes \mathcal{J}_{ts,t'}$$

où :

X_s, Z_s sont les matrices d'incidence (n_s, p) et (n_s, q') de b et u respectivement définies pour un seul caractère,

$\mathcal{J}_{ts,t}$ est une matrice de dimension (t_s, t) obtenue à partir d'une matrice identité (t, t) dans laquelle n'ont été conservées que les t_s lignes correspondant aux caractères présents dans la situation s.

En utilisant les propriétés du produit direct, on peut aisément expliciter les termes du système (10) soit :

$$\mathcal{X}'_s \mathcal{R}_s^{-1} \mathcal{X}_s = X'_s X_s \otimes \mathcal{J}_{t,ts} R_s^{-1} \mathcal{J}_{ts,t}$$
$$\mathcal{X}'_s \mathcal{R}_s^{-1} \mathcal{Z}_s = X'_s Z_s \otimes \mathcal{J}_{t,ts} R_s^{-1} \mathcal{J}_{ts,t'}$$
$$\mathcal{Z}'_s \mathcal{R}_s^{-1} \mathcal{Z}_s = Z'_s Z_s \otimes \mathcal{J}_{t',ts} R_s^{-1} \mathcal{J}_{ts,t'}$$

La matrice $\Omega_{t,t'} = \mathcal{J}_{t,ts} R_s^{-1} \mathcal{J}_{ts,t'}$ de dimension (t, t') correspond à la matrice (t_s, t_s) inverse de la matrice des variances et covariances des t_s caractères présents dans la situation s, complétée à leur place de $t - t_s$ lignes de zéros et $t' - t_s$ colonnes de zéros pour les caractères absents décrits dans les données et à prédire respectivement.

De même :

$$\mathscr{X}'_s \mathscr{R}_s^{-1} \mathscr{Y}_s = (X'_s \otimes \mathscr{I}_{t,ts} R_s^{-1}) \mathscr{Y}_s$$

qui s'écrit aussi :

$$\text{vec } [(X'_s Y_{s,1} ; X'_s Y_{s,2} ; \ldots ; X'_s Y_{s,c}, \ldots ; X'_s Y_{s,ts}) R_s^{-1} \mathscr{I}_{ts,t}]'$$

où $Y_{s,c}$ est le vecteur $(n_s, 1)$ des données relatives au caractère c ($c = 1, \ldots, t_s$) mesuré dans la situation s.

Sous forme condensée, le système (10) des équations du modèle mixte peut s'expliciter ainsi :

$$\left[(\sum_s C_s) + \begin{pmatrix} \underline{0} & \underline{0} \\ \underline{0} & \mathscr{G}^{-1} \end{pmatrix} \right] \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \sum_s S_s \qquad (11)$$

où :

$$C_s = \begin{bmatrix} X'_s X_s \otimes \Omega_{tt} & X'_s Z_s \otimes \Omega_{tt'} \\ Z'_s X_s \otimes \Omega_{t't} & Z'_s Z_s \otimes \Omega_{t't'} \end{bmatrix}$$

$$S_s = \begin{bmatrix} \text{vec } [(X'_s Y_{s,1} ; \ldots ; X'_s Y_{s,c} ; \ldots ; X'_s Y_{s,ts}) R_s^{-1} \mathscr{I}_{ts,t}]' \\ \text{vec } [(Z'_s Y_{s,1} ; \ldots ; Z'_s Y_{s,c} ; \ldots ; Z'_s Y_{s,ts}) R_s^{-1} \mathscr{I}_{ts,t'}]' \end{bmatrix}$$

En classant les données et les facteurs par caractère en majeur, on aboutit au système suivant :

$$\left[\sum_s \begin{pmatrix} \Omega_{tt} \otimes X'_s X_s & \Omega_{tt'} \otimes X'_s Z_s \\ \Omega_{t't} \otimes Z'_s X_s & \Omega_{t't'} \otimes Z'_s Z_s \end{pmatrix} + \begin{pmatrix} \underline{0} & \underline{0} \\ \underline{0} & \mathscr{G}^{-1} \end{pmatrix} \right] \begin{bmatrix} \hat{b}_o \\ \hat{u}_o \end{bmatrix}$$

$$= \sum_s \begin{pmatrix} \mathscr{I}_{t,ts} R_s^{-1} \otimes X'_s \\ \mathscr{I}_{t',ts} R_s^{-1} \otimes Z'_s \end{pmatrix} Y_s \qquad (12)$$

où :

$$Y'_s = (Y'_{s,1} ; Y'_{s,2} ; \ldots ; Y'_{s,ts})$$

Dans le cas d'un seul facteur aléatoire u, $\mathscr{G}$ est égal à $A \otimes G$. Pourvu que G et A soient alors définis pour les t' caractères et q' individus à prédire, on aboutit ainsi à une formulation générale des équations permettant d'évaluer des individus et des caractères autres que ceux décrits dans le modèle de décomposition des données. Dans la situation où l'on a recours à une transformation canonique, la prédiction de caractères corrélés nécessite une approche indirecte (cf. annexe) proposée initialement par HENDERSON (1977).

III. Application numérique

L'exemple traité concerne une évaluation individuelle de mâles de race bovine à partir d'informations sur 3 caractères : poids à la naissance (PN), poids à 480 jours (P 480) et note de développement musculaire (DM) dont les matrices de variances et covariances G et R sont données au tableau 1. Cet exemple étant avant tout illustratif

de la méthodologie présentée, il a été pris volontairement très simple avec très peu d'individus (5 animaux contrôlés) et un seul facteur annexe de variation b à 2 niveaux, identifiable par exemple au lot de contrôle en station (tabl. 2). Par ailleurs, on a supposé que les individus (3) et (4) étaient demi-frères paternels issus de (6), ce dernier étant lui-même demi-frère de (5) par le père (7). Le schéma de parenté est donc le suivant :

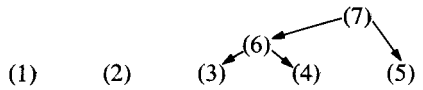

TABLEAU 1

Matrices des variances et covariances génétique (G) et résiduelle (R).
Genetic (G) and residual (R) matrices of variances and covariances.

G R	Poids à la naissance PN	Poids à 480 jours P480	Développement musculaire DM
PN (kg)	3,2000 12,8000	32,5835 921,6000	1,7709 30,0528
P480 (kg)	26,6043 1,7709	1 382,4000 73,6139	24,5000 24,5000
DM (pt)			

TABLEAU 2

Données analysées.
Data analyzed.

Animal	Lot	PN (kg)	P480 (kg)	DM (pt)
1	1	50 (*)	805	39,0
2	2	48	722	39,6
3	1	42	776	34,3
4	2	51	754	39,7
5	1	53 (*)	655	33,0 (*)

(*) Caractère supposé absent dans l'exemple d'information incomplète.

Enfin, on s'est assigné pour objectif d'évaluer ces 7 individus non seulement pour les 3 caractères mentionnés précédemment mais aussi pour un quatrième caractère non contrôlé, le poids adulte (PA) dont on connaît la variance génétique ($G_{\beta\beta} = 35437,5$) et les covariances génétiques avec les autres caractères ($G_{\beta\alpha} = 235,7244$; 2285,9291 ; 0).

1. *Cas d'information complète*

Pour les 5 individus ayant des performances et l'un quelconque des 3 caractères contrôlés, le modèle s'écrit :

$$Y_{ij} = b_j + u_i + e_{ij}$$

où :

b_j est l'effet fixé du $j^{ème}$ lot ($j = 1,2$),
u_i est l'effet aléatoire de la valeur génétique du $i^{ème}$ individu à évaluer,
e_{ij} est l'effet aléatoire résiduel.

La matrice Q des coefficients $\begin{matrix} X'X & X'Z \\ Z'X & Z'Z \end{matrix}$ relatifs aux paramètres $(b_1 b_2 u_1 u_2 u_3 u_4 u_5)$ s'écrit conformément aux données du tableau 2.

$$Q = \begin{bmatrix} 3 & 0 & 1 & 0 & 1 & 0 & 1 \\ & 2 & 0 & 1 & 0 & 1 & 0 \\ & & 1 & 0 & 0 & 0 & 0 \\ & & & 1 & 0 & 0 & 0 \\ & & & & 1 & 0 & 0 \\ & & & & & 1 & 0 \\ & & & & & & 1 \end{bmatrix}$$

Les seconds membres SM (en colonne par caractère) sont :

$$SM = \begin{bmatrix} (PN) & (P480) & (DM) \\ 145 & 2\,236 & 106,3 \\ 99 & 1\,476 & 79,3 \\ 50 & 805 & 39,0 \\ 48 & 722 & 39,6 \\ 42 & 776 & 34,3 \\ 51 & 754 & 39,7 \\ 53 & 655 & 33,0 \end{bmatrix}$$

La transformation canonique est effectuée directement sur les seconds membres en post-multipliant la matrice SM précédente par $(P^{-1})'$. La matrice P^{-1} dont les lignes sont les vecteurs propres gauches R-normés de la matrice GR^{-1} s'écrit :

$$P^{-1} = \begin{bmatrix} 0,007379 & 0,017162 & -0,215219 \\ -0,065186 & -0,021777 & -0,046789 \\ -0,277699 & 0,010942 & 0,010311 \end{bmatrix}$$

D'où les seconds membres transformés $SM^* = SM \,.\, (P^{-1})'$

$SM_1^{*\prime} =$
(16,5662 8,9946 5,7908 4,2224 6,2456 4,7722 4,5299)

$SM_2^{*\prime} =$
(— 63,1202 — 42,3074 — 22,6150 — 20,7052 — 21,2421 — 21,6023 — 19,2632)

$SM_3^{*\prime} =$
(— 14,7044 — 10,5244 — 4,6747 — 5,0213 — 2,8188 — 5,5032 — 7,2109)

Il reste à calculer A^{-1} qui s'obtient directement par la règle d'HENDERSON (1976), soit pour chaque couple (individu i, père j) :

- $a^{ii} = 4/3$; $a^{jj} = 1/3$ et $a^{ij} = a^{ji} = -2/3$ si i et j connus,
- $a^{ii} = 1$ si j inconnu.

D'où, pour les individus 1, 2, 3, 4, 5, 6 et 7 respectivement :

$$A^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ & 1 & 0 & 0 & 0 & 0 & 0 \\ & & 4/3 & 0 & 0 & -2/3 & 0 \\ & & & 4/3 & 0 & -2/3 & 0 \\ & & & & 4/3 & 0 & -2/3 \\ & & & & & 2 & -2/3 \\ & & & & & & 5/3 \end{bmatrix}$$

Le vecteur λ des valeurs propres de GR^{-1} a pour valeur $\lambda = 1{,}187065$; 0,668871 ; 0,158344.

Il suffit alors de faire un BLUP sur chacune (l) des 3 variables canoniques en résolvant le système construit comme suit. Pour évaluer simultanément les individus contrôlés (I) et non contrôlés (J), u' est partitionné en u'_I, u'_J. Si on pose :

$$A^{-1}\lambda_l^{-1} = \begin{bmatrix} T^{(l)}_{II} & T^{(l)}_{IJ} \\ T^{(l)}_{JI} & T^{(l)}_{JJ} \end{bmatrix} \quad Q = \begin{bmatrix} Q_{bb} & Q_{bI} \\ Q_{Ib} & Q_{II} \end{bmatrix} \quad SM^* = \begin{bmatrix} SM^*_b \\ SM^*_I \end{bmatrix}$$

Le système relatif à la l[ième] variable canonique s'écrit :

$$\begin{bmatrix} Q_{bb} & Q_{bI} & 0 \\ Q_{Ib} & Q_{II} + T^{(l)}_{II} & T^{(l)}_{IJ} \\ 0 & T^{(l)}_{JI} & T^{(l)}_{JJ} \end{bmatrix} \begin{bmatrix} \hat{b}^* \\ u^*_I \\ u^*_J \end{bmatrix} = \begin{bmatrix} SM^*_b \\ SM^*_I \\ 0 \end{bmatrix}$$

Les solutions numériques de ces 3 systèmes sont les suivantes :

$\hat{\gamma}^{*\prime}_1$ = 5,5109 ; 4,4807 ; 0,1519 ; — 0,1402 ; 0,3822 ; 0,1735 ; — 0,5005 ; 0,1367 ; — 0,1455,

$\hat{\gamma}^{*\prime}_2$ = — 21,0422 ; — 21,1870 ; — 0,6304 ; 0,1931 ; — 0,0545 ; — 0,1265 ; 0,6913 ; 0,0367 ; 0,2912,

$\hat{\gamma}^{*\prime}_3$ = — 4,8958 ; — 5,2785 ; 0,0302 ; 0,0352 ; 0,2417 ; — 0,0026 ; — 0,2889 ; 0,0475 ; — 0,0966.

Les solutions finales (en colonne par caractère) s'obtiennent aisément en post-multipliant la matrice $(\hat{\gamma}^*_1, \hat{\gamma}^*_2, \hat{\gamma}^*_3)$ de dimension (9,3) par P'. Les résultats figurent au tableau 3.

TABLEAU 3

Estimations et prédictions.

Estimations and predictions.

Paramètres (3)	PN (kg)		P480 (kg)		DM (pt)		PA (kg)	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
b_1	48,316	41,237	745,366	745,373	35,487	34,638		
b_2	49,600	49,525	738,903	738,601	39,803	39,781		
U_1	0,871	0,855	23,726	23,418	1,216	1,655	58,918	55,263
U_2	— 0,427	— 0,412	— 7,648	— 7,487	0,027	0,038	— 32,631	— 31,518
U_3	— 0,638	0,399	7,065	11,015	— 1,234	— 0,602	— 43,241	33,383
U_4	0,227	0,362	5,842	6,284	— 0,332	— 0,300	18,976	29,240
U_5	— 0,182	— 1,207	— 30,888	— 34,553	— 0,144	— 1,273	— 10,805	— 83,734
U_6	— 0,186	0,107	0,212	1,338	— 0,625	— 0,543	— 10,995	11,204
U_7	— 0,147	— 0,440	— 12,270	— 13,286	— 0,307	— 0,726	— 8,720	— 29,012

(1), (2) Cas d'information complète et incomplète respectivement.

(3) Effet du lot de contrôle et valeur génétique individuelle respectivement.

TABLEAU 4

Valeurs du coefficient de détermination.

Values of the coefficient of determination.

Û	PN		P480		DM		PA	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
1	0,1743	0,0949	0,2689	0,2644	0,3339	0,2584	0,0877	0,0454
2	0,1326	0,1326	0,2053	0,2053	0,2561	0,2561	0,0668	0,0668
3	0,1666	0,0947	0,2582	0,2539	0,3228	0,2651	0,0839	0,0457
4	0,1381	0,1351	0,2128	0,2127	0,2639	0,2638	0,0695	0,0679
5	0,1628	0,0899	0,2532	0,2472	0,3178	0,0172	0,0820	0,0401
6	0,0636	0,0493	0,0974	0,0967	0,1199	0,1197	0,0320	0,0245
7	0,0407	0,0259	0,0628	0,0617	0,0780	0,0315	0,0205	0,0121

(1), (2) Cas d'information complète et incomplète respectivement.

Connaissant les valeurs de la matrice VEP* (3,7) des erreurs de prédiction sur les variables canoniques, on en déduit immédiatement par la formule (8) la précision des prédiction sur les variables d'origine. Ici, on trouve VEP* égale à :

0,7656	0,8634	0,7790	0,8542	0,7849	1,0361	1,0886
0,4928	0,5345	0,4999	0,5295	0,5033	0,6050	0,6277
0,1440	0,1475	0,1448	0,1469	0,1452	0,1530	0,1550

Les valeurs de la précision relatives aux 3 caractères contrôlés sont données au tableau 4 sous la forme équivalente du coefficient de détermination :

$$CD = 1 - \text{var}(\hat{u} - u) / \text{var}\, u$$

Les prédictions des 7 individus sur le 4e caractère sont calculées suivant la formule (2) de l'annexe :

$$u_{PA,i} = Cu_i^* \quad \text{pour} \quad i = 1, 2, ..., 7$$

avec :

$$C = G_{\beta\alpha}(\Delta P^{\wedge})^{-1}$$

on trouve ici :

$$C = 34{,}5140\ ;\ -97{,}3998\ ;\ -255{,}4463$$

On en déduit immédiatement les valeurs prédites données au tableau 3 ainsi que les précisions correspondantes (tableau 4) à partir des formules suivantes :

$$\text{Diag}\ [\text{var}(\hat{u}_{PA} - u_{PA})] = \mathbf{1} \otimes 35437{,}5 - \text{vec}(C \circledcirc C)\, D$$

$$D = \mathbf{1}' \otimes \lambda - VEP^*$$

2. *Cas d'information incomplète*

Supposons maintenant que le poids à la naissance de l'animal (1) soit absent et, de même, que l'animal (5) ne présente d'information que pour le poids à 480 j. On distingue alors les 3 sous-ensembles suivants d'information :

E_1 animal (1),

E_2 animaux (2), (3) et (4)

E_3 animal (5).

Les matrices d'incidence $Q_s = \begin{pmatrix} X'_sX_s & X'_sZ_s \\ Z'_sX_s & Z'_sZ_s \end{pmatrix}$ et les seconds membres $SM_s = (X'_sY_{s,1}\ ;\ \ldots\ ;\ X'_sY_{s,ts})$ correspondant à chacune de ces situations s'écrivent alors :

$$Q_1 = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ & 0 & 0 & 0 & 0 & 0 & 0 \\ & & 1 & 0 & 0 & 0 & 0 \\ & & & 0 & 0 & 0 & 0 \\ & & & & 0 & 0 & 0 \\ & & & & & 0 & 0 \\ & & & & & & 0 \end{bmatrix} \quad Q_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ & 2 & 0 & 1 & 0 & 1 & 0 \\ & & 0 & 0 & 0 & 0 & 0 \\ & & & 1 & 0 & 0 & 0 \\ & & & & 1 & 0 & 0 \\ & & & & & 1 & 0 \\ & & & & & & 0 \end{bmatrix} \quad Q_3 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ & 0 & 0 & 0 & 0 & 0 & 0 \\ & & 0 & 0 & 0 & 0 & 0 \\ & & & 0 & 0 & 0 & 0 \\ & & & & 0 & 0 & 0 \\ & & & & & 0 & 0 \\ & & & & & & 1 \end{bmatrix}$$

$$SM'_1 = \begin{pmatrix} 805 & 0 & 805 & 0 & 0 & 0 & 0 \\ 39{,}0 & 0 & 39{,}0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$SM'_2 = \begin{pmatrix} 42 & 99 & 0 & 48 & 42 & 51 & 0 \\ 776 & 1\,476 & 0 & 722 & 776 & 754 & 0 \\ 34{,}3 & 79{,}3 & 0 & 39{,}6 & 34{,}3 & 39{,}7 & 0 \end{pmatrix}$$

$$SM'_3 = \begin{pmatrix} 655 & 0 & 0 & 0 & 0 & 0 & 655 \end{pmatrix}$$

De même, à chacun de ces sous-ensembles correspond une matrice R^{-1} soit :

$$10^4 R_1^{-1} = \begin{pmatrix} 8{,}6117 & -25{,}8750 \\ & 485{,}9086 \end{pmatrix}$$

$$10^4 R_2^{-1} = \begin{pmatrix} 814{,}2060 & -14{,}9232 & -14{,}0132 \\ & 8{,}8852 & -25{,}6182 \\ & & 486{,}1498 \end{pmatrix}$$

$$10^4 R_3^{-1} = 7{,}2338$$

On calcule ensuite les coefficients du système (11).

A^{-1} est la matrice relative aux 7 individus explicitée précédemment.

On prend pour G la matrice de variances et covariances génétiques correspondant aux 4 caractères à prédire.

Le second membre par exemple s'écrit comme suit :

$S' =$ [2,2136 ; 1,6051 ; — 0,5672 5,7469 ; 0,9606 ; — 0,0648
0 ; 0,5923 ; — 0,1879 ; 0 2,7752 ; 0,4684 ; 0,0083 ; 0
2,2136 ; 0,5389 ; — 0,3793 ; 0 2,9716 ; 0,4921 ; — 0,0731 ; 0
0 ; 0,4738 ; 0 ; 0 0 ; 0 ; 0 ; 0 ;
0 ; 0 ; 0 ; 0]

La résolution du système (11) ainsi obtenu, conduit aux solutions données au tableau 3. Les éléments diagonaux de la matrice inverse des coefficients de ce système fournissent les variances d'échantillonnage des estimateurs et prédicteurs. Pour les valeurs génétiques prédites $\hat{u}$, on peut également exprimer la précision sous la forme du coefficient de détermination CD tel que $CD = 1 - \text{var}(\hat{u} - u) / \text{var } u$ (tableau 4).

IV. Discussion - Conclusion

Cet article n'a abordé qu'un des aspects du BLUP multicaractères à savoir l'écriture des équations de résolution dans la situation où les mêmes facteurs affectent chacun des t caractères considérés. Dans ce cas et, en présence d'un seul facteur aléatoire, la résolution ne présente pas de difficulté majeure si l'information est complète sur les t caractères ; en effet, grâce à une transformation canonique, le pro-

blème revient à calculer t BLUP unicaractère sur chacune des variables transformées. Dans une situation quelconque où l'information sur un ou plusieurs caractères peut faire défaut chez certains individus, on aboutit à un système d'équations général permettant notamment une prédiction « directe » d'individus apparentés et de caractères corrélés et pouvant s'appliquer également à plusieurs facteurs aléatoires. Par contre, la résolution de ce système devient nettement plus complexe puisqu'il faut fractionner le fichier en ensembles d'individus dont les résiduelles sont non corrélées d'un ensemble à l'autre. Il faut alors reconstruire dans chacune de ces situations la matrice des coefficients et les seconds-membres des équations normales, les pondérer par les éléments de la matrice inverse des variances et covariances résiduelles, sommer le tout et résoudre. Cela n'est pas sans poser des problèmes numériques de précision dans la constitution des matrices tout d'abord, puis d'efficacité de résolution de systèmes qui deviennent très vite de grande taille. Il serait donc nécessaire de mettre au point des méthodes numériques adéquates si l'on veut élargir l'application à un nombre global de niveaux relativement important (plusieurs centaines à plusieurs milliers).

C'est pourtant, le traitement de fichiers avec information incomplète sur les caractères qui devrait présenter le plus d'intérêt en sélection animale, notamment pour corriger certains biais de prédiction dus à une sélection séquentielle sur caractères (POLLAK & QUAAS, 1981 ; GOFFINET, 1982, communication en cours de publication). A ces fins et, eu égard aux difficultés numériques de résolution du BLUP multicaractères, on pourrait faire appel à des méthodes approchées plus simples à mettre en œuvre. Une possibilité consiste, par exemple, à combiner des BLUP unicaractère suivant la théorie classique des indices de sélection (ERIKSSON, 1981). Mais, il y aurait lieu d'examiner de façon générale l'efficacité relative d'une telle méthode d'autant qu'elle nécessite nombre d'approximations. En effet, en toute rigueur, cette méthode revient à établir un BLUP de BLUP, ce qui n'est guère plus aisé à calculer qu'un BLUP multicaractères.

Reçu pour publication en décembre 1981.

Remerciements

Les auteurs tiennent à remercier B. GOFFINET (I.N.R.A.-Biométrie) et P. LEROY (Université de Liège) pour leurs utiles critiques et suggestions.

Summary

Alternative expressions of multiple trait BLUP equations

This paper gives expressions for multiple trait BLUP equations when the same factors are influencing all traits. Two cases are discussed depending on whether the information on the traits is complete or not.

When all traits are recorded and with only one random factor considered in the model, the problem is simplified because a canonical transformation can be applied to the data or to the right-band sides of the mixed model equations. The system reduces to single trait BLUP evaluations.

When records on some traits are missing, the sample can be divided into disjoint subsets of individuals having the same traits recorded within a group so that residuals are uncorrelated among subsets. In the system so obtained, the number of equations is a multiple

of the number of traits leading to severe numerical difficulties, thus limiting the size of present applications. Nevertheless, it is easy in every case to get predictions of related individuals and correlated traits not described in the model for the data. A simple example is presented to illustrate the formulae.

Annexe

Prédiction indirecte de caractères corrélés non mesurés

Ce problème a été traité de façon théorique par HENDERSON (1977) pour des individus apparentés non mesurés à prédire sur un seul caractère. Les résultats obtenus s'appliquent aisément au modèle mixte pluricaractères. Qu'il s'agisse d'individus ou de caractères non contrôlés. Rappelons brièvement le raisonnement. Soit à prédire le vecteur de variables aléatoires W dont les éléments n'apparaissent pas dans la décomposition des données selon le modèle $\mathcal{Y} = \mathcal{X}b + \mathcal{Z}u + e$.

En supposant que :

$$E(W) = 0$$

$$\text{var}\begin{pmatrix} U \\ W \end{pmatrix} = \begin{pmatrix} \mathcal{G} & \mathcal{H} \\ \mathcal{H}' & \mathcal{C} \end{pmatrix}$$

on montre aisément que le BLUP de W, soit $\hat{W}$ s'obtient comme une combinaison linéaire du BLUP $\hat{U}$ de U par l'équation suivante :

$$\hat{W} = \mathcal{H}'\mathcal{G}^{-1}\hat{U} \qquad (1)$$

En effet :

$$\text{BLUP}(W) = \text{cov}(W, \mathcal{Y}')\,\mathcal{V}^{-1}(\mathcal{Y} - \mathcal{X}\hat{b}) \quad (\mathcal{V} = \text{var}\,\mathcal{Y})$$
$$\hat{W} = \text{cov}(W, U')\,\mathcal{G}^{-1}\mathcal{G}\mathcal{Z}'\mathcal{V}^{-1}(\mathcal{Y} - \mathcal{X}\hat{b})$$

or :

$$\hat{U} = \mathcal{G}\mathcal{Z}'\mathcal{V}^{-1}(\mathcal{Y} - \mathcal{X}\hat{b})$$

Appliquons maintenant ce résultat à l'évaluation des individus I (considérés dans le modèle de décomposition des données sur les r caractères non mesurés (β) pour lesquels on connaît les variances et covariances génétiques entre eux et avec les t caractères mesurés (x)

On posera, pour l'individu $i = 1, 2, ..., q$

- $U'_i = (U'_{\alpha i}, U'_{\beta i})$
- $\text{var}\ U_i = a_{ii}\begin{pmatrix} G_{\alpha\alpha} & G_{\alpha\beta} \\ G_{\beta\alpha} & G_{\beta\beta} \end{pmatrix}$

Si on applique l'équation (7) aux prédictions canoniques $\hat{U}^*$ on obtient :

$$\hat{U}_\beta = \text{cov}(U_\beta, U_\alpha^{*\prime})\,(\text{var}\,U_\alpha^*)^{-1}\,\hat{U}_\alpha^*$$

où :

$$\hat{U}_\alpha^{*\prime} = \hat{U}_{\alpha 1}^{*\prime}\ ;\ \hat{U}_{\alpha 2}^{*\prime}\ ;\ \ldots\ ;\ \hat{U}_{\alpha i}^{*\prime}\ ;\ \ldots\ \hat{U}_{\alpha q}^{*\prime}$$

Soit, encore, compte tenu de la propriété de non corrélation entre variables canoniques :

$$\hat{U}_\beta = (Iq \otimes C)\, \hat{U}^*_\alpha \quad \text{c.-à-d.} \quad \hat{U}_{\beta i} = C\hat{U}^*_{\alpha i} \tag{2}$$

avec :

$$C = G_{\beta\alpha}(\Delta P')^{-1}$$

Δ : matrice diagonale des t valeurs propres de GR^{-1}

P : matrice de passage telle que $\mathcal{Y} = (I \otimes P)\, \mathcal{Y}^*$

La variance des erreurs de prédiction s'obtient alors classiquement par :

$$\text{Var}\,(\hat{U}_{\beta i} - U_{\beta i}) = \text{var}\; U_{\beta i} - C\, \text{var}\, \hat{U}^*_{\alpha i}\, C'$$

or :

$$\text{Var}\, U_{\beta i} = a_{ii}\, G_{\beta\beta}$$

$\text{Var}\, \hat{U}^*_{\alpha i}$ est une matrice diagonale telle que :

$$\text{Diag var}\, \hat{U}^*_{\alpha i} = D_i = a_{ii}\, \text{Diag}\, \Delta - VEP^*_i$$

On en déduit :

$$\text{Diag var}\,(\hat{U}_{\beta i} - U_{\beta i}) = a_{ii}\, \text{Diag}\, G_{\beta\beta} - (C \circledH C)\, D_i$$

$C \circledH C$ produit carré d'HADAMARD de la matrice C.

En posant D, matrice de dimension (t, q) égale à :

$$(\text{Diag A})' \otimes (\text{Diag}\, \Delta) - VEP^*$$

Il vient :

$$\text{Diag var}\; (\hat{U}_\beta - U_\beta) = (\text{Diag A}) \otimes (\text{Diag}\; G_{\beta\beta}) - \text{vec}\,(C \circledH C)\, D \tag{3}$$

formule qui permet de calculer simplement la précision des prédictions sur les r caractères (β).

Références bibliographiques

ARNASON Th., 1981. Prediction of breeding values for multiple traits in small non-random mating (horse populations). 32nd E.A.A.P. meeting, Zagreb, 31 august-3 september 1981. *Com. Horse prod. anim. Genetics,* 10 p.

ERIKSSON J.A., 1981. Best linear unbiased prediction of breeding values with regard to related contemporaries and selection of records. Thesis. Swed. University of agricul. Sci., Report 50.

HENDERSON C.R., 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics,* **32**, 69-83.

HENDERSON C.R., 1977. Best linear unbiased prediction of breeding values not in the model for records. *J. Dairy Sci.*, **60**, 783-787.

HENDERSON C.R., QUAAS R.L., 1976. Multiple trait evaluation using relatives' records. *J. Anim. Sci.*, **43**, 1188-1197.

LEE A.J., 1979. Mixed model, multiple trait evaluation of related sires when all traits are recorded. *J. Anim. Sci.*, **48**, 1079-1088.

POLLAK E.J., QUAAS R.L., 1981. Monte carlo study of genetic evaluations using sequentially selected records. *J. Anim. Sci.*, **52**, 257-264.