INRA Prod. Anim.,
1990, 3 (1), 3-16

V. DUCROCQ

INRA Station de Génétique Quantitative et Appliquée
78350 Jouy-en-Josas

Les techniques d'évaluation génétique des bovins laitiers

Les méthodes d'évaluation génétique des animaux domestiques ont évolué constamment depuis le début du siècle sous l'influence combinée du progrès des connaissances en génétique et en statistique, de l'extension du contrôle laitier et de l'accroissement des moyens informatiques disponibles. La sophistication progressive de ces méthodes a rendu de plus en plus délicate leur explication, alors même que l'utilisation des index de valeur génétique est de mieux en mieux admise par les éleveurs, surtout de bovins laitiers. Nous tentons de montrer ici que l'évolution des méthodes d'indexation résulte naturellement d'un souci de cohérence et de description aussi fidèle que possible des effets, génétiques ou non, qui influencent la production laitière.

L'évaluation génétique (ou « indexation ») d'un reproducteur peut être définie comme l'estimation de sa valeur génétique à partir de performances mesurées sur lui-même et / ou sur des individus apparentés. En général, on ne s'intéresse qu'à sa valeur génétique « additive », c'est-à-dire à celle qu'il sera susceptible de transmettre à sa descendance. Par opposition, la partie non additive du patrimoine génétique d'un individu est recréée aléatoirement à chaque génération.

Résumé

Les méthodes statistiques utilisées au cours des 50 dernières années pour l'évaluation génétique des taureaux et des vaches laitières sont présentées en insistant sur les raisons ayant motivé leur développement, puis leur abandon. La théorie classique des index de sélection appliquée aux données de base telles que les écarts des performances des vaches à leurs contemporaines d'étable est devenue obsolète car la correction des effets du milieu était imparfaite et le progrès génétique sous-jacent n'était pas pris en compte. La meilleure prédiction linéaire non biaisée (BLUP) permet une évaluation simultanée des effets génétiques et du milieu. Appliquée initialement à des modèles simples pour l'estimation des taureaux (modèles père ou père-grand-père), elle est maintenant de plus en plus utilisée avec un « modèle animal », qui permet l'évaluation conjointe des mâles et des femelles. Certaines propriétés intéressantes sont alors respectées. D'autres aspects, tels que la modélisation des performances, la prise en compte de plusieurs lactations et les difficultés de calcul des index sont également abordés.

L'objectif de l'évaluation génétique est de donner aux sélectionneurs - éleveurs et unités de sélection en bovins laitiers - les outils nécessaires au classement aussi juste que possible des reproducteurs selon leur valeur génétique. Les meilleurs animaux ainsi identifiés pourront contribuer à la procréation de la génération suivante de vaches et de taureaux. C'est cette sélection des meilleurs qui est à l'origine de tout progrès génétique. Secondairement, les valeurs génétiques estimées - les « index » - peuvent aussi servir à éviter ou au contraire à rechercher les accouplements entre animaux génétiquement très différents (accouplements raisonnés). Enfin, un calcul de moyennes d'index par catégorie de reproducteurs et par année de naissance ou par génération permet d'analyser a posteriori l'efficacité d'un programme de sélection, au niveau d'un élevage, d'une unité de sélection, d'une race ou d'un pays ainsi que d'en prédire l'évolution dans le proche avenir.

Les premières tentatives d'estimation de la valeur génétique d'un taureau à travers la production de ses filles datent du début du siècle. Très grossières à l'origine - la valeur d'un taureau était appréciée par exemple à partir de la moyenne brute des meilleures lactations de ses filles - les techniques d'évaluation génétique se sont peu à peu améliorées au cours des trois dernières décennies sous l'influence simulta-

née de progrès en statistique, en informatique et en génétique quantitative et grâce au développement du contrôle laitier et de schémas rationnels de sélection. « BLUP », « modèle père », « modèle animal », « approche multicaractère », « écriture matricielle » : l'indexation a atteint un tel niveau de sophistication que sa compréhension semble réservée à quelques spécialistes alors même que l'utilisation pratique de ses résultats ne cesse de s'étendre. Ainsi, la valeur commerciale d'un animal est maintenant plus souvent une fonction de son index que de son aspect extérieur ou de son pedigree. Pour bien percevoir les fondements de l'évaluation génétique et la portée des changements passés et à venir dans ce domaine, il est nécessaire de faire le point.

1 / Les données et le modèle de base

Les performances sur lesquelles est basée toute évaluation laitière sont des mesures, le plus souvent mensuelles, de quantité de lait et de taux butyreux et protéique. Mais celles-ci ne sont pas actuellement utilisées directement dans les calculs. Elles sont combinées en mesures brutes par lactation qui sont ensuite transformées en production « standard ». Par exemple, une production standard adulte en 305 jours (Q_{305}) est calculée à partir d'une production brute (Q) comme :

$$Q_{305} = c_1 . c_2 . Q \qquad [1]$$

où c_1 est un coefficient dépendant du numéro de lactation et de l'âge de la vache et c_2 est une fonction de la durée de la lactation DL ($c_2 = 385 / (80 + DL)$; Poutous *et al* 1981).

Cette première transformation a pour but de corriger les effets d'échelle non liés à la valeur génétique de la vache. Ces corrections multiplicatives ramènent à des niveaux plus comparables les moyennes et les variances de l'ensemble des performances brutes et, à travers le coefficient c_2, réduisent la liaison positive (et donc défavorable pour la fertilité femelle) qui existe entre longueur de lactation et production en 305 jours. Les constantes multiplicatives sont définies une fois pour toutes et ne doivent pas varier sensiblement avec le temps.

L'étape préliminaire à l'évaluation génétique proprement dite est la *modélisation* des performances brutes, que l'on peut définir comme une représentation mathématique des effets génétiques et non génétiques (conditions d'élevage entre autres) influençant le caractère étudié. Cette étape est tout à fait déterminante : de façon évidente, il n'est pas raisonnable d'utiliser des techniques très sophistiquées d'évaluation génétique sur des données décrites selon un modèle d'analyse déficient.

En pratique, le modèle comprend toujours deux parties résumées par l'équation classique de la génétique quantitative :

Phénotype = Génotype + Environnement [2]

Le phénotype est représenté par les mesures de production brutes standardisées (y). Les effets des gènes (génotype) et de l'environnement sont supposés indépendants et additifs. De façon plus précise, on écrit pour une performance y_i d'une vache i :

$$y_i = a_i + (\mu + m_i) + e_i \qquad [3]$$

où : .a_i désigne la valeur génétique additive de i. Sous l'hypothèse d'un déterminisme polygénique du caractère, classique en génétique quantitative, les gènes influençant le caractère considéré sont supposés très nombreux et ayant chacun un effet fiable. En application de la loi des grands nombres, la somme a_i de ces gènes suit une distribution normale. Il est d'autre part important de remarquer que dans le modèle décrit par l'équation [3], a_i est la valeur génétique de l'animal *auteur* de la performance mesurée et non pas celle d'individus apparentés (le père, la mère...) comme il sera étudié plus loin. Un modèle possédant cette caractéristique est appelé *modèle individuel* ou *modèle animal.*

.μ représente une moyenne, par exemple la moyenne des performances brutes standardisées des vaches adultes de la race considérée ayant vêlé une année donnée ;

.m_i est la somme d'effets du milieu *identifiables* influençant la performance y_i, tels que l'âge au vêlage de la vache i, ses mois et année de vêlage, son étable (qui caractérise une alimentation et un type de conduite du troupeau particuliers), etc. Chacun de ces effets est exprimé en écart à μ. Aucune hypothèse n'est faite concernant la distribution statistique de ces effets. On les appelle « effets fixés » : en théorie, ils peuvent prendre n'importe quelle valeur avec la même probabilité ;

.e_i est la « résiduelle » ou l' « erreur » du modèle et englobe l'ensemble des facteurs non encore pris en compte, qu'ils soient d'origine génétique (valeur génétique non additive) ou liés à l'environnement (facteurs non identifiés tels que l'état reproductif de l'animal, le temps séparant 2 traites, la position dans la salle de traite, etc, ou facteurs non identifiables tels que l'erreur commise en calculant la production totale à partir de contrôles mensuels). Là encore, ces facteurs sont nombreux et sont supposés avoir chacun un effet faible sur y_i. En conséquence, on considère que la résiduelle e_i suit une distribution normale de moyenne nulle, indépendante de la distribution de a_i.

Le choix du modèle n'est pas un acte innocent. Considérons par exemple l'effet du numéro de lactation sur la production laitière. Cet effet peut être pris en compte à travers une correction *multiplicative, a priori* comme en [3], qui conduira par exemple à ajouter systématiquement 1 200 kg à la production de cette même vache. Dans bon nombre de pays, seul le premier type de correction est réalisé. En France, Poutous *et al* (1981) ont choisi de combiner les deux types de corrections afin d'affiner cet effet du numéro de lactation en fonction de l'année, la région, etc.

Le choix des effets du milieu est une étape délicate. Ignorer certains effets ayant une influence réelle sur la performance conduit à des estimations biaisées des autres effets, en particulier de la valeur génétique. Par exemple, si on omet de corriger les performances pour

l'effet du mois de vêlage, une vache vêlant en été sera pénalisée : on imputera sa plus faible production laitière à une valeur génétique moins élevée. Le classement des animaux candidats à la sélection sera modifié, conduisant à des erreurs dans les opérations de sélection. Une autre conséquence de ces biais est de surestimer la précision de l'évaluation. Celle-ci est en effet calculée sous l'hypothèse d'absence de biais systématique dans l'évaluation.

Inversement, l'inclusion dans l'analyse de facteurs de variations inutiles, c'est-à-dire dont les effets sont sans importance réelle, est à éviter : elle ne biaise pas l'estimation des autres facteurs mais diminue la précision des valeurs génétiques. Elle accroît aussi les risques de disconnexion dans l'évaluation. La disconnexion est l'absence d'informations permettant une comparaison fiable des animaux. C'est le cas par exemple lorsque l'on tente d'évaluer les valeurs génétiques respectives d'animaux de troupeaux différents fonctionnant exclusivement en circuit fermé (en n'utilisant que la monte naturelle) et n'ayant aucune ascendance commune. Comment alors séparer la supériorité des animaux qui est d'origine génétique de celle due à la technicité de l'éleveur ? La disconnexion est un phénomène bien connu pour les évaluations des bovins à viande (Foulley *et al* 1984) mais peut très bien exister à un autre niveau (entre unités de sélection, entre mois de vêlage dans un même troupeau, etc.) pour les bovins laitiers.

Le dernier élément à considérer dans le choix des effets fixés du modèle est l'existence fréquente d'interactions : l'effet du mois de vêlage, par exemple, ne sera *a priori* pas le même pour une année de sécheresse que pour une année humide, pour une vache en première lactation que pour une vache en troisième lactation. On doit définir alors cet effet du mois de vêlage pour chaque année et pour chaque lactation. Là encore, la prise en compte d'interactions non justifiées diminue la précision et peut créer des disconnexions.

2 / La comparaison aux contemporaines et les index de sélection

La difficulté essentielle de l'évaluation génétique réside dans le besoin de séparer objectivement les effets génétiques des effets de l'environnement. Vers le milieu des années 50, apparaîssait au Royaume-Uni, en Finlande, aux Etats-Unis et en Nouvelle-Zélande, un système d'évaluation génétique des mâles de race laitière appelé « comparaison aux contemporaines ». Son principe est de se débarrasser des effets fixés de [3] en retranchant de la production y_i de la vache i, fille du taureau t, la production moyenne $\bar{y}$ de ses contemporaines d'étable, c'est-à-dire des vaches dont la lactation est soumise exactement aux mêmes effets fixés au même moment.

Avec :

$$y_i = \mu + m_i + \frac{1}{2}a_t + e_i$$

et $\bar{y} = \mu + m_i + \bar{a} + \bar{e}$

on obtient un écart :

$$d_i = y_i - \bar{y} = (\frac{1}{2}a_t - \bar{a}) + (e_i - \bar{e})$$

(e_i inclut ici la partie génétique additive qui n'est pas d'origine paternelle).

On calcule ainsi la moyenne $\bar{w}_t$ des écarts aux contemporaines d_i de toutes les filles du taureau t qui, moyennant certaines hypothèses (niveau génétique des pères et des mères des contemporaines égal à la moyenne de la population), peut être considérée comme une mesure de la moitié de la valeur génétique additive du taureau t (en fait, ce serait la mesure qui serait utilisée comme estimation de la valeur génétique de t si on ignorait l'appartenance de a_t à une distribution normale, c'est-à-dire si on considérait la valeur génétique additive comme un effet fixé). $\bar{w}_t$ est une donnée de base pour l'estimation de la valeur génétique a_t à partir de la théorie des *index de sélection* initialement développée par Smith (1936) et Hazel (1943).

Pour pouvoir relier cette théorie aux méthodes plus modernes d'évaluation, elle peut être présentée comme une méthode statistique de *prédiction* d'une variable aléatoire : la meilleure prédiction linéaire (« Best Linear Prediction », BLP) telle que l'a définie Henderson (1973). Le point de départ est la recherche d'un prédicteur $\hat{a}_i$ d'une variable aléatoire a_i non observable (ici la valeur génétique additive d'un individu i) à partir d'observations w_i. Ces observations sont regroupées dans un vecteur $\mathbf{w}$ [1] et proviennent d'animaux ou de groupes d'animaux apparentés à l'individu i. Pour cela, on suppose connues a priori les espérances $E(a_i)$ de a_i et $E(\mathbf{w})$ de $\mathbf{w}$ (c'est-à-dire leur « moyenne », que l'on supposera nulle sans perte de généralité : $E(a_i) = 0$; $E(\mathbf{w}) = \mathbf{0}$) ainsi que les variances et covariances entre les observations incluses dans $\mathbf{w}$ d'une part et entre celles-ci et a_i d'autre part :

$$\mathrm{Var}(\mathbf{w}) = \mathbf{V} \; ; \; \mathrm{Cov}(a_i, \mathbf{w}) = \mathbf{c}_i$$

$\mathbf{c}_i$ et $\mathbf{V}$ sont des fonctions de la variance génétique et de la variance de la résiduelle du modèle [3]. Ce sont donc en particulier des fonctions de l'héritabilité du caractère. Le prédicteur retenu sera celui dont l'espérance de l'erreur quadratique $E[(\hat{a}_i - a_i)^2]$ - c'est-à-dire la valeur « moyenne » du carré de la différence entre la valeur vraie et la valeur prédite - est minimale (*meilleur* prédicteur) parmi l'ensemble des prédicteurs de la forme

$$\hat{a}_i = \alpha + \beta'\mathbf{w} = \alpha + \beta_1 w_1 + \ldots + \beta_n w_n,$$

c'est-à-dire des combinaisons linéaires des observations (prédicteurs *linéaires*). Selon un résultat classique d'algèbre, la valeur minimale de $E[(\hat{a}_i - a_i)^2]$ est obtenue lorsque ses dérivées partielles par rapport aux inconnus α, $\beta_1, \ldots \beta_n$ sont nulles. On montre ainsi que l'index de sélection pour l'animal i est

$\hat{a}_i = \hat{\beta}'\mathbf{w} = \hat{\beta}_1 w_1 + \hat{\beta}_2 w_2 + \ldots + \hat{\beta}_n w_n$ où $\hat{\alpha} = 0$ et $\hat{\beta}$ est le vecteur solution du système d'équations linéaires :

$$\mathbf{V}\hat{\beta} = \mathbf{c}_i$$

[1] Une lettre minuscule en caractère gras désigne un vecteur ; une lettre majuscule en caractère gras représente une matrice.

Si l'animal i est évalué à partir d'une seule performance propre w_i, le calcul conduit à un index de sélection de la forme $\widehat{a_i} = h^2 w_i$, où h^2 est l'héritabilité du caractère. En ce qui concerne l'évaluation du taureau t précédemment considéré à partir de la moyenne $\overline{w_t}$ des écarts aux contemporaines de ses n filles, la résolution de [4] permet de définir un index de sélection à peine plus compliqué : $\widehat{a_t} = \widehat{\beta_t}\overline{w_t}$ où $\widehat{\beta_t} = 0{,}5\ n\ h^2/[1 + 0{,}25(n-1)h^2]$.

La théorie des index de sélection s'applique exactement selon les mêmes principes, que les observations aient été faites sur un seul individu (ou groupe d'individus) ou sur plusieurs, ou bien qu'elles concernent un seul caractère ou plusieurs (Rouvier 1969). Si toutes les hypothèses faites sont adéquates, l'index de sélection possède un certain nombre de propriétés intéressantes : c'est un prédicteur non biaisé ($E(\widehat{a_i}) = E(a_i)$), c'est-à-dire que si on calculait un grand nombre de fois $\widehat{a_i}$ à partir d'observations indépendantes, *en moyenne* cette valeur génétique estimée serait égale à la valeur vraie. L'index de sélection conduit à l'estimée $\widehat{a_i}$ qui maximise la corrélation entre valeur prédite et valeur vraie ($\rho(\widehat{a_i}, a_i)$ maximum). Enfin, il correspond à la prédiction qui maximise la probabilité d'un classement correct des valeurs génétiques des animaux pris 2 à 2.

Mais en pratique, les hypothèse faites sont très fréquemment mises à défaut, tout particulièrement celle prétendant la connaissance *a priori* de l'espérance des observations ($E(\mathbf{w}) = \mathbf{0}$). Celle-ci suppose une correction parfaite des données pour les effets fixés, ce qui n'est réalisable que lorsque les contemporaines sont nombreuses et représentatives de l'ensemble de la population. Or, le nombre de contemporaines « vraies » d'une vache, c'est-à-dire soumises exactement aux mêmes effets de l'environnement, est en général très réduit. De plus, une constatation fondamentale met en cause la représentativité « génétique » de ces contemporaines. Dans la comparaison aux contemporaines, les taureaux de testage sont systématiquement désavantagés : leurs filles sont comparées à des filles de meilleurs taureaux - car sélectionnés - taureaux qui en outre sont plus vieux et ont donc eux-mêmes été évalués à une période où le niveau moyen des contemporaines d'étable de leurs filles était plus faible (Harville et Henderson 1969, Hargrove et Legates 1971, Everett et Quaas 1979). Ce problème n'existait évidemment pas à l'origine, lorsque les schémas de sélection ont été mis en place et que le progrès génétique de l'ensemble de la population était très faible. C'est ce principal facteur de distorsion dans les évaluations qui a conduit les généticiens de l'Université Cornell (USA) à abandonner la comparaison aux contemporaines en 1970 et ceux de l'USDA (Ministère de l'Agriculture des USA) en 1974 et de France en 1978 à la modifier en introduisant une correction des écarts pour le niveau génétique moyen du père des contemporaines.

Il existe d'autres problèmes non pris en compte par la théorie des index de sélection, au moins dans sa version simple utilisée en pratique, comme l'existence de biais dus aux différences de niveau génétique moyen des vaches accouplées aux divers taureaux (Everett 1974) ou l'absence de prise en compte du fait que la variance génétique des caractères varie au cours des générations sous l'effet de la sélection et des accouplements raisonnés. Enfin, on peut reprocher à cette méthode son manque de souplesse : si pour chaque animal, on a un nombre variable de performances propres et / ou d'individus apparentés, il faut à chaque fois recalculer les coefficients de l'index. Cette contrainte a conduit les généticiens - en particulier dans les espèces porcine et avicoles - à se limiter à un nombre relativement faible de sources d'information différentes (parents, frères ou demi-frères, descendants de première génération).

3 / Le BLUP, les équations du modèle mixte

Pour contourner l'ensemble des difficultés liées à la comparaison aux contemporaines et définir un cadre d'évaluation plus général que la théorie des index de sélection, Henderson (1959, 1963 et 1973) a développé une méthode statistique d'évaluation *simultanée* des effets fixés et aléatoires du « modèle mixte » défini en (3) : la meilleure prédiction linéaire non biaisée (« Best Linear Unbiased Prediction » = BLUP). La différence essentielle avec la théorie des index de sélection est que l'on abandonne l'hypothèse de connaissance de *l'espérance* des observations, précédemment représentée par $E(\mathbf{w}) = \mathbf{0}$.

Considérons de nouveau les performances élémentaires y_i, regroupées dans un vecteur de performances **y**. Pour l'ensemble des animaux, on peut réécrire le modèle [3] sous forme matricielle :

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \qquad [5]$$

où **b**, **a** et **e** sont des vecteurs regroupant respectivement l'ensemble des effets fixés (les éléments inclus dans m_i dans l'équation [3]), des valeurs génétiques et des résiduelles. **X** et **Z** sont des matrices dites « d'incidence » qui relient les observations aux effets fixés et aléatoires qui les ont influencées. Par exemple, pour le pedigree et les performances des animaux schématisés à la figure 1, on écrira :

$$\begin{bmatrix} y_2 \\ y_4 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} 5000 \\ 4500 \\ 5500 \\ 5000 \\ 6000 \end{bmatrix} = \begin{bmatrix} 100 \\ 100 \\ 010 \\ 010 \\ 001 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} 10000 \\ 01000 \\ 00100 \\ 00010 \\ 00001 \end{bmatrix} \begin{bmatrix} a_2 \\ a_4 \\ a_6 \\ a_7 \\ a_8 \end{bmatrix} + \begin{bmatrix} e_2 \\ e_4 \\ e_6 \\ e_7 \\ e_8 \end{bmatrix} \qquad [6]$$

où b_1, b_2 et b_3 correspondent aux effets « année » 1, 2 et 3. On retrouve bien, par exemple, pour la vache 6 : $y_6 = 5500 = b_2 + a_6 + e_6$. En fait, pour pouvoir inclure dans l'évaluation des animaux qui n'ont pas de performance propre mais qui nous intéressent, tels que les taureaux 1, 3 et 5, on modifie la matrice d'incidence **Z** en lui ajoutant des colonnes nulles.

Ainsi :

$$\begin{bmatrix} y_2 \\ y_4 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} 100 \\ 100 \\ 010 \\ 010 \\ 001 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} 01000000 \\ 00010000 \\ 00000100 \\ 00000010 \\ 00000001 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \end{bmatrix} + \begin{bmatrix} e_2 \\ e_4 \\ e_6 \\ e_7 \\ e_8 \end{bmatrix} \quad [7]$$

Les espérances de **a** et **e** sont supposées connues (E(**a**) = E(**e**) = **0**) ainsi que les matrices de variances et covariances des effets aléatoires : Var(**a**) = **G**, Var(**e**) = **R**, Cov(**a**, **e'**) = **0**. On a alors Var(**y**) = **V** = **Z G Z'** + **R** et cov(a_i, **y**) = $\mathbf{c}_i$ est la i^e colonne de **Z G** .Par construction, E (**y**) = **X b** où **b** est *inconnu*.

Comme pour les index de sélection, le prédicteur $\widehat{a_i}$ de a_i que l'on retiendra sera celui pour lequel E [($\widehat{a_i}$ - a_i)^2] est minimum (*meilleur* prédicteur) parmi ceux de la forme

$$\widehat{a_i} = \alpha + \beta'\mathbf{y} = \alpha + \beta_1 y_1 + .. + \beta_n y_n$$

(prédicteurs *linéaires*). Pour estimer **b**, une condition supplémentaire est introduite : le prédicteur doit être *sans biais*, soit E ($\widehat{a_i}$) = E (a_i). Cette absence de biais était une propriété des index de sélection. C'est ici une contrainte que l'on choisit. Elle revient à imposer que la prédiction de a_i ne soit en moyenne pas affectée par les effets fixés. Cette contrainte est choisie par souci d'objectivité : il semble raisonnable d'imposer dans une évaluation nationale qu'en moyenne, une vache ne soit pas systématiquement surévaluée ou sous-évaluée suivant son élevage, sa région, sa période de vêlage, etc.

La recherche du minimum d'une fonction avec contrainte passe par l'annulation des dérivées partielles de E[($\widehat{a_i}$-a_i)^2]-λ [E($\widehat{a_i}$) - E (a_i)] par rapport à α, β_1, ... β_n et λ.λ est une variable fonctionnelle appelée multiplicateur de Lagrange. La résolution du système des dérivées partielles conduit à un prédicteur tel que :

$$\widehat{a_i} = \widehat{\beta}' (\mathbf{y} - \mathbf{X}\widehat{\mathbf{b}}) =$$

$$\widehat{\beta_1} (y_1 - \Sigma_j \widehat{b_{j\,[1]}}) + ... + \widehat{\beta_n} (y_n - \Sigma_j \widehat{b_{j\,[n]}}) \quad [8]$$

où :

.$\widehat{\beta_1}$, ... $\widehat{\beta_n}$ sont solutions du système $\mathbf{V}\widehat{\beta} = \mathbf{c}_i$

.$\Sigma_j \widehat{\beta_{j[i]}} = \widehat{m_i}$ est la somme de solutions des « moindres carrés généralisés » des effets fixés affectant la performance de l'animal i (l'estimation par moindres carrés généralisés des effets fixés d'un modèle M est la recherche des solutions de ces effets telles que la somme *pondérée* des carrés des erreurs de prédiction sur l'ensemble des observations, $\Sigma\, \omega_k [y_k - \widehat{y_k}(M)]^2$, soit minimale. Les pondérations ω_k prennent alors en compte la structure non homogène de variance et covariance des performances). En fait, il peut arriver qu'il existe une infinité de solutions $\widehat{b_{j[i]}}$ du système des moindres carrés généralisés, mais on peut montrer que le choix de n'importe laquelle de ces solutions ne modifie pas la valeur de $\widehat{a_i}$.

Sous cette forme, l'évaluation BLUP des valeurs génétiques a_i peut être considérée comme celle résultant de l'application de la théorie des index de sélection sur des données corrigées à partir de la meilleure estimation possible (au sens des moindres carrés généralisés) des effets fixés. Elle possède les mêmes propriétés que les index de sélection mais sous des hypothèses de départ moins strictes. Goffinet et Elsen (1984) ont également montré qu'une sélection sur les estimées BLUP des animaux maximise l'espérance de la valeur génétique vraie des sélectionnés. D'autres justifications du BLUP ont été données, par exemple par Gianola *et al* (1986).

Malgré tous ces avantages, une difficulté majeure pourrait laisser penser que ce type d'évaluation est irréalisable en pratique : le système d'équations à résoudre pour obtenir des solutions des moindres carrés généralisés des effets fixés **b** est de la forme :

$$(\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})\, \mathbf{b} = \mathbf{X}' \mathbf{V}^{-1} \mathbf{y} \quad [9]$$

Mais **V** = **Z G Z'** + **R** est une matrice dont le nombre de lignes et de colonnes est égal au nombre d'observations. Or, même sur les plus gros ordinateurs, on ne peut inverser directement une matrice dont la taille dépasse quelques centaines ou quelques milliers de lignes. Sous cette forme, l'obtention des estimées BLUP des valeurs génétiques pour des populations de grande taille paraît donc impossible.

En 1963, Henderson a montré qu'on pouvait alléger considérablement les calculs. Supposons que dans le modèle [5] :

$$\mathbf{y} = \mathbf{X b} + \mathbf{Z a} + \mathbf{e}$$

les valeurs génétiques additives soient considérées de façon provisoire comme fixées et non plus comme aléatoires. On a alors Var(**y**) = Var (**e**) = **R**. Les équations des moindres carrés généralisés pour l'estimation de **b** et **a** s'écrivent, par extension de l'équation [9] :

$$\begin{bmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{bmatrix} [\mathbf{R}]^{-1} [\mathbf{X} \quad \mathbf{Z}] \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X}' \\ \mathbf{Z}' \end{bmatrix} [\mathbf{R}]^{-1} \mathbf{y}$$

soit :

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \quad [10]$$

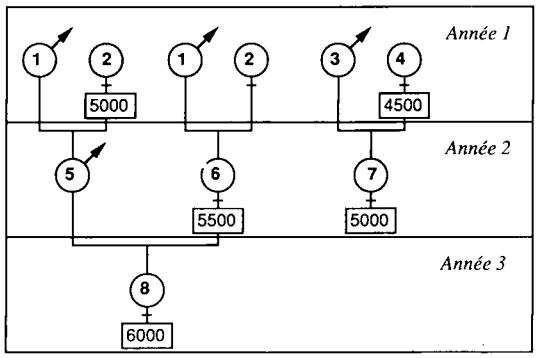
Henderson a découvert qu'en modifiant légèrement ces équations des moindres carrés généralisés et plus précisément en ajoutant l'inverse de la matrice Var (**a**) = **G** au bloc **Z'R^{-1} Z** de la matrice de gauche de [10], le système obtenu fournit *simultanément* les solutions $\widehat{\mathbf{b}}$ des moindres carrés généralisés des effets fixés et les solutions BLUP $\widehat{\mathbf{a}}$ des valeurs génétiques :

[11]

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

Ce système d'équations s'appelle *équations du modèle mixte* (mixte parce que l'on considère à la fois des effets fixés et des effets aléatoires). Elles présentent un intérêt considérable. En effet, les matrices inverses **R^{-1}** et **G^{-1}** sont beaucoup plus simples à calculer que **V^{-1}**. Par exemple, on fait très fréquemment l'hypothèse que les résiduelles du modèle [5] sont distribuées selon une loi normale de moyenne nulle

Figure 1. *Généalogie et performances utilisées dans l'exemple (voir texte).*



et de variance ∂_e^2 et sont indépendantes entre elles. On a alors Var (**e**) = **R** = σ_e^2 **I** où **I** est la matrice identité (matrice carrée dont les éléments de la diagonale descendante sont égaux à 1 et dont tous les autres éléments sont nuls), et donc $\mathbf{R}^{-1} = (\sigma_e^2)^{-1}$ **I**.

Le calcul de $\mathbf{G}^{-1}$ est moins immédiat mais ne pose pas de difficultés majeures (cf. encadré).

4 / Le modèle animal

Connaissant $\mathbf{G}^{-1}$ et $\mathbf{R}^{-1}$, il est possible d'écrire les équations du modèle mixte correspondant à l'exemple de la figure 1. Pour simplifier les notations, on peut multiplier toutes les équations du système [11] par σ_e^2, ce qui revient à supprimer la matrice $\mathbf{R}^{-1}$. A partir des valeurs des matrices **X** et **Z** correspondant à l'exemple (cf. équation [7]), on obtient :

La matrice de parenté et son inverse

Cette partie est largement inspirée des présentations de Quaas (1984) et Kennedy *et al* (1988)

Lors de la définition du modèle [5], il a été supposé que la matrice Var (**a**) = **G** était connue. Pour bien comprendre comment celle-ci et son inverse sont obtenues, il n'est pas inutile de revenir aux principes de base de la génétique. On sait que les gènes d'un individu i proviennent pour une moitié de son père p et pour l'autre moitié de sa mère m, mais qu'ils ne sont réellement identiques ni à l'un ni à l'autre à cause des recombinaisons des gènes ayant lieu au cours de la méïose. On écrira :

$$a_i = 0{,}5\, a_p + 0{,}5\, a_m + \phi_i \qquad [12]$$

où a_i, a_p et a_m sont les valeurs génétiques additives (la somme des effets moyens des gènes) de l'animal i, de son père et de sa mère et ϕ_i correspond à la « contribution » de l'aléa de la méïose à la valeur finale a_i. Compte tenu des hypothèses classiques concernant le nombre et le mode d'actions des gènes, on a :

$$E(\phi_i) = 0 \text{ et } Var(\phi_i) = \frac{1}{2}\left(1 - \frac{F_p + F_m}{2}\right)\sigma_a^2 \qquad [13]$$

(Foulley et Chevalet 1981 ; Verrier 1989)

où F_p et F_m sont les coefficients de consanguinité du père et de la mère de i et σ_a^2 est la variance génétique additive du caractère étudié. Pour pouvoir introduire dans l'analyse des animaux dont les parents sont inconnus, on peut modifier et étendre l'équation [12] :

• Lorsque les parents de i sont connus :

$$a_i = 0{,}5\, a_p + 0{,}5\, a_m + \psi_i$$

$$\text{avec } E(\psi_i) = 0 \text{ et } Var(\psi_i) = Var(\phi_i) \qquad [14]$$

• Si un des parents, par exemple la mère, est inconnu et non sélectionné :

$$a_i = 0{,}5\, a_p + \psi_i$$

$$\text{avec } E(\psi_i) = 0 \text{ et } Var(\psi_i) = \left(\frac{3}{4} - \frac{1}{4} F_p\right)\sigma_a^2 \qquad [15]$$

• Lorsque les parents de i sont inconnus et non sélectionnés (animaux « de base ») :

$$a_i = \psi_i$$

$$\text{avec } E(\psi_i) = 0 \text{ et } Var(\psi_i) = \sigma_a^2 \qquad [16]$$

ψ_i est ici une sorte de résiduelle regroupant tout ce qui, dans la valeur génétique additive de i, ne peut être expliqué par l'ascendance. L'absence de sélection des parents inconnus est indispensable pour avoir E (ψ_i) = 0.

On aura pour l'exemple de la figure 1 :

$a_8 = 0{,}5\, a_5 + 0{,}5\, a_6 + \psi_8$

mais comme : $a_5 = 0{,}5\, a_1 + 0{,}5\, a_2 + \psi_5$

$a_6 = 0{,}5\, a_4 + 0{,}5\, a_2 + \psi_6$

$a_1 = \psi_1$

$a_2 = \psi_2$

on a : $a_8 = 0{,}5\, \psi_1 + 0{,}5\, \psi_2 + 0{,}5\, \psi_5 + 0{,}5\, \psi_6 + \psi_8$

Sous forme matricielle, on obtient pour l'ensemble des animaux :

$$\mathbf{a} = \mathbf{T}\,\psi \qquad [17]$$

avec :

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0{,}5 & 0{,}5 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0{,}5 & 0{,}5 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0{,}5 & 0{,}5 & 0 & 0 & 1 & 0 \\ 0{,}5 & 0{,}5 & 0 & 0 & 0{,}5 & 0{,}5 & 0 & 1 \end{bmatrix} \qquad \psi = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \psi_3 \\ \psi_4 \\ \psi_5 \\ \psi_6 \\ \psi_7 \\ \psi_8 \end{bmatrix}$$

La matrice **T** est très simple à construire. Il suffit pour cela de calculer systématiquement les lignes correspondant aux descendants *après* celles de leurs parents selon la formule : $t_{ij} = 0{,}5\,(t_{pj} + t_{mj})$ où p et m sont les parents de i et $t_{ij} = 0$ si i n'a pas de parents connus. Par exemple : $t_{81} = 0{,}5\,(t_{51} + t_{61}) = 0{,}5\,(0{,}5 + 0{,}5) = 0{,}5$. L'équation [17] montre que chaque valeur génétique peut être exprimée comme combinaison linéaire des valeurs génétiques des animaux sans parents connus et des aléas de méïose créés à chaque génération. Sous le modèle polygénique, ces aléas de méïose sont bien sûr indépendants des valeurs génétiques des parents et sont donc non affectés par la sélection et la dérive génétique. Les animaux sans parents connus constituent une population de fondateurs, *supposés non consanguins et non sélectionnés.*

$$\mathbf{X'X} = \begin{bmatrix} 200 \\ 020 \\ 001 \end{bmatrix} \qquad \mathbf{X'Z} = \begin{bmatrix} 01010000 \\ 00000110 \\ 00000001 \end{bmatrix} \qquad \mathbf{X'y} = \begin{bmatrix} y_2+y_4 \\ y_6+y_7 \\ y_8 \end{bmatrix}$$

$$\mathbf{Z'Z} = \begin{bmatrix} 00000000 \\ 01000000 \\ 00000000 \\ 00010000 \\ 00000000 \\ 00000100 \\ 00000010 \\ 00000001 \end{bmatrix} \qquad \mathbf{Z'y} = \begin{bmatrix} 0 \\ y_2 \\ 0 \\ y_4 \\ 0 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix}$$

Ces matrices et ces vecteurs ont une interprétation simple : l'élément (j, j) de la matrice **X'X** représente le nombre d'observations y_i recueillies l'année j. L'élément j du vecteur **X'y** est la somme de ces observations. De même, l'élément (k, k) de **Z'Z** est égal au nombre de performances réalisées par l'animal k et la somme de ces performances se trouve à la k-ième ligne de **Z'y**. Enfin, l'élément (j, k) de la matrice **X'Z** est égal au nombre de performances de l'animal k réalisées l'année j.

En posant $\alpha = \sigma_e^2 / \sigma_a^2$, on obtient le système des équations du modèle mixte correspondant au modèle animal décrit par l'équation [7] :

[19]

$$\begin{bmatrix} 2 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 2\alpha & \alpha & 0 & 0 & -\alpha & -\alpha & 0 & 0 \\ 1 & 0 & 0 & \alpha & 1+2\alpha & 0 & 0 & -\alpha & -\alpha & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1{,}5\alpha & 0{,}5\alpha & 0 & 0 & -\alpha & 0 \\ 1 & 0 & 0 & 0 & 0 & 0{,}5\alpha & 1+1{,}5\alpha & 0 & 0 & -\alpha & 0 \\ 0 & 0 & 0 & -\alpha & -\alpha & 0 & 0 & 2{,}5\alpha & 0{,}5\alpha & 0 & -\alpha \\ 0 & 1 & 0 & -\alpha & -\alpha & 0 & 0 & 0{,}5\alpha & 1+2{,}5\alpha & 0 & -\alpha \\ 0 & 1 & 0 & 0 & 0 & -\alpha & -\alpha & 0 & 0 & 1+2\alpha & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & -\alpha & -\alpha & 0 & 1+2\alpha \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \end{bmatrix} = \begin{bmatrix} y_2+y_4 \\ y_6+y_7 \\ y_8 \\ 0 \\ y_2 \\ 0 \\ y_4 \\ 0 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix}$$

A partir de l'équation [17], on tire :

$$\mathbf{G} = \text{Var}(\mathbf{a}) = \text{Var}(\mathbf{T}\psi) = \mathbf{T}\,\text{Var}(\psi)\,\mathbf{T'} = \mathbf{T}(\mathbf{D}\,\sigma_a^2)\,\mathbf{T'} = \mathbf{A}\,\sigma\frac{2}{a} \qquad [18]$$

avec $\mathbf{A} = \mathbf{T}\,\mathbf{D}\,\mathbf{T'}$ et :

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0{,}5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0{,}5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0{,}5 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0{,}5 \end{bmatrix}$$

D est une matrice diagonale dont l'élément (i, i) est égal à 1, 3/4 ou 1/2 suivant que l'animal i a 0, 1 ou 2 parents connus. **A** est appelée *matrice de parenté* : l'élément (i, j) de cette matrice est égal à 2 fois le coefficient de parenté entre i et j, tel que défini par Malécot (1948).

Compte tenu des remarques précédentes, il apparaît que cette manière de construire **G** ou **A** prend en compte les changements de variance génétique dus à la sélection, à la dérive génétique et même à la consanguinité. L'animal 8 de l'exemple est consanguin et on peut vérifier, à partir de l'équation [18], que :

$$\text{var}(a_8)$$
$$= \text{var}(0{,}5\,\psi_1 + 0{,}5\,\psi_2 + 0{,}5\,\psi_5 + 0{,}5\,\psi_6 + \psi_8)$$
$$= [(0{,}5)^2(1) + (0{,}5)^2(1) + (0{,}5)^2(0{,}5) + (0{,}5)^2(0{,}5) + (1)^2(0{,}5)]\,\sigma_a^2$$
$$= 1{,}25\,\sigma_a^2 \neq \sigma_a^2$$

L'équation [18] est également très utile pour obtenir l'inverse de G. En effet :

$$\mathbf{G} = \mathbf{A}\,\sigma_a^2 = \mathbf{T}\,\mathbf{D}\,\mathbf{T'}\,\sigma_a^2 \text{ donc } \mathbf{G}^{-1} = \mathbf{A}^{-1}(\sigma_a^2)^{-1}$$
et $\mathbf{A}^{-1} = (\mathbf{T'})^{-1}\,\mathbf{D}^{-1}\,\mathbf{T}^{-1}$

Or l'inverse de $\mathbf{D}^{-1}$ est très simple : c'est une matrice diagonale dont l'élément (i,i) est égal à 1, 4/3 ou 2 suivant que i a 0, 1 ou 2 parents connus. $\mathbf{T}^{-1}$ a aussi une structure très simple car chaque ligne i a au plus 3 éléments non nuls : 1 sur la diagonale et − 0,5 dans les colonnes correspondant aux parents de i :

$$\mathbf{T}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -0{,}5 & -0{,}5 & 0 & 0 & 1 & 0 & 0 & 0 \\ -0{,}5 & -0{,}5 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -0{,}5 & -0{,}5 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & -0{,}5 & -0{,}5 & 0 & 1 \end{bmatrix}$$

et donc : $\mathbf{A}^{-1} = \begin{bmatrix} 2 & 1 & 0 & 0 & -1 & -1 & 0 & 0 \\ 1 & 2 & 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1{,}5 & 0{,}5 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0{,}5 & 1{,}5 & 0 & 0 & -1 & 0 \\ -1 & -1 & 0 & 0 & 2{,}5 & 0{,}5 & 0 & -1 \\ -1 & -1 & 0 & 0 & 0{,}5 & 2{,}5 & 0 & -1 \\ 0 & 0 & -1 & -1 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & 0 & 2 \end{bmatrix}$

Cette présentation permet de retrouver les *règles de Henderson* (1976) de calcul direct de l'inverse de la matrice de parenté $\mathbf{A}^{-1}$. Elles peuvent être résumées de la façon suivante :

- Si p et m sont le père et la mère de i et ne sont pas eux-mêmes consanguins :
  - ajouter 2 à l'élément (i,i) de $\mathbf{A}^{-1}$
  - ajouter − 1 à (i,p), (p,i), (i,m), (m,i)
  - ajouter 1/2 à (p,p), (p,m), (m,p), (m,m)
- Si l'un des parents est inconnu, par exemple la mère :
  - ajouter 4/3 à (i,i)
  - ajouter − 2/3 à (i,p), (p,i)
  - ajouter 1/3 à (p,p)
- Si les 2 parents de i sont inconnus :
  - ajouter 1 à (i,i)

Compte tenu de la définition $h^2 = \sigma_a^2 / (\sigma_e^2 + \sigma_a^2)$ de l'héritabilité, α s'exprime très simplement en fonction de celle-ci : $\alpha = (1 - h^2) / h^2$. On peut remarquer également que les valeurs génétiques estimées solutions du système [19] sont obtenues directement, sans l'étape intermédiaire de calcul des coefficients $\hat{\beta}_i$ de l'expression [8].

A ce stade, il est particulièrement instructif de considérer la manière par laquelle chaque effet peut être logiquement estimé si tous les autres effets étaient connus et de montrer le lien entre l'estimée ainsi obtenue et la solution correspondante du système [19]. Par exemple, il paraît naturel d'estimer chaque effet fixé par une moyenne des performances influencées par cet effet, après correction pour tous les autres effets. Le premier effet « année » de l'exemple de la figure 1 serait ainsi estimé comme la moyenne des lactations des vaches 2 et 4 traites pendant l'année 1, après prise en compte de leur valeur génétique :

$$\hat{b}_1 = \tfrac{1}{2}[(y_2 - \hat{a}_2) + (y_4 - \hat{a}_4)] \qquad [20]$$

En réarrangeant ces termes, on obtient la première équation du système [19] :

$$2\hat{b}_1 + \hat{a}_2 + \hat{a}_4 = y_2 + y_4 \ (= 9500) \qquad [21]$$

Pour l'estimation des valeurs génétiques, 3 sources d'information sont en théorie disponibles : pour chaque animal i, a_i peut être obtenu à partir des performances propres de i, à partir de la connaissance de la valeur génétique de ses parents ou à partir de celle de ses descendants. Notons que l'information provenant des collatéraux de i (frères, sœurs, cousins, etc.) n'est pas considérée directement ici car elle est logiquement prise en compte indirectement, à travers son influence sur la valeur génétique estimée des ascendants communs. Pour calculer chacune des contributions de ces différentes sources d'information, il est utile de revenir aux modèles de base des équations [5] et [14] à [16] (cf encadré). Comme pour l'obtention des estimées des moindres carrés généralisés, l'estimée finale est une moyenne pondérée pour laquelle le poids donné à chaque source d'information est égal à l'inverse de la variance de la résiduelle de l'équation lui correspondant. La contribution à la prédiction de a_i d'une performance propre y_i est ainsi obtenue à partir de l'équation [5] comme :

$$(\hat{a}_i) = y_i - \sum_k \hat{b}_k$$

avec un poids de $1/\text{var}(e_i) = 1/\sigma_e^2$ [22]

Pour la prise en compte de l'ascendance, les équations [14] à [16] conduisent aux contributions suivantes :

• Si les deux parents p et m sont connus et non consanguins :

$$(\hat{a}_i) = \frac{\hat{a}_p + \hat{a}_m}{2}$$

$$\text{avec un poids de } 1/\text{var}(\psi_i) = 1 / \left[\frac{\sigma_a^2}{2}\right] \qquad [23]$$

• Si un seul parent est connu, par exemple le père :

$$(\hat{a}_i) = \frac{\hat{a}_p}{2}$$

$$\text{avec un poids de } 1/\text{var}(\psi_i) = 1 / \left[\frac{3\sigma_a^2}{4}\right] \qquad [24]$$

• Si aucun parent n'est connu :

$(\hat{a}_i) = 0$

avec un poids de $1/\text{var}(\psi_i) = 1/\sigma_a^2$ [25]

Enfin, la contribution de chaque descendant f de i à l'estimation de a_i est également obtenue à partir des équations [14] et [15] :

Si le conjoint c de i est connu (équation [14]) :

$$(\hat{a}_i) = 2\left(\hat{a}_f - \frac{\hat{a}_c}{2}\right) \qquad [26]$$

$$\text{avec un poids de } 1/\text{var}(2\psi_i) = 1 / [2\sigma_a^2]$$

Si le conjoint est inconnu (équation [15]) :

$(\hat{a}_i) = 2\hat{a}_f$

avec un poids de $1/\text{var}(2\psi_i) = 1/[3\sigma_a^2]$ [27]

La correction de a_f pour la valeur génétique a_c du conjoint c de i dans l'équation [26] est fondamentale car elle montre que l'évaluation de i ne sera pas biaisée par des accouplements raisonnés (non aléatoires) avec des animaux meilleurs ou moins bons que la moyenne de la population.

La moyenne pondérée de ces différentes contributions peut s'écrire de façon générale :

$$\hat{a}_i = \frac{1}{1 + \alpha d_i + \sum_f \frac{\alpha}{4} d_f} \left\{ 1 \left[y_i - \sum_k \hat{b}_k \right] + \alpha d_i \left[\frac{\hat{a}_p + \hat{a}_m}{2} \right] + \sum_f \frac{\alpha}{4} d_f \cdot 2 \left(\hat{a}_f - \frac{\hat{a}_c}{2} \right) \right\} \qquad [28]$$

où d_x (x = i ou f) est égal à 2, 4/3 ou 1 suivant que x a 2, 1 ou aucun parent connu et $\hat{a}_p$, $\hat{a}_m$ et $\hat{a}_c$ sont pris égaux à 0 lorsque p, m ou c sont inconnus.

L'équation [28] permet de mieux raisonner les facteurs influençant le calcul d'index : on sait par exemple qu'une forte production y_i peut éventuellement provenir d'un traitement préférentiel de l'animal i, que l'on définira comme une conduite (nourriture, rythme de reproduction, âge au premier vêlage, etc) de cet animal délibérément différente de celle de ses contemporains. Dans un tel cas, l'écart $[y_i - \sum \hat{b}_k]$ est surévalué et la contribution de sa performance propre à son index $\hat{a}_i$ conduira à une évaluation biaisée de celui-ci, comme le montre l'équation [28]. Il n'existe malheureusement pas à l'heure actuelle de moyens efficaces de détection et de correction de ces biais.

Pour illustrer la démarche ayant permis d'obtenir l'équation [28], considérons l'animal 8 de l'exemple de la figure 1, qui a 2 parents connus et une performance propre y_8. La contribution de cette performance à son évaluation est $y_8 - \hat{b}_3$ avec un poids de 1 ; celle de son ascendance est $\frac{\hat{a}_5 + \hat{a}_6}{2}$ avec un poids de 2 α et par conséquent :

$$\hat{a}_8 = \frac{1}{1 + 2\alpha} \left\{ 1 (y_8 - \hat{b}_3) + 2\alpha \left(\frac{\hat{a}_5 + \hat{a}_6}{2} \right) \right\} \qquad [29]$$

Résolution des équations du modèle mixte

Les équations du modèle mixte constituent un système d'équations linéaires classiques, dont la caractéristique principale est sa grande taille, qui rend difficile sa résolution. Considérons, par exemple, la résolution du système suivant de deux équations à deux inconnues x et y :

$$\begin{cases} 3x + y = 2 \\ x + 2y = -1 \end{cases}$$

soit sous forme matricielle :

$$\begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

La technique consistant à inverser la matrice des coefficients permet d'obtenir les solutions suivantes :

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \frac{1}{5} \begin{bmatrix} 2 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

Mais cette technique est très lourde. Le nombre d'opérations à réaliser pour inverser une matrice de n lignes et n colonnes est de l'ordre de n^3, soit de l'ordre de 10^{18} opérations pour un millions d'équations ! En fait, l'inversion directe n'est pas envisageable pour des systèmes comprenant plus de quelques milliers d'inconnues, même sur les ordinateurs les plus puissants.

Pour réduire la taille du système, on peut éliminer certaines inconnues en les exprimant en fonction des autres. C'est la technique dite d'*absorption*. Par exemple, pour l'exemple ci-dessus, la deuxième équation peut s'écrire y = 0,5 (-1-x), et en remplaçant y par cette expression dans la première équation, on se ramène à un système de taille plus réduite (une seule équation) : 3 x + 0,5 (-1 - x) = 3. Par conséquent, x = 1 et en remplaçant x par sa valeur dans l'équation de y : y = 0,5 (-1 -1) = -1. Compte tenu de la taille du système, cette technique n'est réellement intéressante que lorsqu'il est possible « d'absorber » un grand nombre d'équations simplement et en même temps. En pratique, là encore, l'absorption seule n'est pas une solution envisageable dans le cas des équations du modèle animal (sauf pour les effets d'environnement permanent), car la plupart des inconnues ne s'exprime pas simplement en fonction des autres.

Pour contourner ces difficultés, la technique de choix est l'utilisation d'une méthode *itérative* de résolution (Golub et Van Loan 1983). Celle-ci consiste à résoudre séquentiellement chaque équation du système, en supposant que la valeur de toutes les inconnues sauf une a déjà été obtenue. Pour l'exemple retenu ici, on considérera la première équation en supposant que y est connu et égal par exemple à 0. En résolvant pour x, on obtient x = 2/3. En remplaçant x par cette solution dans la deuxième équation, on a y = -5/6. A partir de cette nouvelle valeur de y, la première équation donnera x = 17/18, et ainsi de suite. Moyennant certaines conditions sur la nature du système, conditions toujours respectées par les équations du modèle mixte, on obtient une suite de solutions (x ; y) qui, à chaque « cycle » appelé *itération*, se rapprochent (elles *convergent*) de la solution vraie (1 ; -1) : (2/3 ; 5/6), (17/18 ; 35/36), (107/108 ; -215/216), (647/648 ; -1295/1296), etc. Ces méthodes ont le très grand avantage de la simplicité et de nombreuses variantes existent. Certaines méthodes itératives ne nécessitent pas la création ni le stockage de la matrice des coefficients du système (Schaeffer et Kennedy 1986). Le calcul se fait alors en lisant simplement un fichier incluant données et généalogie. Par contre, dans le cas du modèle animal, il apparaît que la convergence des solutions, c'est-à-dire l'obtention de solutions du système [11] ou [47] satisfaisantes, nécessite un très grand nombre d'itérations (plusieurs dizaines à plusieurs centaines). Lire un tel nombre de fois un fichier incluant plusieurs millions d'enregistrements n'est envisageable qu'en disposant d'un ordinateur ayant une mémoire interne particulièrement vaste (Wiggans *et al* 1988). Une approche différente, également itérative, mais faisant appel à certaines caractéristiques des techniques d'inversion directe (Poivey 1986) et d'absorption a été retenue pour l'évaluation française (Ducrocq *et al* 1990). Sa vitesse de convergence est un peu meilleure que les méthodes itératives classiques mais surtout elle est réalisable sur des ordinateurs de capacité nettement plus restreinte.

ce qui peut s'écrire :

$$\widehat{b_3} - \alpha\widehat{a_5} - \alpha\widehat{a_6} + (1 + 2\alpha)\widehat{a_8} = y_8 \qquad [30]$$

On retrouve la dernière équation du système [19]. On peut aussi écrire [29] sous la forme :

$$\widehat{a_8} = \frac{\widehat{a_5} + \widehat{a_6}}{2} + \frac{1}{1+2\alpha}\left[(y_8 - \widehat{b_3}) - \frac{\widehat{a_5} + \widehat{a_6}}{2} \right] \qquad [31]$$

La valeur génétique $\widehat{\alpha}_8$ comprend alors deux parties. La première est liée à l'espérance de a_8 au moment de l'accouplement : c'est la moyenne des valeurs génétique des parents. La deuxième est une estimation de la contribution de l'aléa de méïose : l'expression entre crochets mesure en effet la différence entre la performance corrigée et la valeur génétique moyenne des parents. Le terme 1/ (1+ 2α) est l'héritabilité intra-famille, mesurant la variabilité génétique parmi les descendants d'un couple père-mère donné. En l'absence de consanguinité :

$$h^2_{intra} = \frac{0{,}5\,h^2}{0{,}5\,h^2 + (1 - h^2)} = \frac{1}{1 + 2\alpha}$$

De même, le mâle 1 est évalué à partir de 2 descendants, un mâle (5) et une femelle (6).Leurs contributions à l'estimation de $\widehat{a_1}$ sont égales respectivement à $2\left(\widehat{a_5} - \frac{\widehat{a_2}}{2}\right)$ et $2\left(\widehat{a_6} - \frac{\widehat{a_2}}{2}\right)$

En pondérant chacune de ces contributions par $\frac{\alpha}{2}$ et en ajoutant celle de l'ascendance (= 0 avec un poids de α car les parents de 1 sont inconnus), on obtient :

$$\widehat{a_1} = \frac{1}{2\alpha}\left\{ \alpha \cdot 0 + \frac{\alpha}{2}\left\{ 2\left(\widehat{a_5} - \frac{\widehat{a_2}}{2}\right)\right\} + \frac{\alpha}{2}\left\{ 2\left(\widehat{a_6} - \frac{\widehat{a_2}}{2}\right)\right\}\right\} \qquad [32]$$

On retrouve donc la quatrième équation de [19] :

$$2\alpha\hat{a}_1 + \alpha\hat{a}_2 - \alpha\hat{a}_5 - \alpha\hat{a}_6 = 0 \qquad [33]$$

Enfin, l'animal 6 regroupe les 3 sources possibles d'information : une performance propre corrigée ($= y_6 - \hat{b}_2$ avec un poids de 1), une contribution des parents $\hat{a}_1 + \hat{a}_2$ avec un poids de 2 α) et une contribution d'une fille ($(8)_{(= 2(\hat{a}_8 - \frac{1}{2}\hat{a}_5)}$ avec un poids de $\frac{\alpha}{2}$). En combinant ces 3 sources, on obtient :

$$\hat{a}_6 = \frac{1}{1+2,5\alpha}\left\{1(y_6 - \hat{b}_2) + 2\alpha\left(\frac{\hat{a}_1+\hat{a}_2}{2}\right) + \frac{\alpha}{2}\,2(\hat{a}_8 - \frac{1}{2}\hat{a}_5)\right\} \qquad [34]$$

Encore une fois, après manipulation, on retrouve l'équation correspondante du système [19] :

$$\hat{b}_2 - \alpha\hat{a}_1 - \alpha\hat{a}_2 + 0,5\,\alpha\hat{a}_5 + (1+2,5\alpha)\hat{a}_6 - \alpha\hat{a}_8 = y_6 \qquad [35]$$

Ces exemples illustrent clairement la « logique interne » des équations du modèle animal.

La résolution du système [19] en faisant l'hypothèse d'une héritabilité égale à $h^2 = 0,25$ ($\alpha = 3$) conduit aux solutions suivantes :

$$\begin{pmatrix}\hat{b}_1\\ \hat{b}_2\\ \hat{b}_3\end{pmatrix} = \begin{pmatrix}4750\\ 5250\\ 5930\end{pmatrix} \quad \text{et} \quad \begin{pmatrix}\hat{a}_1\\ \hat{a}_2\\ \hat{a}_3\\ \hat{a}_4\\ \hat{a}_5\\ \hat{a}_6\\ \hat{a}_7\\ \hat{a}_8\end{pmatrix} = \begin{pmatrix}28\\ 83\\ -28\\ -83\\ 56\\ 83\\ -83\\ 70\end{pmatrix} \qquad [36]$$

On peut noter qu'il a été possible d'estimer l'effet « année 3 » (b_3) alors qu'une seule observation était disponible. Cette estimée $\hat{b}_3$ est différente de y_8 car elle tient compte de la valeur génétique supérieure de l'animal 8 ($\hat{a}_8 = 70$). La prise en compte des conséquences de la sélection dans l'évaluation est illustrée par le calcul de la moyenne des valeurs génétiques estimées par classe d'âge, qui, de fait, mesure le progrès génétique réalisé. Pour la première génération, cette moyenne est nulle : $0,25\,(\hat{a}_1 + \hat{a}_2 + \hat{a}_3 + \hat{a}_4) = 0$. Elle est par contre égale à 18 en deuxième génération et à 70 en troisième génération.

5 / L'évaluation idéale et ses approximations

Des trois paragraphes précédents, il ressort que la méthodologie BLUP présente des propriétés particulièrement intéressantes : les effets du milieu sont estimés en même temps que les effets aléatoires (génétiques), de façon à ne pas biaiser l'évaluation de ces derniers. Lorsqu'un modèle animal est retenu, l'ensemble des performances et des généalogies est utilisée simultanément sans qu'il soit nécessaire d'identifier toutes les sources d'information en fonction de leur degré d'apparentement avec l'animal à évaluer, et en corrigeant « automatiquement » pour tout accouplement non aléatoire. Les « coefficients de l'index » en [8] n'apparaîssent nulle part explicitement. Enfin, l'évolution de la variance génétique au cours des générations, suite à la sélection, à la dérive génétique et à la consanguinité est directement prise en compte à travers l'inverse de la matrice de variance-covariance **G** de valeurs génétiques additives.

Néanmoins, pour que ces propriétés soient effectivement vérifiées, certaines conditions assez restrictives doivent être respectées. En particulier, nous avons déjà souligné la nécessité impérative d'une description correcte des effets génétiques et environnementaux lors du choix d'un modèle. Il est également essentiel pour un calcul exact de $\mathbf{G}^{-1}$ de connaître (à un facteur de proportionnalité près) la variance génétique additive σ_a^2 dans la *population de base* supposée non consanguine et non sélectionnée ainsi que *l'ensemble* des relations de parentés entre les animaux depuis la population de base jusqu'aux animaux les plus jeunes inclus dans l'évaluation. Enfin, en toute rigueur, pour un suivi correct de l'évolution de la variance génétique, *toutes les données* ayant servi aux opérations de sélection depuis cette population de base et sur *l'ensemble des caractères liés* au caractère auquel on s'intéresse et sur lesquels a pu porter la sélection doivent être incluses dans l'analyse (la prise en compte de plusieurs caractères dans une évaluation BLUP ne modifie pas fondamentalement les équations du modèle mixte présentées en [11] ; seuls les calculs de $\mathbf{G}^{-1}$ et $\mathbf{R}^{-1}$ sont plus complexes car ils font intervenir la connaissance des corrélations génétiques et résiduelles entre caractères). Seule une telle évaluation BLUP multicaractères complète basée sur un modèle animal adéquat possède la totalité des propriétés précitées. Toutes les autres approches, y compris la théorie de index de sélection, ne sont que des approximations de cette évaluation idéale. Par exemple, si les parentés entre femelles d'élevage différents sont ignorés, on ajoute (implicitement) l'hypothèse que les différences génétiques entre élevage sont dues aux mâles, et aux mâles seuls.

En fait, ces approximations sont le plus souvent nécessaires et inévitables car la taille du système à résoudre est, pour le cas idéal, égale à $C\,(n_f + n_a)$ où C est le nombre de caractères considérés simultanément, n_f est le nombre d'effets fixés et n_a le nombre d'effets aléatoires (c'est-à-dire ici le nombre total d'animaux à évaluer). De fait, une approximation très courante consiste à négliger les effets corrélés de la sélection ayant porté sur des caractères autres que celui auquel on s'intéresse, ce qui conduit à se ramener à une évaluation indépendante pour chaque caractère. Même dans ce cas, pour l'évaluation nationale en race Française Frisonne et en utilisant l'information collectée sur une période de 10 à 20 ans, le système à résoudre comprend plusieurs millions d'équations. Cette résolution est particulièrement délicate et n'a pu être menée a bien que très récemment (cf encadré). Ceci explique que, jusqu'à il y a peu, d'autres hypothèses simplificatrices concernant la structure de parenté et la nature des données prises en compte étaient couram-

ment retenues pour diminuer le nombre d'équations à résoudre simultanément. On peut, par exemple, ignorer les performances propres des individus dans leur évaluation, celle-ci n'étant alors basée que sur les performances mesurées sur les descendants. Le modèle descriptif qui sera le point de départ de l'évaluation ne correspondra plus à l'équation [3] mais à :

$$y_i = \{\mu + m_i\} + \frac{1}{2}a_p + \frac{1}{2}a_m + e_i^* \qquad [37]$$

où a_p et a_m sont les valeurs génétiques du père p et de la mère m de l'animal i. On parle alors de « modèle père-mère » par opposition au « modèle animal » car cette fois, la valeur génétique de i n'apparaît pas directement dans l'équation [37]. La résiduelle e^*_i du modèle est différente de e_i en [3] car elle inclut la contribution de l'aléa de méïose ϕ_i . La méthodologie BLUP s'applique de la même façon au modèle père-mère qu'au modèle animal. Pour l'ensemble des performances, on pourra en effet écrire [37] sous la forme :

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{Z}\,\mathbf{u} + \mathbf{e}^* \qquad [38]$$

avec $\mu = (\frac{1}{2}a_i)$. Les équations du modèle mixte correspondant à [38] sont obtenues comme en [11] mais leur nombre est plus réduit car seuls les animaux ayant des descendants sont inclus dans le vecteur **u**.

En allant plus loin, on peut également ignorer les relations de parenté entre femelles. Les mères considérées dans le modèle père-mère sont supposées apparentées uniquement à travers les grand-pères maternels et le modèle [37] peut alors être encore simplifié :

$$y_i = \{\mu + m_i\} + \frac{1}{2}a_p + \frac{1}{4}a_{gpm} + e_i^{**} \qquad [39]$$

où a_{gpm} est la valeur génétique du grand-père maternel de i. On obtient alors un « modèle père-grand-père maternel », qui, traité avec la méthodologie BLUP, aboutit à un système d'équations du modèle mixte dont la taille ne dépend que du nombre d'effets fixés et du nombre de mâles dans la population. Le modèle « père-grand-père maternel » a été utilisé à l'université Cornell pour l'évaluation laitière des taureaux du nord-est des Etats-Unis de 1979 à 1989 (avec une approche « multicaractères » à partir de 1982). Les valeurs génétiques des femelles sont ensuite estimées en combinant à travers un index de sélection simple leurs performances corrigées pour les effets fixés considérés dans le modèle [39] et la valeur génétique estimée de leur père (et éventuellement de leur grand-père maternel).

Enfin, si l'on fait l'hypothèse extrême que les mères des individus contrôlés ne sont pas apparentées et n'ont jamais plus d'un descendant chacune dans l'évaluation, on peut définir un « modèle père » qui sera décrit par l'équation :

$$y_i = \{\mu + m_i\} + \frac{1}{2}a_p + e_i^+ \qquad [40]$$

La résiduelle e_i^+ comprend à la fois la résiduelle du modèle [3], la moitié de la valeur génétique additive de la mère de i et la contribution de l'aléa de méïose. La taille du système des équations BLUP correspondant au modèle père est égale à la somme du nombre d'effets fixés et du nombre de pères à évaluer. Le modèle père a été utilisé à l'Université Cornell de 1970 (initialement sous forme « simplifiée », en négligeant les parentés entre taureaux) à 1979.

Pour illustrer les avantages et les inconvénients de ces modèles « approchés » du modèle animal, considérons l'utilisation d'un modèle père pour l'évaluation des mâles de l'exemple de la figure 1. Le modèle complet, sous forme matricielle, s'écrit :

$$\begin{bmatrix} y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} u_1 \\ u_3 \\ u_5 \end{bmatrix} + \begin{bmatrix} e_6^+ \\ e_7^+ \\ e_8^+ \end{bmatrix} \qquad [41]$$

où $u_i = \frac{1}{2}a_i$. Compte-tenu des hypothèses, on a d'autre part :

$$\mathbf{G} = \mathrm{Var}(u_i) = \mathrm{Var}(\tfrac{1}{2}a_i) = \tfrac{1}{4}\sigma_a^2\,\mathbf{A}$$

$$\text{et } \sigma_e^{2+} = \sigma_e^2 + \tfrac{3}{4}\sigma_a^2$$

En appliquant les règles d'Henderson pour la construction de A^{-1}, et en notant que le seul apparentement connu se limite ici à la relation père-fils entre 1 et 5, on obtient :

$$\mathbf{A^{-1}} = \begin{bmatrix} 1+\frac{1}{3} & 0 & -\frac{2}{3} \\ 0 & 1 & 0 \\ -\frac{2}{3} & 0 & \frac{4}{3} \end{bmatrix}$$

En définissant $\alpha = \sigma_e^{2+}/\sigma_u^2 = (4\sigma_e^2 + 3\sigma_a^2)/\sigma_a^2$, les équations du modèle mixte [11] appliquées au modèle père [41] s'écrivent :

$$\begin{bmatrix} 2 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1+\frac{4}{3}\alpha & 0 & -\frac{2}{3}\alpha \\ 1 & 0 & 0 & 1+\alpha & 0 \\ 0 & 1 & -\frac{2}{3}\alpha & 0 & 1+\frac{4}{3}\alpha \end{bmatrix} \begin{bmatrix} b_2 \\ b_3 \\ u_1 \\ u_3 \\ u_5 \end{bmatrix} = \begin{bmatrix} y_6 + y_7 \\ y_8 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} \qquad [42]$$

Là encore, il serait possible de construire chacune de ces équations par un raisonnement indirect, illustrant la manière naturelle utilisée pour combiner les différentes sources d'information disponibles pour l'évaluation de chaque taureau ou chaque effet fixé.

Avec une héritabilité $h^2 = 0{,}25$ ($\alpha = 15$), la résolution du système [42] conduit aux solutions suivantes :

$$\begin{bmatrix} \widehat{b_2} \\ \widehat{b_3} \end{bmatrix} = \begin{bmatrix} 5250 \\ 5992 \end{bmatrix} \quad \begin{bmatrix} \widehat{u_1} \\ \widehat{u_3} \\ \widehat{u_5} \end{bmatrix} = \begin{bmatrix} 16 \\ -16 \\ 8 \end{bmatrix} \quad \Longrightarrow \quad \begin{bmatrix} \widehat{a_1} \\ \widehat{a_3} \\ \widehat{a_5} \end{bmatrix} = \begin{bmatrix} 32 \\ -32 \\ 16 \end{bmatrix}$$

Sur cet exemple très simple, la différence avec les solutions du modèle animal est frap-

pante. L'effet « année 3 » est surestimé dans le modèle père car la supériorité génétique de l'animal 8 est imparfaitement reflétée par la moitié de la valeur génétique de son père. Le mâle 1 est surévalué dans le modèle père parce qu'une partie de la supériorité de sa fille 6, due à la femelle 2, lui est incorrectement attribuée. Le mâle 5 est sous-estimé parce qu'en faisant l'hypothèse que la vache 6 est d'un niveau moyen (= 0), on attribue la plus grande partie de la bonne performance de la fille 8 à des conditions de milieu favorables (excellent effet année). Ces exemples illustrent parfaitement l'intérêt du modèle animal par rapport à un modèle plus simple, qui accumule certaines hypothèses souvent oubliées.

A titre de comparaison, et sans vouloir entrer ici dans les détails, il n'est pas inutile de rappeler les caractéristiques générales du calcul d'index de la méthode française IF2 utilisée de 1978 à 1989 (Poutous *et al* 1981). Cette méthode est essentiellement une version améliorée de la comparaison aux contemporaines mais certaines caractéristiques la rapprochent d'une évaluation de type « modèle animal ». La formulation générale des index mâles dans la méthode IF2 s'écrit très schématiquement :

$$\hat{a}_i = \omega_1 R_f + \omega_2 [2\ \overline{y_f - \sum_k \hat{b}_k}] \qquad [43]$$

où $\overline{y_f - \sum_k \hat{b}_k}$ est une moyenne (pondérée) des productions des filles du taureau i corrigées pour tous les effets fixés. Ceux-ci ont été estimés précédemment et individuellement par calcul de moyenne de performances, corrigées pour tous les autres effets, y compris génétiques. R_f est une « référence », qui joue (surtout) le rôle d'index sur ascendance, faisant intervenir l'index du père et du grand-père maternel du taureau. La valeur des index des fils du taureau i est aussi introduite dans cette référence. ω_1 et $\omega_2 = 1-\omega_1$ sont les poids accordés à chacune des deux informations. ω_2 est égal au coefficient de détermination de l'index sur descendance, fonction de l'héritabilité du caractère et du nombre de filles dont les lactations sont connues. Pour les femelles, l'index s'écrit :

$$\hat{a}_i = \omega_i I_A + \omega_2 [\overline{y_i - \sum_k \hat{b}_k}] \qquad [44]$$

où cette fois, I_A est réellement un index sur ascendance et $\overline{y_i - \sum_k \hat{b}_k}$ est la moyenne des productions de la vache i, corrigées pour les effets fixés. Comme précédemment, ω_1 et $\omega_2 = 1-\omega_1$ sont des facteurs de pondération. ω_2 dépend du nombre de lactations de l'animal, de l'héritabilité et de la répétabilité du caractère (voir ci-après).

Il apparaît donc un grand nombre de similitudes entre les index de IF2 et les équations du modèle animal. Les équations [43] et [44] sont d'une forme voisine de l'équation [28], combinant plusieurs sources d'information. Effets fixés et aléatoires sont estimés simultanément. Par contre, le modèle animal est beaucoup plus clair (la définition de la « référence » est explicite et identique pour les mâles et les femelles), plus complet (la descendance des femelles est intégrée dans le calcul des index vaches), plus précis (les pondérations accordées à chaque type d'information apparaissent naturellement, sans l'intermédiaire de coefficients de détermination, qui ne sont que des approximations de la précision des éléments de l'index), plus rigoureux (le modèle et ses propriétés sont connus parfaitement), ce qui en rend l'explication et l'interprétation des résultats plus aisées.

6 / Modèles plus complexes

6.1 / Modèle animal avec effet de groupes génétiques

Les données de terrain sont nécessairement incomplètes : il n'est pas possible de connaître l'ensemble des généalogies jusqu'à la « population de base » non sélectionnée, ni l'ensemble des performances ayant servi aux opérations de sélection. Pour certains animaux, on pourra remonter 5 ou 10 générations en arrière. Pour d'autres, par exemple pour ceux d'élevages entrant au contrôle laitier, aucune information ne sera disponible sur leurs parents. Or les modèles présentés jusqu'ici supposent que les animaux dont les ascendants sont inconnus font partie d'une même population non sélectionnée, « d'espérance nulle » ($E(a_i) = 0$). Cette hypothèse n'est plus justifiable dans le contexte actuel : une vache Holstein de 1989, même issue de parents inconnus, n'a certainement pas la même espérance de valeur génétique qu'une vache frisonne du début des années 70. Pour surmonter cette contradiction, il a été proposé de définir des *groupes d'animaux issus de parents inconnus* qui prennent en compte leurs différents degrés de sélection depuis la population de base, en fonction de leur type génétique (par exemple, Holstein ou Frisonne), de leur date de naissance (par exemple par classe de 2 ou 3 ans), leur sexe (par exemple, père inconnu de taureaux, ou père inconnu de vaches). De cette façon et par une légère modification du système des équations du modèle mixte [11] (Westell 1984, Quaas 1988), il est possible d'estimer a posteriori l'espérance de la valeur génétique de ces animaux dont les parents sont inconnus.

6.2 / Modèle animal avec répétabilité

Une autre modification fréquemment considérée est l'extension du modèle [3] permettant la prise en compte de plusieurs lactations par vache. Ce modèle est en effet trop simple car les résiduelles e_i de [3] sont supposées indépendantes. Or un grand nombre de facteurs du milieu (conditions de naissance, de croissance et d'élevage propres à l'animal) et génétiques (effets génétiques non additifs) contribuent à rendre 2 lactations d'une même vache plus semblables que deux lactations de deux vaches différentes, même contemporaines. Pour se rapprocher de cette réalité, on ajoute au modèle [3] un « effet vache » ou « effet d'environnement permanent » p_i regroupant tous les facteurs autres que la valeur génétique qui affectent la production laitière d'une vache

pendant toute sa carrière. Pour la jème performance de la vache i, on écrira :

$$y_{ij} = \{\mu + m_i\} + a_i + p_i + e_{ij} \qquad [45]$$

Comme pour les autres effets aléatoires, l'effet d'environnement permanent p_i est supposé être la résultante de nombreux effets élémentaires, chacun d'importance réduite. Pi est donc un effet aléatoire suivant une distribution normale de moyenne nulle et de variance σ_p^2. Les résiduelles e_{ij} sont distribuées normalement, avec une moyenne nulle et une variance σ_e^2. Le rapport $r = (\sigma_a^2 + \sigma_p^2)/(\sigma_a^2 + \sigma_p^2 + \sigma_e^2)$ est la *répétabilité* du caractère et représente la fraction de la variabilité totale correspondant à la variabilité des performances d'un même animal.

La méthodologie BLUP s'applique sans modification majeure pour ce modèle animal avec répétabilité. Sous forme matricielle :

$$\mathbf{y} = \mathbf{X}\,\mathbf{b} + \mathbf{Z}\,\mathbf{a} + \mathbf{Z}\,\mathbf{p} + \mathbf{e} \qquad [46]$$

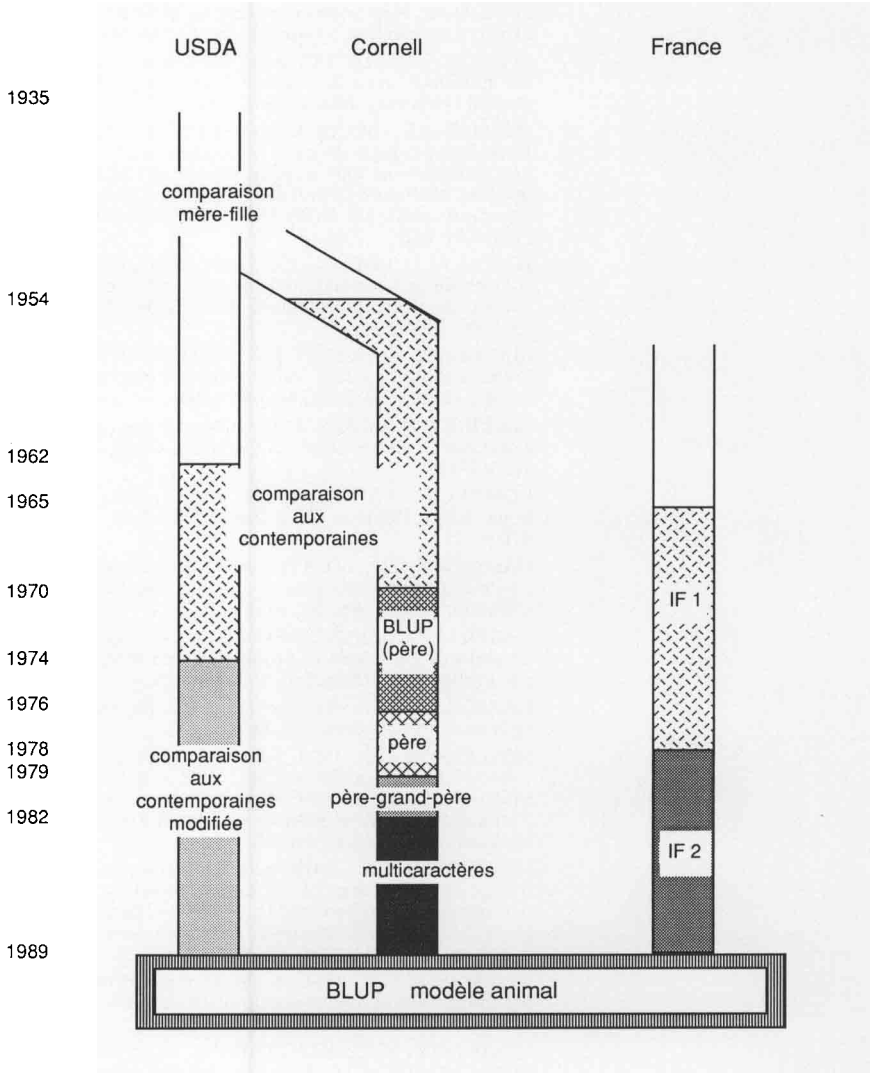
Les équations du modèle mixte pour l'obtention des solutions de **b**, **a** et **p** s'écrivent par extension de [11] :

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + \sigma_a^2 A^{-1} & Z'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z & Z'R^{-1}Z + \sigma_p^2 I \end{bmatrix} \begin{bmatrix} b \\ a \\ p \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \\ Z'R^{-1}y \end{bmatrix} \qquad [47]$$

$\widehat{a_i} + \widehat{p_i}$ est une mesure de la *capacité productive* de la vache i. Cette somme caractérise mieux la valeur réelle de i que $\widehat{a_i}$. En effet, $\widehat{a_i} + \widehat{p_i}$ est une prévision de la production future de i, à mois de vêlage, numéro de lactation, etc, fixés. C'est donc un critère important à considérer lors du choix des vaches à réformer. Par contre, la sélection des descendants de i ne doit faire intervenir que la partie transmissible du patrimoine de i, c'est-à-dire $\widehat{a_i}$.

Des modèles incorporant d'autres effets aléatoires tels qu'une interaction taureau x troupeau reflétant un traitement commun aux filles d'un même taureau dans chaque troupeau ont été également suggérés (Wiggans *et al* 1988).

Figure 2. *Schéma résumant la succession chronologique des méthodes d'évaluation génétique des bovins laitiers utilisées aux Etats-Unis (Ministère de l'agriculture (USDA) et Université Cornell) et en France.*
(La méthode française IF1 repose sur le principe de comparaison aux contemporaines. La méthode IF2 est une comparaison aux contemporaines modifiée, dont certaines caractéristiques sont proches des évaluations de type « modèle animal »).



Conclusion

Les recherches sur de nouvelles méthodes d'indexation laitière ont toujours répondu au même objectif : fournir aux praticiens des estimations de valeur génétique de leurs animaux de plus en plus fiables et précises. Dans un contexte international des animaux reproducteurs de plus en plus concurrentiel, il est impératif de rechercher sans cesse l'obtention d'un classement aussi correct que possible de taureaux dont certains auront plusieurs dizaines de milliers de filles et de vaches dont certaines fourniront les élites mâles et femelles de demain. La méthodologie BLUP appliquée au « modèle animal » permet de se rapprocher d'une évaluation idéale, en s'affranchissant des nombreuses hypothèses implicites ou explicites des méthodes précédentes. Mais un regard sur le passé (figure 2) appelle à la modestie. L'évolution des méthodes s'est faite de manière de plus en plus rapide, à tel point que l'adoption dans l'ensemble des pays développés d'une évaluation basée sur le modèle animal n'aura pris que quelques mois. Le modèle animal tel qu'il a été présenté ici n'est probablement qu'une étape vers d'autres évaluations plus sophistiquées, qui devront utiliser encore mieux l'information de base (contrôles mensuels), intégrer les apports de la génétique moléculaire et des biotechnologies (sélection assistée par marqueurs, transferts de gènes), s'adapter au développement de nouvelles techniques de reproduction (« splitting », clonâge), et suivre l'évolution des schémas de sélection et des techniques d'élevage (Colleau 1989 , Colleau et Mocquot 1989).

Références bibliographiques

COLLEAU J.J., 1989. Impact of the use of bovine somatotropin in dairy cattle. Genet. Sel. Evol. (sous presse).

COLLEAU J.J., MOCQUOT J.C., 1989. Using embryo transfer in cattle breeding. Fifth annual meeting of the European embryo transfer association. 8-9 Sept. Lyon.

DUCROCQ V., BOICHARD D., BONAITI B., BARBAT A., BRIEND M., 1990. A pseudo-absorption strategy for solving animal model equations for large data files. J. Dairy Sci., 73 (sous presse).

EVERETT R.W., 1974. Problems in evaluating dairy sires. in Ed. Garsi. First world congress on genetics applied to livestock production. Volume II. pp 99-103. Madrid.

EVERETT R.W., QUAAS R.L., 1979. Sire evaluation in the northeast. Animal Science Mimeograph Series, 44. Cornell University, Ithaca, New-York.

FOULLEY J.L., BOUIX J., GOFFINET B., ELSEN J.M., 1984. Comparaison de pères et connexions. in Insémination artificielle et amélioration génétique : bilan et perspectives critiques. Toulouse-Auzeville, France, 23-24 novembre 1983. Ed. INRA Publ. 1984 (Les colloques de l'INRA, n° 29).

FOULLEY J.L., CHEVALET C., 1981. Méthode de prise en compte de la consanguinité dans un modèle simple de simulation des performances. Ann. Génét. Sél. Anim., 13, 189 - 196.

GIANOLA D., FOULLEY J.L., FERNANDO R.L., 1986. Prediction of breeding values when variances are not known. Génét. Sél. Evol., 18, 485 - 498.

GOFFINET B., ELSEN J.M., 1984. Critère optimal de sélection. Quelques résultats généraux. Génét. Sél. Evol., 16, 307-317.

GOLUB G.H., VAN LOAN C.F., 1983. Matrix computations. Johns Hopkins Univ. Press, Baltimore, Maryland. 476 p.

HARGROVE G.L., LEGATES J.E., 1971. Biases in dairy sire evaluation attribuable to genetic trend and female selection. J. Dairy Sci., 54, 1041 - 1051.

HARVILLE D.A., HENDERSON C.R., 1967. Environmental and genetic trends in production and their effects on sire evaluation. J. Dairy Sci., 50, 870 - 875.

HAZEL L.N., 1943. The genetic basis for constructing selection indexes. Genetics, 28, 476-490.

HENDERSON C.R., 1963. Selection index and expected genetic advance. in Hanson W.D. et H.F. Robinson (Eds.) Statistical genetics and plant breeding. pp.141-163. National Academy of Science - National Research Council, Publ. 982, Washington, DC.

HENDERSON C.R., 1973. Sire evaluation and genetic trends. in Proceedings of the Animal Breeding and Genetics symposium in honor of Dr J.L. Lush, Blacksburg, Virginia. August 1972, pp.10-41. American Society of Animal Science, Champaign, Illinois.

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

HENDERSON C.R., KEMPTHORNE O., SEARLE S.R., VON KROSIGK C.M.1959. The estimation of environmental and genetic trends from records subject to culling. Biometrics, 15, 192 -218.

KENNEDY B.W., SCHAEFFER L.R., SORENSEN D.A., 1988. Genetic properties of animal models. J. Dairy Sci., 71 (Suppl. 2.) 17 - 26.

MALECOT G., 1948. Les mathématiques de l'hérédité. Masson, Paris. 60 p.

POIVEY J.P., 1986. Méthode simplifiée de calcul des valeurs génétiques des femelles tenant compte de toutes les parentés. Génét. Sél. Evol., 18, 321-332.

POUTOUS M., BRIEND M., CALOMITI S., DOAN D., FELGINES C., STEIER G., 1981. Méthode de calcul des index laitiers. Bases générales. Bul. Tech. Inf., 361, 433-446.

QUAAS R.L., 1984. Used of Mixed model for prediction. in Blup School handbook. R.L. Quaas, R. D. Anderson et A.R. Gilmour, pp 1-77. Animal Genetics and Breeding Unit. University of New England. Australia.

QUAAS R.L., 1988. Additive genetic model with groups and relationships. J. Dairy Sci., 71 (Suppl. 2.) 91 - 98.

ROUVIER R., 1969. Pondération des valeurs génotypiques dans la sélection par index sur plusieurs caractères. Biometrics., 25, 295-307.

SCHAEFFER L.R., KENNEDY B.W., 1986. Computing solution to mixed model equations. in Dickerson G.E. et Johnson R.K. (eds.) Third world congress on genetics applied to livestock production. Volume XII. pp 382-393. University of Nebraska, Lincoln, Nebraska.

SMITH H.F. , 1936. A discriminant function for plant selection. Annals of Eugenics,7, 240 -250.

VERRIER E., 1989. Prédiction de l'évolution de la variance génétique dans les populations animales d'effectif limité soumises à sélection. Thèse de docteur-ingénieur INA-PG. 259 p.

WESTELL R.A., 1984. Simultaneous genetic evaluation of sires and cows for a large population of dairy cattle. Thèse de Ph.D., Cornell University, Ithaca, New-York. 70 p.

WIGGANS G.R., MISZTAL I., VAN VLECK L.D., 1988. Implementation of an animal model for genetic evaluation of dairy cattle in the United States. J. Dairy Sci., 71 (Suppl. 2) 54-69.

Summary

Genetic evaluation techniques in dairy cattle.

The statistical methods used during the past 50 years for genetic evaluation of dairy bulls and cows are presented with special emphasis on the reasons which motivated their development and then their withdrawal. The selection index theory applied to data such as deviation of the cows' records from their herdmates became obsolete because environmental effects were imperfectly corrected and the underlying genetic trend was not accounted for. The Best Linear Unbiased Prediction (BLUP) enables simultaneous estimation of genetic and environmental effects. Initially applied to simple models for sire evaluation (sire or sire-grand-sire models), BLUP is increasingly used with an « animal model » nowadays, allowing a joint evaluation of males and females. Using this method, some interesting properties are obtained. Other aspects, such as the modelling of performance records, the inclusion of several lactations and the computation difficulties involved are also tackled.

DUCROCQ V., 1990. Les techniques d'évaluation génétique des bovins laitiers. INRA Prod. Anim., 3(1), 3-16.