*D. BOICHARD, B. BONAÏTI, Anne BARBAT, Michèle BRIEND*

*INRA Station de Génétique Quantitative et Appliquée 78352 Jouy-en-Josas Cedex*

L'évaluation des reproducteurs

# Le modèle sous-jacent à l'évaluation des valeurs génétiques

***Résumé.*** *La valeur génétique additive peut être prédite en modélisant d'une part les performances, d'autre part le déterminisme génétique des caractères. Les performances sont décomposées en effets génétiques, en effets de milieu identifiés et en une résiduelle du modèle, constituée de multiples effets génétiques ou de milieu non identifiés et non maîtrisés. Si le déterminisme du caractère est polygénique additif, la corrélation entre les valeurs génétiques de deux individus est proportionnelle à leur coefficient de parenté. Toute l'information (performances, facteurs de variation, relations de parenté) est combinée en un système d'équations unique qui permet d'estimer simultanément les effets génétiques et les effets de milieu. L'adoption d'un modèle "animal" permet de combiner toute cette information de façon optimale et de prendre en compte l'effet de la sélection et des accouplements non au hasard dans la population. Grâce à la structure de l'inverse de la matrice de parentés, le modèle animal fournit des équations très simples, facilitant l'explication et la diffusion de son principe : un effet de milieu est estimé par une moyenne de performances ajustées ; l'index d'un individu combine trois informations, sur ascendance, sur descendance et sur performances propres. Le modèle est souple et peut facilement être modifié pour prendre en compte des situations complexes. L'évaluation génétique est bien sûr un outil de sélection mais, compte tenu de ses propriétés, représente aussi un outil puissant de diagnostic et de prévision.*

La performance d'un individu (ou valeur phénotypique) est déterminée par des effets génétiques et des effets de milieu. La part d'origine génétique peut elle-même être décomposée en une part dûe aux effets additifs individuels de chaque gène, dite valeur génétique additive, et une composante liée aux interactions entre gènes au même locus (dominance) et entre locus (épistasie). Seule la valeur génétique additive se transmet d'une génération à l'autre, tandis que les interactions sont recréées aléatoirement à chaque génération. Dans un programme de sélection intra population, on cherche donc à augmenter la valeur génétique additive, en retenant comme reproducteurs à chaque génération les individus à valeur génétique additive la plus élevée.

Toutefois, la valeur génétique additive n'est pas une donnée observable. L'évaluation génétique, ou "indexation", a pour objectif d'estimer au mieux la valeur génétique des animaux reproducteurs potentiels. Elle est donc un outil primordial d'aide à la sélection puisqu'elle fournit en pratique le critère optimal pour réaliser le choix des reproducteurs. Elle permet aussi de mesurer *a posteriori* l'efficacité des programmes de sélection. Cet article vise à présenter les principes de l'évaluation des reproducteurs dans le cadre particulier des caractères laitiers.

## 1 / Le modèle de description des données

L'évaluation génétique laitière découle directement du modèle génétique et statistique de description des données, présenté par Ducrocq (1992), et en constitue une application importante. Elle repose sur deux types d'informations : d'une part les performances laitières, d'autre part les généalogies. Ces informations sont disponibles grâce à une gestion organisée de l'information et à un système d'identification des animaux fiable et permanent : depuis la loi de l'élevage de 1966, tous les bovins ont un numéro permanent et unique à 10 caractères.

Les performances (quantité de lait produite le jour du contrôle, taux butyreux et taux protéique) sont mesurées mensuellement sur les animaux inscrits au Contrôle laitier, qui constituent la base de sélection. Actuellement, 2,5 millions de vaches sont au contrôle laitier, soit près de la moitié du cheptel français.

Ces données brutes ne sont pas analysées directement. La première étape vise à définir les caractères à analyser. Les productions par lactation sont calculées à partir des contrôles mensuels pour 5 caractères : les quantités de lait, de matière grasse, de matière protéique, le taux butyreux et le taux protéique. La production par lactation n'est connue exactement que lorsque la lactation est terminée. Toutefois, on peut l'extrapoler avec une précision rai-

sonnable pour les lactations en cours. Les productions extrapolées sont donc aussi analysées, permettant une évaluation des animaux plus précoce.

L'évaluation consiste ensuite à décomposer la performance observée en un effet génétique et des effets de milieu. On considère en général trois types d'effets de milieu :

- ceux dont on connait à la fois la cause et l'amplitude. Les performances sont alors corrigées *a priori* pour ces effets. Par exemple, les performances dépendant de la durée de la lactation, sont standardisées pour une durée de 305 jours. Les animaux adultes produisant plus que les primipares, les productions sont exprimées en équivalent-adulte.

- ceux dont on a identifié la cause mais dont on ne connait pas l'amplitude. L'effet doit alors être estimé dans le modèle d'analyse. Par exemple, on sait que le troupeau est un facteur influençant les performances, mais on ne connait pas *a priori* l'effet d'un troupeau particulier.

- ceux qu'on ne maîrise pas du tout. Ils constituent l'erreur du modèle, qu'on cherche à minimiser.

Il est important de remarquer que les effets que l'on estime sont en général supposés additifs : ils affectent de façon identique toutes les performances, quels que soient leurs niveaux. Par contre, les effets connus *a priori* peuvent être corrigés de façon additive ou multiplicative, et ce choix n'est pas neutre. Par une correction multiplicative, on change à la fois le niveau et la variabilité des performances. En pratique, les corrections *a priori* sont généralement multiplicatives, de façon à homogénéiser la variabilité des performances, tandis que le modèle d'analyse vise ensuite à en corriger le niveau de façon additive. Ainsi, les lactations de primipares sont transformées en équivalent adulte par un coefficient de 1,3. Leur niveau comme leur variabilité sont donc augmentés de 30 %.

Après cette standardisation préliminaire, chaque performance est ensuite décomposée dans le modèle d'analyse en au minimum trois différentes composantes additives :

$$y_{ikl} = m_k + a_i + e_{ikl} \quad (1)$$

* $y_{ikl}$ est la lème performance de l'animal i, réalisée dans les conditions de milieu k,

* $m_k$ est la somme des effets de milieu identifiés auxquels est soumise la performance. Elle inclut au minimum la moyenne de la population $\mu$. Elle inclut généralement d'autres effets enregistrés et connus pour influencer la production. Dans le cas des bovins laitiers, le principal effet pris en compte est celui du troupeau, qui résume l'ensemble des conditions (climat, région, niveau technique, alimentation, conditions sanitaires, horaires de traites...) auxquelles sont soumises toutes les vaches d'un même troupeau. Un autre effet important est celui de l'année, traduisant les variations climatiques et économiques. Du fait de la forte interaction entre ces deux facteurs, on les combine en une entité synthétique, le troupeau-année, qui constitue la cellule de base pour la comparaison des performances. Le numéro de lactation est aussi inclus dans le modèle. Il est intéressant de noter qu'il est donc pris en compte deux fois : d'abord lors des ajustements préliminaires multiplicatifs *a priori*, en vue d'homogénéiser la variance génétique entre performances de différentes lactations ; ensuite dans le modèle, de façon additive, en vue d'ajuster pour les différences de niveau. La production varie aussi en fonction du mois de mise bas : en France, les lactations initiées en automne sont généralement plus productives que celles commencées en été. La production augmente avec l'âge à la mise bas (en première lactation) et avec l'intervalle entre vêlages (pour les lactations suivantes) qui sont donc également pris en compte. Ces quatre derniers facteurs pouvant varier selon les conditions, ils sont donc considérés intra année et intra région. Ce modèle complexe contraste avec celui de nombreux autres pays, où les corrections *a priori* sont souvent plus importantes et raffinées et le modèle d'analyse au contraire beaucoup plus simple.

A ce stade de la définition du modèle, dépendant d'une analyse zootechnique préliminaire, il est crucial de prendre en compte tous les facteurs affectant les performances et, si possible, seulement eux. Si un facteur n'est pas pris en compte, cet oubli sera en général à l'origine de biais dans l'évaluation et donc d'un mauvais classement des animaux. Inversement, la prise en compte dans le modèle de facteurs inutiles est à éviter. D'une part, cela diminue le nombre de données auxquelles la performance est comparée et, par conséquent, la précision de l'évaluation. D'autre part, les risques de disconnexion du dispositif augmentent, c'est-à-dire les risques de ne pas pouvoir estimer tous les paramètres du modèle. A titre d'exemple, imaginons un troupeau complètement isolé génétiquement du reste de la population, n'utilisant aucun reproducteur de l'extérieur, et n'en vendant aucun. Il n'est alors pas possible de dissocier le niveau génétique moyen de ce troupeau de son niveau de conduite. Le niveau génétique de ce troupeau, par rapport au reste de la population, n'est pas estimable et le dispositif n'est pas connecté. Ducrocq (1992) présente plus de détails sur cet aspect de définition du modèle.

* $a_i$ est la valeur génétique additive de l'animal i. Elle est toujours une valeur relative, exprimée en déviation à la population , c'est-à-dire aux autres individus. Par exemple, on ne peut pas dire que le potentiel d'une vache est de 8000 kg de lait, mais seulement qu'il est à +500 kg au-dessus de la moyenne de la population. Contrairement aux effets de milieu, on a une certaine connaissance *a priori* de la distribution des effets génétiques. Parce que la valeur génétique est supposée être la somme des petits effets d'un grand nombre de gènes, sa distribution est normale. Comme elle ne représente qu'une déviation à la population, sa distribution a une espérance arbitrairement fixée à 0. Par définition de la valeur génétique additive, un parent transmet en espérance la moitié de sa valeur à son produit. Il en résulte une structure de covariance (c'est-à-dire des relations statistiques) entre les valeurs génétiques additives d'individus apparentés, proportionnelle au coefficient de parenté. Enfin, dans une population non sélectionnée et non consanguine, la distribution des valeurs génétiques additives a pour variance $\sigma_a^2$, représentant seulement une fraction, dite héritabilité $h^2$ de la variance phénotypique des performances $\sigma^2$. En résumé, la distribution du vecteur **a** des effets $a_i$ peut être notée $\mathbf{N(0, A}\ \sigma_a^2\mathbf{)}$, où **A** est la matrice de parenté de la population. Le terme (p,q) de **A** est égal au coefficient de parenté entre les individus p et q. Ce qui signifie que pour deux individus p et q, la corrélation entre leurs valeurs génétiques $a_p$ et $a_q$ est égale au coefficient de parenté entre p et q.

* $e_{ikl}$ est la résiduelle du modèle et englobe tout ce que le modèle ne peut pas expliquer, c'est-à-dire des effets de milieu non systématiques, l'erreur de mesure de la performance, la valeur génétique non additive... On peut la comparer à un bruit de fond, que l'on souhaite le plus faible possible. Comme précédemment, elle est supposée être la somme de nombreux petits effets de sorte que sa distribution est supposée normale, indépendante de celle de **a** et d'espérance nulle. Les erreurs sont de plus supposées indépendantes les unes des autres. Dans le cas le plus simple, elles sont supposées de même variance $\sigma_e^2$. En résumé, la distribution du vecteur e des erreurs est notée $\mathbf{N(0, I\,\sigma_e^2)}$, **I** étant la matrice identité.

Soulignons dans ce qui précède que deux types d'effets apparaissent dans le modèle. Certains, les effets de milieu, sont dit "fixés". Il n'est pas fait d'hypothèse sur leur distribution et leur estimation ne dépend que des données. Ainsi, par exemple, l'effet d'un troupeau est estimé par la moyenne des performances réalisées dans ce troupeau, corrigées pour les différences génétiques et les autres effets de milieu (mois, âge...). Au contraire, la valeur génétique est dite "aléatoire", car on fait des hypothèses sur sa distribution. Son estimation n'est pas seulement fonction des données, mais aussi de la connaissance de la valeur génétique *a priori*, découlant de ces hypothèses. Ainsi, la valeur génétique d'un animal n'est pas estimée simplement par la moyenne $\bar{y}_c$ des performances réalisées par cet animal et corrigées pour les effets de milieu. Elle est obtenue en combinant l'information "performances" et l'information *a priori*. Qu'est-ce que l'information *a priori* ? Nous reviendrons sur ce concept mais considérons un exemple simple pour fixer les idées. Soit un animal i, de père p et de mère m. Supposons que p et m soient évalués par ailleurs, et $\hat{a}_p$ et $\hat{a}_m$ sont leurs index respectifs. Avant même que i ne réalise des performances, on a une première estimation de $a_i$, basée sur l'information parentale : par définition, chaque parent a transmis en espérance la moitié de sa valeur génétique à i et $\hat{a}_i = \hat{a}_a = 1/2\ (\hat{a}_p + \hat{a}_m)$ constitue donc l'information *a priori*. Lorsque i réalise une performance, cette information supplémentaire ne remplace pas l'information *a priori*, elle la complète, et cette information supplémentaire est combinée avec l'information *a priori*, avec les poids appropriés $p_a$ et $p_y$, tels que $p_a + p_y = 1$ :

$\hat{a}_i = p_y\,\bar{y}_c + p_a\,\hat{a}_a$

## 2 / Le modèle génétique

Les systèmes d'évaluation les plus modernes actuellement reposent sur un modèle génétique très précis, dit "modèle animal". Du fait de son importance, cette notion, déjà développée par Ducrocq (1992), est reprise ici sous un angle historique.

Un parent transmettant la moitié de ses gènes, il transmet en espérance la moitié de sa valeur génétique à son produit. Cependant, en raison des aléas de méiose liés au tirage au hasard des chromosomes transmis et aux recombinaisons, une partie de ce qui est transmis ne peut pas être prédit à partir de la valeur des parents et varie de façon aléatoire d'un gamète à l'autre. La valeur génétique d'un individu i peut donc s'écrire en fonction de celle de ses parents p et m :

$a_i = 1/2\ a_p + 1/2\ a_m + \phi_i$ (2)

où $\phi_i$ est l'aléa de méiose, indépendant de la valeur des parents (et donc de la sélection), d'espérance nulle et de variance $1/2\ \sigma_a^2$ dans une population non consanguine.

La formule (2) permet de présenter simplement différents modèles et leurs relations. Historiquement, l'évaluation génétique a été mise en place en même temps que le développement des programmes de sélection fondés sur l'insémination artificielle, en vue d'évaluer les mâles mis en testage sur descendance. L'individu évalué est donc le père, tandis que les performances sont réalisées par ses filles. Le modèle considéré est dit modèle père. L'individu évalué n'est pas celui qui réalise la performance. Le modèle s'écrit de la façon suivante :

$y_{ikl} = m_k + 0{,}5\ a_p + e^*_{ikl}$ (3)

Les composantes de la mère $(1/2\ a_m)$ et de l'aléa de méiose $\phi_i$ dans la valeur de i, qui n'apparaissent pas dans (3), sont donc inclus dans la résiduelle, qui est donc égale à

$e^*_{ikl} = e_{ikl} + 0{,}5\ a_m + \phi_i$

Les hypothèses, explicites ou implicites, d'un tel modèle sont nombreuses. Les résiduelles sont supposées indépendantes entre elles et indépendantes des valeurs génétiques des pères. Pour que cette hypothèse soit respectée, les mâles sont donc supposés accouplés à un échantillon aléatoire et non sélectionné de femelles, qui n'ont qu'une fille chacune. Dans le modèle père, les femelles n'existent pas en tant que telles, et les apparentements entre femelles ne peuvent donc pas être pris en compte. Le modèle père peut considérer les apparentements entre mâles $\mathbf{[Var(a) = A\,\sigma_a^2]}$. Dans ce cas, l'index d'un mâle combine la performance moyenne de ses filles $\bar{y}_c$ et la valeur des autres mâles apparentés. Simplification supplémentaire, les parentés entre mâles peuvent être ignorées $\mathbf{[Var(a) = I\sigma_a^2]}$. Dans ce cas, les mâles sont tous indépendants les uns des autres et reçoivent tous la même valeur *a priori* : 0. Leur index s'écrit donc :

$\hat{a}_i = p_a\ 0 + p_y\ 2\ \bar{y}_c = 2\ p_y\ \bar{y}_c$

Le modèle père a pour avantage d'être relativement facile à résoudre, puisque le nombre d'inconnues du système d'équations est limité au nombre de mâles et d'effets de milieu. Mais il a plusieurs inconvénients majeurs. D'abord, il ne fournit aucune évaluation des autres individus et en particulier des vaches candidates pour procréer la nouvelle génération de mâles. Or il s'agit d'une phase très importante du programme. Ensuite, il est de moins en moins valide au fur et à mesure que le programme de sélection est plus efficace et que les hypothèses sur lequel il repose sont moins respectées. Le niveau des mères n'est pas constant dans le temps ni entre troupeaux. Les accouplements n'ont pas lieu au hasard mais au contraire sont systématiquement raisonnés (homogamie). En pratique, le modèle père n'est valide que lors de la mise en place d'un programme.

Plusieurs autres modèles, plus précis et reposant sur des hypothèses moins restrictives et moins nombreuses, ont été envisagés. Pour chacun, le principe est le même : compléter le modèle par de nouveaux paramètres à estimer, et ainsi réduire l'importance de la résiduelle. Le modèle le plus précis, et utilisé actuellement, est le modèle animal. La valeur génétique introduite dans le modèle d'analyse est celle de l'animal réalisant la performance.

$y_{ikl} = m_k + a_i + e_{ikl}$ (4)

La résiduelle, qui ne contient plus aucune composante génétique additive, a donc une variance minimale. Un tel modèle fournit des estimations de valeurs génétiques qui, sous certaines conditions sur lesquelles nous reviendrons, sont insensibles à l'évolution de la variabilité génétique, à la sélection et à l'homogamie. Toutes les parentés sont prises en compte, c'est-à-dire que l'index d'un individu combine l'information de tous ses apparentés, qui peuvent être très nombreux. En revanche, le nombre d'équations à résoudre, toujours supérieur au nombre d'animaux évalués, est généralement très élevé.

## 3 / Introduction aux notations matricielles ; modèle à effets fixés

Pour une présentation rigoureuse, il est indispensable d'utiliser les notations matricielles. Nous rappelons ici, sans les démontrer, les quelques notions essentielles pour la suite. Considérons le modèle très simple suivant, à un facteur m où $y_{ij}$ est la jème performance réalisée dans le milieu i :

$y_{ij} = m_i + e_{ij}$ (5)

Le facteur m comptant k modalités, ce modèle peut se réécrire :

$y_{ij} = 0\ m_1 + \ldots\ 1\ m_i \ \ldots + 0\ m_k + e_{ij}$ (6)

Une ligne de ce type peut être écrite pour chaque performance, soit n lignes au total. Ce tableau de n lignes se réécrit sous forme matricielle :

$\mathbf{y = X m + e}$ (7)

avec **y** le vecteur colonne des n performances (connues), **e** le vecteur colonne des n erreurs (inconnues, à minimiser), **m** le vecteur colonne des k effets inconnus (à estimer), **X** une matrice n x k, connue, dite matrice d'incidence. Chaque ligne de **X** correspond à une performance. Elle est constituée des coefficients 0 et 1 affectant les effets dans la formule (6). **X** est illustrée dans l'exemple suivant, où 4 performances ($y_1 = 10$, $y_2 = 11$, $y_3 = 13$, $y_4 = 14$) sont expliquées par un facteur à deux niveaux ($m_1$ et $m_2$). $y_1$ et $y_2$ sont soumis à $m_1$ tandis que $y_3$ et $y_4$ sont soumis à $m_2$. On peut alors écrire les 4 équations :

$y_1 = 1\ m_1 + 0\ m_2 + e_1$
$y_2 = 1\ m_1 + 0\ m_2 + e_2$
$y_3 = 0\ m_1 + 1\ m_2 + e_3$
$y_4 = 0\ m_1 + 1\ m_2 + e_4$

ou, selon (7), de façon équivalente sous forme matricielle, avec

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

Le modèle est d'autant plus précis que l'erreur e est plus réduite. L'estimation de **m** est donc obtenue en minimisant **e**, ou plus exactement la somme des carrés des erreurs, soit **e'e**. En annulant la dérivée de **e'e** par rapport à **m**, on obtient la solution suivante, dite solution des moindres carrés :

$\hat{\mathbf{m}} = \mathbf{(X'X)^- X'y}$ (8)

**X'** signifie transposée de **X**. Son élément (i,j) est l'élément (j,i) de **X. (X'X)**$^-$ signifie inverse généralisée de **X'X**. Développons cette expression (8) :

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

**X'X** est une matrice contenant l'effectif de données dans chaque combinaison i x j de facteurs : il y a deux données affectées par $m_1$, deux données affectées par $m_2$ et aucune donnée affectée à la fois par $m_1$ et par $m_2$. De même,

$$\mathbf{X'y} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 10 \\ 11 \\ 13 \\ 14 \end{bmatrix} = \begin{bmatrix} 21 \\ 27 \end{bmatrix}$$

**X'y** est donc la somme des performances par niveau de facteur. **m** est alors estimé par :

$$\hat{\mathbf{m}} = \begin{bmatrix} 0,5 & 0 \\ 0 & 0,5 \end{bmatrix} \begin{bmatrix} 21 \\ 27 \end{bmatrix} = \begin{bmatrix} 21 / 2 \\ 27 / 2 \end{bmatrix} = \begin{bmatrix} 11,5 \\ 13,5 \end{bmatrix}$$

Dans ce cas très simple, l'estimation des moindres carrés de **m** est la moyenne des données par niveau de facteur. La précision de cette estimation est donnée par l'inverse de la variance d'erreur. La variance d'erreur d'une moyenne est la variance résiduelle divisée par l'effectif, soit dans le cas présent $\sigma_e^2/2$. Cette expression est retrouvée et généralisée de façon matricielle :

$\mathbf{Var\,[\,(X'X)^-\,X'y\,] = (X'X)^-\,X'X\,(X'X)^-\,\sigma_e^2 = (X'X)^-\,\sigma_e^2}$ (9)

Ces principaux résultats vont être utilisés dans la partie suivante.

## 4 / La combinaison des différentes informations

Il existe plusieurs méthodes pour présenter comment les différentes informations sont combinées entre elles pour obtenir un index. Elles sont toutes assez complexes. Celle que nous avons choisie est la plus intuitive. Le modèle général (1) peut être écrit sous forme matricielle :

$\mathbf{y = X m + Z a + e}$ (10)

avec **y** le vecteur des n performances, **m** le vecteur des $n_m$ niveaux de facteurs de milieu, **a** le vecteur des $n_a$ valeurs génétiques, **X** la matrice de dimension (n, $n_m$), constituée de 0 et de 1, le terme (p,q) étant égal à 1 si la performance p est soumise au niveau de facteur q, **Z** la matrice de dimension (n, $n_a$), constituée de 0 et de 1, le terme (p,q) étant égal à 1 si la performance p est soumise à l'effet $a_q$, et **e** le vecteur des n résiduelles.

### 4.1 / Effets génétiques

Dans un premier temps, on suppose pour simplifier la présentation que les effets de milieu sont connus. Comme les effets génétiques sont des effets

aléatoires, on dispose donc de deux types d'information les concernant, les données corrigées pour les effets de milieu ($\mathbf{y - X m = y_c}$) et l'information *a priori* sur la distribution de **a**.

Considérant les données seulement, on a le modèle
$\mathbf{y_c = Z a + e}$ (11)
et on en déduit une première estimation de **a**, notée $\hat{\mathbf{a}}_1$, en utilisant (8).
$\hat{\mathbf{a}}_1 = \mathbf{(Z'Z)^- Z'y_c}$ (12)

**a** est estimé par la moyenne des performances de chaque individu, corrigées pour les effets de milieu. D'après (9), la variance d'erreur de cet estimateur est $\mathbf{V_1 = (Z'Z)^-} \sigma_e^2$.

On sait par ailleurs, d'après les règles de génétique quantitative, que les valeurs génétiques sont distribuées normalement, avec une espérance nulle et une variance $\mathbf{A}\,\sigma_a^2$. On sait donc qu'à défaut d'autre information, **E(a)=0**. Cette information *a priori* fournit un deuxième estimateur de **a**, son espérance :
$\hat{\mathbf{a}}_2 = \mathbf{0}$ (13)

Comme cet estimateur n'est basé que sur l'information *a priori*, sa variance d'erreur est égale à la variance de **a**, $\mathbf{V_2 = A}\,\sigma_a^2$. $\hat{\mathbf{a}}_1$ **et** $\hat{\mathbf{a}}_2$ étant deux estimations indépendantes, basées sur aucune information commune, on peut les combiner de façon optimale en les pondérant par leur précision, c'est-à-dire par l'inverse de leur variance d'erreur.
$\hat{\mathbf{a}} = [\mathbf{V_1^{-1} + V_2^{-1}}]^{-1}\,[\mathbf{V_1^{-1}}\,\hat{\mathbf{a}}_1 + \mathbf{V_2^{-1}}\,\hat{\mathbf{a}}_2]$ (14)

Le développement de ces expressions donne les résultats suivants :
$\mathbf{V_1^{-1}}\,\hat{\mathbf{a}}_1 = \mathbf{Z'y_c} / \sigma_e^2$
$\mathbf{V_2^{-1}}\,\hat{\mathbf{a}}_2 = \mathbf{0}$
$\mathbf{V_1^{-1} + V_2^{-1} = Z'Z} / \sigma_e^2 + \mathbf{A^{-1}} / \sigma_a^2$

Donc finalement
$\hat{\mathbf{a}} = [\mathbf{Z'Z} / \sigma_e^2 + \mathbf{A^{-1}} / \sigma_a^2]^{-1}\,\mathbf{Z'y_c} / \sigma_e^2$

ou encore, en posant $\alpha = \sigma_e^2 / \sigma_a^2$ :
$\hat{\mathbf{a}} = [\mathbf{Z'Z} + \mathbf{A^{-1}}\,\alpha]^{-1}\quad \mathbf{Z'\,(y - X\,m)}$ (15)

## 4.2 / Effets de milieu

En fait, les effets de milieu sont inconnus et doivent donc être estimés simultanément. En supposant les valeurs génétiques connues, on peut écrire
$\mathbf{y - Z\,a = X\,m + e}$ **(16)**

Comme **m** est considéré comme fixé, aucune hypothèse n'est faite sur sa distribution et on ne dispose comme information que des données ajustées pour les valeurs génétiques. **m** est estimé par la moyenne des données corrigées pour les valeurs génétiques, soit en appliquant (8) à **(y - Z a)**
$\hat{\mathbf{m}} = \mathbf{(X'X)^-\,X'\,(y - Z\,a)}$ (17)

## 4.3 / Equations du modèle mixte

On dispose donc d'un double système d'équations (15) et (17), connu sous le nom d'**équations du modèle mixte** :

$$\begin{aligned} \mathbf{X'X}\quad & \hat{\mathbf{m}} = \mathbf{X'\,(y - Z\,\hat{a})} \\ [\mathbf{Z'Z + A^{-1}}\,\alpha]\quad & \hat{\mathbf{a}} = \mathbf{Z'\,(y - X\,\hat{m})} \end{aligned}$$

Ce système est présenté habituellement sous la forme suivante :

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z + A^{-1}}\,\alpha \end{bmatrix} \begin{bmatrix} \hat{\mathbf{m}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix} \quad (18)$$

Cette combinaison d'informations *a priori* et de données peut être illustrée simplement dans le cadre d'un modèle père sans relations de parenté (donc **A = I**). Supposons les effets de milieu connus, donc seule l'équation (15) doit être résolue.
$\mathbf{1/2\,\hat{a} = [Z'Z + I}\,\alpha]^{-1}\,\mathbf{Z'\,y_c}$

Explicitons chacun des termes impliqués. **Z'Z** est une matrice carrée de dimension égale au nombre de père ($n_p$), dont tous les termes hors diagonaux sont nuls (on dit que la matrice est diagonale), et dont le ième terme diagonal, correspondant au père i, contient le nombre de performances $n_i$ des filles de i. **Z' y**$_c$ est un vecteur de dimension $n_p$, dont le ième terme, correspondant au père i, contient la somme des performances des filles de i , $\Sigma\, y_c$, (ajustées pour les effets de milieu). $\alpha$ est le rapport (constant) de la variance résiduelle sur la variance des effets pères. La variance des effets pères est $V(1/2\,a_i) = 0{,}25\,\sigma_a^2 = 0{,}25\,h^2\,\sigma^2$, et la variance résiduelle est $\sigma^2 - 0{,}25\,h^2\,\sigma^2 = \sigma^2\,(1 - 0{,}25\,h^2)$. Donc $\alpha = (4 - h^2)/h^2$ et vaut par exemple 15 si $h^2 = 0{,}25$. Le terme (i,i) de **Z'Z + I** $\alpha$ est donc $n_i + \alpha$, et on en déduit que

$$1/2\,\hat{a}_i = \Sigma\, y_c / (n_i + \alpha)$$

$$\hat{a}_i = 2\,[\,n_i / (n_i + \alpha)\,]\,\bar{y}_c$$

Lorsque le nombre de filles est élevé, l'index tend vers $2\,\bar{y}_c$, soit deux fois la moyenne de ses produits $\bar{y}_c$ (en déviation à la population), ce qui est exactement la définition de la valeur génétique additive. Par contre, si $n_i$ est faible, l'index ne vaut qu'une fraction de la supériorité moyenne des filles, et cette fraction est d'autant plus faible que le père a peu de produits. L'index est alors régressé vers son espérance 0.

Dans les équations du modèle mixte (18), les effets génétiques et les effets du milieu sont estimés simultanément. Cette propriété très importante est caractéristique du BLUP (=Best Linear Unbiased Prediction) d'Henderson (1973). On peut au contraire envisager une estimation séquentielle, plutôt que simultanée, des effets de milieu, puis des effets génétiques. Les données sont alors ajustées pour les effets de milieu préalablement estimés (dans un modèle n'incluant pas les effets génétiques), avant d'être utilisées pour estimer les effets génétiques. C'était le principe des premières méthodes d'évaluation, dites de "comparaison aux contemporaines", utilisées jusqu'au milieu des années 70. Cette évaluation séquentielle, bien que beaucoup plus simple, n'est en général pas acceptable dans les conditions actuelles, car elle repose sur des hypothèses clairement non vérifiées. Par exemple, les différences de performances entre années sont dues partiellement à des effets de milieu, partiellement au progrès génétique. Ajuster préalablement les données pour les différences entre années revient à supposer que ces différences entre années sont uniquement dues au milieu, et les différences génétiques entre années sont gommées. Les index calculés sont alors des index intra année, non comparables dans le temps. Il en est de même entre troupeaux : une correction préliminaire des effets troupeau suppose qu'il n'existe pas de variation de niveau génétique entre troupeaux. Les index calculés sont des index intra troupeau, non comparables dans l'espace. Au contraire, si le modèle de description des données est correct, les index BLUP sont comparables dans le temps et l'espace, au niveau de l'ensemble de la population. Ainsi, l'index d'une vache née en 1989 en Bretagne peut être comparé à l'index d'une autre vache née en 1975 dans le Nord.

## 5 / La matrice de parentés

La construction du système d'équations (18) requiert l'inverse de la matrice de parentés $A^{-1}$. Une méthode pour construire simplement cette inverse a été développée par Henderson (1976). L'aléa de méiose $\phi_i$ est défini en (2) comme la différence entre la valeur génétique d'un individu i et la moyenne de celles de ses parents p et m. Si $w_i$ est défini comme la part de $a_i$ non explicable par la généalogie, $w_i$ prend les valeurs suivantes :

$w_i = \phi_i = a_i - 0,5\ a_p - 0,5\ a_m$ si p et m sont connus,
$w_i = \phi_i + 0,5\ a_m = a_i - 0,5\ a_p$ si p est connu et m inconnue,
$w_i = \phi_i + 0,5\ a_p = ai - 0,5\ a_m$ si m est connue et p inconnu,
$w_i = \phi_i + 0,5\ a_p + 0,5\ a_m = a_i$ si p et m sont inconnus.

Sous les hypothèses d'absence de consanguinité dans la population et que les parents inconnus sont non sélectionnés, l'espérance de $w_i$ est nulle et sa variance vaut respectivement $0,5\ \sigma_a^2$, $0,75\ \sigma_a^2$, $0,75\ \sigma_a^2$ et $\sigma_a^2$ dans les 4 situations précédentes. Tous les $w_i$ sont indépendants les uns des autres. Donc le vecteur **w** des éléments $w_i$ a pour espérance **0** et pour variance $\mathbf{D}\ \sigma_a^2$. **D** est une matrice diagonale dont l'élément diagonal (i,i) vaut 1, 0,75 ou 0,5 selon que l'individu i a 0, 1 ou 2 parents connus.

D'après la définition donnée plus haut, le vecteur **w** peut s'écrire **w = L a**, avec **L** une matrice triangulaire inférieure, dont les termes diagonaux sont égaux à 1 et contenant 0, 1 ou 2 termes hors diagonaux non nuls, égaux à -0,5, et reliant un individu à son père et à sa mère. On peut utiliser cette relation pour calculer $\mathbf{A}^{-1}$.

$\mathbf{Var(a) = A}\ \sigma_a^2$
et $\mathrm{Var}(\mathbf{a}) = \mathrm{Var}(\mathbf{L}^{-1}\ \mathbf{w}) = \mathbf{L}^{-1}\ \mathrm{Var}(\mathbf{w})\ \mathbf{L}^{-1\prime} = \mathbf{L}^{-1}\ \mathbf{D}\ \mathbf{L}^{-1\prime}\ \sigma_a^2$

Donc $\mathbf{A = L}^{-1}\ \mathbf{D}\ \mathbf{L}^{-1\prime}$ (19)
et finalement $\mathbf{A}^{-1} = \mathbf{L'}\ \mathbf{D}^{-1}\ \mathbf{L}$ (20)

Grâce à la structure très simple de **L** et **D**, cette expression peut être développée algèbriquement, et Henderson a fourni les règles pour construire $\mathbf{A}^{-1}$ en l'absence de consanguinité (en présence de consanguinité, les éléments diagonaux de $\mathbf{D}^{-1}$ doivent être calculés préalablement). Ces règles ne nécessitent pas de calculer **A** et s'appliquent directement à une liste de généalogies de type (individu i, père p, mère m). Pour chaque triplet (i, p, m), elles consistent à :

- déterminer $d_{ii} = d$, le terme diagonal de $\mathbf{D}^{-1}$ correspondant à i, égal à 1, 4/3 ou 2, selon que 0, 1 ou les 2 parents de i sont connus.

- ajouter d au terme (i,i) de $\mathbf{A}^{-1}$
ajouter -d/2 à (i,p) et (p,i), d/4 à (p,p) si p est connu
ajouter -d/2 à (i,m) et (m,i), d/4 à (m,m) si m est connue
ajouter d/4 à (p,m) et (m,p) si p et m sont connus

Grâce à ces règles, les équations du modèle mixte peuvent être facilement construites. Illustrons ces équations à l'aide d'un petit exemple, constitué des quatre généalogies suivantes :

| Individu | Père | Mère |
|---|---|---|
| 1 | inc | inc |
| 2 | inc | inc |
| 3 | 1 | 2 |
| 4 | 1 | inc |

Les matrices **D** et **L** sont égales à

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0,5 & 0 \\ 0 & 0 & 0 & 0,75 \end{bmatrix} \quad \text{et} \quad \mathbf{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -0,5 & -0,5 & 1 & 0 \\ -0,5 & 0 & 0 & 1 \end{bmatrix}$$

Application des règles d'Henderson :

| Animal | d | Contributions à $\mathbf{A}^{-1}$ | |
|---|---|---|---|
| | | Termes | Valeurs |
| 1 | 1 | (1,1) | 1 |
| 2 | 1 | (2,2) | 1 |
| 3 | 2 | (3,3) | 2 |
| | | (3,1),(1,3),(3,2),(2,3) | -1 |
| | | (1,1),(1,2),(2,1),(2,2) | 0,5 |
| 4 | 4/3 | (4,4) | 4/3 |
| | | (1,4),(4,1) | -2/3 |
| | | (1,1) | 1/3 |

Finalement, la matrice $\mathbf{A}^{-1}$ est égale à :

$$\mathbf{A}^{-1} = \begin{bmatrix} 1,83 & 0,5 & -1 & -0,67 \\ 0,5 & 1,5 & -1 & 0 \\ -1 & -1 & 2 & 0 \\ -0,67 & 0 & 0 & 1,33 \end{bmatrix}$$

## 6 / Illustration des équations du modèle animal

Pour comprendre la signification des équations du modèle mixte, il est utile d'individualiser chacune de ses équations. Dans le système (18), elles sont de deux types, correspondant aux valeurs génétiques et aux effets de milieu.

### 6.1 / Effets de milieu

Un effet de milieu (l'effet $m_k$ du troupeau k par exemple) est estimé par la moyenne des $n_k$ données réalisées dans ce milieu (dans le troupeau k), ajustées pour le niveau génétique des animaux et éventuellement pour les autres effets de milieu (effet $s_l$ de la saison l par exemple) :

$$\hat{m}_k = \Sigma\ (y_{ikl} - \hat{a}_i - \hat{s}_l)/\ n_k \quad (21)$$

### 6.2 / Effets génétiques

Les règles d'Henderson montrent que dans $\mathbf{A}^{-1}$, un individu n'est relié qu'à ses parents, ses produits et ses conjoints. L'équation de l'individu i peut donc s'écrire comme la combinaison de trois sources d'information, l'ascendance, la descendance (ajustée pour le niveau des conjoints) et les performances (ajustées pour les effet de milieu) :

$$\hat{a}_i = \frac{0,5\ d_{ii}\ (\hat{a}_p + \hat{a}_m) + 0,5\ \Sigma\ d_{ff}\ (\hat{a}_f - 0,5\ \hat{a}_c) + (1\ /\ \alpha)\ \Sigma\ (y_{ikl} - \hat{m}_k - \hat{s}_l)}{d_{ii} + 0,25\ \Sigma\ d_{ff} + n_i\ /\ \alpha} \quad (22)$$

Lorsqu'un apparenté est inconnu (parent, conjoint), sa valeur est supposée nulle. Les individus f et c représentent les produits et les conjoints de i, et $n_i$ le nombre de performances de i. Pour une meilleure compréhension, la formule (22), qui apparaît assez complexe, peut être réécrite de façon synthétique en

individualisant les trois sources d'information, l'ascendance, la descendance et les performances propres. Ces trois composantes résument sans aucune approximation toute l'information concernant l'individu i, qui est souvent très volumineuse.

$$\hat{a}_i = \frac{p_1 \hat{a}_a + p_2 \hat{a}_d + p_3 \bar{y}_c}{p_1 + p_2 + p_3} \quad (23)$$

où :

* $\hat{a}_a$ est la valeur de i sur ascendance, c'est-à-dire la moyenne des index des parents. Cette composante, toujours définie, est l'espérance de $a_i$ : c'est sa valeur la plus probable, en l'absence d'autre information. Les autres informations $\hat{a}_d$ et $\bar{y}_c$ (qui n'existent que si i a des produits ou des performances) permettent d'estimer l'aléa de méiose, qui ne peut pas être prédit à partir des parents.
$\hat{a}_a = 0{,}5\ (\hat{a}_p + \hat{a}_m)$ (24)

* $\hat{a}_d$ représente deux fois la valeur moyenne des produits de i, ajustée pour la valeur des conjoints
$\hat{a}_d = 2 \Sigma d_{ff} (\hat{a}_f - 0{,}5\ \hat{a}_c) / \Sigma d_{ff}$ (25)

Dans le cas où tous les conjoints sont identifiés, $\hat{a}_d = 2 \Sigma (\hat{a}_f - 0{,}5\ \hat{a}_c) / n_f$, ce qui s'interprète aisément :

$0{,}5\ \hat{a}_c$ est l'espérance de ce qu'a transmis le conjoint c au descendant f,

$\hat{a}_f - 0{,}5\ \hat{a}_c$ est l'espérance de ce que i a transmis à f,

$2\ (\hat{a}_f - 0{,}5\ \hat{a}_c)$ est donc l'espérance de $a_i$ au vu de la valeur de f,

et donc $\hat{a}_d$ représente l'espérance de $a_i$ connaissant tous les descendants et conjoints.

* $\bar{y}_c$ est la moyenne des performances propres de i, corrigées pour les effets de milieu
$\bar{y}_c = \Sigma (y_{ikl} - \hat{m}_k - \hat{s}_l) / n_i$ (26)

Le poids de ces composantes est d'autant plus élevé que l'information contenue est plus riche. Il est fonction de la précision de l'information apportée par chacune des composantes. Bonaïti *et al* (1990) ont illustré comment chacune des pondérations peut être dérivée de l'inverse de la variance d'erreur de prédiction de la valeur génétique, connaissant chaque type d'information.

* $p_1$ est le poids de la valeur sur ascendance, égal à $d_{ii}$, soit 1, 4/3 ou 2, selon le nombre de parents connus (0,1 ou 2).

* $p_2$ est le poids de la valeur sur descendance, augmentant avec le nombre de descendants et fonction des coefficients $d_{ff}$, égaux à 4/3 ou 2 selon que les conjoints sont connus ou non.
$p_2 = \Sigma d_{ff} / 4$ (27)

* $p_3$ est le poids des performances propres. C'est une fonction de l'héritabilité qui assure en quelque sorte la conversion des performances en unité d'index. Il augmente avec l'héritabilité du caractère et avec le nombre de performances :
$p_3 = n_i / \alpha = n_i\ h^2 / (1 - h^2)$ (28)

Bien que résumée en trois sources d'information synthétiques (parents, descendants, performances), la valeur de i est estimée sans aucune approximation à partir de toutes les informations de la population. Par exemple, les performances des demi-soeurs de père de i sont intégrées dans leur propre index, donc contribuent à l'index du père en tant qu'information sur descendance. Elles contribuent finalement à l'index de i comme information sur ascendance. De plus, les poids des différentes informations sont les poids optimaux, c'est-à-dire ceux qui fournissent l'estimation globale la plus précise, qui présente la plus forte corrélation avec la valeur génétique vraie. Une propriété intéressante à signaler est l'ajustement de la valeur du descendant pour le conjoint (26), permettant ainsi de s'affranchir de l'hypothèse d'absence d'accouplements raisonnés (ou homogamie), hypothèse jamais respectée dans les programmes de sélection actuels.

Pour résoudre chacune de ces équations, il est nécessaire de connaître les autres effets. Ainsi, pour estimer un effet de milieu, il faut connaître les autres effets de milieu et les valeurs génétiques. Pour estimer la valeur génétique d'un individu, il faut connaître les effets de milieu et les index des apparentés. Cette dépendance entre les inconnues est traduite par le système d'équations. Tous les effets étant estimés simultanément, ils représentent donc une mesure indépendante des autres effets. Par exemple, l'effet du troupeau X n'inclut ni le niveau génétique des animaux de ce troupeau, ni la pyramide des âges, ni la politique de reproduction (mise bas précoces ou tardives, d'automne ou de printemps...), etc, effets qui sont estimés par ailleurs dans le modèle.

# 7 / Variantes du modèle

Le modèle de base (1) ne permet pas de couvrir toutes les situations. Laloë (1992) montre comment le modèle peut et doit être adapté aux différents caractères. Dans le domaine laitier, trois modifications importantes sont imposées par la réalité des programmes de sélection.

## 7.1 / Effet d'environnement permanent

Une femelle laitière peut réaliser plusieurs lactations. Restreindre l'analyse aux données de première lactations engendre donc une perte d'information. Mais si toutes les données sont analysées selon le modèle (1), l'hypothèse d'indépendance des résidus n'est pas respectée. En effet, deux performances d'un même animal tendent à se ressembler, et à se ressembler plus que ne le suppose l'héritabilité. La répétabilité r des performances, c'est-à-dire la corrélation entre performances d'un même animal, est supérieure à l'héritabilité $h^2$. Ceci s'explique par la présence d'un effet propre à l'animal, non transmissible à ses produits mais affectant toutes ses performances au cours de sa carrière. Cet effet, nommé "effet d'environnement permanent", résulte de divers effets de milieu, comme l'élevage en phase jeune ou les conséquences de troubles divers, et intègre aussi (dans le cas de notre modèle) la composante génétique non additive. Le modèle intègre cet effet d'environnement permanent comme facteur supplémentaire, noté **p**.
$y_{ikl} = m_k + a_i + p_i + e_{ikl}$ (29)

ou, en notation matricielle
$\mathbf{y = X m + Z a + W p + e}$ (30)

**p** est supposé indépendant de a et e, normalement distribué, d'espérance nulle. Si la répétabilité des performances est r, la variance liée à l'effet animal (incluant l'effet génétique et l'effet d'environnement permanent) est $r\ \sigma^2$. La variance de l'effet d'environ-

nement permanent seul est donc I (r-$h^2$) $\sigma^2$. En pratique, si cet effet était omis, les animaux réalisant plusieurs lactations, donc sélectionnés, seraient surestimés.

## 7.2 / Groupes

L'existence de progrès génétique, donc de changement de niveau génétique au cours du temps, n'est pas en contradiction avec l'hypothèse que la distribution des valeurs génétiques a une espérance nulle. Cette espérance n'est relative qu'à la population de base, c'est-à-dire à l'ensemble des ancêtres, supposés non sélectionnés, non apparentés et non consanguins, fondateurs de la population actuelle. L'élévation du niveau génétique est traduite par l'intermédiaire de la matrice de parentés : d'après (22), connaissant les parents p et m d'un animal i, l'espérance de $a_i$ n'est pas 0 mais 0,5 ($a_p$ + $a_m$). Si p et m sont sélectionnés, 0,5 ($a_p$ + $a_m$) n'est pas nul.

La population de base, d'espérance nulle, est constituée des animaux évalués qui n'ont pas de parents connus. Or la connaissance du pedigree des animaux actuels est très variable. Pour certains, tous les ancêtres sont connus sur plusieurs générations, alors que pour d'autres, l'arbre généalogique est beaucoup plus court, voire inconnu. Dans un programme de sélection efficace, un parent actuel inconnu a un niveau génétique *a priori* plus élevé qu'un parent inconnu ancien, du fait du progrès génétique. D'autres raisons peuvent aussi être à l'origine d'une hétérogénéité de niveau génétique des fondateurs, comme le croisement entre lignées ou l'intensité de sélection *a priori* des fondateurs. Pour prendre en compte ce phénomène, plusieurs populations de base, appelées groupes, sont supposées coexister, différant entre elles par leur espérance de niveau génétique. Cette espérance, *a priori* inconnue, doit être estimée dans le modèle, ce qui rajoute des inconnues supplémentaires (Quaas 1988).

## 7.3 / Variance résiduelle

Dans le modèle de base, les erreurs du modèle sont supposées indépendantes et de même variance. En pratique, la variance d'erreur n'est pas toujours constante. Par exemple les lactations extrapolées sont connues avec une précision moindre que les lactations terminées. Leur variance d'erreur est donc plus grande. La distribution de e est donc supposée $\mathbf{N}(\mathbf{0}, \mathbf{W}\,\sigma_e^2)$, **W** étant une matrice diagonale où chaque terme diagonal reflète les variations de variance résiduelle entre types de performance. Plus la variance résiduelle d'une performance est élevée, plus son poids dans l'analyse, égal à l'inverse de sa variance résiduelle, est faible. D'autres facteurs sont susceptibles de créer des hétérogénéités de variance et nécessitent, pour certains, des analyses très fouillées. En pratique, en l'absence d'ajustement, les meilleurs (et les plus mauvais) index seraient trouvés dans les conditions générant les plus grandes variations résiduelles.

# 8 / Propriétés et limites du système d'évaluation

Les limites des modèles simplifiés, autres que le modèle animal, ont été précisées par ailleurs. Elles résident essentiellement dans le non-respect des hypothèses posées dans ces modèles. Nous ne développons donc que les propriétés et limites du BLUP appliqué à un modèle animal. Ses propriétés théoriques regroupent à la fois celles du BLUP et celles du modèle animal. Dans le BLUP, les effets de milieu et les effets génétiques sont estimés simultanément. Les valeurs génétiques sont estimées dans le cadre d'une population et sont donc comparables dans le temps et l'espace. D'autre part, le modèle animal fournit des estimations non biaisées par la sélection, ni par l'homogamie, ni par la réduction de la variabilité génétique sous l'effet de la sélection et de la dérive génétique. Toute l'information est prise en compte, en particulier celle de tous les apparentés, et de façon optimale. Il semble par ailleurs que les résultats soient relativement robustes à un écart au modèle infinitésimal.

Toutefois, ces résultats théoriques ne sont obtenus en pratique que si certaines conditions sont remplies. Une condition importante est l'analyse de la population complète à chaque évaluation, incluant toutes les données et toutes les généalogies connues. Ce n'est qu'à cette condition que les résultats peuvent être comparés dans toute la population. De même, les résultats sont non biaisés par la sélection seulement si l'ensemble de l'information expliquant cette sélection est incluse dans l'analyse. Le modèle animal est supposé fournir des estimations de valeur génétique indépendantes de l'homogamie, grâce à l'ajustement de la valeur du produit pour la contribution du conjoint (26). Cet ajustement n'est correct que si la valeur du conjoint est bien estimée, grâce à ses performances et son pedigree. Si ces informations ne sont pas prises en compte dans l'analyse, la valeur du conjoint est mal estimée, et donc celles de tous les autres individus aussi. Autre exemple, si les moins bonnes performances n'apparaissent pas dans le fichier analysé, les meilleurs individus sont sous-estimés, manquant de contemporains pour mettre en évidence leur supériorité. Cet exemple montre la nécessité de contrôler et d'analyser tous les individus d'un même groupe de contemporains, donc d'un même élevage.

Même quand toutes les données sont analysées, il reste deux principaux problèmes potentiels :

- le modèle de description des données peut être incorrect. On a déjà dit que tout facteur affectant les données et non inclus dans le modèle engendre des biais dans l'évaluation génétique. Par exemple, une mise bas d'hiver est plus favorable qu'une mise bas d'été. Si l'effet saison de mise bas n'est pas inclus dans le modèle, la supériorité des performances des animaux mettant bas en hiver se traduit de façon incorrecte par une supériorité de leur index.

- la comparaison des résultats dans le temps et l'espace n'est possible que si le dispositif est connecté, c'est-à-dire si ces différences dans le temps et l'espace sont estimables. Par exemple, il est impossible d'estimer le niveau génétique moyen d'un troupeau isolé génétiquement, c'est-à-dire n'utilisant pas de reproducteurs ou d'animaux du reste de la population, car le niveau génétique moyen du troupeau ne peut pas être distingué de l'effet milieu troupeau. L'insémination artificielle apparaît comme un outil de choix pour créer des liens génétiques entre troupeaux et entre années, donc pour améliorer la connexion. Les échanges d'animaux sont un autre outil efficace, mais dont l'effet est plus difficile à quantifier.

Enfin, il faut garder à l'esprit que les index ne sont que des estimations de la valeur génétique vraie. Une estimation est d'autant plus précise que la quantité d'information est importante. En toute rigueur, on ne connaît la valeur génétique vraie d'un individu que si l'on mesure les performances d'un très grand nombre de produits, dans des élevages de grande taille. Dans tous les autres cas, l'index n'est que la valeur génétique la plus probable. La précision d'un index est mesurée par son coefficient de détermination, compris entre 0 (aucune information disponible) et 1 (on connaît la valeur génétique vraie). Un index sur ascendance seule a un coefficient de détermination inférieur à 0,5, un index basé sur une seule performance a un coefficient de détermination inférieur ou égal à l'héritabilité du caractère, tandis que le coefficient de détermination d'un index sur descendance peut atteindre théoriquement 1 si les produits sont très nombreux. A partir de l'index â et de son coefficient de détermination CD, on peut construire un intervalle de confiance de la valeur génétique vraie (a), pour un risque donné $\alpha$ :

$$\hat{a} - 1{,}96\sqrt{CD}\,\sigma_g < a < \hat{a} + 1{,}96\sqrt{CD}\,\sigma_g \qquad (\alpha < 5\%)$$

# 9 / Résolution du système d'équations

Le système d'équations (18), même modifié pour prendre en compte les remarques du paragraphe précédent, et même adapté à un modèle complexe, comptant de nombreux facteurs et interactions, garde une forme relativement simple. Toutefois, dans le cadre d'un modèle animal, il est caractérisé par une très grande taille, avec souvent plusieurs millions d'équations et autant d'inconnues : un index par animal, un effet d'environnement permanent par animal avec performance, plusieurs centaines de milliers de troupeau-année, plusieurs milliers d'autres effets de milieu. Cette très grande taille est due à la nécessité de réanalyser toutes les données connues et pas seulement les plus récentes.

La résolution d'un tel système est restée longtemps impossible, imposant alors de poser des modèles plus simples ou de réaliser des approximations numériques. Il est possible maintenant, grâce aux progrès de l'informatique et au développement d'algorithmes efficaces, de résoudre exactement ce sytème dans le cadre d'un modèle animal, mais cela reste encore extrêmement lourd. Le but de cet article n'est pas de présenter les techniques numériques utilisées mais il faut cependant être conscient que la mise en pratique de la théorie est une difficulté majeure de la génétique quantitative, qui requiert à la fois beaucoup de capacités de calcul et beaucoup de travail d'analyse et de programmation informatique.

Comme les procédures d'inversion matricielle deviennent prohibitives au delà de quelques centaines à quelques milliers d'équations, elles sont donc totalement inadaptées aux problèmes d'évaluation génétique. Les algorithmes les plus simples généralement utilisés sont de deux types : la résolution itérative et l'absorption. Ces deux méthodes sont illustrées sur l'exemple suivant. Soit le système de deux équations à deux inconnues x et y :

$$ax + by = c \quad (31a)$$
$$dx + ey = f \quad (31b)$$

## 9.1 / Absorption

L'absorption, aussi connue sur le nom de substitution ou d'élimination Gaussienne, consiste à exprimer y en fonction de x à l'aide d'une équation et remplacer y par son expression dans l'autre équation. Le système n'a plus alors qu'une équation à une inconnue. De l'équation (31b), on obtient $y = (f - d\,x)/e$. A l'aide de cette expression, (31a) se réécrit alors : $(a - b\,d/e)\,x = c - b\,f/e$, équation à une inconnue x que l'on résout facilement. L'inconnue y est ensuite déterminée en reportant la valeur de x dans la deuxième équation.

## 9.2 / Résolution itérative

Le système peut se réécrire :

$$x = (c - b\,y)/a \quad (32a)$$
$$y = (f - d\,x)/e \quad (32b)$$

Avec (32a), en supposant y connu, on peut déterminer x, puis avec cette nouvelle valeur de x, on peut calculer y à l'aide de (32b). Cette nouvelle valeur de y permet de recalculer x, et ainsi de suite. L'ensemble de ce processus constitue une itération. Il requiert le choix de valeurs initiales des inconnues et il est poursuivi jusqu'à la convergence, c'est-à-dire jusqu'à ce que les estimations ne varient plus d'une itération à l'autre. Il existe de multiples variantes sur le même principe, ainsi que des procédés permettant d'accélérer la convergence, c'est-à-dire diminuer le nombre d'itérations nécessaires. En pratique, une procédure d'évaluation est souvent constituée de différentes étapes incluant des absorptions, des phases itératives, parfois des résolutions directes ou d'autres types d'algorithmes.

# 10 / Utilisation des index

L'utilisation principale des index est bien sûr le classement des animaux candidats à la sélection, puis leur sélection proprement dite. Cette tâche est grandement facilitée par les propriétés intéressantes du BLUP. En effet, comme les effets de milieu et les effets génétiques sont estimés simultanément, les valeurs génétiques sont comparables dans le temps et l'espace. Il est donc possible de sélectionner de façon optimale en choisissant un seuil unique au sein de la population, sans avoir à considérer l'âge ou l'origine des animaux. En pratique, l'évaluation laitière est réalisée tous les trimestres et les résultats sont diffusés à l'ensemble des organismes concernés, en charge de la sélection : Institut de l'Elevage (lui-même en charge de la diffusion aux centres d'insémination, aux médias et au public par l'édition d'un catalogue), UPRA (en charge de la diffusion aux éleveurs), fichiers nationaux et régionaux...

Pour la même raison, les différences génétiques entre troupeaux, entre régions, ainsi que le progrès génétique, peuvent être simplement mesurés par la moyenne des index des animaux par troupeau, région ou année de naissance. Ceci constitue un outil puissant d'analyse rétrospective, par exemple pour vérifier l'efficacité des programmes de sélection. Un bilan, publié annuellement par l'INRA et l'Institut de l'Elevage, fournit une photographie de l'évolution génétique récente en France et des différences régionales. Les figures 1 et 2 illustrent l'évolution génétique pour les femelles des trois principales races bovines laitières françaises. Elles montrent que plus de la moitié du progrès phénotypique, de l'ordre de

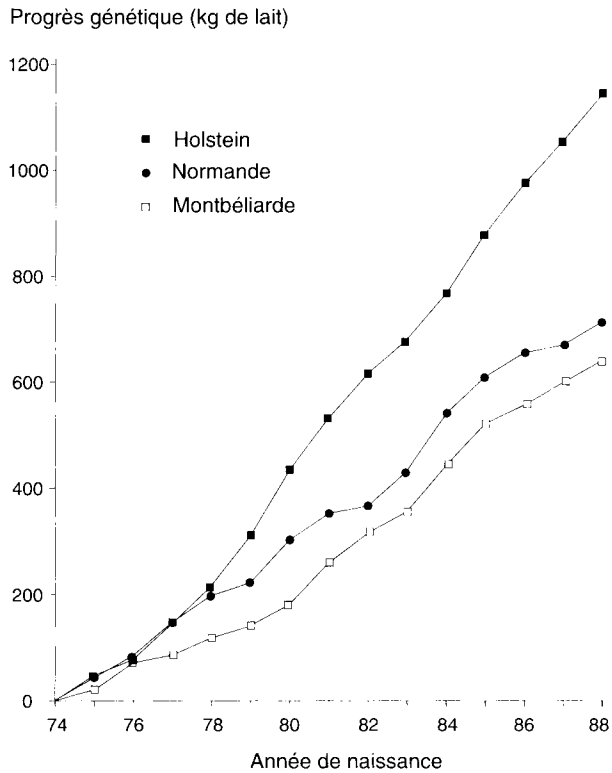


**Figure 1.** *Evolution du niveau génétique des femelles pour la quantité de lait dans les trois principales races françaises.*

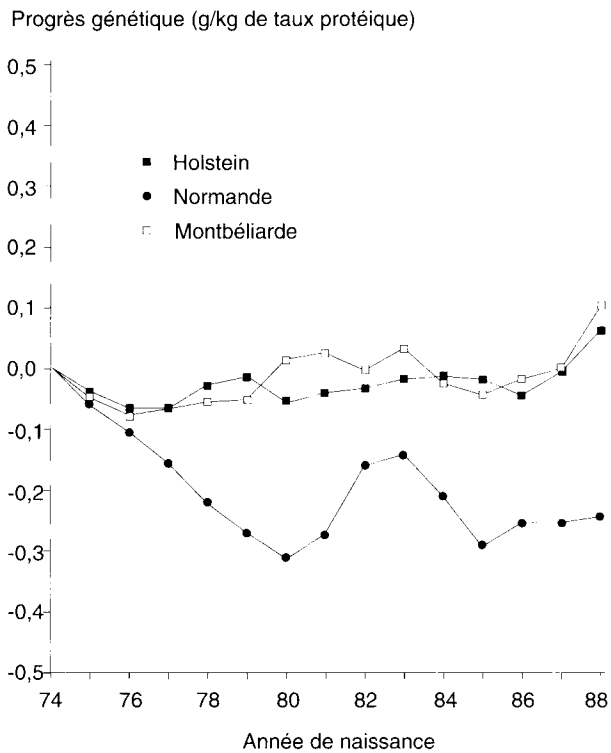


**Figure 2.** *Evolution du niveau génétique des femelles pour le taux protéique dans les trois principales races françaises.*

100 kg de lait par an, est directement imputable à la génétique. Le progrès laitier apparaît plus élevé en Holstein que dans les autres races, du fait du croisement d'absorption de la Frisonne européenne par la Holstein nord-américaine : dans ce cas, l'évolution génétique correspond à la somme des effets de la migration et de la sélection. En moyenne, l'augmentation du niveau laitier a été réalisée avec une augmentation du taux butyreux et un maintien du taux protéique, à l'exception de la race Normande, qui présente une dégradation légère du taux protéique, du fait d'une sélection plus axée sur la quantité de lait.

De la même façon, en fonction des inséminations pratiquées et des choix de sélection réalisés, on peut prédire très précisement le niveau génétique futur des troupeaux ou du cheptel national. Cette prédiction est réalisée dans le bilan zootechnique des inséminations, publié annuellement par l'INRA et l'Institut de l'Elevage, et qui récapitule le niveau génétique des inséminations pratiquées en France.

L'évaluation fournit enfin des estimations d'effets de milieu utilisables à des fins d'appui technique et de prévision de production. Elle permet de cartographier le niveau technique comme le niveau génétique des éleveurs. Elle peut aussi fournir des indicateurs de la valeur économique de chaque animal, à bien distinguer de sa valeur génétique : tandis que l'index représente la meilleure estimation de la valeur génétique d'un animal, c'est-à-dire ce qu'il est capable de transmettre à ses produits, la somme index + effet d'environnement permanent prédit au mieux l'aptitude d'une femelle à produire au cours de ses lactations futures. Autrement dit, l'index doit servir pour le choix des reproducteurs, tandis que les réformes doivent plutôt être basées sur l'aptitude à produire.

## Conclusion

L'évaluation génétique a beaucoup évolué dans sa théorie et dans sa mise en application depuis son origine dans les années 60. Les premières méthodes étaient des modèles pères basés sur la comparaison aux contemporaines (CC). Le BLUP a progressivement remplacé la CC dans les années 70. La prise en compte des parentés est postérieure à 1975. Le principe du modèle animal n'a été clairement énoncé qu'en 1980 (Quaas et Pollack 1980), et ses possibilités d'utilisation, liées au développement de l'informatique, sont encore très récentes (Wiggans *et al* 1988).

Le BLUP appliqué à un modèle animal constitue actuellement la référence internationale, grâce à ses intéressantes propriétés théoriques et son aspect unificateur, facilitant les comparaisons et les échanges internationaux. Toutefois, il faut rester conscient que ces propriétés théoriques ne sont réalisées en pratique que si les données le permettent, c'est-à-dire si elles sont fiables et de bonne qualité.

Si le BLUP appliqué à un modèle animal apparaît actuellement comme un aboutissement, il faut garder à l'esprit qu'une méthode d'évaluation est un système en constante évolution. Dans le futur, on peut prévoir deux grands types d'adaptation. L'une concerne la définition des caractères évalués, qui sera de plus en plus fine. Par exemple, il est possible qu'on analyse bientôt les productions par contrôle, plutôt que les productions par lactation comme maintenant. L'autre est la définition plus fine du déterminisme génétique. Alors que les méthodes actuelles sont basées sur le modèle infinitésimal, purement statistique et opérationnel, l'évaluation future devra prendre en compte les acquis de la génétique moléculaire dans la connaissance du déterminisme génétique des caractères d'importance économique. Elle intègrera donc probablement à la fois un déterminisme polygénique et les QTL (=Quantitative Trait Locus, gène individualisé gouvernant un caractère quantitatif) marqués.

## Références bibliographiques

Bonaïti B., Boichard D., Verrier E., Ducrocq V., Barbat A., Briend M., 1990. La méthode française d'évaluation génétique des reproducteurs laitiers. INRA Prod. Anim., 3, 83-92.

Ducrocq V., 1992. Du modèle génétique au modèle statistique. INRA Prod. Anim., hors série "Eléments de génétique quantitative et application aux populations animales", 75-81.

Henderson C.R., 1973. Sire evaluation and genetic trend. In "Proceeding of Animal Breeding and Genetics in Honor of Dr J.L. Lush", Blackburgh, Virginia, August 1972, p10-41. American Society of Animal Science, Champaign, Illinois.

Henderson C.R., 1976. A simple method to compute the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics, 32, 69-83.

Laloë D., 1992. Complications des modèles d'évaluation : exemples des performances répétées et des effets maternels. INRA Prod. Anim., hors série "Eléments de génétique quantitative et application aux populations animales", 197-199.

Quaas R.L., 1988. Additive genetic model with groups and relationships. J. Dairy Sci., 71, 1338-1345.

Quaas R.L., Pollack E.J., 1980. Mixed model methodology for farm and ranch beef cattle testing program. J. Anim. Sci., 51, 1277-1287.

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.