

Critère optimal de sélection : quelques résultats généraux

B. GOFFINET et J.M. ELSÉN*

I.N.R.A., Laboratoire de Biométrie

* *Station d'Amélioration génétique des Animaux, Centre de Recherches de Toulouse,
B.P. 12, F 31320 Castanet Tolosan*

Résumé

Dans l'axe du travail de COCHRAN (1951) appliqué à l'amélioration génétique des animaux, on donne les méthodes de sélection qui maximisent l'espérance de la valeur des retenus, dans diverses situations expérimentales : informations inégales et variables sur les candidats, nombre de mâles mis en testage aléatoire, existence d'effets élevage ou année, etc. On montre que, dans toutes ces situations, la méthode classique consistant à sélectionner les individus dont l'indice dépasse un seuil se généralise. Ce seuil est unique pour tous les candidats et l'indice obtenu est l'espérance conditionnelle de la valeur génétique sachant l'information sur le candidat ; cette dernière étant « orthogonalisée » par rapport aux effets fixes.

Mots-clés : Critère de sélection, progrès génétique, BLUP.

Summary

Optimal criteria for selection : some general results

Following the approach of COCHRAN (1951) concerning the genetic improvement of animals, methods of selection are presented which maximize the expectation of the genetic value of retained individuals in various experimental situations, the latter include unequal or variable information concerning the candidates, random number of males tested, existence of herd or year effects, etc. It is shown that the classical method of selecting individuals whose index exceeds a certain threshold can be generalized to cover all these situations. This generalized threshold is the same for all candidates, and is relative to an index which is the conditional expectation of the genetic value, given the information concerning the candidate, where that information is "orthogonalized" with respect to the fixed effects. An example is given to illustrate the results.

Key-words : Selection criteria, genetic change, BLUP.

I. Introduction

Le premier problème auquel est confronté un sélectionneur d'animaux est la définition précise de son objectif. Tout candidat i a une valeur u_i (valeur génétique additive, par exemple) et les caractéristiques de la distribution de ces valeurs u_i pour les individus retenus dépendent de la méthode de sélection.

On peut avoir pour objectif de maximiser la probabilité d'être retenu pour une espérance de la valeur à améliorer minimum fixée (RAO, 1952), ou de minimiser la probabilité que la valeur des retenus soit inférieure à une limite. Nous nous plaçons ici dans la situation la plus fréquente où l'objectif est de maximiser l'espérance de la valeur pour un caractère donné des individus sélectionnés (SMITH, 1936 ; HAZEL, 1943 ; COCHRAN, 1951).

Pour atteindre cet objectif la méthode de sélection devra être raisonnée compte tenu des contraintes sur le nombre de retenus et en fonction des informations disponibles. Selon la situation, le sélectionneur est soumis à des contraintes plus ou moins fortes pour le renouvellement de ses reproducteurs : par exemple, un éleveur disposant d'une étable de taille donnée devra choisir un nombre fixé de génisses de remplacement (contrainte C1 : nombre de retenus fixé). A l'opposé, un centre d'insémination artificielle a des impératifs de renouvellement moins stricts qui lui permettent de répartir éventuellement le renouvellement sur plusieurs années (contrainte C2 : espérance du nombre de retenus fixée) ; avec cette contrainte, moins stricte, on peut espérer une valeur moyenne des retenus supérieure à ce qui est possible quand le nombre de retenus est fixé.

L'essentiel des résultats publiés concernant cette deuxième contrainte. Historiquement, la complexité des conditions expérimentales étudiées a été croissante. Les premiers résultats (SMITH, 1936 ; HAZEL, 1943) ont été obtenus dans le cas où les prédictes et les valeurs des candidats suivent des lois multinormales et identiques d'un candidat à l'autre. COCHRAN (1951) généralise ces résultats au cas où les variables suivent une loi quelconque.

HENDERSON (1963, 1977) affirme que cette généralisation n'est directement utilisable pour la sélection parmi plusieurs candidats que sous les restrictions suivantes : candidats non apparentés et sur lesquels on possède la même information. ROUVIER (1969) lève cette dernière restriction dans le cas de la multinormalité. Sous la première contrainte, BULMER (1980) donne un critère de sélection optimal en levant ces restrictions.

Nous ne nous intéresserons ici qu'à la deuxième contrainte. Nous allons montrer que les restrictions évoquées par HENDERSON (1977) ne sont pas justifiées. En effet, nous généralisons les résultats de COCHRAN (1951) au cas où les candidats sont apparentés et connus avec des informations différentes. De plus, nous étudions des situations expérimentales plus complexes : nombre de candidats et information sur les candidats variables, espérance des prédictes inconnue (effet élevage).

Nous allons consacrer un premier paragraphe à la formalisation de ces situations. Puis, nous résoudrons les problèmes ainsi posés.

II. Description des situations étudiées

Pour clarifier les idées nous dirons que le problème est de sélectionner des mâles et nous noterons u_i l'espérance de la valeur d'un descendant du mâle i , que nous considérons aussi comme sa valeur génétique. Rappelons la situation étudiée par COCHRAN (1951) : un seul mâle est candidat et la décision de le retenir ou de l'éliminer est prise après observation de la réalisation y d'un vecteur aléatoire Y à valeurs dans D_0 . Une contrainte est fixée sur la probabilité d'être retenue et la loi conjointe de Y et de u_i est connue.

COCHRAN cherche donc la fonction indicatrice $d_1(Y)$ qui maximise :

$$\Omega = E(d_1(Y)u_1)$$

avec $E(d_1(Y)) = \alpha$ probabilité d'être retenu

et $d_1(y) \in \{0,1\} \forall y \in D_0$

$d_1(y) = 1$ si l'individu est retenu et 0 dans le cas contraire.

Nous allons étudier des généralisations croissantes de cette situation, de façon à se rapprocher de plus en plus des cas concrets.

A. Cas de plusieurs candidats

Nous généralisons d'abord au cas où plusieurs (N) individus sont simultanément candidats, avec une information plus ou moins importante sur chacun d'eux, la même prédictrice pouvant être liée aux valeurs génétiques de plusieurs candidats. Par exemple, on peut mesurer les performances d'une femelle qui est à la fois fille d'un 1^{er} mâle, sœur d'un 2^e, mère d'une 3^e.

Dans ces conditions, en regroupant dans le même vecteur toutes les prédictrices, nous définissons autant d'indicatrices que de candidats et nous cherchons les fonctions $d_i(Y)$ qui maximisent :

$$\Omega = E\left(\sum_{i=1}^N d_i(Y)u_i\right) \text{ espérance du gain total de la sélection}$$

avec $E\left(\sum_{i=1}^N d_i(Y)\right) = M$ espérance du nombre d'individus retenus

et $d_i(y) \in \{0,1\}$ pour $i = 1, \dots, N$ et $\forall y \in D_0$

B. Nombre de descendants potentiels inégaux

D'autre part, la décision de sélection peut être plus nuancée que l'alternative garder ou éliminer un animal. On peut décider de lui faire produire plus ou moins de descendants et, si M_i est le nombre maximum de descendants potentiels du $i^{\text{ème}}$ mâle, on considère que d_i est à valeurs dans $\{0,1, \dots, M_i\}$. La contrainte $E\left(\sum_{i=1}^N d_i(Y)\right) = M$ concerne alors le nombre total de descendants produits.

C. Information sur les candidats et nombre de candidats aléatoires

Dans les schémas de sélection sur descendance le plan d'expérience c'est-à-dire le nombre de filles mesurées par candidat, et leur répartition dans les élevages n'est pas fixé. Ainsi, quand on sélectionne des mâles sur des caractères exprimés par les femelles tels que la valeur laitière, on leur fait produire des filles dans un certain nombre d'élevages pour y mesurer leur production de lait. Ce nombre d'élevages varie d'un mâle à l'autre pour diverses raisons telles que l'âge du mâle utilisé.

En pratique, le nombre de mâles soumis à la sélection est lui-même aléatoire. Ainsi, dans beaucoup de schémas de sélection, on regroupe, en station de contrôle

individuel, un certain nombre de mâles pour y mesurer leur croissance ou leur conformation. Même si le nombre de candidats entrés en station est rigoureusement fixé, le nombre de mâles disponibles en fin de contrôle est aléatoire du seul fait de la mortalité.

De façon générale, on peut considérer que le plan d'expérience qui permet d'obtenir les informations sur les candidats, caractérisé par le nombre de candidats et le type d'information obtenu, est le résultat d'une variable aléatoire \mathcal{S} pouvant prendre S valeurs numérotées de 1 à S . Au plan $s \in \{1, \dots, S\}$ correspond le vecteur d'observation Y_s à valeurs dans D_{os} .

Nous définissons pour chaque plan d'expérience s le vecteur $u_s = (u_{s1}, u_{s2}, \dots, u_{sN_s})$ correspondant aux valeurs des N_s mâles mesurés dans le plan d'expérience s . Un même mâle peut se retrouver dans plusieurs plans d'expérience, mais pour éviter de définir une fonction qui associe à chaque mâle les différents plans d'expérience où il est mesuré, nous raisonnons sur l'ensemble $u = (u_1, \dots, u_s, \dots, u_S)$ de $N = \sum_{s=1}^S N_s$ mâles. Si on retrouve un même mâle dans q plans d'expérience, ce mâle sera répété q fois dans le vecteur u , et dans l'éventualité d'une non indépendance entre \mathcal{S} et la valeur génétique (u_{si}) de ce mâle, les lois *a priori* de ces variables u_{si} dépendront de s . C'est en particulier le cas de la sélection sur descendance des caractères de reproduction (fertilité mâle) pour laquelle on aura plus de descendants, donc plus d'informations pour les reproducteurs de plus grande valeur génétique.

Pour toutes les réalisations de s et y_s et pour tous les mâles i ($i = 1, \dots, N_s$) du plan réalisé on cherche les valeurs des indicatrices $d_{si}(y_s)$ qui satisfont les contraintes :

$$d_{si}(y_s) \in \{0, M_{si}\} \quad \forall y_s \in D_{os} \quad (i)_o$$

$$\text{et } \sum_{s=1}^S \text{Prob}(\mathcal{S} = s) E \left(\sum_{i=1}^{N_s} \delta_{si} d_{si}(Y_s) \right) = M \quad (ii)$$

et maximisent :

$$\Omega = \sum_{s=1}^S \text{Prob}(\mathcal{S} = s) E \left(\sum_{i=1}^{N_s} d_{si}(Y_s) u_{si} \right) \quad (o)$$

La maximisation de (o) sous les contraintes (i)_o et (ii) recouvre tous les cas précédemment étudiés.

D. *Espérance des prédictrices inconnue*

Enfin, une situation fréquente en sélection est celle où on connaît pour chacun des plans d'expérience la loi conjointe des prédicteurs centrés $Y_s - E(Y_s)$ et des variables à prédire u_s , mais où les espérances des prédicteurs sont inconnues, mais de la forme $E(Y_s) = X_s \beta_s$ avec X_s une matrice connue de rang inférieur à la dimension de Y_s et β_s un vecteur inconnu. C'est le cas de la sélection en ferme où les descendants d'un mâle sont répartis dans plusieurs élevages, dont les valeurs sont inconnues mais participent additivement aux valeurs des prédictrices.

Dans cette situation, il n'est pas possible de trouver une méthode de sélection optimale quel que soit β_s et on se limitera, pour le choix des indicatrices, aux fonctions invariantes par translation, c'est-à-dire telles que :

$$d_{si}(y_s) = d_{si}(y_s + X_s \beta_s) \quad \text{pour tout } i, s, y_s \text{ et } \beta_s \quad (iii)_o$$

Sous cette restriction, dans l'exemple de la sélection en ferme les valeurs des indicatrices seront les mêmes quels que soient les effets élevages. La contrainte (iii)₀ est équivalente à :

$$d_{si}(y_s) = d_{si}(P_{X_s^\perp}(y_s)) \tag{iii}$$

où $P_{X_s^\perp}$ désigne un projecteur sur l'orthogonal de l'espace engendré par les colonnes de X_s .

Remarque :

Pour résoudre le problème dans le cas particulier où les variables aléatoires Y_s et u_s sont discrètes (prolificité, fertilité...) nous avons constaté qu'il est nécessaire de remplacer la contrainte : (i)₀ par $d_{si}(y_s) \in [0, M_{si}] \forall y_s \in D_{os}$ (i)

Notons que pour $M_{si} = 1$, les valeurs d_{si} seront les probabilités de retenir l'individu concerné.

En conclusion, nous allons rechercher le maximum de Ω défini par (0) sous les 3 contraintes (i) à (iii). Dans le cas où $E(Y_s)$ est connue les résultats s'appliqueront en remplaçant X_s par 0.

III. Résolution du problème

Nous donnons d'abord la résolution mathématique dans le cas général ; nous montrerons ensuite comment utiliser ce résultat général dans certains cas plus simples.

A. Résolution mathématique

Notons $Z_s = P_{X_s^\perp}(Y_s)$. Soit D_s l'ensemble des réalisations de Z_s pour $s = 1, \dots, S$, soit $D = \sum_{s=1}^S D_s$, soit F l'ensemble des fonctions mesurables de D dans \mathbb{R}^N , soit $J = \{(s,i) | s \in (1, \dots, S), i \in (1, \dots, N_s)\}$. Compte tenu de la contrainte (iii) et en utilisant les propriétés de l'espérance conditionnelle, l'objectif est la recherche d'un élément $d = (d_{si})(s,i) \in J$ de F qui maximise :

$$\Omega = \sum_{s=1}^S \text{Prob}(\mathcal{G} = s) E \left[\sum_{i=1}^{N_s} d_{si}(Z_s) E(u_{si} | Z_s) \right] \tag{I}$$

sous les contraintes :

$$d_{si}(Z_s) \in [0, M_{si}] \forall z_s \in D_s \tag{I}$$

$$\sum_{s=1}^S \text{Prob}(\mathcal{G} = s) E \left(\sum_{i=1}^{N_s} d_{si}(Z_s) \right) = M \tag{II}$$

Soit $\{F_{[h,p]}\}_{(h,p) \in \Theta}$ une famille d'éléments de F où Θ est défini par :

$$(h,p) \in \Theta \iff \begin{cases} h \in \mathbb{R} \\ p = (p_{11}, p_{12}, \dots, p_{1N_1}, \dots, p_{si}, \dots, p_{SN_S}) \\ 0 \leq p_{si} \leq M_{si} \forall (s,i) \in J \end{cases}$$

et :

$F_{[h,p]} = \{F_{si[h,p]}(s,i) \}_{(s,i) \in J}$ défini par :

$$F_{si[h,p]}(z_s) = \begin{cases} M_{si} & \text{si } E(u_{si}|Z_s = z_s) > h \\ p_{si} & \text{si } E(u_{si}|Z_s = z_s) = h \quad \forall z_s \in D_s \\ 0 & \text{si } E(u_{si}|Z_s = z_s) < h \end{cases}$$

Théorème : Sous la contrainte $\sum_s \text{Prob}(\mathcal{S} = s) \sum_i M_{si} \geq M$

1) Il existe au moins un élément (λ, P) de Θ telle que $F_{[\lambda, P]}$ satisfasse aux contraintes I et II.

2) Dans la classe des fonctions satisfaisant aux contraintes I et II, $F_{[\lambda, P]}$ maximise Ω .

On trouvera la démonstration en annexe.

B. Cas des lois continues

Nous allons décrire la stratégie optimale de sélection dans le cas où les lois de répartition des couples (u_{si}, Y_s) sont continues : la valeur du seuil λ est calculée *a priori* au moyen de la contrainte (II). On utilise, pour ce faire, les lois de répartition de toutes les espérances conditionnelles $E(u_{si}|P_{X_s^\perp}(Y_s))$. Rappelons que la projection $P_{X_s^\perp}(Y_s)$ est un ensemble de combinaisons linéaires de Y_s , d'espérance nulle et de rang $= \dim Y_s - \text{rang } X_s$ (si $X_s = 0$ on conserve y_s).

Ce seuil λ est le même pour tous les candidats, quelle que soit l'information sur leur valeur et pour toutes les situations possibles s' . On lève ainsi les restrictions énoncées par HENDERSON en 1977.

Un plan d'expérience s est alors réalisé et on mesure les valeurs y_s prises par les variables aléatoires Y_s . On calcule $P_{X_s^\perp}(y_s)$ projection de y_s sur l'orthogonal de l'espace engendré par les colonnes de X_s .

Enfin, on fait produire le nombre maximum de descendants aux individus pour lesquels les valeurs $E(u_{si}|P_{X_s^\perp}(Y_s))$ dépassent le seuil λ (seuil unique pour tous les individus).

Remarque :

Dans le cadre de la multinormalité, $E(u_{si}|P_{X_s^\perp}(y_s))$ est parmi les prédicteurs linéaires sans biais $T_{si}(y_s)$ le meilleur pour le risque quadratique $E(u_{si} - T_{si}(Y_s))^2$ (sigle anglo-saxon B.L.U.P.).

En effet, tout prédicteur linéaire sans biais de u_{si} est une fonction linéaire de $P_{X_s^\perp}(y_s)$.

Or, $E(u_{si}|P_{X_s^\perp}(y_s))$ (fonction de $P_{X_s^\perp}(y_s)$ qui minimise le risque quadratique), est une fonction linéaire de $P_{X_s^\perp}(y_s)$ dans le cadre de la multinormalité.

Pour illustrer cette stratégie, nous allons la mettre en œuvre sur un petit exemple faisant appel à un modèle volontairement simplifié.

IV. Illustration

Nous nous plaçons dans le cas où des mâles sont sélectionnés après mesure de leur croissance en station de contrôle individuel. Nous supposons que cette station a une capacité maximale de P places. On fait 2 séries successives par an et la sélection a lieu après la réalisation de ces 2 séries de mesures (donc sur $2 \times P$ candidats maximum). Le nombre de mâles effectivement contrôlé est aléatoire.

TABLEAU 1

Description des situations expérimentales étudiées.

Description of experimental situations studied.

Nom de la situation	Nombre de places	Nombre de candidats par série	Espérance du nombre de retenus par an (M)
C1	6	2, 3, 4, 5 ou 6	0,5 à 4
C2	15	11, 12, 13, 14 ou 15	1 à 22
C3	60	40, 45, 50, 55 ou 60	1 à 80

Diverses combinaisons du nombre de places de station et de l'espérance du nombre de retenus ont été étudiées. Pour chacune d'entre elles le nombre de candidats peut prendre 5 valeurs équiprobables (ces valeurs et l'équiprobabilité ont été choisies pour des raisons de simplicité) (tabl. 1). Une situation expérimentale s est caractérisée par le couple $(N_{s1}, N_{s2}) = N_s$ d'effectifs de mâles à contrôler en 1^{re} et en 2^e série.

Vingt-cinq situations sont donc possibles, chacune ayant la probabilité $1/25^e$ de se réaliser.

On suppose que les prédicteurs $Y_s : \{Y_{skj}, k = 1, 2, j = 1, \dots, N_{sk}\}$ suivent le modèle :

$$Y_{skj} = \beta_k + u_{skj} + e_{skj}$$

où :

- β_k est l'effet du mois de contrôle ($k = 1, 2$ est le numéro de la série) ;
- u_{skj} est la valeur génétique du $j^{\text{ème}}$ mâle de la $k^{\text{ème}}$ série de la $s^{\text{ème}}$ situation. u_{skj} est une variable aléatoire normale d'espérance nulle et telle que :

$$\begin{aligned} \text{cov}(u_{skj}, u_{sk'j'}) &= \gamma_u \text{ si } k = k' ; j = j' \\ &= 0 \text{ sinon ;} \end{aligned}$$

- e_{skj} est une variable aléatoire normale, indépendante des variables aléatoires u_{skj} , d'espérance nulle et telle que :

$$\begin{aligned} \text{cov}(e_{skj}, e_{sk'j'}) &= \gamma_e \text{ si } k = k' ; j = j' \\ &= 0 \text{ sinon} \end{aligned}$$

Sous ces hypothèses, $E(u_{skj} | P_{X_s^1}(Y_s)) = (\gamma_u / (\gamma_u + \gamma_e)) (Y_{skj} - Y_{sk.})$, $Y_{sk.}$ étant la moyenne des Y_{skj} .

Cette espérance conditionnelle suit une loi normale d'espérance nulle et de variance :

$$\sigma_{sk}^2 = (\gamma_u^2 / (\gamma_u + \gamma_e)) (N_{sk} - 1) / N_{sk}$$

Nous nous plaçons dans le cas où l'alternative pour un candidat est d'être éliminé ou retenu ($M_{si} = 1$ dans la contrainte (ii)).

L'espérance de la somme des valeurs des retenus est ici :

$$\Omega = (1/25) \sum_{s=1}^{25} \sum_{k=1}^2 N_{sk} (1/\sqrt{2\pi}) \exp(- (1/2) (\lambda/\sigma_{sk})^2) \sigma_{sk}$$

où λ satisfait la contrainte :

$$M = (1/25) \sum_{s=1}^{25} \sum_{k=1}^2 N_{sk} (1 - \pi(\lambda/\sigma_{sk}))$$

π étant la fonction de répartition de la loi normale (0,1).

Remarque :

Dans le cas où le selectionneur ne connaît pas la loi de \mathcal{S} son objectif est de maximiser l'espérance Ω_s de la somme de la valeur des retenus conditionnellement au plan d'expérience réalisé.

$$\Omega_s = \sum_{k=1}^2 N_{sk} (1/\sqrt{2\pi}) \exp(- 1/2) (\lambda_s/\sigma_{sk})^2 \sigma_{sk}$$

où λ_s est donné par :

$$M = \sum_{k=1}^2 N_{sk} (1 - \pi(\lambda_s/\sigma_{sk}))$$

Si la loi de \mathcal{S} est celle supposée au début, le gain Ω' global, obtenu dans ce cas est :

$$\Omega' = (1/25) \sum_{s=1}^{25} \Omega_s$$

On trouvera sur la figure 1 des valeurs prises par la supériorité relative $\delta = 100 (\Omega - \Omega')/\Omega$ en fonction du rapport du nombre de retenus au nombre moyen de candidats $M/E(N)$ (4, 13 ou 50). Cette supériorité est indépendante du rapport des variances γ_u/γ_e et donc de l'héritabilité du caractère sélectionné.

On constate qu'elle est généralement faible, sauf dans les cas où beaucoup de candidats sont retenus. Dans l'exemple étudié, on peut donc, sans grande perte, déterminer le seuil de sélection conditionnellement au nombre de candidats effectivement mesurés.

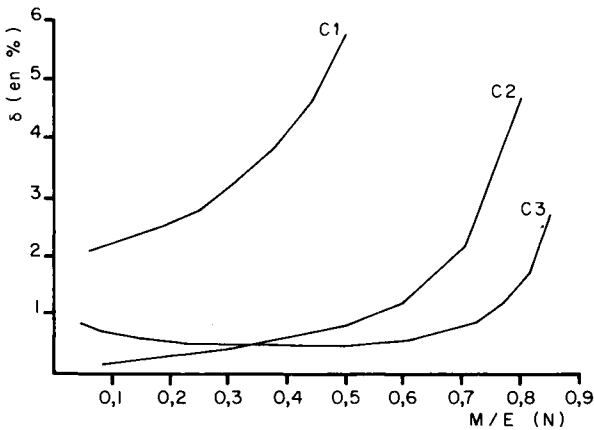


FIG. 1

Supériorité relative δ en p. 100 de la méthode optimale par rapport à la méthode suboptimale en fonction du rapport $M/E(N)$ du nombre de retenus au nombre moyen de candidats. (C1, C2, C3 ont été définis dans le tableau 1).

Relative superiority δ (in p. 100) of the optimal method relative to the suboptimal method as a function of $M/E(N)$, i.e. of the number of candidates selected relative to the average number of candidates. (C1, C2 and C3 are described in table 1).

Conclusions

Il est classique, en sélection, d'attribuer à tout candidat une valeur (indice) : on utilise généralement celle qui se rapproche le plus possible de la valeur génétique du candidat (on minimise l'erreur quadratique). Nous avons montré ici que cette technique maximise aussi l'espérance de la valeur génétique des retenus et ceci dans des situations tout à fait générales. Ceci complète donc les résultats préalablement obtenus par COCHRAN (1951), ROUVIER (1969), HENDERSON (1977) et BULMER (1980), et en particulier montre l'unicité du seuil de sélection.

Dans le cas où les espérances de prédictrices sont inconnues nous nous sommes limités aux indicatrices invariants par translation des effets fixes. Lorsque ces effets représentent des valeurs additives d'élevage, cette limitation généralise la pratique courante des sélectionneurs (comparaison aux contemporaines, BLUP...) et permet donc de ne pas faire dépendre la sélection des élevages dans lesquels sont placés les candidats.

Dans le cas où on ne connaît que les moments d'ordre 2 des différentes variables aléatoires, et contrairement à la recherche de critères optimaux pour le risque quadratique, la limitation des indicatrices de sélection aux fonctions linéaires ne semble pas fournir de résultats simples pour la maximisation de notre objectif de sélection.

Remarquons enfin que, pour cette étude, nous avons généralisé l'expression de l'espérance du gain dû à la sélection, à des situations qui ne sont pas prises en compte dans les formules classiques. Ceci donne en particulier un critère unique de mesure de

la qualité d'un schéma de sélection, critère utilisable pour classer *a priori* différents schémas. Il est à noter que les résultats obtenus ici ne se généralisent pas au cas où il existe des différences génétiques systématiques entre des groupes de candidats.

Remerciements

Nous tenons à remercier H. CAUSSINUS, Professeur à l'Université Paul-Sabatier, Toulouse, pour l'aide apportée à la formalisation de ces problèmes.

Annexe

Démonstration du théorème

Nous notons G_{si} et H_s les fonctions de répartition des variables aléatoires $E(u_{si}|Z_s)$ et Z_s .

Remarque préliminaire :

Dans le cas particulier où un seul plan d'expérience est possible ($S = 1$), où $M_{si} = M_o$, où on dispose de la même information sur tous les individus et où Z_s admet une densité, le théorème est démontré par l'application du lemme de Neyman-Pearson.

En effet, dans ce cas, on cherche la fonction $d_{si}(Z_s)$ qui maximise $\Omega = N_s \int d_{si}(Z_s) E(u_{si}|Z_s) dH_s$

$$\text{avec } d_{si}(z_s) \in [0, M_o] \quad (I)_o \quad z_s \in D_s$$

$$\text{et } N_s \int d_{si}(Z_s) dH_s = M \quad (II)_o$$

et on sait que la solution est donnée par :

$$d_{si}(z_s) = M_o \quad \text{si } E(u_{si}|Z_s = z_s) > \lambda$$

$$d_{si}(z_s) = 0 \quad \text{sinon}$$

où λ est une fonction de M donnée par $(II)_o$.

1) Existence

$F_{[h,p]}$ vérifie la contrainte I. Il suffit de vérifier qu'il existe h et p tels que :

$$G[h,p] = \sum_{s=1}^S \text{Prob}(\mathcal{F} = s) E\left(\sum_{i=1}^{N_s} F_{si}[h,p](Z_s)\right) = M$$

Notons :

$$G_{si}(h + o) = \lim_{\epsilon \rightarrow 0} G_{si}(h + \epsilon)$$

$$\epsilon \rightarrow 0 \quad \epsilon > 0$$

$$G[h,p] = \sum_{s=1}^S \text{Prob}(\mathcal{Y} = s) \left[\sum_{i=1}^{N_s} M_{si} \text{Prob}(E(u_{si}|Z_s) > h) + \sum_{i=1}^{N_s} p_{si} \text{Prob}(E(u_{si}|Z_s) = h) \right]$$

Notons : $M'_{si} = \text{Prob}(\mathcal{Y} = s) M_{si}$ et $p'_{si} = \text{Prob}(\mathcal{Y} = s) p_{si}$

$$G_{[h,p]} = \sum_{s=1}^S \sum_{i=1}^{N_s} [M'_{si} (1 - G_{si}(h)) + (p'_{si} - M'_{si}) (G_{si}(h+0) - G_{si}(h))]$$

a) S'il existe λ tel que $\sum_{s=1}^S \sum_{i=1}^{N_s} M'_{si} (1 - G_{si}(\lambda + 0)) = M$

(ce qui est le cas en particulier si les fonctions de répartition G_{si} sont continues), alors $P = (0, 0, \dots, 0)$, $F_{[\lambda,p]}$ satisfait à la contrainte II.

b) Dans le cas contraire, prenons :

$$\lambda = \sup \{h \mid \sum_{s=1}^S \sum_{i=1}^{N_s} M'_{si} (1 - G_{si}(h)) \geq M\}$$

ceci entraîne que :

$$\sum_{s=1}^S \sum_{i=1}^{N_s} M'_{si} (1 - G_{si}(\lambda + 0)) < M$$

et donc que :

pour $p_{si} = M_{si} \forall (s,i) \in J$ $G_{[\lambda,p]} \geq M$

pour $p_{si} = 0 \forall (s,i) \in J$ $G_{[\lambda,p]} < M$

puisque la fonction $G_{[\lambda,p]}$ est continue en chacun des p_{si} , il existe au moins un ensemble de valeurs :

$$P = (P_{11}, P_{12}, \dots, P_{si}, \dots, P_{SN_S}) \text{ tel que } G_{[\lambda,p]} = M$$

et donc tel que $F_{[\lambda,p]}$ satisfait à la contrainte II ;

2) Optimalité de $F_{[\lambda,p]}$

soit $g : \{g_{si}\} si \in J$ une fonction quelconque satisfaisant aux contraintes I et II.

Etudions la différence d :

$$\begin{aligned} d &= \sum_{s=1}^S \text{Prob}(\mathcal{Y} = s) E \left[\sum_{i=1}^{N_s} (F_{si[\lambda,p]}(Z_s) - g_{si}(Z_s)) u_{si} \right] \\ &= \sum_{s=1}^S \text{Prob}(\mathcal{Y} = s) \sum_{i=1}^{N_s} \left[\int_{D_s} (F_{si[\lambda,p]}(z_s) - g_{si}(z_s)) \cdot E(u_{si}|Z_s = z_s) dH_s(z_s) \right] \\ &= I_1 + I_2 + I_3 + I_4 \end{aligned}$$

pour $\ell = 1, 2, 3$

$$I_\ell = \sum_{s=1}^S \text{Prob}(\mathcal{Y} = s) \sum_{i=1}^{N_s} \left[\int_{D_{s\ell}} (F_{si[\lambda,p]}(z_s) - g_{si}(z_s)) (E(u_{si}|Z_s) - \lambda) dH_s(z_s) \right]$$

D_{s1} : domaine défini par $E(u_{si}|z_s) > \lambda$

D_{s2} : domaine défini par $E(u_{si}|z_s) = \lambda$

D_{s3} : domaine défini par $E(u_{si}|z_s) < \lambda$

$$I_4 = \lambda \sum_{s=1}^S \text{Prob}(\mathcal{G} = s) \sum_{i=1}^{N_i} \int (F_{si[\lambda, P]} - g_{si}(z_s)) dH_s(z_s)$$

on a :

$I_1 \geq 0$ puisque les $F_{si[\lambda, P]}$ sont maximum sur D_{s1}

$I_2 = 0$

$I_3 \geq 0$ puisque les $F_{si[\lambda, P]}$ sont minimum sur D_{s3}

$I_4 = 0$ puisque $F_{[\lambda, P]}$ et g sont soumis à la contrainte I.

Ainsi, d est positif ou nul quelle que soit la fonction f satisfaisant aux différentes contraintes et Ω est maximum pour $F_{[\lambda, P]}$.

Reçu le 7 décembre 1983.

Accepté le 26 janvier 1984.

Références bibliographiques

- COCHRAN W.G., 1951. Improving by means of selection. *Proceedings of the second Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, 1951*. Neyman J. (ed.), 449-470., University of California Press, Berkeley.
- BULMER M.G., 1980. *The mathematical theory of quantitative genetics*, 254 pp., Oxford, University Press, Oxford.
- HAZEL L.N., 1943. The genetic basis for constructing selection indexes. *Genetics*, **28**, 476-490.
- HENDERSON C.R., 1963. *Selection index and expected genetic advance*. Statistical genetics in plant breeding. NAS-NCR Publication, **982**, 141-163.
- HENDERSON C.R., 1977. Prediction of future records. *Proceedings of the International Conference on Quantitative Genetics, Ames, Iowa, August 16-21, 1976*, POLLACK E., KEMPTHORNE O., BAILEY T.B. Jr. (ed.), Iowa State University Press, Ames, 615-638.
- RAO C.R., 1952. *Advanced statistical methods in biometric research*, 522 pp., Wiley Press, New York.
- ROUVIER R., 1969. *Contribution à l'étude des index de sélection sur plusieurs caractères*. Thèse de 3^e cycle, Faculté des Sciences, Paris.
- SMITH H.F., 1936. A discriminant function for plant selection. *Ann. Eugen.*, **7**, 240-250.