Conception d'index pour la sélection réciproque récurrente : aspects génétiques, statistiques et informatiques

Philippe BARADAT
Unité de recherche biométrie, CIRAD-CP, Montpellier, France.
Thierry LABBÉ
Station de recherches forestières de Bordeaux-Cestas, INRA, France.
Jean-Marc BOUVET
CIRAD-Forêt, Pointe Noire, République du Congo.

Résumé. Les index permettent un traitement unifié de tous les types de sélection mis en oeuvre dans un programme d'amélioration génétique. Ils autorisent une combinaison libre de plusieurs sources d'information sur les génotypes (valeurs propres et performances d'apparentés) et de plusieurs caractères. On peut ou non hiérarchiser les observations en « caractères cibles » que l'on cherche à améliorer et « caractères prédicteurs » qui servent, par exemple, à affiner la prédiction des valeurs génétiques des caractères cibles ou à raccourcir le cycle de sélection (sélection précoce de caractères « adultes » sur la base de caractères juvéniles qui leur sont génétiquement corrélés). La notion de caractères cibles et de caractères prédicteurs permet également de traiter l'interaction génotype-environnement. Il peut s'agir d'une sélection pour les performances sur un nombre limité d'années, l'objectif étant l'amélioration dans des caractéristiques climatiques moyennes ou, au contraire, extrêmes. La même approche est valable pour les sites de test ou les façons culturales ou pour associer des gènes marqueurs à la sélection de caractères d'intérêt agronomique. Par ailleurs, en fonction des objectifs (sélection de la population d'amélioration ou type de création variétale), les index permettent de prendre en compte la valeur génétique additive seule ou la valeur génétique totale. Le présent article passe en revue les différentes approches utilisées pour la construction des index de sélection ainsi que les outils nécessaires pour leur utilisation. Il donne ensuite et justifie les modèles adaptés au cas de la sélection réciproque récurrente en prenant comme exemple l'amélioration de l'Eucalyptus au Congo. L'architecture et le fonctionnement des programmes de calcul sont enfin décrits brièvement.

La théorie des index de sélection multicaractères a été formulée pour la première fois par HAZEL (1943) dans le cas d'une sélection massale et étendue par ROUVIER (1969) à des cas plus compliqués tels que la sélection combinée sur plans de croisements hiérarchiques.

Les index de sélection sont traditionnellement calculés dans le cadre d'un modèle BLP « best linear predictors » où effets fixés et effets aléatoires sont traités de façon séquentielle. Ils sont de plus en plus abordés à la suite des travaux sur le modèle mixte (HENDERSON, 1973) et de leur adaptation à la génétique animale (QUAAS et POLLAK, 1980) selon un modèle intégré où les effets génétiques aléatoires (BLUP « best linear unbiased predictors ») sont traités simultanément avec les effets fixés, en général de nature environnementale (BLUE « best linear unbiased estimators »). Parallèlement, les méthodes traditionnelles d'analyse de variance non-

orthogonale : HENDERSON type 3 essentiellement (HENDERSON, 1953) et accessoirement MINQUE « minimum norm quadratic unbiased estimator » (RAO, 1971) sont remplacées par des méthodes du maximum de vraisemblance, ML « maximum likelihood » (ANDERSON et BANCROFT, 1952) ou REML « restricted maximum likelihood » où le maximum de vraisemblance est limité aux effets aléatoires (THOMPSON, 1962).

Toutefois, l'approche BLP conserve l'avantage d'une plus grande souplesse et d'une grande rapidité de calcul. Son caractère modulaire fait que l'on peut implémenter les programmes correspondants sur micro-ordinateurs. Le fait de dissocier prédiction des effets génétiques et ajustement des données aux facteurs du milieu donne la possibilité d'utiliser pour l'ajustement des méthodes distinctes de l'analyse de variance. Ce type de modèle reste compétitif dans la plupart des situations rencontrées chez les plantes pérennes (effectifs importants, effets du milieu en général bien contrôlés). Cela est d'autant plus vrai que l'on peut profiter de sa souplesse pour utiliser des modèles génétiques bien adaptés. C'est ainsi que la méthode BLP est actuellement utilisée pour le programme coopératif d'amélioration des pins du Sud-Est des USA, *Pinus taeda* et *Pinus elliottii* qui représente un enjeu économique considérable (WHITE et HODGE, 1988).

JEFFERSON (1989) a montré expérimentalement que, dans la pratique, les index calculés selon la méthode BLP pour une sélection combinée portant sur quelques centaines d'individus aboutissent à un classement des génotypes pratiquement identique à ceux construits à partir de la technique BLUP (corrélation de rang voisine de 1).

Les index de sélection multicaractères sont définis comme des combinaisons linéaires de valeurs génétiques (non-observables), $\hat{g}$, prédites par régression sur des « prédicteurs phénotypiques » (observables). Leur forme générale est : $I = \sum_i b_i \hat{g}_i$.

En amélioration des plantes, ces prédicteurs sont le plus souvent obtenus à partir de dispositifs expérimentaux qui ont un double rôle :

a) assurer l'absence ou la faible incidence des effets d'environnement commun, de façon à ne pas biaiser les estimations des variances et des covariances phénotypiques et génétiques ainsi que des prédicteurs phénotypiques. Cet objectif est atteint en fractionnant les unités génétiques mises en comparaison (clones, familles, provenances, espèces...) en « répétitions » (parcelles unitaires) ;

b) réduire la variabilité due au milieu et donc augmenter l'héritabilité des caractères mis en jeu, par regroupement des parcelles unitaires en « blocs » de petite taille.

Par ailleurs, une même collection de génotypes se trouve souvent implantée dans plusieurs sites ou « stations » qui diffèrent par des caractéristiques pédologiques et climatiques.

Les effets bloc et station sont en, règle générale, considérés comme fixés et se pose alors le problème de leur prise en compte dans les modèles de prédiction des valeurs génétiques aléatoires. Il existe souvent un effet d'interaction génotype-station qui peut aboutir à l'extrême à des changements de classement des unités génétiques. Le modèle global d'analyse de variance de cet exemple s'écrit alors pour une observation correspondant à l'individu l de l'unité génétique i, située dans le bloc k de la station j, μ désignant la moyenne générale :

$$y_{ijkl} = \mu + a_i + \gamma_j + \beta_{jk} + (a\gamma)_{ij} + e_{ijkl}$$

où $(a\gamma)_{ij}$ est l'effet d'interaction génotype-station (aléatoire).

Comme dans tout ce qui suit, les lettres romaines désignent un effet aléatoire et les lettres grecques un effet fixé.

La méthode BLP suppose les effets fixés connus ou, du moins, estimés de façon précise. Telle que la décrit HENDERSON (1977) et appliquée à ce cas de figure, elle procèdera en deux étapes :

a) ajustement des valeurs observées aux effets fixés, station et bloc/station, estimés par la méthode des moindres carrés. Les valeurs ajustées seront alors :

$$y^*_{il} = y_{ijkl} - \hat{\gamma}_j - \hat{\beta}_{jk}$$

b) calcul des index de sélection sur les effets aléatoires, a^*_i et e^*_{il}, suivant le modèle :

$$y^*_{il} = \mu + a^*_i + e^*_{il}$$

où l'effet du génotype ajusté, a^*_i , sera éventuellement décomposé en effets élémentaires correspondant, par exemple, à un plan de croisements. La deuxième partie illustrera une telle décomposition. Elle montrera également que des effets génétiques peuvent être considérés comme fixés, par exemple les effets génétiques additifs de testeurs.

La méthode BLUP, formalisée pour la première fois par HENDERSON (1973) traite directement le modèle global de la première équation en prenant en compte simultanément les effets fixés pour lesquels elle donnera des estimations non biaisées et de variance minimum et les effets aléatoires qui permettront de prédire les valeurs génétiques. Les effets fixés ne sont pas supposés connus, contrairement à la méthode BLP. En revanche, les deux méthodes considèrent que les matrices de variances-covariances des effets aléatoires sont connues. Elles possèdent des propriétés optimales si la distribution des caractères pris en compte dans le modèle est multi-normale. Par ailleurs les prédictions de gains génétiques sont non-biaisées seulement si cette condition est réalisée.

La présente contribution concerne une approche BLP d'index adaptés à la sélection réciproque récurrente. Elle traite également du problème de la précision de l'estimation des gains génétiques. Elle indique enfin la conception et le mode de fonctionnement des programmes de calcul mis au point à l'INRA et au CIRAD. Les variantes de modèles d'index sont étudiées selon les objectifs : sélection de clones *Gn* sur test de descendance, des meilleures combinaisons hybrides ou des meilleurs clones *Gn+1*.

La conception des programmes permet de déterminer facilement les conséquences du choix de la méthode de sélection et de la pondération des caractères tant sur les espérances des gains génétiques que sur la diversité de la sous-population sélectionnée.

Principe de construction, propriétés et utilisation des index

Les index de sélection, outre la prédiction optimale des effets génétiques, posent un certain nombre de problèmes qui ne seront pas traités ici mais simplement cités.

Il s'agit par exemple du calcul des intensités de sélection dans des populations d'effectif limité (LINGREN et NILSSON, 1985), de la prise en compte de caractères discrets (FOULLEY *et al.*, 1983 ; FOULLEY et MANFREDI, 1991 ; GIANOLA et FOULLEY, 1983) ou de l'estimation optimale des composantes de la variance ou de la covariance nécessaires pour la construction des index (DOLIGEZ, 1992 ; FOULLEY, 1992 ; HARVILLE, 1977 ; HUBER *et al.*, 1992 ; SEARLE, 1971 ; SEARLE *et al.*, 1992 ; THOMPSON, 1962). Il s'agit également des conséquences de mauvaises estimations de matrices de variances-covariances sur la faisabilité des calculs d'index (FOULLEY et OLLIVIER, 1986), ainsi que de la troncature par sélection des populations de référence sur l'estimation des paramètres génétiques (OLLIVIER et DERRIEN, 1981).

Ne seront abordés ici que les aspects plus directement liés à la conception des modèles et des programmes ou à leur mode d'utilisation qui résulte de cette conception.

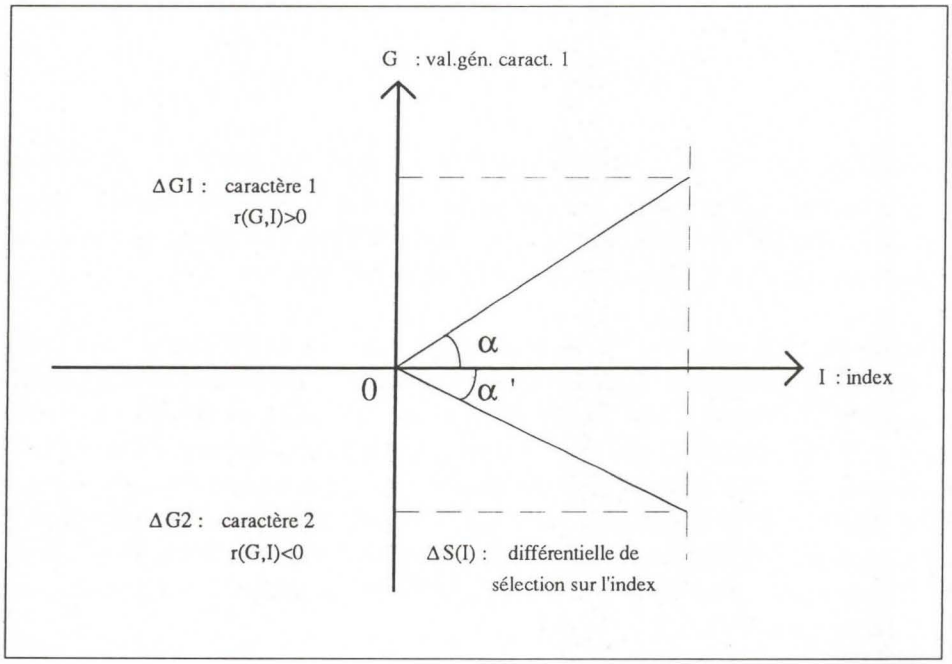


Figure 1. Réalisation de gains génétiques partiels sur deux caractères par troncature de la population pour un index corrélé à leurs valeurs génétiques.

La valeur génétique du caractère 1 (gain génétique ΔG1) est positivement corrélée à l'index ; celle du caractère 2 (gain génétique ΔG2) est corrélée négativement. Les deux gains génétiques, ΔG1 et ΔG2, sont déterminés par la différentielle de sélection sur l'index : $\Delta S(I) = i\sigma_I$ où i est l'intensité de sélection, et par le coefficient de régression de chaque valeur génétique sur l'index : $b = cov(G,I) / \sigma_I^2$. On a : $b1 = tg(\alpha)$ et $b2 = tg(\alpha')$.

La régression du génotype sur des prédicteurs phénotypiques

Il est commode de dissocier la présentation des index de sélection en deux étapes distinctes :
- prédiction des valeurs génétiques (obtention d'un vecteur colonne de q valeurs) ;
- réalisation d'une combinaison linéaire de ces prédictions par un jeu de coefficients.

La deuxième phase donne un scalaire, l'index de sélection I, dont la corrélation avec les valeurs génétiques G est fonction des rapports et du signe de ces coefficients. La figure 1 montre comment la corrélation entre l'index et les valeurs génétiques permet de réaliser des gains génétiques sur chaque caractère en tronquant la population pour une valeur-seuil de l'index.

La première étape peut être schématisée par la formule générale :

$$\left[\hat{G}\right] = \left[\Sigma GP\right]\left[\Sigma PP\right]^{-1}\left[\hat{p}\right]$$

Cette formule exprime que le vecteur colonne (q, 1) des q valeurs génétiques est prédit par régression linéaire sur des prédicteurs phénotypiques concernant q' caractères. Nous appelerons par la suite « caractères cibles » les q caractères qui font l'objet de la sélection et auxquels on applique une pondération et « caractères prédicteurs » les q' variables qui permettent la prédiction des q valeurs génétiques des caractères cibles.

Ces prédicteurs phénotypiques sont, d'une façon générale, sauf pour le cas trivial de la sélection massale, des effets aléatoires estimés par analyse de variance et apurés des effets fixés du modèle, s'il en existe. Le nombre de « sources d'information » disponibles, s, n'est autre que le nombre de facteurs génétiques du modèle d'analyse de variance plus, éventuellement, certaines de leurs interactions ainsi que les effets individuels intra-unité génétique.
Dans ces conditions, la matrice de covariances entre effets génétiques et prédicteurs phénotypiques, $\left[\Sigma GP\right]$, est de dimensions (q, sq') et $\left[\hat{p}\right]$, le vecteur des prédicteurs phénotypiques, un vecteur colonne (sq',1) composé de s sous-vecteurs de dimension (q',1).

$\left[\Sigma PP\right]$ est la matrice de variances-covariances entre prédicteurs. Elle est symétrique et de dimensions (sq', sq').

Cette écriture, tout à fait générale, permet d'envisager trois cas de figure :
a) Les q caractères cibles sont identiques aux q' caractères prédicteurs. On a alors affaire à un modèle symétrique où chaque variable sert à la fois à prédire sa propre valeur génétique et les q-1 valeurs génétiques des autres caractères ;

b) Les q' caractères prédicteurs et les q caractères cibles correspondent à deux ensembles d'observations complètement disjoints. Il s'agit d'un modèle de sélection indirecte d'un type très utilisé chez les plantes pérennes et, en particulier, les arbres forestiers, pour prédire les performances adultes à partir de caractères juvéniles. Ce type d'approche est illustré par WHITE et HODGE (1991) ;

c) Les deux groupes de caractères ne sont que partiellement disjoints. Un certain nombre d'observations jouent à la fois le rôle de caractères prédicteurs et de caractères cibles.

Ce troisième type de modèle s'impose lorsque l'on veut améliorer la précision de l'estimation des valeurs génétiques de caractères économiquement intéressants mais relativement peu héritables. On leur adjoint alors des « caractères satellites » qui n'ont d'autre intérêt que d'être plus héritables et de présenter de fortes corrélations génétiques avec les caractères cibles.

La deuxième étape permet d'obtenir les index de sélection à partir de $[\hat{G}]$ par :

$$I = [b]'[\hat{G}]$$

où $[b]'$ est le transposé du vecteur colonne (q, 1) des coefficients b.

La figure 1 montre comment la troncature des valeurs de l'index dans une population permet de réaliser un gain génétique sur des caractères qui présentent une corrélation génétique avec au moins un des prédicteurs. Elle illustre le fait que les index constituent une généralisation des différentes méthodes de sélection en traitant comme des cas particuliers la sélection directe ou la sélection indirecte. Le coefficient de régression de chaque valeur génétique sur l'index joue le même rôle que l'héritabilité dans le cas d'une sélection massale sur un seul caractère.

La mise en oeuvre des index utilise leurs propriétés statistiques qui sont les suivantes :

– la variance de l'index est donnée par :

$$\sigma_I^2 = [b]'[\Sigma GP][\Sigma PP]^{-1}[\Sigma GP]'[b] \ ;$$

– le vecteur colonne (q, 1) des covariances des valeurs génétiques des caractères cibles Y et de l'index I est donné par :

$$[\mathrm{cov}(G,I)] = [\Sigma GP][\Sigma PP]^{-1}[\Sigma GP]'[b].$$

Ces deux expressions permettent de calculer le progrès génétique attendu pour chaque caractère intégré dans l'index, connaissant l'intensité de sélection, i.

Pour le caractère Y^1, $b(G^1 / I)$ désignant le coefficient de régression de sa valeur génétique G^1 sur l'index I, on aura :

$$\Delta G^1 = i\sigma_I b(G^1 / I) = i\sigma_I \mathrm{cov}(G^1, I) / \sigma_I^2 = (i / \sigma_I)\mathrm{cov}(G^1, I) .$$

Il est par ailleurs utile de mesurer la fiabilité d'un calcul d'index par sa corrélation r(H,I) avec le « mérite » correspondant (index théorique calculé à partir de valeurs génétiques connues et non prédites).

– La variance du mérite, $H = [b]'[G]$, est donnée par :

$$\sigma_H^2 = [b]'[\Sigma GG][b]$$

où $[\sum GG]$ est la matrice carrée (q, q) des variances et des covariances entre caractères cibles.

On montre par ailleurs que $cov(H,I) = \sigma^2_I$. On a donc :

$$r(H,I) = \sigma^2_I / (\sigma_I \sigma_H) = \sigma_I / \sigma_H .$$

Le carré de cette corrélation est le coefficient de détermination génétique, CD(I), qui est également utilisé comme critère de fiabilité d'un calcul d'index.

La différence entre la variance du mérite et celle de l'index, $\sigma^2_H - \sigma^2_I = \sigma^2_H \left[1 - r^2_{(H,I)}\right]$, est la « variance d'erreur » de la prédiction, par analogie avec une régression multiple usuelle. Toutefois, elle ne saurait être considérée comme telle que si les matrices de variances-covariances et les prédicteurs phénotypiques étaient connus et non estimés.
Pour évaluer l'erreur standard sur les estimations de gains génétiques, il faudra donc avoir recours à d'autres techniques qui tiennent compte de l'ensemble des erreurs d'échantillonnage sur les paramètres du modèle utilisé.

Remarque : La variance du mérite est estimée sur l'ensemble de la population soumise à sélection. En revanche, celle de l'index est estimée pour chaque unité génétique puisque les matrices $[\ \sum \mathbf{GP}\ ]$ et $[\ \sum \mathbf{PP}\ ]$ dépendent de son effectif ainsi qu'éventuellement de ceux d'individus apparentés (familles de demi-frères de mère commune ou de père commun dans le cas des exemples développés dans la deuxième partie).

On est alors conduit à estimer la valeur de r(H,I) ou de CD(I) et les gains génétiques sur l'ensemble de la population à partir d'estimations partielles portant sur chaque unité génétique.
L'option prise dans le logiciel OPEP est de calculer une moyenne pondérée de ces estimations, le facteur de pondération étant l'effectif de chaque unité génétique. Par exemple, en posant : $E(\Delta G) = (1/N) \sum_i n_i E(\Delta G_i)$, on calcule l'espérance de progrès génétique sur un échantillon d'individus choisis au hasard dans la population, c'est-à-dire indépendamment des unités génétiques auxquelles ils appartiennent.
Cette estimation *a priori* est non biaisée seulement s'il n'y a pas de lien entre les effectifs et les performances des unités génétiques. Le problème ne se pose évidemment pas dans le cas d'une expérimentation orthogonale ou simplement équilibrée.

On pourrait envisager de réaliser une estimation bayésienne du gain génétique en utilisant l'information *a posteriori* sur la sous-population sélectionnée en terme d'effectifs par unité génétique.

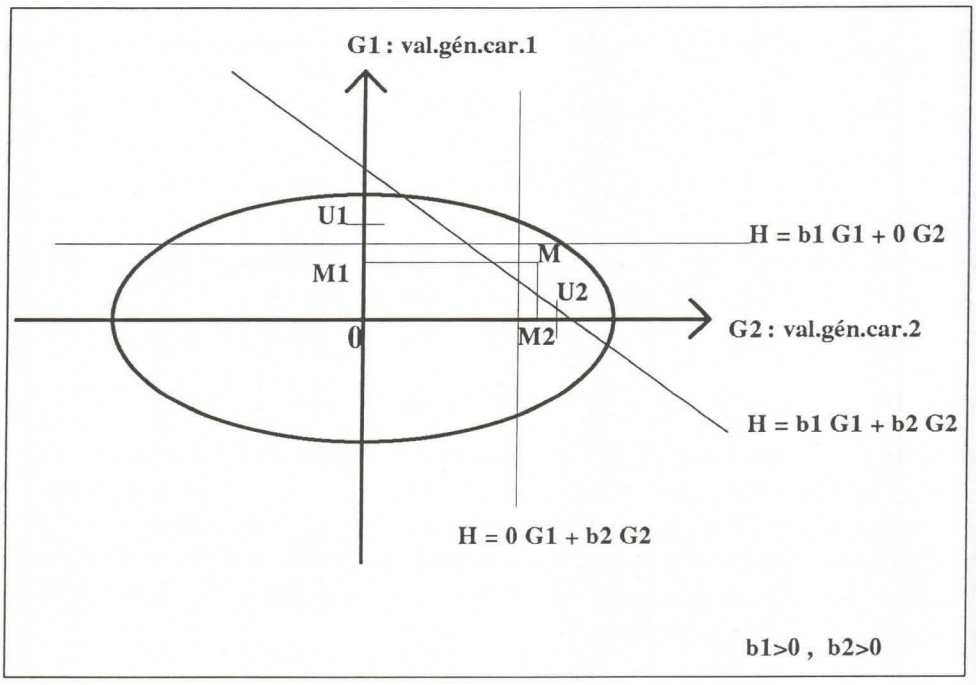


Figure 2. Gain génétique par caractère en fonction de son coefficient dans l'index.

Le gain génétique maximum pour chaque caractère est obtenu pour un coefficient de ce caractère dans l'index infini par rapport à celui des autres caractères. Ceci peut être obtenu en faisant, par exemple, b = 1 pour le caractère dont on veut maximiser le gain génétique (-1 si on fait une contre-sélection pour ce caractère) et 0 pour les autres. H est le « mérite » correspondant à l'index de sélection I (c'est l'index théorique que l'on calculerait si les valeurs génétiques étaient connues et non prédites). Cette figure montre, dans le cas de 2 caractères supposés génétiquement non corrélés, que l'index qui rend le gain génétique maximum pour un caractère, le taux de sélection étant fixé, est celui qui tronque la population perpendiculairement à l'axe des valeurs génétiques du caractère. Les valeurs génétiques moyennes correspondent aux points U1 et U2. Les points M1 et M2 sont les projections sur l'axe des valeurs génétiques des caractères des moyennes des valeurs génétiques de la sous population des individus de mérite H>b1G1+b2G2. Ils correspondent à des gains génétiques inférieurs à ceux qui sont donnés par U1 et U2. Cette propriété est valable également pour des caractères génétiquement corrélés (directement ou par l'intermédiaire de prédicteurs communs).

Pondération des caractères

La figure 2 montre sur un exemple simple (2 caractères cibles) comment la pondération affecte les gains génétiques partiels sur chaque caractère :

Sur cet exemple, le gain génétique pour un caractère est maximum lorsque son coefficient dans l'index vaut 1 et celui de l'autre caractère 0.
Les conclusions intuitives tirées de l'examen de ce graphique se généralisent au cas de 3 caractères. L'index est alors un plan, sous-espace de l'espace à 3 dimensions généré par les 3 axes des valeurs génétiques, qui recoupe l'ellipsoïde de leur distribution conjointe. Elles sont vraies aussi pour q>3 (l'index est un hyperplan). La propriété générale demeure que le gain génétique maximum sur un caractère est atteint lorsque l'index coupe perpendiculairement l'axe de ses valeurs génétiques. Cela se réalise pour un coefficient b non nul pour ce caractère et b=0 pour les autres.

Les choix de coefficients pour la réalisation d'une combinaison donnée de progrés génétiques par caractère sont obtenus par trois grands groupes de méthodes :

Maximisation d'une fonction économique

Cette façon de procéder suppose que l'on soit capable de traduire les gains génétiques par caractère en une unité monétaire commune. Chaque coefficient est alors proportionnel au profit supplémentaire généré par un accroissement d'une unité de la valeur génétique de chaque caractère (donc, de sa valeur phénotypique moyenne). Un tel mode de détermination des coefficients, qui sont alors des poids économiques, est à peu près le seul qui soit utilisé dans le monde anglo-saxon. Il est conforme à la première conception des index de sélection (HAZEL, 1943). TALBERT (1984) en fournit des exemples dans le domaine de l'amélioration des arbres forestiers.

Une telle approche possède à notre avis deux inconvénients majeurs :

– elle suppose que la fonction économique est linéaire par rapport aux gains génétiques, ce qui est rarement le cas. Pour ne prendre qu'un exemple, dans le cas de caractères de qualité, il suffira d'atteindre une certaine valeur-seuil pour que le prix de vente soit stable, au moins pour une certaine plage de variation. Cette loi « en marches d'escalier » se rencontre pour les caractères de branchaison des arbres forestiers ;
– elle ne tient pas compte de l'incertitude qui peut exister sur la fiabilité des prédicteurs des divers caractères cibles ni sur la robustesse des gains génétiques espérés vis-à-vis de certaines combinaisons de coefficients.

Recherche de progrès génétiques par caractère fixés *a priori*

Ce procédé prend le plus souvent la forme de contraintes exercées sur les gains génétiques de certains caractères : par exemple, pour un arbre forestier, choisir un jeu de coefficients tel que les gains sur la croissance et la forme soient associés à une valeur génétique constante pour la densité du bois (caractère dont la corrélation génétique avec la croissance est le plus souvent négative).

Les aspects théoriques de cette approche ont été traités par BRASCAMP (1984), CUNNINGHAM *et al.* (1970) et MALLARD (1972).

Choix de coefficients *a posteriori* au vu de l'évolution des progrés génétiques

Cette attitude empirique est largement pratiquée en France pour l'amélioration des arbres forestiers. Elle est fondée sur la connaissance préalable des progrès génétiques maximaux par caractère. La figure 2 montre que ces valeurs maximales sont faciles à déterminer. Le but de cette démarche est de trouver un jeu de coefficients réalisant le meilleur compromis entre l'écart de chaque gain génétique par rapport au maximum réalisable et la fiabilité du calcul d'index : recherche d'un r(H,I) élevé.

De fait, cette troisième méthode n'est pas antinomique des deux premières. En effet, connaissant la loi de variation des gains génétiques en fonction des coefficients b, il est possible de construire une fonction simple qui tienne compte de l'importance économique relative des différents caractères cibles. Par ailleurs, on peut intégrer des contraintes sur les valeurs génétiques de certains caractères. En l'absence d'idées précises sur leurs valeurs économiques

relatives, on peut appliquer d'autres critères. Par exemple, réaliser sur chacun la même proportion du gain génétique maximum. On peut aussi choisir les coefficients b de façon à obtenir le même gain génétique par caractère, mesuré en écarts-types phénotypiques : méthode citée par COTTERILL et JACKSON (1985) qui, du point de vue de ces auteurs, est une façon d'attribuer une importance économique égale aux différents caractères cibles.

C'est cette dernière méthode qui a été choisie pour le logiciel OPEP. Un exemple d'utilisation sera donné dans la deuxième partie dans le cas d'une sélection indirecte dans un schéma de sélection réciproque récurrente.

Evaluation de l'erreur standard sur les prédictions de gains génétiques

Comme cela a été précisé ci-dessus, la « variance d'erreur » de l'index, ne prend pas en compte l'erreur d'échantillonnage sur l'estimation des variances-covariances et des prédicteurs phénotypiques qui ont servi à les calculer.
Le calcul de la variance d'erreur des estimations de gains génétiques, dont une valeur, au moins approximative, est indispensable pour une prise de décision, doit donc prendre en compte tous les facteurs de variation qui interviennent dans le calcul des index. TAI (1979) propose une méthode approchée dans le cas de la sélection sur descendance mais seulement pour une analyse de variance à un facteur, des effectifs constants et avec un seul caractère. Dans le cas d'un mode de sélection plus complexe et avec plusieurs caractères, il n'existe pas de formules analytiques qui permettent d'estimer l'erreur standard sur l'espérance de gain génétique.
Il est alors nécessaire d'avoir recours à des techniques utilisant la simulation ou le rééchantillonnage qui ont un champ d'application très général et que les moyens de calcul actuels rendent facilement applicables.

Méthodes de simulation

Les méthodes de simulation dites de Monte-Carlo sont maintenant utilisées en routine, notamment en génétique animale. Elles sont toutefois assez laborieuses. Le principe est de générer une population modèle pour laquelle la distribution conjointe des caractères, les valeurs des effets fixés et des composantes de la variance et de la covariance soient les mêmes que dans la population étudiée.
On peut alors en extraire des sous-échantillons et réestimer les paramètres dont on veut déterminer la variance d'échantillonnage. La description de cette méthode est donnée par ROBERT (1992). Une variante de ces procédures, le « Gibbs sampler », est particulièrement bien adaptée à la génération de données multivariables.

Méthodes de rééchantillonnage

Contrairement aux précédentes, ces méthodes utilisent les données de l'expérimentation elle-même et non des données générées par simulation. Elles appartiennent à deux groupes :

■ La technique du « bootstrap »

Il s'agit d'un rééchantillonnage avec remise, qui inclut donc la possibilité d'avoir les mêmes données dans des sous-échantillons différents. Toutefois, cette méthode s'applique lorsque

l'autocorrélation entre les sous-échantillons est réduite et donc lorsque la proportion de données communes est faible. Elle est très utilisée en génétique des populations car les faibles effectifs de sous-groupes qu'elle entraîne implique une structuration simple et robuste (en général, il s'agit d'une population unique ou de hiérarchies à 1 ou 2 niveaux). Pour une description détaillée du « bootstrap », voir l'ouvrage d'EFRON (1982).

■ La méthode du « Jacknife »

Cette méthode, mise au point dans les années 50 par QUENOUILLE et TUKEY a été largement perfectionnée, diffusée et utilisée depuis. C'est la seule que nous développerons puisqu'elle s'applique bien à l'estimation des variances d'échantillonnage des espérances de gain génétique. En effet, elle procède par sous-échantillonnage exhaustif et opère sur des sous-échantillons dont l'effectif est proche de celui de l'échantillon total (les sous-échantillons sont donc très autocorrélés). Il en résulte qu'elle s'accomode de classifications complexes et « fragiles », comme par exemple les classifications croisées où des effectifs trop faibles poseraient des problèmes de cohérence des paramètres estimés (FOULLEY et OLLIVIER, 1986). La technique du « Jacknife » est exposée dans LEBART *et al.* (1979). Quelques compléments doivent lui être apportés pour qu'elle soit pleinement utilisable. C'est cette méthode complétée que nous exposons ci-dessous.

❑ Définition des sous-échantillons

Soit une expérimentation concernant N individus. On peut réaliser k sous-groupes chacun de taille $(k\text{-}1)u$, où u=int(N/k) : ce sont des échantillons tronqués de la $k$ième partie de l'échantillon total d'effectif ku (c'est-à-dire de u individus).
Ce sous- échantillonnage exhaustif est réalisé sur les ku premiers individus parmi les N en éliminant tour à tour les individus de rangs 1 à u, u+1 à $2u$ .....$(k\text{-}1)u$+1 à ku.

Un cas particulier, « all but one » chez les anglo-saxons, est celui où l'on n'élimine qu'un individu par sous-échantillon : k=N, u=1.

Les sous-échantillons doivent être représentatifs de l'ensemble de la population (c'est-à-dire, en général, de tous les niveaux de facteurs faisant l'objet de l'expérience). Cela peut être réalisé facilement en créant, à partir d'un ordre de succession initial quelconque des individus, quelques permutations aléatoires : il n'existe alors plus aucun lien entre le rang de chaque individu et les niveaux de facteurs auxquels il appartient. On peut donc, par la procédure d'échantillonnage décrite plus haut, obtenir k sous-échantillons non biaisés.

❑ Calcul de la variance d'erreur de paramètres estimés

Chaque individu est représenté par n observations : $y1$, $y2$ ...yn et l'on calcule sur la population un paramètre quelconque, $F(y1,y2,...yn)$.

Cette fonction des observations peut être recalculée sur chaque sous-échantillon, mais il est évident que, compte tenu de la forte autocorrélation positive entre les sous-échantillons, qui ont $(k\text{-}2)u$ individus en commun, une variance ordinaire des valeurs du paramètre donnerait une estimation très optimiste de la variance d'erreur. L'estimateur non biaisé de cette variance d'erreur (estimateur de Quenouille-Tukey) est donné par :

$$\hat{S}^2 = [\ 1/k(k-1)] \left[\Sigma_{i=1}^{k} F_i^2 - (\Sigma_{i=1}^{k} F_i)^2 / k \right]$$

où :

$F_i = k\hat{F} - (k-1)F_i^*$ (« pseudo-valeur » de TUKEY) ;

F_i^* est la valeur du paramètre calculée sur le sous-échantillon de rang i amputé des individus de rangs $u(i\text{-}1)+1$ à ui ;

$\hat{F}$ est la valeur du paramètre calculée sur l'échantillon total (ku individus).

❑ Conditions de validité de la méthode

Les « pseudo-valeurs » peuvent être considérées comme des variables approximativement indépendantes et la statistique $[\hat{F} - E(F)\ ]/\hat{S}$ suit une distribution très voisine du t de Student à k-1 degrés de liberté. Cette propriété est relativement robuste vis-à-vis de la distribution des estimations du paramètre pour laquelle il faut simplement éviter des discontinuités ou une forte asymétrie.

Approche BLP en sélection réciproque récurrente

La sélection réciproque récurrente permet de sélectionner des géniteurs appartenant à deux populations différentes, aux caractéristiques complémentaires ou dont l'intercroisement conduit à des effets d'hétérosis (GALLAIS, 1990).

Elle combine à chaque génération une phase de test des parents de chaque population pour leur aptitude générale à la combinaison avec l'autre population et une phase de recombinaison intra-population. La phase de test est reconduite à la génération suivante avec le matériel recombiné.

Cette procédure conduit à sélectionner d'abord pour l'aptitude générale à la combinaison et à tirer partie à chaque génération, au niveau des sorties variétales, de l'aptitude spécifique à la combinaison (sélection des meilleures combinaisons hybrides) et des effets de dominance et d'épistasie individuels. Ces derniers effets seront valorisés au mieux si la multiplication végétative peut être utilisée pour la production des variétés améliorées.

Chez l'Eucalyptus, l'exploitation de la valeur génétique totale dans les sorties variétales s'impose particulièrement puisque le matériel amélioré sera habituellement constitué de boutures.

Exemple de sélection réciproque récurrente : *E. grandis* et *E. urophylla*

Le schéma de sélection réciproque récurrente de l'Eucalyptus au Congo met en jeu trois espèces :

- *E. grandis*, originaire d'Australie ;
- *E. urophylla,* provenant des iles de la Sonde ;
- *E. pellita,* originaire d'Australie.

La première espèce est mal adaptée au climat tropical du Congo mais présente des caractéristiques de croissance et de forme très favorables. Les deux dernières ont une croissance moindre mais sont bien adaptées et présentent une bonne rectitude du fût.

Le programme utilise les combinaisons : *E. grandis-E. urophylla* et *E. urophylla-E. pellita.*

La figure 3 décrit les deux premières générations du programme de sélection réciproque récurrente entre *E. urophylla* et *E. grandis.*

Le tableau I représente le plan de croisements factoriel réalisé en 1990 entre les deux espèces.

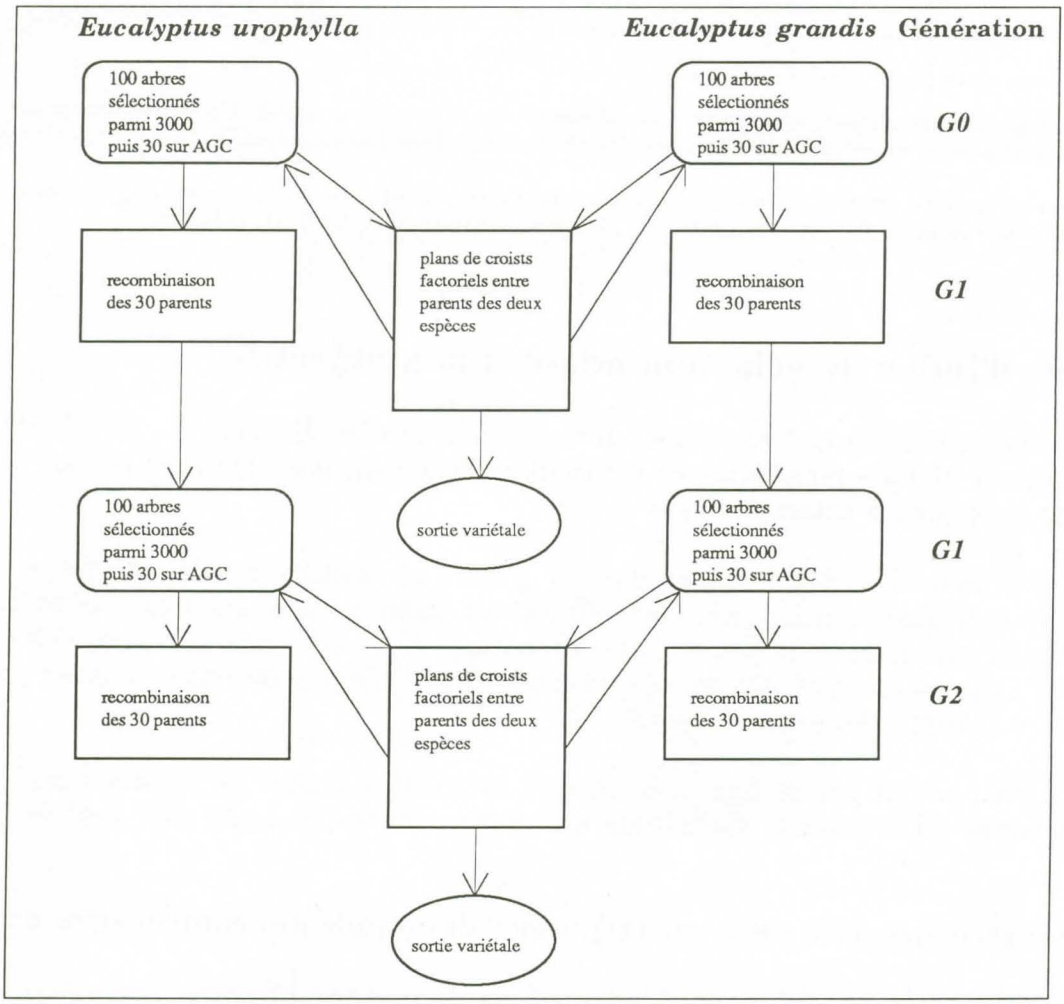


Figure 3. Schéma du programme d'amélioration de l'Eucalyptus au Congo par sélection réciproque récurrente des deux espèces *E. urophylla* et *E. grandis.*

Tableau I. Plan de croisements factoriel entre 15 mères *E. urophylla* et 9 pères *E. grandis.*

mère \ père	1	2	3	4	5	6	7	8	9
1		8 (27)		9 (35)					
2		11 (9)		12 (26)	13 (8)		14 (8)		15 (25)
3								16 (27)	
4				17 (8)		18 (7)	19 (6)	20 (29)	
5				21 (12)	22 (8)			23 (6)	
6			24 (17)		25 (26)		26 (5)	27 (20)	
7			28 (27)		29 (6)				
8	30 (35)	31 (32)		32 (36)	33 (26)	34 (35)	35 (33)	36 (33)	37 (17)
9	38 (35)	39 (26)	40 (9)				41 (15)		42 (30)
10		43 (32)						44 (9)	45 (18)
11				46 (9)	47 (18)			48 (9)	
12	49 (33)						50 (9)	51 (9)	
13	1 (21)								
14				2 (36)	3 (33)	4 (33)	5 (33)	6 (17)	
15				7 (34)					

Nombre de familles de pleins-frères : 50. Le dispositif de terrain comprend de 1 à 4 répétitions par famille sans regroupement en blocs utilisables pour des ajustements. Parcelles unitaires de 16 arbres à l'installation.

Le premier nombre figurant dans chaque case est le code de famille. Le deuxième, entre parenthèses, est le nombre d'arbres mesurés pour l'ensemble des caractères utilisés dans les calculs d'index.

Modèles d'index de sélection adaptés aux objectifs

Le programme esquissé ci-dessus fait intervenir à la fois la sélection sur test de descendance des parents de chaque population et la sélection combinée des individus ou des familles obtenus par hybridation interspécifique.

A notre connaissance, des modèles d'index généraux traitant ces cas de figure et tenant compte de toute l'information, avec des effectifs de familles de pleins-frères et de demi-frères très déséquilibrés, n'existent pas dans la littérature. Nous décrirons donc en détail toutes les étapes du raisonnement et des calculs en partant des notions de base utilisées pour établir les modèles statistiques et génétiques.

Comme dans tous les cas de figure, construire un modèle d'index de sélection revient à identifier les termes d'un modèle statistique aux termes correspondants d'un modèle génétique.

Le modèle génétique n'est autre que l'expression des covariances entre groupes d'apparentés.

Cette démarche permet d'obtenir les éléments de la matrice $\left[\Sigma GP\right]$. Par ailleurs, le calcul

des éléments de la matrice $[\sum]$ n'utilise qu'un modèle purement statistique qui ne prend en compte que les effectifs et les éléments des matrices de variances-covariances des effets aléatoires.

Nous exposerons donc brièvement les modèles statistiques et génétiques, renvoyant en annexe le détail des calculs.

Modèles statistiques

■ **Sélection sur test de descendance**

Suivant que l'on considère une sélection des mères ou des pères, le modèle s'écrit :

Index maternel : $y_{ijk}=\mu+m_i+\beta_j+(m\beta)_{ij}+e_{ijk}$

Index paternel : $y_{ijk}=\mu+\alpha_i+p_j+(\alpha p)_{ij}+e_{ijk}$

Où α_i et β_j sont respectivement les effets de la mère et du père considérés comme fixés

alors que m_i et p_j sont les mêmes effets considérés comme aléatoires.

Les variances correspondantes sont σ^2_m et σ^2_p (variances mère et père), $\sigma^2_{(\alpha p)}$ et $\sigma^2_{(m\beta)}$

(variances d'interaction) et σ^2_e (variance intra famille de pleins-frères). Par la suite, les suffixes « *e* » ou « *E* » désigneront les composantes intra famille et en aucun cas les composantes environnementales que les modèles utilisés n'ont pas besoin d'isoler.
Les prédicteurs phénotypiques sont :

$\hat{m}_i=\bar{y}^*_{i..}-\mu$ (effet de la mère i ajusté au père j) ;

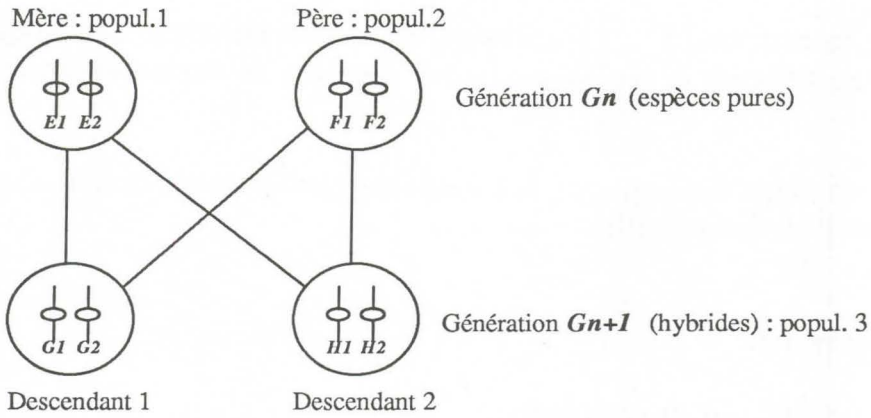
$\hat{p}_j=\bar{y}^*_{.j.}-\mu$ (effet du père j ajusté à la mère i).

L'annexe donne la démarche utilisée pour calculer les effets aléatoires ajustés, les covariances entre valeurs génétiques additives et prédicteurs phénotypiques et les variances-covariances entre prédicteurs.

■ **Sélection combinée**

Le modèle statistique est purement aléatoire et s'écrit : $y_{ijk}=\mu+m_i+p_j+(mp)_{ij}+e_{ijk}$

Les variances correspondantes sont : σ^2_m, σ^2_p, $\sigma^2_{(mp)}$ et σ^2_e.



SIm : simple identité de l'allèle maternel SIp : simple identité de l'allèle parternel
DI : double identité des allèles maternel et paternel

	Relation d'identité	Probabilité	Composante de la covariance
Mère-desc. :	$(G1 \equiv E1) \cup (G1 \equiv E2)$	1	$(1/2)\,\mathrm{cov}\,A_m = \mathrm{cov}\,A_1$
	ou		
	$(H1 \equiv E1) \cup (H1 \equiv E2)$		
Père-desc. :	$(G2 \equiv F1) \cup (G2 \equiv F2)$	1	$(1/2)\,\mathrm{cov}\,A_p = \mathrm{cov}\,A_2$
	ou		
	$(H2 \equiv F1) \cup (H2 \equiv F2)$		
Entre desc. :			
SIm	$(G1 \equiv H1) \mid (G2 \neq H2)$	$(1+F_1)/2$	$(1/2)\,\mathrm{cov}\,A_m = \mathrm{cov}\,A_1$
SIp	$(G2 \equiv H2) \mid (G1 \neq H1)$	$(1+F_2)/2$	$(1/2)\,\mathrm{cov}\,A_p = \mathrm{cov}\,A_2$
DI	$(G1 \equiv H1) \cap (G2 \equiv H2)$	$(1+F_1)(1+F_2)/4$	$\mathrm{cov}\,A_1 + \mathrm{cov}\,A_2 + \mathrm{cov}\,D$

Figure 4. Modèle génétique pour le calcul des covariances entre apparentés.

*G*1 et *H*1 sont les allèles reçus par un descendant *Gn+1* de la mère *Gn*
*G*2 et *H*2 sont les allèles homologues hérités du père *Gn*
Les 3 populations de référence sont :

- la population *Gn* des mères (ici, *E. urophylla*)
- la population *Gn* des pères (ici, *E. grandis*)
- la population *Gn+1* des hybrides

$\mathrm{cov}\,A_m, \mathrm{cov}\,A_p$ sont les composantes additives en *Gn* chez les mères et les pères

$\mathrm{cov}\,A_1, \mathrm{cov}\,A_2$ sont les valeurs de ces composantes en *Gn+1*, chez les hybrides

Les quatre prédicteurs phénotypiques (sources d'information) sont :

$\hat{m}_i = \bar{y}_{i..} - \mu$ (effet de la mère i) ;

$\hat{p}_j = \bar{y}_{.j.} - \mu$ (effet du père j) ;

$(\hat{mp})_{ij} = \bar{y}_{ij.} - \bar{y}_{.j.} - \bar{y}_{.j} + \mu$ (effet d'interaction entre mère i et père j) ;

$\hat{e}_{ijk} = y_{ijk} - \bar{y}_{ij.}$ (écart individuel intra-famille).

L'annexe précise les modes de calcul des covariances entre valeurs génétiques additives et de dominance et prédicteurs phénotypiques ainsi que des variances-covariances entre prédicteurs.

Nous considérerons par la suite la moyenne générale, μ , comme connue et ne la ferons donc pas intervenir dans les calculs de variances et de covariances. Cela est logique dans le cadre de la méthode BLP où les effets fixés sont supposés connus. Par ailleurs, conformément aux notations utilisées ci-dessus, nous désignerons respectivement par $(n_{i.}, \bar{y}_{i.})$, $(n_{.j}, \bar{y}_{.j})$ et $(n_{ij}, \bar{y}_{ij})$ les effectifs et moyennes par mère, par père et par famille de pleins-frères.

❑ Modèles génétiques

La figure 4 donne les éléments nécessaires pour le calcul des covariances entre apparentés dans le cas d'une sélection sur test de descendance ou d'une sélection combinée.

Dans les modèles usuels de génétique quantitative, la population de référence pour laquelle on estime les effets et les composantes de la variance et de la covariance est unique. C'est la population *Gn* (infinie) dont la génération *Gn+1* (de taille finie) est issue à partir d'un échantillon représentatif de parents.

En ce qui concerne le cas traité, il est évident que l'on est en présence d'une situation complètement différente. En effet, les valeurs génétiques additives concernent deux pools gamétiques disjoints et la possession d'un allèle commun par deux individus entraîne une covariance additive différente suivant la provenance de cet allèle :

(1/2) $\mathrm{cov}\, A_m$ si cet allèle est d'origine maternelle ;

(1/2) $\mathrm{cov}\, A_p$ si cet allèle est d'origine paternelle.

D'une façon générale, $\mathrm{cov}\, A_m \neq \mathrm{cov}\, A_p$.

Cette situation est traduite par les six premières lignes des relations d'identité de la figure 4. Il s'agit d'une simple identité (un allèle en commun et un seul pour un locus donné).

La simple identité est, en effet, le seul cas de figure possible entre parent et descendant ou entre demi-frères.

Les probabilités correspondantes se calculent aisément :

Pour la relation parent-descendant, la probabilité vaut 1 puisque, si l'on considère un couple d'allèles homologues d'un descendant *Gn+1*, l'un proviendra forcément de sa mère et l'autre de son père. Cette probabilité est donc indépendante du coefficient de consanguinité dans la population dont proviennent les parents.

Pour la relation entre demi-frères de mère commune (ligne 5), la simple identité résulte de deux couples d'évènements élémentaires qui s'excluent mutuellement :

– a) la mère commune n'est pas consanguine, avec la probabilité $1-F_1$ et elle transmet le même allèle aux deux demi-frères, avec la probabilité conditionnelle 1/2 ;

– b) la mère commune est consanguine, avec la probabilité F_1 et elle transmet le même allèle aux deux demi-frères avec la probabilité conditionnelle 1.

De façon plus formelle, on a donc :

$$\Pr(G1 \equiv H1) | (G2 \neq H2) = (1-F_1)/2 + F_1 = (1+F_1)/2$$

Dans cette formule, « $|(G2 \neq H2)$ » exprime la condition d'apparentement uniquement par la mère (les deux individus considérés doivent avoir des pères différents).

Pour des demi-frères de père commun, le raisonnement est parfaitement symétrique et l'on obtient :

$$\Pr(G2 \equiv H2) | (G1 \neq H1) = (1-F_2)/2 + F_2 = (1+F_2)/2$$

La double identité (ligne 7) ne peut évidemment être réalisée que chez deux pleins-frères. Sa probabilité est le produit des deux probabilités de simple identité maternelle et paternelle. Les composantes additives s'ajoutent et, par ailleurs, cette situation entraine une identité de l'effet d'interaction entre allèles homologues. On additionne donc la composante de dominance.

Il faut, à ce niveau, faire trois remarques :

– si les mères et les pères sont issus de la même population, on retrouve bien les relations classiques exprimées habituellement par rapport aux variances génétiques additives (*VA*) et de dominance (*VD*) : cov *HS*= (1+*F*)/4 *VA* et cov *FS*= (1+*F*)/2 *VA*+(1+*F*)²/4 *VD*.

– manifestement, la composante de dominance n'est définie ici que dans la génération *Gn+1* où elle met en jeu des interactions alléliques différentes de celles qui s'expriment dans chaque population de génération *Gn*.

– les composantes additives maternelle et paternelle sont également liées à l'intercroisement. Il s'agit donc de caractéristiques propres au couple de populations considérées.

On peut tirer deux conclusions de ces remarques :

a) La population de référence qui s'impose pour l'utilisation d'un modèle aléatoire est la po-

pulation des hybrides *Gn+1*. Mais on sait que les pools gamétiques maternel et paternel sont issus de deux populations différentes et il est nécessaire de les distinguer. On peut toujours imaginer que les familles dont on dispose sont un échantillon représentatif de toutes les combinaisons possibles entre parents de génération *Gn*.

b) Il faut considérer que cette population hybride a une variabilité phénotypique qui s'exprime, en ignorant les effets d'épistasie, par :

$$VP = VA+VD+VE$$

Mais, par ailleurs, la variance additive, *VA*, est décomposée en contributions maternelle et paternelle qui correspondent à la population d'origine du parent et l'on a :

$$VP = VA1+VA2+VD+VE$$

Cette relation implique que l'on divise par 2 les composantes additives maternelle et paternelle, VA_m et VA_p. En suivant les notations générales que nous avons adoptées on pose donc :

$$\mathrm{cov}\, A_1 = (1/2)\,\mathrm{cov}\, A_m$$

$$\mathrm{cov}\, A_2 = (1/2)\,\mathrm{cov}\, A_p$$

Par ailleurs, si l'on définit les deux coefficients des composantes additives dans les populations *Gn* dont proviennent les mères et les pères :

$$C_1=(1+F_1)/4$$

$$C_2=(1+F_2)/4$$

Il faut les multiplier par 2 pour obtenir les coefficients des composantes additives maternelle et paternelle dans la génération *Gn+1*. Dans cette génération, les covariances entre demi-frères de mère commune ou de père commun et entre pleins-frères ont donc les valeurs suivantes :

$$\mathrm{cov}\, HS_m = 2c_1\,\mathrm{cov}\, A_1$$

$$\mathrm{cov}\, HS_p = 2c_2\,\mathrm{cov}\, A_2$$

$$\mathrm{cov}\, FS = 2(c_1\,\mathrm{cov}\, A_1 + c_2\,\mathrm{cov}\, A_2) + 4c_1c_2\,\mathrm{cov}\, D$$

De ces définitions découlent les modèles génétiques des trois types d'index de sélection concernant la sélection sur descendance ou la sélection combinée. On aurait pu raisonner autrement et définir les composantes additives totales comme la moyenne des contributions maternelle et paternelle. C'est ce que fait implicitement le modèle classique où les pools gamétiques des deux parents proviennent de la même population de référence. Il aurait alors fallu multiplier par 2 les composantes additives et diviser par deux leurs coefficients ce qui donnerait évidemment les mêmes valeurs pour les covariances entre apparentés. Il est toutefois plus rigoureux d'un point de vue statistique de ne pas construire un modèle utilisant la

moyenne de composantes *a priori* hétérogènes. Le point de vue adopté pour la construction des index conduirait logiquement à diviser par 2 les héritabilités au sens strict maternelle et paternelle (si l'on calcule une héritabilité au sens strict dans le cas d'une seule population de référence, on la définit implicitement pour la somme des contributions gamétiques des deux parents). Toutefois, pour ne pas trop changer les usages, les valeurs des héritabilités maternelles et paternelles données plus loin sont obtenues en multipliant par 4 (F est supposé égal à 0) la variance mère ou la variance père. Dans ces conditions, l'héritabilité globale au sens strict est la moyenne des héritabilités maternelle et paternelle.

Une dernière remarque : les formules présentées pour le calcul des covariances entre apparentés supposent que les parents appariés ne soient pas eux-mêmes apparentés et ne produisent donc pas de descendants consanguins. En effet, dans ce cas, on ne peut calculer la probabilité de double identité comme le produit de deux probabilités de simple identité, les deux évènements correspondants n'étant pas indépendants. Pour le calcul des covariances entre apparentés consanguins, on se reportera à HARRIS (1964). Dans le cas de la sélection réciproque récurrente, il est bien évident que ce problème ne se pose pas au niveau des hybrides, puisque les populations intercroisées sont maintenues séparées d'un cycle de sélection à l'autre. Il peut en revanche se poser dans la phase de recombinaison intra- population. Pour la discussion de l'approche classique avec une seule population de référence en modèle aléatoire, on se référera à BECKER (1984), bien que cet auteur confonde l'hypothèse d'absence de consanguinité dans les descendances étudiées avec celle de la population de référence, ce qui n'est pas du tout la même chose. On trouvera chez ce deuxième auteur l'expression générale des covariances entre divers types d'apparentés en fonction des composantes additives, de dominance et d'épistasie, toujours dans l'hypothèse d'une population de référence unique. Les covariances calculées pour une sélection réciproque peuvent s'en déduire facilement et s'exprimer à partir des coefficients c_1 et c_2 définis ci-dessus. COCKERHAM (1980) a étudié le problème de l'estimations des composantes de la variance génétique dans des populations hybrides issues de plans de croisements factoriels. Il aboutit à la même conclusion que celle développée plus haut : la population de référence du modèle aléatoire est la population hybride, mais il est nécessaire de distinguer les apports gamétiques maternels et paternels pour le calcul des covariances entre apparentés.

Sélection de parents *Gn* sur test de descendance *Gn+1*

Ce mode de sélection est, comme le montre la figure 3, adapté à la gestion des populations d'amélioration des deux espèces d'Eucalyptus impliquées dans le programme de sélection réciproque récurrente. Il s'agit d'un modèle à une source d'information. Toutefois, sa mise en oeuvre nécessite la prise en compte d'un modèle mixte à 2 facteurs croisés.

■ Covariances entre valeurs génétiques et prédicteurs phénotypiques

Il s'agit là, bien sûr, uniquement des valeurs génétiques additives.

Selon que l'on prédit les valeurs génétiques additives des mères ou des pères, les matrices de covariances entre ces valeurs et les performances moyennes des descendants s'expriment de façon différente.

a) Index maternel :

$$\left[\sum \mathrm{GPm}\right] = \left[\sum \mathrm{APm}\right] = \left[\mathrm{cov}(A_i^l, \hat{m}_i^{l'})\right]$$

b) Index paternel :

$$\left[\sum \mathrm{GPp}\right] = \left[\sum \mathrm{APp}\right] = \left[\mathrm{cov}(A_j^l, \hat{p}_j^{l'})\right]$$

Les éléments de ces matrices de dimensions (q, q') sont donnés pour tout couple de caractères par $\mathrm{cov}\, A_1$ (index maternel) ou $\mathrm{cov}\, A_2$ (index paternel) : cf. figure 4 et annexe.

■ Variances et covariances entre prédicteurs phénotypiques

Les éléments de ces matrices de dimensions (q', q') sont déterminés en annexe. Il faut noter qu'ils intègrent les composantes d'interaction mère-père. En conséquence (ce qui est intuitif), l'accroissement de l'importance des composantes de dominance diminue l'efficacité de la sélection en augmentant la variance des prédicteurs phénotypiques sans influer sur leurs covariances avec les valeurs génétiques additives à prédire.
Pour alléger les formules, nous ne donnerons par la suite que la définition des éléments des diverses matrices des modèles, en supposant implicitement que celle-ci est valable pour tout couple de caractères l et l', en incluant le cas où $l = l'$.

a) Index maternel :

$$\left[\sum \mathrm{PPm}\right] = \left[\mathrm{cov}(\hat{m}_i^l, \hat{m}_i^{l'})\right] = \left[2c_1 \,\mathrm{cov}\, A_1 + 4c_1 c_2 \sum_j (n_{ij}^2 / n_{i.}^2)\,\mathrm{cov}\, D + (1/n_{i.})\,\mathrm{cov}\, E\right]$$

b) Index paternel :

$$\left[\sum \mathrm{PPp}\right] = \left[\mathrm{cov}(\hat{p}_j^l, \hat{p}_j^{l'})\right] = \left[2c_2 \,\mathrm{cov}\, A_2 + 4c_1 c_2 \sum_i (n_{ij}^2 / n_{.j}^2)\,\mathrm{cov}\, D + (1/n_{.j})\,\mathrm{cov}\, E\right]$$

Sélection des hybrides *Gn+1* aux niveaux individuel et familial

Ce type de sélection est avant tout adapté à la création variétale, sous forme de boutures de clones sélectionnés ou de production de graines correspondant aux combinaisons hybrides les plus performantes. Il concerne à la fois les valeurs génétiques additives et les valeurs génétiques de dominance. Il est cependant intéressant d'être en mesure de calculer les index en

intégrant ou non les composantes de dominance. En effet, cela peut permettre, par exemple, la création d'une variété en disjonction *Gn+2* issue de l'interpollinisation libre ou en « polycross » d'individus *Gn+1* sélectionnés pour leurs seules valeurs génétiques additives. Des considérations sur le coût de la création variétale, la rapidité de diffusion du progrès génétique ou le polymorphisme génétique pourraient conduire à une telle solution même en présence de composantes de dominance importantes.

■ Covariances entre valeurs génétiques et prédicteurs phénotypiques

Les valeurs de ces covariances sont données par une matrice de dimensions (q, 4q') ou (q, 3q') suivant que le modèle est adapté à la sélection des individus ou des familles de pleins-frères représentant les combinaisons hybrides les plus performantes.
Leur calcul est légèrement différent d'un cas à l'autre puisque ce n'est que dans la première situation qu'intervient la covariance génétique entre un individu et lui-même, covIND. Les formules correspondantes sont justifiées en annexe.

a) Index pour la sélection individuelle :

$$[\Sigma \mathbf{GP1}] = \left[\mathrm{cov}(G^{l}_{ijk}, \hat{e}^{l'}_{ijk}) \quad \mathrm{cov}(G^{l}_{ijk}, (\hat{mp})^{l'}_{ij}) \quad \mathrm{cov}(G^{l}_{ijk}, \hat{p}^{l'}_{j}) \quad \mathrm{cov}(G^{l}_{ijk}, \hat{m}^{l'}_{i})\right]$$

b) Index pour la sélection de familles de pleins-frères hybrides :

$$[\Sigma \mathbf{GP2}] = \left[\mathrm{cov}(G^{l}_{ij}, (\hat{mp})^{l'}_{ij}) \quad \mathrm{cov}(G^{l}_{ij}, \hat{p}^{l'}_{j}) \quad \mathrm{cov}(G^{l}_{ij}, \hat{m}^{l'}_{i}) \quad \right]$$

Nous donnons ci-dessous l'expression des covariances entre valeurs génétiques additives ou de dominance et prédicteurs phénotypiques.

❑ Valeurs génétiques additives

La démonstration des formules de ce paragraphe est donnée en annexe.

Index pour la sélection individuelle :

$$\mathrm{cov}(A_{ijk}, \hat{e}_{ijk}) = \left\{1 - \left[2c_1(n_{ij}-1)+1\right]/n_{ij}\right\}\mathrm{cov}\, A_1 + \left\{1 - \left[2c_2(n_{ij}-1)+1\right]/n_{ij}\right\}\mathrm{cov}\, A_2$$

$$\mathrm{cov}(A_{ijk}, \hat{m}_{i}) = \left\{\left[2c_1(n_{i.}-1)+1\right]/n_{i.}\right\}\mathrm{cov}\, A_1 + \left\{\left[2c_2(n_{ij}-1)+1\right]/n_{i.}\right\}\mathrm{cov}\, A_2$$

$$\mathrm{cov}(A_{ijk}, \hat{p}_{j}) = \left\{\left[2c_1(n_{ij}-1)+1\right]/n_{.j}\right\}\mathrm{cov}\, A_1 + \left\{\left[2c_2(n_{.j}-1)+1\right]/n_{.j}\right\}\mathrm{cov}\, A_2$$

$$\operatorname{cov}\left[A_{ijk},(\hat{mp})_{ij}\right]=\left\{\left[2c_1\ (n_{ij}-1)+1\right]/n_{ij}-[\ 2c_1(n_{i.}-1)+1\ ]/n_{i.}-\left[2c_1(n_{ij}-1)+1\right]/n_{.j}\right\}\operatorname{cov}A_1$$

$$+\left\{\left[2c_2(n_{ij}-1)+1\right]/n_{ij}-\left[2c_2(n_{.j}-1)+1\right]/n_{.j}-\left[2c_2(n_{ij}-1)+1\right]/n_{i.}\right\}\operatorname{cov}A_2$$

Index pour la sélection de familles de pleins-frères hybrides :

$$\operatorname{cov}(A_{ij},\hat{m}_i)=2c_1\operatorname{cov}A_1+2c_2(n_{ij}/n_{i.})\operatorname{cov}A_2$$

$$\operatorname{cov}(A_{ij},\hat{P}_j)=2c_1(n_{ij}/n_{.j})\operatorname{cov}A_1+2c_2\operatorname{cov}A_2$$

$$\operatorname{cov}\left[A_{ij},(\hat{mp}_{ij})\right]=-2n_{ij}[(c_1/n_{.j})\operatorname{cov}A_1+(c_2/n_{i.})\operatorname{cov}A_2\ ]$$

❑ Valeurs génétiques de dominance

La démonstration des formules de ce paragraphe est donnée en annexe

Index pour la sélection individuelle :

$$\operatorname{cov}(D_{ijk},\hat{e}_{ijk})=\left\{1-\left[4c_1c_2(n_{ij}-1)+1\right]/n_{ij}\right\}\operatorname{cov}D$$

$$\operatorname{cov}(D_{ijk},\hat{m}_i)=\left\{\left[4c_1c_2(n_{ij}-1)+1\right]/n_{i.}\right\}\operatorname{cov}D$$

$$\operatorname{cov}(D_{ijk},\hat{P}_j)=\left\{\left[4c_1c_2(n_{ij}-1)+1\right]/n_{.j}\right\}\operatorname{cov}D$$

$$\operatorname{cov}\left[D_{ijk},(\hat{mp})_{ij}\right]=\left[4c_1c_2(n_{ij}-1)+1\right](1/n_{ij}-1/n_{i.}-1/n_{.j})\operatorname{cov}D$$

Index pour la sélection de familles de pleins-frères hybrides :

$$\operatorname{cov}(D_{ij},\hat{m}_i)=4c_1c_2(n_{ij}/n_{i.})\operatorname{cov}D$$

$$\operatorname{cov}(D_{ij},\hat{P}_j)=4c_1c_2(n_{ij}/n_{.j})\operatorname{cov}D$$

$$\operatorname{cov}[\ D_{ij},(\hat{mp})_{ij}\ ]=4c_1c_2\left[1-n_{ij}(1/n_{i.}+1/n_{.j})\right]\operatorname{cov}D$$

■ Variances et covariances des prédicteurs phénotypiques

Les deux matrices suivantes concernent la sélection individuelle puis la sélection des combinaisons hybrides. Contrairement aux matrices du paragraphe précédant, les valeurs de leurs éléments ne dépendent pas du type de sélection car seul le modèle statistique est concerné. La justification des formules est présentée en annexe.

Ces matrices étant symétriques, nous ne donnons que leurs éléments situés dans leur partie triangulaire basse.

❑ Index pour la sélection individuelle

$$[\Sigma \mathbf{PP1}] = \begin{bmatrix} \operatorname{cov}\left(\hat{e}^{l}_{ijk}, \hat{e}^{l'}_{ijk}\right) & 0 & 0 & 0 \\ 0 & \operatorname{cov}\left[(\hat{mp})^{l}_{ij}, (\hat{mp})^{l'}_{ij}\right] & \operatorname{cov}\left[(\hat{mp})^{l}_{ij}, \hat{p}^{l'}_{j}\right] & \operatorname{cov}\left[(\hat{mp})^{l}_{ij}, \hat{m}^{l'}_{i}\right] \\ 0 & \operatorname{cov}\left[\hat{p}^{l}_{j}, (\hat{mp})^{l'}_{ij}\right] & \operatorname{cov}(\hat{p}^{l}_{j}, \hat{p}^{l'}_{j}) & \operatorname{cov}(\hat{p}^{l}_{j}, \hat{m}^{l'}_{i}) \\ 0 & \operatorname{cov}\left[\hat{m}^{l}_{i}, (\hat{mp})^{l'}_{ij}\right] & \operatorname{cov}(\hat{m}^{l}_{i}, \hat{p}^{l'}_{j}) & \operatorname{cov}(\hat{m}^{l}_{i}, \hat{m}^{l'}_{i}) \end{bmatrix}$$

❑ Index pour la sélection de familles de pleins-frères hybrides

$$[\Sigma \mathbf{PP2}] = \begin{bmatrix} \operatorname{cov}\left[(\hat{mp})^{l}_{ij}, (\hat{mp})^{l'}_{ij}\right] & \operatorname{cov}\left[(\hat{mp})^{l}_{ij}, \hat{p}^{l'}_{j}\right] & \operatorname{cov}\left[(\hat{mp})^{l}_{ij}, \hat{m}^{l'}_{i}\right] \\ \operatorname{cov}\left[\hat{p}^{l}_{j}, (\hat{mp})^{l'}_{ij}\right] & \operatorname{cov}(\hat{p}^{l}_{j}, \hat{p}^{l'}_{j}) & \operatorname{cov}(\hat{p}^{l}_{j}, \hat{m}^{l'}_{i}) \\ \operatorname{cov}\left[\hat{m}^{l}_{i}, (\hat{mp})^{l'}_{ij}\right] & \operatorname{cov}(\hat{m}^{l}_{i}, \hat{p}^{l'}_{j}) & \operatorname{cov}(\hat{m}^{l}_{i}, \hat{m}^{l'}_{i}) \end{bmatrix}$$

- Calcul de $\operatorname{cov}(\hat{m}^{l}_{i}, \hat{m}^{l'}_{i})$

$$2\left[c_1 \operatorname{cov} A_1 + c_2 \sum_j (n^2_{ij}/n^2_{i.}) \operatorname{cov} A_2\right] + 4 c_1 c_2 \sum_j (n^2_{ij}/n^2_{i.}) \operatorname{cov} D + (1/n_{i.}) \operatorname{cov} E$$

- Calcul de $\mathrm{cov}(\hat{p}^{l}_{j}, \hat{p}^{l'}_{j})$

$$2\left[c_1 \sum_i (n_{ij}^2/n_{.j}^2)\,\mathrm{cov}\,A1 + c_2\,\mathrm{cov}\,A2\right] + 4\,c_1\,c_2 \sum_i (n_{ij}^2/n_{.j}^2)\,\mathrm{cov}\,D + (1/n_{.j})\,\mathrm{cov}\,E$$

- Calcul de $\mathrm{cov}\left[(\hat{mp})^{l}_{ij}, (\hat{mp})^{l'}_{ij}\right]$

$$2c_1\left[\sum_i (n_{ij}^2/n_{.j}^2) + 2(1 - n_{ij}/n_{i.} - n_{ij}/n_{.j} + n_{ij}^2/(n_{i.}n_{.j}))\right]\mathrm{cov}\,A1$$

$$+2c_2\left[\sum_j (n_{ij}^2/n_{i.}^2) + 2(1 - n_{ij}/n_{i.} - n_{ij}/n_{.j} + n_{ij}^2/(n_{i.}n_{.j}))\right]\mathrm{cov}\,A2$$

$$+4c_1c_2\left[\sum_j n_{ij}^2/n_{i.}^2 + \sum_i n_{ij}^2/n_{.j}^2 - 2n_{ij}/n_{i.} - 2n_{ij}/n_{.j} + 2n_{ij}^2/(n_{i.}n_{.j}) + 1\right]\mathrm{cov}\,D$$

$$+\left[1/n_{ij} - 1/n_{i.} - 1/n_{.j} + 2n_{ij}/(n_{i.}n_{.j})\right]\mathrm{cov}\,E$$

- Calcul de $\mathrm{cov}\,(\hat{e}^{l}_{ijk}, \hat{e}^{l'}_{ijk})$

$$\left[(n_{ij} - 1)/n_{ij}\right]\mathrm{cov}\,E$$

- Calcul de $\mathrm{cov}\left[\hat{p}^{l}_{j}, (\hat{mp})^{l'}_{ij}\right]$

$$2c_1\left[n_{ij}/n_{.j} - n_{ij}^2/(n_{i.}n_{.j}) - \sum_i (n_{ij}^2/n_{.j}^2)\right]\mathrm{cov}\,A1 + 2c_2\left[n_{ij}/n_{i.} - n_{ij}^2/(n_{i.}n_{.j}) - 1\right]\mathrm{cov}\,A2$$

$$+4c_1c_2\left[n_{ij}/n_{.j} - n_{ij}^2/(n_{i.}n_{.j}) - \sum_i n_{ij}^2/n_{.j}^2\right]\mathrm{cov}\,D - n_{ij}/(n_{i.}n_{.j})\,\mathrm{cov}\,E$$

- Calcul de $\mathrm{cov}(\hat{m}^{l}_{i'} \hat{p}^{l'}_{j})$

$$2\left[n_{ij}^2/(n_{i.}n_{.j})\right]\left[c_1 \mathrm{cov}\, A_1 + c_2 \mathrm{cov}\, A_2 + 2c_1c_2 \mathrm{cov}\, D\right] + \left[n_{ij}/(n_{i.}n_{.j})\right]\mathrm{cov}\, E$$

- Calcul de $\mathrm{cov}\left[\hat{m}^{l}_{i'}(\hat{mp})^{l'}_{ij}\right]$

$$2c_1\left[n_{ij}/n_{.j} - n_{ij}^2/(n_{i.}n_{.j}) - 1\right]\mathrm{cov}\, A_1 + 2c_2\left[n_{ij}/n_{i.} - n_{ij}^2/(n_{i.}n_{.j}) - \sum_j n_{ij}^2/n_{i.}^2\right]\mathrm{cov}\, A_2$$

$$+4c_1c_2\left[n_{ij}/n_{i.} - n_{ij}^2/(n_{i.}n_{.j}) - \sum_j n_{ij}^2/n_{i.}^2\right]\mathrm{cov}\, D$$

Mise en oeuvre informatique des modèles d'index en sélection réciproque

Les modèles d'index décrits ci-dessus sont intégrés, avec tout un ensemble d'autres modèles adaptés à la sélection des plantes pérennes, dans le logiciel OPEP, en cours de développement à l'INRA et au CIRAD. Les programmes correspondants sont associés avec d'autres qui permettent notamment la gestion des fichiers, différents types d'ajustement aux effets du milieu, l'intégration de critères de stabilité (analyse de l'interaction génotype-environnement) ou la vérification de la normalité de distribution des caractères élémentaires et des index. Nous décrirons dans un premier temps l'architecture des modules qui interviennent directement dans les calculs d'index. Nous y situerons les chaines de traitement réalisables à partir des modèles adaptés à la sélection réciproque récurrente. Nous indiquerons également les contraintes à respecter pour utiliser un modèle donné (en particulier, identification du jeu de caractères prédicteurs et des caractères cibles et correction des matrices de base).
Dans un deuxième temps, nous nous intéresserons au traitement des « effets de bord » (cas extrêmes de déséquilibre et de non-orthogonalité) qui est nécessaire pour que ces programmes aient un domaine d'utilisation tout à fait général. Ces effets de bord ne concernent que les modèles d'index en sélection combinée qui sont les plus complexes et les plus « fragiles ».

La figure 5 situe les principaux choix possibles dans l'organigramme des modules OPEP impliqués dans les calculs d'index. Pour des raisons de lisibilité, ces choix ne sont détaillés que pour la sélection d'individus dans les familles de pleins-frères hybrides les plus performantes.

Les étapes concernant l'ajustement aux effets du milieu et la prédiction des valeurs génétiques sont complètement dissociées (caractéristique de la méthode BLP). Par ailleurs, le choix de la méthode de détermination des coefficients des caractères cibles par calcul des gains génétiques partiels implique que l'on puisse réaliser de façon interactive un nombre élevé,

non connu *a priori*, de combinaisons de ces coefficients. Ceci implique de dissocier également les calculs concernant l'évaluation des gains génétiques et des r(H,I) du calcul et de l'édition des index. Les conditions de normalité des index et de diversité génétique de la sous-population sélectionnée ne peuvent être vérifiées qu'après cette dernière étape. Si la diversité génétique après sélection est trop faible, on pourra l'augmenter par des contraintes sur le nombre minimum de familles à sélectionner, le nombre de parents représentés et le nombre maximum d'individus retenus par famille. La façon de corriger les prédictions de gains génétiques par caractère pour tenir compte de ces restrictions est indiquée à la fin de cet article.

La figure 6 montre comment sont pris en compte les caractères cibles et les caractères prédicteurs au niveau de la structure des matrices de base (solution valable quel que soit le modèle d'index).
La figure 7 indique comment sont traités les quatre cas de déséquilibre ou de non-orthogonalité extrêmes qui peuvent être rencontrés :

- famille de pleins-frères représentée par un seul individu ;
- mère représentée par une seule famille de pleins-frères ;
- père représenté par une seule famille de pleins-frères ;
- famille de pleins-frères déconnectée des autres familles.

Identification du type de caractère, corrections et structure des sous-matrices

Les caractères cibles et les caractères prédicteurs sont identifiés par deux paramètres :

p = nombre de caractères étudiés (défini dans le programme d'analyse de variance) ;
q' = nombre de caractères prédicteurs.

Le nombre de caractères cibles, q, est défini par :

q = p-q' si p>q' (traitement dissymétrique où l'on distingue les caractères prédicteurs) ;
q = q' = p si p = q' (traitement symétrique où l'on ne distingue pas les prédicteurs).

Dans la liste des p caractères étudiés, les q' caractères prédicteurs doivent être définis avant les q caractères cibles au niveau du premier programme d'analyse de variance (ajustement aux effets fixés ou estimation des variances-covariances). Si le traitement est dissymétrique, les caractères qui sont à la fois prédicteurs et cibles doivent être définis deux fois pour figurer danschaque groupe. Par contre, si le traitement est symétrique, chaque caractère ne nécessite d'être défini qu'une seule fois (ce qui est permis par une valeur de q conditionnelle à p-q').

Les calculs d'index sont précédés par une vérification de validité et une correction éventuelle d'éléments des matrices de variances-covariances (p, p) issues du programme d'analyse de variance multivariable non-orthogonale (méthode Henderson 3). La valeur absolue maximum admise pour une corrélation est introduite comme paramètre (*Mco*). Le test est alors pour un couple de caractères l et l' :

$$\left|\mathrm{cov}^{(l,l')}\right| \geq Mco\sqrt{\sigma^{2(l)}\sigma^{2(l')}}$$

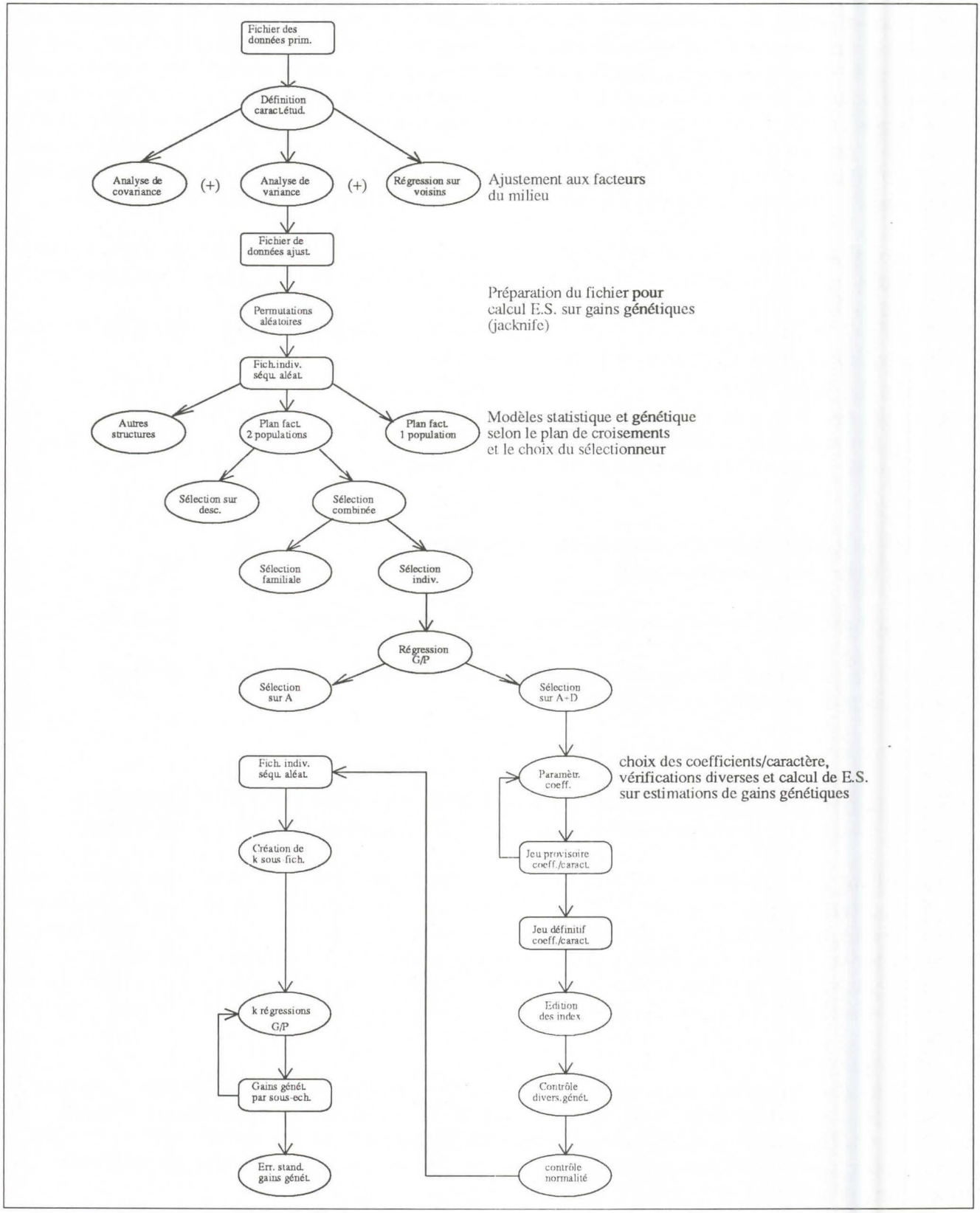


Figure 5. Chaine de traitement permettant le calcul d'index en plan factoriel.

Pour la clarté du schéma ne sont représentées que les principales options en ne détaillant que le calcul d'index de sélection individuels. Il est par exemple possible d'estimer les matrices de variances-covariances sur une série de plans de croisements déconnectés (estimations de variance d'échantillonnage minimum) et de caculer les index à partir des effets et des effectifs correspondant à chaque plan de croisements. Les dispositifs peuvent par ailleurs être multistationnels et l'on pourra alors prendre en compte l'effet station et l'interaction famille-station.

Si la condition n'est pas remplie, le programme remplace alors, selon son signe, chaque covariance par :

$$Mco\sqrt{\sigma^{2(l)}\sigma^{2(l')}} \quad \text{ou} \quad -Mco\sqrt{\sigma^{2(l)}\sigma^{2(l')}}$$

Sous-matrice globale

[PP] (q', q')	[GP]' (q', q)
[GP] (q, q')	[GG] (q, q)

Sous-matrice [PP]

[PP] (q', q')				0	0	0
				0	0	0
				0	0	0
				0	0	0
0	0	0	0	1	0	0
0	0	0	0	1	0	0
0	0	0	0	0	0	1

Sous-matrice [GP]

[GP] (q, q')				0	0	0
				0	0	0
				0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

Sous-matrice [GG]

[GG] (q, q)			0	0	0	0	0
			0	0	0	0	0
			0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Figure 6. Structure des sous-matrices élémentaires.

La structure des sous-matrices assure la compatibilité des programmes qui calculent la régression des valeurs génétiques sur les prédicteurs phénotypiques, les gains génétiques et les corrélations r(H,I) entre eux et avec les programmes éditeurs d'index. Ces derniers opèrent avec des matrices dont les dimensions sont indépendantes de la distinction entre caractères cibles et caractères prédicteurs : toutes les sous-matrices sont carrées et de dimensions (p, p) . La partie « utile » de chaque sous-matrice occupe les premières lignes et les premières colonnes. La partie restante est mise à 0, sauf pour la sous-matrice [PP] qui doit être inversible et dont les éléments diagonaux sont mis à 1, générant une matrice identité. On vérifie facilement que les produits [ΣGP][ΣPP]-1 donnent les mêmes résultats que s'ils étaient réalisés avec les seules parties utiles de ces sous-matrices.

La valeur prise pour *Mco* est de 1 pour les sous-matrices [GP] et [GG] ; elle n'est paramétrable que pour la sous-matrice [PP], la seule pour laquelle se pose un problème d'inversibilité. Il est souhaitable que, dans ce cas, elle soit aussi proche de 1 que possible. Ce principe de correction ne garantit pas que les matrices $\left[\sum \text{PP}\right]$ seront définies positives. La condition nécessaire et suffisante est, en effet, que toutes les corrélations partielles calculées à partir des variances-covariances estimées soient dans l'espace de définition des paramètres [-1, +1] (FOULLEY et OLLIVIER, 1986). Toutefois, la méthode permet de réduire les cas de non-inversibilité et ne corrige (de façon minimale) que des estimations de covariances incohérentes. On aurait pu songer à corriger à l'inverse les estimations des variances, en les augmentant pour les rendre cohérentes avec les estimations de covariances. Toutefois, cette méthode conduit à des modifications plus importantes des matrices de variances-covariances, puisque

Modèle complet

	$\hat{e}_{ijk}$	$(\hat{mp}_{ij})$	$\hat{p}_j$	$\hat{m}_i$
$\hat{e}_{ijk}$	P4	0	0	0
$(\hat{mp}_{ij})$	0	P3	P5	P7
$\hat{p}_j$	0	P5	P2	P6
$\hat{m}_i$	0	P7	P6	P1

Modèle avec 1 individu/famille

	$\hat{e}_{ijk}$	$(\hat{mp}_{ij})$	$\hat{p}_j$	$\hat{m}_i$
$\hat{e}_{ijk}$	0	0	0	0
$(\hat{mp}_{ij})$	0	P3	P5	P7
$\hat{p}_j$	0	P5	P2	P6
$\hat{m}_i$	0	P7	P6	P1

Mère ou père représentés par 1 famille

	$\hat{e}_{ijk}$	$(\hat{mp}_{ij})$	$\hat{p}_j$	$\hat{m}_i$
$\hat{e}_{ijk}$	P4	0	0	0
$(\hat{mp}_{ij})$	0	0	0	0
$\hat{p}_j$	0	0	P2	P6
$\hat{m}_i$	0	0	P6	P1

Famille de pleins-frères déconnectée

	$\hat{e}_{ijk}$	$(\hat{mp}_{ij})$	$\hat{p}_j$	$\hat{m}_i$
$\hat{e}_{ijk}$	P4	0	0	0
$(\hat{mp}_{ij})$	0	0	0	0
$\hat{p}_j$	0	0	0	0
$\hat{m}_i$	0	0	0	P1

Figure 7. Traitement des effets de bord.

Dans le cas du modèle complet (4 sources d'information), on inverse la sous-matrice P4 et le bloc (P1, P2, P3, P5, P6, P7) puisque les 4 effets du modèle sont définis (variances-covariances d'une catégorie d'effets et covariances entre effets).
Si une famille de pleins-frères n'est représentée que par 1 individu, l'effet individuel intra-famille, $\hat{e}_{ijk}$, n'est pas défini. La sous-matrice P4 est donc mise à 0.
Si une mère n'est représentée que par 1 famille connectée à au moins 2 familles de même père, on a 3 sources d'information: $\hat{e}_{ijk}$, $\hat{p}_j$ et $\bar{y}_{ij.} - \mu$. Dans le cas symétrique où un père n'est représenté que par une famille, on dispose également de 3 sources d'information: $\hat{e}_{ijk}$, $\hat{m}_i$ et $\bar{y}_{ij.} - \mu$.
Dans l'un et l'autre cas, les éléments de P1, P2 et P6 sont définis. On inverse donc P4 et le bloc (P1, P2, P6)

Pour une famille de pleins-frères déconnectée, on ne dispose que de deux sources d'information : $\hat{e}_{ijk}$ et $\bar{y}_{ij.} - \mu$. On peut arbitrairement affecter les variances-covariances correspondantes à P1 ou à P2. C'est la première solution qui a été choisie. On inverse donc P4 et P1.

Le programme traite évidemment les cas où une famille de pleins-frères n'a qu'un seul individu et où elle relève de l'un de ces 3 cas de non-orthogonalité.
Les éléments de [ΣGP] sont correctement calculés dans tous les cas particuliers. On n'a donc pas besoin de traiter des exceptions à leur niveau.

la correction d'un élément diagonal modifie p-1 corrélations. Par ailleurs, il est difficile dans ce cas de trouver un critère objectif de correction. En effet, si une corrélation est hors limites, on ne sait quelle variance corriger ou s'il faut corriger simultanément les deux estimations. D'où une infinité de solutions.

Traitement des effets de bord, optimisation de la vitesse de calcul et résultats fournis

Le programme traite les quatre situations de déséquilibre/non-orthogonalité comme des sous-modèles correspondant à l'absence ou à la modification de la nature de sources d'information. Cela implique une gestion adéquate de la matrice $\left[\sum PP\right]$ qui doit être inversée par blocs correspondant à une sous-matrice ou des sous-matrices adjacentes après que les situations aient été identifiées par comparaison des effectifs par famille de pleins-frères et par familles de demi-frères. Ces mécanismes sont détaillés dans la figure 7. Le fonctionnement du programme éditeur des index n'est pas détaillé. Il faut simplement signaler que, dans le cas de la sélection combinée, les individus sont repérés dans le dispositif, au choix, par coordonnées cartésiennes ou à l'intérieur de parcelles unitaires (caractérisées par une combinaison famille-bloc). Par ailleurs, pour économiser le temps de calcul et s'il s'agit d'une sélection d'individus, chaque index est scindé en deux parties :

– partie « familiale », qui ne dépend que de la famille de pleins-frères considérée et des deux familles de demi-frères qui s'y rattachent éventuellement ;

– partie « individuelle » qui utilise comme source d'information l'écart de chaque individu à la moyenne de famille de pleins-frères.

Le premier terme est calculé une fois pour toutes et stocké à une adresse correspondant au code de la famille. On l'additionne au terme individuel qui doit seul être actualisé pour une famille de pleins-frères donnée.

Ce principe permet de calculer plusieurs dizaines de milliers d'index en quelques secondes avec un serveur ou une station SUN sous UNIX.

Enfin, on peut n'imprimer que les valeurs d'index correspondant à un taux de sélection maximum. Le programme indique, pour chaque index imprimé, le taux de sélection qui lui correspond. L'édition se termine par la liste des index moyens par famille de pleins-frères et par parent ainsi que par le nombre d'individus retenus à chacun de ces niveaux pour 5 taux de sélection de référence (50%, 20%, 10%, 5% et 1%). Ce sont ces mêmes taux de sélection qui sont utilisés pour caractériser chaque index imprimé.

Tableau II. Héritabilités et corrélations génétiques des caractères utilisés pour la sélection.

Héritabilités

	h^2ss.patern.	h^2ss.matern.	h^2ss.globale	h^2sl.globale	% var.dom.
haut.18 mois	0,558	0,030	0,294	0,589	50,12
circ.18 mois	0,323	0,010	0,167	0,338	50,58
pilod.18 mois	0,326	0,137	0,231	0,356	34,97
vol.48 mois	0,411	0,124	0,268	0,308	13,04
pilod.48 mois	0,096	0,194	0,145	0,220	34,17

• Définition des héritabilités au sens strict :

$$h^2\text{ss. maternelle} = 4\sigma^2_m / [\ \sigma^2_m + \sigma^2_p + \sigma^2_{(mp)} + \sigma^2_e\ ]$$

$$h^2\text{ss. paternelle} = 4\sigma^2_p / [\ \sigma^2_m + \sigma^2_p + \sigma^2_{(mp)} + \sigma^2_e\ ]$$

$$h^2\text{ss. globale} = 2[\sigma^2_m + \sigma^2_p\ ] / [\ \sigma^2_m + \sigma^2_p + \sigma^2_{(mp)} + \sigma^2_e\ ]$$

• Définition de l'héritabilité au sens large :

$$h^2\text{ sl. globale} = [2(\sigma^2_m + \sigma^2_p) + 4\sigma^2_{(mp)}\ ] / [\ \sigma^2_m + \sigma^2_p + \sigma^2_{(mp)} + \sigma^2_e\ ]$$

Corrélations génétiques au sens large

	haut.18 mois	circ.18 mois	pilod.18 mois	vol.48 mois
circ.18 mois	0,952			
pilod.18 mois	0,105	0,242		
vol.48 mois	0,825	0,809	-0,073	
pilod.48 mois	-0,094	0,033	0,698	-0,080

• Définition du coefficient de corrélation génétique au sens large :

$$[\ \mathrm{cov}^{(l,l')}_m + \mathrm{cov}^{(l,l')}_p + 2\,\mathrm{cov}^{(l,l')}_{(mp)}\ ] / [\ \sqrt{\sigma^{2(l)}_m + \sigma^{2(l)}_p + 2\sigma^{2(l)}_{(mp)}}\sqrt{\sigma^{2(l')}_m + \sigma^{2(l')}_p + 2\sigma^{2(l')}_{(mp)}}\ ]$$

Noter les valeurs très différentes des héritabilités au sens strict paternelles et maternelles qui soulignent la nécessité de séparer les composantes génétiques additives des deux parents. Par ailleurs, les corrélations génétiques sont très fortes entre volume à 48 mois et circonférence ou hauteur à 18 mois (prédicteurs de ce caractère cible). Cette corrélation est forte entre les mesures au pilodyn (fonction décroissante de la densité) à 18 et à 48 mois. En revanche, les corrélations entre les deux groupes de caractères (croissance et mesure indirecte de la densité du bois) sont faibles. On est donc dans une situation *a priori* favorable pour concilier des gains génétiques à la fois sur le volume et sur la densité.

Exemple de paramétrage des coefficients des caractères cibles

Cet exemple illustre la situation où l'on veut prédire des caractères « adultes » à partir de prédicteurs relativement juvéniles. Le tableau II présente les principaux paramètres génétiques. Les caractères cibles sont ici au nombre de deux : le volume à 48 mois et une mesure indirecte de la densité du bois à 48 mois réalisée au pilodyn (l'enfoncement de l'aiguille calibrée dans le bois est corrélé négativement à la densité et l'on doit donc exercer une contre-sélection sur ce caractère). Il y a trois caractères prédicteurs : la hauteur, la circonférence et la mesure au pilodyn à 18 mois. La matrice des corrélations génétiques montre clairement que les prédicteurs utiles sont, pour le volume à 48 mois, la hauteur et la circonférence à 18 mois, pour la mesure au pilodyn à 48 mois, la même mesure à 18 mois. On constate une forte proportion de variance de dominance dans la variance génétique ; mais cette proportion tend à diminuer avec l'âge. Les autres commentaires sont reportés au bas du tableau II.

La figure 8 présente l'évolution des espérances de gains génétiques relatifs sur les deux caractères cibles lorsque le coefficient du volume à 48 mois dans l'index est constant et égal à 1 et pour une variation de -0,3 à +0,3 du coefficient de la mesure au pilodyn.

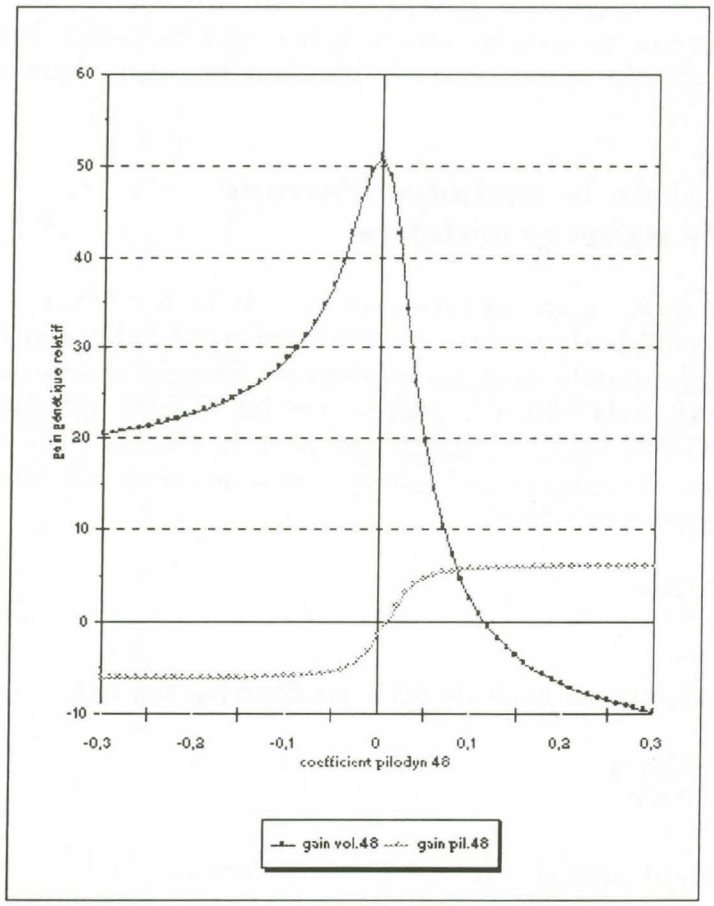


Figure 8. Courbes de paramétrage des coefficients des caractères cibles dans l'index.

Le coefficient du volume, b1, est constant (b1=1) et le coefficient du pilodyn, b2, varie de -0.3 à + -0.3. Noter la très forte variation induite sur le gain génétique relatif pour le volume par une faible variation du coefficient du pilodyn autour de la valeur b2 = 0. Par ailleurs, la courbe des gains génétiques sur le pilodyn donne une valeur légèrement négative pour b2 = 0. Ceci traduit la légère corrélation génétique négative entre volume et pilodyn à 48 mois (-0,08).

On constate que le gain maximum sur le volume est bien atteint pour un coefficient de 0 de la mesure au pilodyn (51,13%). Par ailleurs, un gain génétique maximum sur la densité du bois, de -6,11% sur la mesure au pilodyn, se traduirait par une perte inacceptable de gain génétique sur le volume (qui passe de plus de 50% à environ 20%). En revanche, un coefficient de -0,02 sur la mesure au pilodyn donne une espérance de gain de -4,12% sur le pilodyn (environ 67% du maximum) et de 43,34% pour le volume (environ 85% du maximum). Un tel coefficient est donc « raisonnable ». Pour une détermination plus fine du coefficient du pilodyn, il faudrait connaître la loi de variation qui relie cette mesure à la densité du bois. S'il s'agit d'une sélection pour le « bois d'industrie », la fonction économique à optimiser serait alors le volume de matière sèche, donc le produit du volume par la densité. Si l'on n'introduisait que ce critère dans la décision, sans y ajouter par exemple la valeur de r(H,I) ou la diversité génétique de la sous-population sélectionnée, cet objectif serait atteint par une transformation log sur les deux caractères, pour linéariser la fonction économique. Il suffirait ensuite de donner le même poids (b>0) au volume et à la densité.

Ce qu'il faut retenir d'essentiel dans cette figure est la forte non-linéarité de la variation des espérances de progrès génétiques par caractère en fonction des valeurs relatives de leurs coefficients dans l'index. La connaissance de ces lois de variation est indispensable pour le sélectionneur. En l'absence de cette information, il risque de perdre beaucoup en valeur économique globale s'il cherche à maximiser le gain sur un ou quelques caractères.

Exemple de calcul de la variance d'erreur de prédictions de gains génétiques

La technique du « Jacknife » a été utilisée avec N = 1050, k = 50 et u = 21, donc 49 degrés de liberté. Le jeu de coefficients adopté est celui qui a été indiqué ci-dessus comme présentant un compromis raisonnable pour les espérances de gain génétique sur le volume et la valeur du pilodyn à 48 mois : b1 = 1 (volume) et b2 = -0,02 (pilodyn) : Gains génétiques relatifs attendus de 43,34% (85% du maximum) pour le volume et de -4,12% (67% du maximum) pour le pilodyn. On obtient les valeurs suivantes pour les erreurs standard sur les prédictions de gain génétique relatif :

ES (ΔG volume) = 15,02%
ES (ΔG pilodyn) = 1,73%

Les intervalles de confiance au seuil de 5% sont donc les suivants :

ΔG volume : [13,13 ; 73,50]
ΔG pilodyn : [-7,59 ; -0,65]

Il est donc pratiquement assuré (plus de 97,5 chances sur 100) que ce jeu de coefficients donnera un gain génétique relatif apppréciable pour le volume (supérieur à 13%). En ce qui concerne la densité du bois, corrélée négativement à la valeur du pilodyn, il y a très peu de risque (moins de 2,5 chances sur 100) pour qu'elle baisse et plus de 97,5 chances sur 100 d'obtenir une évolution favorable, vis-à-vis de ce caractère, de la variété sélectionnée.

On aurait *a priori* prévu des erreurs standard beaucoup plus importantes et des estimations de gains génétiques non significativement différentes de 0. En fait, l'intuition que l'on peut avoir au vu du plan de croisements ne tient pas compte de l'accroissement de précision dû à une sélection multicaractère. Par ailleurs, les héritabilités au sens large sont élevées ce qui

augmente la précision des estimations de gain génétique.Toutefois, l'importance des intervalles de confiance des estimations de gains génétiques montre que l'on peut intégrer l'incertitude sur les progrès génétiques comme critère de choix des coefficients des caractères (par exemple, critère du minimax-regret utilisant une approche bayésienne, qui minimise l'espérance du coût maximal d'un choix erroné : JEFFERSON, 1989 ; ROBERT, 1992).

Contrôle de la diversité génétique de la sous-population sélectionnée

Comme cela a été indiqué plus haut, les programmes d'édition des index de sélection donnent pour un certain nombre de taux de sélection « standard » (50%, 20%, 10%, 5% et 1%) les nombres d'individus retenus par unité génétique élémentaire (ici les familles de pleins-frères) et par unité génétique apparentée (ici, les familles de demi-frères de mère commune ou de père commun). Si l'on veut connaître les effectifs pour des taux de sélection plus diversifiés, un programme annexe, commun à tous les modèles d'index de sélection, permet d'obtenir cette information par relecture du fichier où sont stockés les index de sélection.

Cela permet au sélectionneur d'intégrer dans sa décision finale le critère de diversité génétique à la création variétale qu'il réalisera. La même démarche peut s'appliquer à la population d'amélioration. La correction à réaliser sur les espérances de gains génétiques est très simple puisque les index sont exprimés en valeurs centrées-réduites. On applique alors la formule suivante :

$$\Delta Gc = \Delta Gs \; Im/Is$$

Où ΔGc est l'espérance de gain génétique corrigée pour une sous-population sélectionnée d'index moyen Im. ΔGs est l'espérance de progrès génétique correspondant à un taux de sélection quelconque, 10% par exemple, conduisant à un index moyen Is (identique à l'intensité de sélection si les index sont centrés-réduits).

Conclusion

La construction et l'utilisation des index de sélection posent donc des problèmes qui nécessitent souvent de revenir aux bases des modèles génétiques et statistiques mis en oeuvre.

La grande plasticité de cet outil a en effet pour contrepartie une grande facilité d'utilisation erronée dès que l'on aborde des domaines qui nécessitent une approche spécifique.
Ainsi, dans le cas de la sélection réciproque, l'existence de trois populations de référence pour les effets génétiques rend très approximative l'utilisation de modèles d'index établis pour une sélection intra-population. Il est alors indispensable de repenser la construction des index en fonction des problèmes pratiques posés et des sources d'information disponibles. Cette approche pragmatique conduit entre autres à constater que la définition d'effets génétiques ou environnementaux comme effets fixés ou aléatoires correspond beaucoup plus à une stratégie d'utilisation raisonnée des informations qu'à la nature intrinsèque d'une expérimentation.

Une telle stratégie d'optimisation dépend à son tour des objectifs, très divers, que l'on peut assigner à un même ensemble de résultats.

Ainsi, il n'est pas absurde de considérer comme aléatoires ou fixés des effets parentaux suivant que l'on cherche à tirer partie de l'ensemble de la variabilité génétique (cas de la sélection combinée) ou, au contraire, à détecter de la façon la plus précise la variabilité génétique additive d'une catégorie de parents (cas de la sélection sur descendance). Dans ce deuxième cas, les effets des mêmes parents sont considérés tour à tour comme aléatoires ou fixés.

Une limite difficile à contourner est qu' il sera toujours difficile de considérer comme aléatoires des effets parentaux représentés par un trop faible nombre de niveaux, pour des raisons de précision des estimations des variances ou des covariances correspondantes.

En fait, la statistique n'est qu'un outil parmi d'autres, qui n'a d'autre but, dans le cas de la génétique quantitative, que d'approcher très grossièrement, par des modèles linéaires, une réalité complexe non directement accessible à l'expérimentation. En d'autres termes, elle n'a d'autres rôles que descriptif et prévisionnel, pour une plage de variation où ces modèles restent acceptables.

Cet article a pour but principal d'expliquer le mode de raisonnement qui permet de construire un modèle d'index de sélection répondant à des objectifs particuliers et de l'utiliser au mieux, en distinguant éventuellement un jeu de caractères cibles et un jeu de caractères prédicteurs. Une telle distinction permet d'étendre facilement ces modèles à la sélection assistée par marqueurs : l'utilisation de marqueurs moléculaires ne constitue qu'un cas particulier de sélection indirecte. Les principes mis en oeuvre pour la sélection réciproque peuvent être reconduits pour n'importe quelle autre situation. En particulier, les mêmes notions permettent de reconstruire des modèles d'index à partir d'un plan de croisements factoriel, si pères et mères appartiennent à la même population (cas de la sélection récurrente intra-population). Il n'y a plus alors lieu de distinguer les composantes maternelles ou paternelles de la variance ou de la covariance ainsi que la covariance entre demi-frères de mère commune ou de père commun. En revanche, il faudra utiliser une estimation commune des composantes maternelles ou paternelles de variance d'échantillonnage minimum. Le logiciel OPEP traite cette situation comme une option particulière du même programme.
Dans le cadre des programmes d'amélioration génétique du CIRAD-CP, les modèles d'index de sélection développés par cet article peuvent être utilisés sur le caféier, le cacaoyer et le palmier à huile. Il est par ailleurs très facile d'introduire dans les modèles de sélection combinée des termes d'épistasie pour autant que l'on dispose d'estimations des matrices de variances-covariances correspondantes.
Le problème de l'estimation des variances et des covariances nécessaires pour utiliser un modèle d'index particulier est totalement indépendant de la construction du modèle, c'est pourquoi il n'a pas été abordé ici.

Annexe

Calcul des covariances entre valeurs génétiques et prédicteurs et des variances-covariances entre prédicteurs

Les conventions générales dans les notations utilisées sont les suivantes :

$$c_1=(1+F_1)/4$$

$$c_2=(1+F_2)/4$$

où F_1 est le coefficient de consanguinité moyen dans la population de référence des mères et F_2 celui qui correspond à la population de référence des pères.

Par ailleurs, les composantes génétiques additives de la variance ou de la covariance dans la population des hybrides $Gn+1$ sont : $\mathrm{cov}\, A_1 = (1/2)\,\mathrm{cov}\, A_m$ et $\mathrm{cov}\, A_2 = (1/2)\,\mathrm{cov}\, A_p$ où $\mathrm{cov}\, A_m$ et $\mathrm{cov}\, A_p$ sont respectivement les composantes additives de la variance ou de la covariance génétique dans les populations de référence Gn des mères et des pères.

Enfin, $\mathrm{cov}\, D$ désigne les composantes de dominance, qui ne sont définies que dans la population hybride.

La justification des principes généraux utilisés dans les modèles génétiques et statistiques et la définition des éléments statistiques des modèles sont données dans le corps de cet article. La démarche utilisée pour le calcul des espérances de variance d'effets estimés est expliquée par GRAYBILL (1961) pour des carrés moyens de modèles d'analyse de variance non-orthogonaux. Le calcul des espérances de covariances entre deux caractères suit exactement le même principe en remplaçant les sommes de carrés par les sommes de coproduits entre caractères correspondantes. Nous ne raisonnerons ici que sur les variances en sachant que le raisonnement a une valeur générale.

Les formules données ci-dessus concernent la constitution des éléments des sous-matrices de variances-covariances entre prédicteurs phénotypiques à partir des composantes génétiques et non des composantes statistiques, sauf pour la composante intra-famille, $\mathrm{cov}\, E$. Cette écriture permet d'homogénéiser ces modèles d'index avec ceux où mères et pères appartiennent à la même population. On retrouve ces éléments à partir des égalités suivantes où par souci d'homogénéité avec les formules de cette section, le terme général « cov » est remplacé par « var » :

$$\sigma^2_m = 2\,c_1\,\mathrm{var}\, A_1$$

$$\sigma^2_p = 2\,c_2\,\mathrm{var}\, A_2$$

$$\sigma^2_{(mp)} = 4\,c_1\,c_2\,\mathrm{var}\, D$$

Modèle d'index pour la sélection sur test de descendance

Ce modèle est utilisable à la fois pour sélectionner des mères ou des pères pour leurs performances en croisement avec des testeurs appartenant à une autre population.

Covariances entre valeurs génétiques additives et prédicteurs phénotypiques

Suivant le sens du croisement, il faut prédire les valeurs génétiques additives des mères ou des pères en fonction des performances moyennes de leurs descendants, ajustées par rapport à l'autre parent, dont l'effet est considéré comme fixé : α_i pour la mère et β_j pour le père.

On a respectivement en désignant par $\bar{y}^*$ une moyenne ajustée :

$$\operatorname{cov}(A_i, \hat{m}_i) = \operatorname{cov}\left[A_i, (\bar{y}^*_{i..} - \mu)\right] = (1/2)\operatorname{cov} A_m = \operatorname{cov} A_1$$

$$\operatorname{cov}(A_j, \hat{P}_j) = \operatorname{cov}\left[A_j, (\bar{y}^*_{.j.} - \mu)\right] = (1/2)\operatorname{cov} A_p = \operatorname{cov} A_2$$

En effet, la covariance entre l'effet moyen de la famille de demi-frères et la valeur additive du parent commun n'est autre que la covariance entre la valeur phénotypique de l'un quelconque de ces descendants et la valeur additive du parent. Elle est donc indépendante de l'effectif de la famille.

On pose :

$$\bar{y}^*_{i..} = \left[\sum_j \sum_k (y_{ijk} - \hat{\beta}_j)\right] / n_{i.} \text{ avec la condition : } \sum_i n_{i.}\, \hat{\beta}_j = 0$$

$$\bar{y}^*_{.j.} = \left[\sum_i \sum_k (y_{ijk} - \hat{\alpha}_i)\right] / n_{.j} \text{ avec la condition : } \sum_j n_{.j}\, \hat{\alpha}_i = 0$$

Variances et covariances des prédicteurs phénotypiques

La variance de l'estimation de l'effet de la mère ajusté, $\hat{m}_i$, ne dépend plus de l'effet fixé du père, β_j, mais elle est toujours fonction de l'interaction mère-père, $(m\beta)_{ij}$, qui est une variable aléatoire. On a donc :

$$E(\hat{m}_i) = E\left[(\bar{y}^*_{i..} - \mu)\right] = \left[n_{i.} m_i + \sum_j n_{ij} (m\beta)_{ij} + \sum_j \sum_k e_{ijk} \right] / n_{i.}$$

et, en élevant au carré :

$$E(\hat{m}^2_i) = \left[n^2_{i.} E(m^2_i) + \sum_j n^2_{ij} E((m\beta)^2_{ij}) + ni.\ E(e^2_{ijk}) \right] / n^2_{i.}$$

Soit, en remplaçant les espérances de carrés d'effets par les variances correspondantes :

$$\sigma^2_{\hat{m}_i} = \left[n^2_{i.} \sigma^2_m + \sum_j n^2_{ij} \sigma^2_{(m\beta)} \right] / n^2_{i.} + \sigma^2_e / n_{i.} = \sigma^2_m + \sum_j (n^2_{ij} / n^2_{i.}) \sigma^2_{(m\beta)} + \sigma^2_e / n_{i.}$$

La variance de l'estimation de l'effet du père, $\hat{p}_j$, ajusté à l'effet fixé α_i de la mère i, se calcule facilement par symétrie et l'on obtient :

$$\sigma^2_{\hat{p}_j} = \left[n^2_{.j} \sigma^2_p + \sum_i n^2_{ij} \sigma^2_{(\alpha p)} \right] / n^2_{.j} + \sigma^2_e / n_{.j} = \sigma^2_p + \sum_i (n^2_{ij} / n^2_{.j}) \sigma^2_{(\alpha p)} + \sigma^2_e / n_{.j}$$

Modèle d'index pour la sélection combinée d'hybrides

Ce modèle est valable pour sélectionner les têtes de clone dans les différentes familles hybrides ou les combinaisons hybrides les plus performantes.

Covariances entre valeurs génétiques et prédicteurs phénotypiques

Les conventions adoptées dans cette annexe pour désigner les covariances génétiques entre apparentés sont les suivantes et concernent, suivant le contexte, l'additivité ou la dominance :

cov*FS* = covariance entre pleins-frères = $2(c_1 \operatorname{cov} A_1 + c_2 \operatorname{cov} A_2)$ ou $4c_1c_2 \operatorname{cov} D$;

cov*HSm* = covariance additive entre demi-frères de mère commune = $2c_1 \operatorname{cov} A_1$;

cov*HSp* = covariance additive entre demi-frères de père commun = $2c_2 \operatorname{cov} A_2$;

cov*IND* = covariance entre un individu et lui-même = $\operatorname{cov} A_1 + \operatorname{cov} A_2$ ou $\operatorname{cov} D$.

Valeurs génétiques additives

Nous calculerons tour à tour la covariance de chaque prédicteur phénotypique avec la valeur génétique additive individuelle, puis familiale.

Si le prédicteur est l'effet de la mère, on a la relation :

$$\operatorname{cov}\left[A_{ijk},(\bar{y}_{i..}-\mu)\right]=\left[(n_{i.}-n_{ij})\operatorname{cov} HSm+(n_{ij}-1)\operatorname{cov} FS+\operatorname{cov} IND\right]/n_{i.}$$

$$=\left[2c_1\operatorname{cov}A_1(n_{i.}-n_{ij})+2(c_1\operatorname{cov}A_1+c_2\operatorname{cov}A_2)(n_{ij}-1)+\operatorname{cov}A_1+\operatorname{cov}A_2\right]/n_{i.}$$

$$=\operatorname{cov}A_1\left[\ 2c_1(n_{i.}-1)+1\ \right]/n_{i.}+\operatorname{cov}A_2\left[2c_2(n_{ij}-1)+1\right]/n_{i.}$$

$$\operatorname{cov}\left[A_{ij},(\bar{y}_{i..}-\mu)\right]=2\left[c_1\operatorname{cov}A_1(n_{i.}-n_{ij})+(c_1\operatorname{cov}A_1+c_2\operatorname{cov}A_2)n_{ij}\right]/n_{i.}$$

$$=2\left[c_1\operatorname{cov}A_1+c_2\operatorname{cov}A_2(n_{ij}/n_{i.})\right]$$

On obtient de façon symétrique si le prédicteur est l'effet du père :

$$\operatorname{cov}\left[A_{ijk},(\bar{y}_{.j.}-\mu)\right]=\left[(n_{.j}-n_{ij})\operatorname{cov} HSp+(n_{ij}-1)\operatorname{cov} FS+\operatorname{cov} IND\right]/n_{.j}$$

$$=\left[2c_2\operatorname{cov}A_2(n_{.j}-n_{ij})+2(c_1\operatorname{cov}A_1+c_2\operatorname{cov}A_2)(n_{ij}-1)+\operatorname{cov}A_1+\operatorname{cov}A_2\right]/n_{.j}$$

$$=\operatorname{cov}A_2\left[\ 2c_2(n_{.j}-1)+1\ \right]/n_{.j}+\operatorname{cov}A_1\left[2c_1(n_{ij}-1)+1\right]/n_{.j}$$

$$\operatorname{cov}\left[A_{ij},(\bar{y}_{.j.}-\mu)\right]=2\left[c_2\operatorname{cov}A_2(n_{.j}-n_{ij})+(c_1\operatorname{cov}A_1+c_2\operatorname{cov}A_2)n_{ij}\right]/n_{.j}$$

$$=2\left[c_2\operatorname{cov}A_2+c_1\operatorname{cov}A_1(n_{ij}/n_{.j})\right]$$

Pour obtenir la covariance des valeurs additives avec l'effet d'interaction estimé, il faut calculer la covariance avec la moyenne de famille de pleins-frères et lui retrancher celles qui ont été calculées pour les moyennes de demi-frères. Le résultat est alors :

$$\mathrm{cov}(A_{ijk}, \bar{y}_{ij.}) = \left[2(c_1 \mathrm{cov}\, A_1 + c_2 \mathrm{cov}\, A_2)(n_{ij}-1) + \mathrm{cov}\, A_1 + \mathrm{cov}\, A_2\right] / n_{ij}$$

$$= \mathrm{cov}\, A_1 \left[2c_1(n_{ij}-1)+1\right] / n_{ij} + \mathrm{cov}\, A_2 \left[2c_2(n_{ij}-1)+1\right] / n_{ij}$$

Pour la valeur génétique additive moyenne d'une famille, on a :

$$\mathrm{cov}(A_{ij}, \bar{y}_{ij.}) = 2(c_1 \mathrm{cov}\, A_1 + c_2 \mathrm{cov}\, A_2)$$

Ce qui donne, en tenant compte des covariances de A_{ijk} avec les moyennes de familles de demi-frères calculées ci-dessus :

$$\mathrm{cov}\left[A_{ijk}, (\hat{mp}_{ij})\right] = \mathrm{cov}\, A_1 \left\{ [\ 2c_1(n_{ij}-1)+1\ ]/n_{ij} - [\ 2c_1(n_{i.}-1)+1\ ]/n_{i.} - \left[2c_1(n_{ij}-1)+1\right]/n_{.j} \right\}$$

$$+ \mathrm{cov}\, A_2 \left\{ [\ 2c_2(n_{ij}-1)+1\ ]/n_{ij} - [\ 2c_2(n_{.j}-1)+1\ ]/n_{.j} - \left[2c_2(n_{ij}-1)+1\right]/n_{i.} \right\}$$

On procède de la même façon pour les valeurs génétiques additives moyennes de familles de pleins-frères, ce qui donne :

$$\mathrm{cov}\left[A_{ij}, (\hat{mp}_{ij})\right] = -2[\ c_1 \mathrm{cov}\, A_1 n_{ij}/n_{.j} + c_2 \mathrm{cov}\, A_2 n_{ij}/n_{i.}\ ]$$

Enfin, la covariance entre l'écart individuel à la moyenne de famille de pleins-frères et la valeur génétique additive individuelle se calcule par différence :

$$\mathrm{cov}(A_{ijk}, y_{ijk}) = \mathrm{cov}\, A_1 + \mathrm{cov}\, A_2$$

On a donc :

$$\mathrm{cov}\,(A_{ijk}, \hat{e}_{ijk}) = \mathrm{cov}\left[A_{ijk}, (\bar{y}_{ijk} - y_{ij})\right] = \mathrm{cov}\, A_1 \left\{ 1 - \left[\ 2_{c1}(n_{ij}-1)+1\ \right]/\mathrm{n}_{ij} \right\}$$

$$+ \mathrm{cov}\, A_2 \left\{ 1 - [\ 2c_2(n_{ij}-1)+1\ ]/n_{ij} \right\}$$

Valeurs génétiques de dominance

La démarche est exactement la même que pour le calcul des covariances entre prédicteurs phénotypiques et valeurs génétiques additives. Nous donnons donc simplement ci-dessous le détail des calculs.

$$\mathrm{cov}(D_{ijk},\hat{m}_i)=\mathrm{cov}\left[D_{ijk},(\bar{y}_{i..}-\mu)\right]=\left[(n_{ij}-1)\,\mathrm{cov}\,FS+\mathrm{cov}\,IND\right]/n_{i.}$$

$$=\left[(n_{ij}-1)4c_1c_2\,\mathrm{cov}\,D+\mathrm{cov}\,D\right]=\left\{\left[4c_1c_2(n_{ij}-1)+1\right]/n_{i.}\right\}\mathrm{cov}\,D$$

Pour la valeur de dominance moyenne d'une famille, on a :

$$\mathrm{cov}(D_{ij},\hat{m}_i)=(n_{ij}/n_{i.})\,\mathrm{cov}\,FS=4c_1c_2(n_{ij}/n_{i.})\,\mathrm{cov}\,D$$

De façon symétrique, on obtient d'une part :

$$\mathrm{cov}(D_{ijk},\hat{b}_i)=\left\{\left[4c_1c_2(n_{ij}-1)+1\right]/n_{.j}\right\}\mathrm{cov}\,D$$

et,

$$\mathrm{cov}(D_{ij},\ \hat{b}_i)=4c_1c_2(n_{ij}/n_{.j})\,\mathrm{cov}\,D$$

D'autre part,

$$\mathrm{cov}(D_{ijk},\bar{y}_{ij.})=\left[(n_{ij}-1)\,\mathrm{cov}\,FS+\mathrm{cov}\,IND\right]/n_{ij}=\left\{\left[4c_1c_2(n_{ij}-1)+1\right]/n_{ij}\right\}\mathrm{cov}\,D$$

et,

$$\mathrm{cov}(D_{ij},\bar{y}_{ij.})=4c_1c_2\,\mathrm{cov}\,D$$

On a donc :

$$\mathrm{cov}\left[D_{ijk},(\hat{mp})_{ij}\right]=\left\{\left[4c_1c_2(n_{ij}-1)+1\right]/n_{ij}-\left[4c_1c_2(n_{ij}-1)+1\right]/n_{i.}-\left[4c_1c_2(n_{ij}-1)+1\right]/n_{.j}\right\}\mathrm{cov}\,D$$

soit,

$$\mathrm{cov}[\ D_{ijk},(\hat{mp})_{ij}\ ]=\left[4c_1c_2(n_{ij}-1)+1\right](1/n_{ij}-1/n_{i.}-1/n_{.j})\,\mathrm{cov}\,D$$

et, pour la valeur de dominance moyenne d'une famille :

$$\mathrm{cov}\left[D_{ij},(\hat{mp})_{ij}\right]=4c_1c_2\,\mathrm{cov}\,D-4c_1c_2(n_{ij}/n_{i.})\,\mathrm{cov}\,D-4c_1c_2(n_{ij}/n_{.j})\,\mathrm{cov}\,D$$

soit,

$$\mathrm{cov}\left[D_{ij},(\hat{mp})_{ij}\right]=4c_1c_2\left[1-n_{ij}(1/n_{i.}-1/n_{.j})\right]\mathrm{cov}\,D$$

Enfin, puisque $\text{cov}(D_{ijk}, y_{ijk}) = \text{cov}\, D$,

on obtient, en combinant cette valeur avec la covariance entre D_{ijk} et la moyenne de famille, déjà calculée :

$$\text{cov}(D_{ijk}, \hat{e}_{ijk}) = \text{cov}\, D - \text{cov}\, D\left[4c_1c_2(n_{ij}-1)+1\right]/n_{ij} = \left\{1-\left[4c_1c_2(n_{ij}-1)+1\right]/n_{ij}\right\}\text{cov}\, D$$

Variances et covariances des prédicteurs phénotypiques

Comme cela a été précisé en introduction de ces annexes, nous ne présentons ici que le calcul des variances, celui des covariances entre deux caractères différents suivant exactement la même démarche.

- Calcul de $\sigma^2_{\hat{m}_i}$

On a :

$$E(\hat{m}_i) = E\left[(\bar{y}_{i..} - \mu)\right] = \left[n_{i.} m_i + \sum_j n_{ij}(p_j + (mp)_{ij}) + \sum_j \sum_k e_{ijk} \right] / n_{i.}$$

Soit, en élevant au carré :

$$E(\hat{m}_i^2) = \left[n_{i.}^2 E(m_i^2) + \sum_j n_{ij}^2 E(p_j^2) + \sum_j n_{ij}^2 E((mp)_{ij}^2) + n_{i.} E(e_{ijk}^2) \right] / n_{i.}^2$$

Et, en remplaçant les espérances de carrés d'effets par les variances correspondantes :

$$\sigma^2_{\hat{m}_i} = \left[n_{i.}^2 \sigma^2_m + \sum_j n_{ij}^2 \sigma^2_p + \sum_j n_{ij}^2 \sigma^2_{(mp)} \right] / n_{i.}^2 + \sigma^2_e / n_{i.}$$

$$= \sigma^2_m + \sum_j (n_{ij}^2 / n_{i.}^2)\left[\sigma^2_p + \sigma^2_{(mp)}\right] + \sigma^2_e / n_{i.}$$

- Calcul de $\sigma^2_{\hat{p}_j}$

L'espérance de l'estimation de l'effet du père, $E(\hat{p}_j)$, s'obtient de façon symétrique à celle de l'effet de la mère, $E(\hat{m}_i)$ et l'on a :

$$\sigma^2_{\hat{p}_j} = \left[n^2_{.j}\sigma^2_p + \sum_i n^2_{ij}\sigma^2_m + \sum_i n^2_{ij}\sigma^2_{(mp)} \right] / n^2_{.j} + \sigma^2_e / n_{.j}$$

$$= \sigma^2_p + \sum_i (n^2_{ij} / n^2_{.j})\left[\sigma^2_m + \sigma^2_{(mp)}\right] + \sigma^2_e / n_{.j}$$

Par la suite, nous calculerons directement l'espérance des variances.

-Calcul de $\sigma^2_{\hat{(mp)}_{ij}}$

On a la relation :

$$E\left[(\hat{mp})^2_{ij}\right] = E\left[(\bar{y}_{ij.}-\mu)^2 + (\bar{y}_{i..}-\mu)^2 + (\bar{y}_{.j.}-\mu)^2 - 2(\bar{y}_{ij.}-\mu)(\bar{y}_{i..}-\mu) - 2(\bar{y}_{ij.}-\mu)(\bar{y}_{.j.}-\mu) \right]$$

$$+ E\left[2(\bar{y}_{i..}-\mu)(\bar{y}_{.j.}-\mu) \right]$$

ce qui donne en termes de variances d'effets et de covariances entre effets :

$$\sigma^2_{\hat{(mp)}_{ij}} = \sigma^2_{\bar{y}_{ij.}} + \sigma^2_{\bar{y}_{i..}} + \sigma^2_{\bar{y}_{.j.}} - 2\,\mathrm{cov}(\bar{y}_{ij.}, \bar{y}_{i..}) - 2\,\mathrm{cov}(\bar{y}_{ij.}, \bar{y}_{.j.}) + 2\,\mathrm{cov}(\bar{y}_{i..}, \bar{y}_{.j.})$$

On a, en suivant la même démarche que pour le calcul des éléments de $\sigma^2_{\hat{m}_i}$:

$$\sigma^2_{\bar{y}_{ij}} = \left[n^2_{ij}(E(m^2_i) + E(m^2_i) + E(mp^2_{ij})) + n_{ij}E(e^2_{ijk}) \right] / n^2_{ij} = \sigma^2_m + \sigma^2_p + \sigma^2_{(mp)} + \sigma^2_e / n_{ij}$$

par ailleurs,

$$\mathrm{cov}(\bar{y}_{ij.}, \bar{y}_{i..}) = \mathrm{cov}\left[\bar{y}_{ij.}, (n_{ij}\bar{y}_{ij.})/n_{i.} \right] = (n_{ij} / n_{i.})\sigma^2_{\bar{y}_{ij.}}$$

En effet, la corrélation entre ces deux moyennes est due uniquement à la famille de pleins-frères qui leur est commune.

On a immédiatement par symétrie :

$$\mathrm{cov}(\bar{y}_{ij.}, \bar{y}_{.j.}) = (n_{ij} / n_{.j})\sigma^2_{\bar{y}_{ij.}}$$

Par ailleurs, et selon le même raisonnement :

$$\mathrm{cov}(\bar{y}_{i..},\bar{y}_{.j.}) = \mathrm{cov}\left[(n_{ij}\bar{y}_{ij.})/n_{i.},(n_{ij}\bar{y}_{ij.})/n_{.j}\right]$$

$$= \left[n_{ij}^2 \sigma^2_{\bar{y}_{ij.}} + n_{ij}(n_{i.}-n_{ij})\sigma^2_m + n_{ij}(n_{.j}-n_{ij})\sigma^2_p \right] / (n_{i.}n_{.j})$$

En remplaçant $\sigma^2_{\bar{y}_{ij}}$ par sa valeur et en regroupant, dans la formule concernant les éléments de $\sigma^2_{\hat{(mp)}_{ij}}$, les coefficients de σ^2_m, σ^2_p et $\sigma^2_{(mp)}$, on obtient :

$$\sigma^2_{\hat{(mp)}_{ij}} = \left[\sum_i n_{ij}^2/n_{.j}^2 + 2(1 - n_{ij}/n_{i.} - n_{ij}/n_{.j} + n_{ij}^2/(n_{i.}n_{.j})) \right] \sigma^2_m$$

$$+ \left[\sum_j n_{ij}^2/n_{i.}^2 + 2(1 - n_{ij}/n_{i.} - n_{ij}/n_{.j} + n_{ij}^2/(n_{i.}n_{.j})) \right] \sigma^2_p$$

$$+ \left[\sum_i n_{ij}^2/n_{.j}^2 + \sum_j n_{ij}^2/n_{i.}^2 - 2n_{ij}/n_{i.} - 2n_{ij}/n_{.j} + 2n_{ij}^2/(n_{i.}n_{.j}) + 1 \right] \sigma^2_{(mp)}$$

$$+ \left[1/n_{ij} - 1/n_{i.} - 1/n_{.j} + 2n_{ij}/(n_{i.}n_{.j}) \right] \sigma^2_e$$

- Calcul de $\sigma^2_{\hat{e}_{ijk}}$

On voit facilement que :

$$\sigma^2_{\hat{e}_{ijk}} = \sigma^2_{(y_{ijk}-\bar{y}_{ij.})} = \sigma^2_{y_{ijk}} + \sigma^2_{\bar{y}_{ij.}} - 2\,\mathrm{cov}(y_{ijk},\bar{y}_{ij.}) = \sigma^2_e - (1/n_{ij})\sigma^2_e = \left[(n_{ij}-1)/n_{ij}\right]\sigma^2_e$$

En effet,

$$\mathrm{cov}(y_{ijk},\bar{y}_{ij.}) = (1/n_{ij})\,\mathrm{cov}(y_{ijk},y_{ijk}) = \sigma^2_{\bar{y}_{ij.}} = \sigma^2_m + \sigma^2_p + \sigma^2_{(mp)} + (1/n_{ij})\sigma^2_e$$

- Calcul de $\mathrm{cov}\left[\hat{p}_j\ ,(\hat{mp})_{ij}\right]$

On a la relation :

$$\mathrm{cov}\left[\hat{p}_j\ ,(\hat{mp})_{ij}\right]=\mathrm{cov}(\bar{y}_{ij.},\bar{y}_{.j.})-\mathrm{cov}(\bar{y}_{i..},\bar{y}_{.j.})-\mathrm{cov}(\bar{y}_{.j.},\bar{y}_{.j.})$$

$$=\mathrm{cov}(\bar{y}_{ij.},\bar{y}_{.j.})-\mathrm{cov}(\bar{y}_{i..},\bar{y}_{.j.})-\sigma^2_{\bar{y}_{.j.}}$$

Les deux covariances et la variance ont déjà été calculées. En ordonnant les coefficients par rapport à chacune des variances, on obtient :

$$\mathrm{cov}\left[\hat{p}_j\ ,(\hat{mp})_{ij}\right]=\left[n_{ij}/n_{.j}-n^2_{ij}/(n_{i.}n_{.j})-\sum_i n^2_{ij}/n^2_{.j}\right]\left[\ \sigma^2_m+\sigma^2_{(mp)}\ \right]+\left[n_{ij}/n_{i.}-n^2_{ij}/(n_{i.}n_{.j})-1\right]\sigma^2_p$$

$$-\left[n_{ij}/(n_{i.}n_{.j})\right]\sigma^2_e$$

- Calcul de $\mathrm{cov}(\hat{m}_i\ ,\hat{p}_j)$

Ces éléments sont identiques à $\mathrm{cov}(\bar{y}_{i..},\bar{y}_{.j.})$ qui a été calculée plus haut pour obtenir les éléments de $\sigma^2_{(\hat{mp})_{ij}}$. En exprimant cette covariance en fonction de σ^2_m, σ^2_p, $\sigma^2_{(mp)}$ et σ^2_e, on obtient :

$$\mathrm{cov}(\hat{m}_i,\ \hat{p}_j)=\left[n^2_{ij}/(n_{i.}n_{.j})\right]\left[\sigma^2_m+\sigma^2_p+\sigma^2_{(mp)}\ \right]+\left[n_{ij}/(n_{i.}n_{.j})\right]\sigma^2_e$$

- Calcul de $\mathrm{cov}\left[\hat{m}_i,(\hat{mp})_{ij}\right]$

Ces éléments peuvent être obtenus directement par symétrie à partir de ceux de $\mathrm{cov}\left[\hat{p}_J\ ,(\hat{mp})_{ij}\right]$. Cette transposition donne :

$$\mathrm{cov}\left[\hat{m}i,(\hat{mp})_{ij}\right]=\left[n_{ij}/n_{i.}-n^2_{ij}/(n_{i.}n_{.j})-\sum_j n^2_{ij}/n^2_{i.}\right]\left[\ \sigma^2_p+\sigma^2_{(mp)}\ \right]+\left[n_{ij}/n_{.j}-n^2_{ij}/(n_{i.}n_{.j})-1\right]\sigma^2_m$$

Compte tenu du caractère laborieux et un peu « croûtard » de ces calculs, il est nécessaire de disposer d'un outil pour vérifier leur exactitude. Nous avons utilisé la méthode suivante qui procède par identification de la variance phénotypique calculée à partir des variances vraies des quatre effets à celle déterminée en utilisant les variances des effets estimés et de leurs trois covariances, telles que nous les avons établies. On doit donc avoir :

$$\sigma^2_{y_{ijk}} = \sigma^2_m + \sigma^2_p + \sigma^2_{(mp)} + \sigma^2_e$$

$$= \sigma^2_{\hat{m}_i} + \sigma^2_{\hat{p}_j} + \sigma^2_{(\hat{mp})_{ij}} + \sigma^2_{\hat{e}_{ijk}} + 2\{\mathrm{cov}(\hat{m}_i, \hat{p}_j) + \mathrm{cov}[\hat{m}_i, (\hat{mp})_{ij}] + \mathrm{cov}[\hat{p}_j, (\hat{mp})_{ij}]\}$$

Tableau III. Vérification des formules de variances et covariances des prédicteurs.

	σ^2_m	σ^2_p	$\sigma^2_{(mp)}$	σ^2_e
$\sigma^2_{\hat{m}}$	1	$1/P$	$1/P$	$1/nP$
$\sigma^2_{\hat{p}}$	$1/M$	1	$1/M$	$1/nM$
$\sigma^2_{\hat{p}}$	$2 - 1/M - 2/P + 2/MP$	$2 - 1/P - 2/M + 2/MP$	$1 - 1/M - 1/P + 2/MP$	$1/n - 1/nM - 1/nP + 2/nMP$
$\sigma^2_{\hat{e}}$	0	0	0	$1 - 1/n$
$\mathrm{cov}(\hat{m}, \hat{p})$	$1/MP$	$1/MP$	$1/MP$	$1/nMP$
$\mathrm{cov}[\hat{m}, (\hat{mp})]$	$- 1/MP$	$1/M - 1/MP - 1$	$- 1/MP$	$- 1/nMP$
$\mathrm{cov}[\hat{m}, (\hat{mp})]$	$1/P - 1/MP - 1$	$- 1/MP$	$- 1/MP$	$- 1/nMP$
$\sum c.(\mathrm{var}) + 2c.(\mathrm{cov})$	1	1	1	1

Le tableau croise les quatre variances vraies (colonnes) avec les quatre variances des effets estimés et les trois covariances entre ces effets (lignes). Dans chaque cellule figure la valeur du coefficient de la variance vraie dans la variance de l'effet estimé ou la covariance entre effets estimés. Le plan de croisements est supposé orthogonal et équilibré avec MP familles et n individus par famille. M est le nombre de mères et P est le nombre de pères. Si les variances et les covariances des effets estimés sont correctes, la somme des coefficients par colonne doit être égale à 1. La covariance entre effets estimés n'est pas supprimée par l'orthogonalité car elle est due aux individus communs qui participent conjointement à ces estimations.

Pour plus de clarté, on peut réaliser cette vérification à partir des valeurs prises par les formules dans le cas d'un plan de croisements orthogonal et équilibré, avec M mères, P pères, MP familles et n individus par famille. Le tableau III donne les valeurs des coefficients des 4 variances vraies des effets dans la variance de leur estimation et dans la covariance entre estimations.

L'orthogonalisation du modèle ne présente aucune difficulté. Il suffit, en effet, de reprendre les formules générales et de remplacer les sommes indicées d'effectifs, de produits ou de carrés d'effectifs par des produits, puisque les effectifs sont constants. On a par exemple :

$$\sigma^2_{\hat{m}_i} = \sigma^2_m + \sum_j (n^2_{ij} / n^2_{i.}) \left[\sigma^2_p + \sigma^2_{(mp)} \right] + \sigma^2_e / n_{i.}$$

$$= \sigma^2_m + P \left[n^2 / (nP)^2 \right] \left[\sigma^2_p + \sigma^2_{(mp)} \right] + \sigma^2_e / n = \sigma^2_m + \left[\sigma^2_p + \sigma^2_{(mp)} \right] / P + \sigma^2_e / n$$

On constate que, pour chacune des quatre colonnes correspondant à une variance vraie, on a bien : $\sum coeff(\text{var}) + 2 \sum coef_j(\text{cov}) = 1$, ce qui vérifie la condition énoncée ci-dessus.

Références bibliographiques

ANDERSON R.L., BANCROFT T.A. (1952) Statistical theory in research. Mac Graw Hill, 339 p.

BECKER W. B. (1984) Manual of quantitative genetics, 4e éd. Academic Enterprise (Washington), 188 p.

BRASCAMP E.W. (1984) Selection indices with constraints. *Anim. Breed. Abstr.* **52** (9) : 645-654.

COCKERHAM C.C. (1980) Random and fixed effects in plant genetics. *Theor. Appl. Genet.* **56** : 119-131.

COTTERILL P.P., JACKSON N. (1985) On index selection. -1- Methods of determining economic weight. *Silvae Genet.* **34** (2-3) : 56-63.

CUNNINGHAM E.P., MOEN K.A., GJEDREM T. (1970) Restriction of selection indexes. *Biometrics* **26** : 67-74.

DOLIGEZ A. (1992) Analyse des composantes de la variance génétique dans des populations d'hybrides interspécifiques d'Eucalyptus. Mémoire de D.E.A. INA-PG, CIRAD-Forêt , 48 p.

EFRON B. (1982) The Jacknife, the bootstrap and other resampling plans. Society for industrial and applied. mathematics, Philadelphie.

FOULLEY J.L. (1992) Estimation des composantes de la variance en modèle linéaire. Cours D.E.A.de statistiques et santé, Univ. Paris 11-Inserm, 108 p.

FOULLEY J.L., GIANOLA D., THOMPSON R. (1983) Prediction of genetic merit from data on binary and quantitative variates with an appreciation to calving difficulty birth weight and pelvic opening. *Génét. Sél. Evol.* **15** (3) : 401-424.

FOULLEY J.L., MANFREDI E. (1991) - Approches statistiques de l'évaluation génétique des reproducteurs pour des caractères binaires à seuils. *Génét. Sél. Evol.* **23** : 309-338.

FOULLEY J.L., OLLIVIER L. (1986) Criteria of coherence for the parameters used to construct a selection index. 35th Annual National Breeders' Roundtable, 1- 2 May 1986, St Louis (USA), p. 40-57.

GALLAIS A. (1990) - Théorie de la sélection en amélioration des plantes. Masson, 588 p.

GIANOLA D., FOULLEY J.L. (1983) Sire evaluation for ordered categorical data with a threshold model. *Génét. Sél. Evol.* **15** (2) : 201-224.

GRAYBILL F.A. (1961) An Introduction to linear statistical models.Vol.1. Mac Graw, 463 p.

HARRIS D.L. (1964) Genotypic covariances between inbred relatives. *Genetics* **50** : 1319-1348.

HARVILLE D.A. (1977) Maximum likelihood approaches to variance components estimation and to related problems. *J. Am. Stat. Assoc.* **72** : 320-340.

HAZEL L.M. (1943) The genetic basis for constructing selection indexes. *Genetics* **28** : 476-490.

HENDERSON C.R. (1953) Estimation of variance and covariance components. *Biometrics* **9** : 226-252.

HENDERSON C.R. (1973) Sire evaluation and genetic trends. Proceedings of the animal breeding and genetics symposium in honor of Dr J. Lush. American Society of Animal Science - American Dairy Science Association, p. 10-41.

HENDERSON C.R. (1977) Prediction of future records. Proc. Intern.Conf. on Quantitative Genetics, 16-21/08/1976. The Iowa State Univ. Press, p. 615-638.

HUBER D.A., WHITE T.L, LITTELL R.C., HODGE G.R. (1992) Ordinary least squares estimation of general and specific combining abilities from half-diallel mating designs. *Silvae Genet.* **41** (4-5) : 263-273.

JEFFERSON P.A. (1989) Discriminant functions in tree breeding. Thèse, University of Alberta, 271 p.

LEBART L., MORINEAU A., FÉNELON J.P. (1979) Traitement des données statistiques. Dunod, 510 p.

LINGREN D., NILSSON J.E. (1985) Calculations concerning selection intensity. Document interne Svedish University of Agricultural Sciences, 28 p.

MALLARD J. (1972) La théorie et le calcul des index de sélection avec restrictions : synthèse critique. *Biometrics* **28** (3) : 713-735.

OLLIVIER L. (1981) Eléments de génétique quantitative. Masson (Paris), 152 p.

OLLIVIER L., DERRIEN A.(1981) Une méthode générale d'estimation des paramètres génétiques dans un échantillon sélectionné, avec une application à une sélection sur un indice à trois caractères. *Ann.Génét. Sél. Anim.* **13** (3) : 281-292.

QUAAS R.L., POLLAK E.J. (1980) Mixed model methodology for farm and ranch beef cattle testing programs. *J. Anim. Sci.* **51** : 1277-1287.

RAO C.R. (1971) Estimation of variance components - MINQUE theory. *J. MultivarAnal.* **1** : 257-275.

ROBERT C.(1992) L'analyse statistique bayésienne. Economica (Paris) : 393 p.

ROUVIER R., 1969 Contribution à l'étude des index de sélection sur plusieurs caractères. Thèse, Faculté des Sciences de Paris, 92 p.

SEARLE S.R. (1971) Linear models. Wiley, 532 p.

SEARLE S.R. (1978) The value of indirect selection .-2- Progeny testing. *Theor. Appl. Genet.* **51** : 289-296

SEARLE S.R., CASELLA G., Mac CULLOCH CH.E. (1992) Variance components. Wiley, 501 p.

TAI G.C.C. (1979) An interval estimation of expected response to selection. *Theor. Appl. Genet.* **54** : 273-275.

TALBERT C.B. (1984) An analysis of several approaches to multiple traits index selection in loblolly pine. Thèse University of North Carolina, 106 p.

THOMPSON W.A. (1962) The problem of negative estimates of variance components. *Ann. Math.Stat.* **33** : 273-289.

WHITE T.L., HODGE G.R. (1988) Best linear prediction of breeding values in a forest tree improvement program. *Theor. Appl. Genet.* **76** : 719-727.

WHITE T.L., HODGE G.R. (1991) Indirect prediction of genetic values. *Silvae Genet.* **40** (1) : 20-28.