Optimisation du rééchantillonnage dans un logiciel d'Amélioration des Plantes

Baradat P.
INRA-Département EFPA
UMR AMAP
34398 Montpellier Cedex 5
FRANCE
baradat@ensam.inra.fr
Labbé T.
INRA-Département EFPA
Service Forêt-Bois, Unité EPHYSE
33612 Cestas Cedex
FRANCE
labbe@pierroton.inra.fr

RÉSUMÉ. DIOGENE, un logiciel d'Amélioration des Plantes conçu et développé au sein du Département EFPA de l'INRA, fonctionne sous Solaris et Linux. C'est un logiciel libre (licence GPL), en évolution constante, qui traite de modèles de Biométrie générale, de Génétique Quantitative et de Génétique des Populations. Il fait largement appel aux techniques de rééchantillonnage (jackknife et bootstrap) pour tester des hypothèses nulles et déterminer les intervalles de confiance de paramètres estimés (héritabilités, coefficients de corrélation génétique, gains génétiques prédits…). Pour des raisons de simplicité d'utilisation et de rapidité d'exécution, l'architecture du logiciel (incluant le fichier de données) et les algorithmes de rééchantillonnage ont été optimisés. Cet article décrit brièvement les points originaux.

ABSTRACT. DIOGENE, a Plant Breeding software conceived and developed in the EFPA Department of INRA, runs under Solaris and Linux OS. It is a free software (GPL licence) which deals with models of General Biometry as well as Quantitative and Population Genetics. It intensively makes use of resampling techniques (jackknife and bootstrap) to test null hypotheses and to compute confidence intervals of estimated parameters (heritabilities, genetic correlations, predicted genetic gains…). For sake of simplicity of use and of processing speed, software architecture (including data file) and resampling algorithms were optimized. Our paper shortly describes the involved original features.

MOTS-CLÉS : rééchantillonnage, jackknife, bootstrap, accès direct, fichier binaire, fichier paramètre.

KEYWORDS: resampling, jackknife, bootstrap, direct access, binary file, parameter file.

1. Introduction

Depuis leurs premières applications, au début des années 80, à l'étude des marges de tolérance dans des processus industriels [8], les techniques de rééchantillonnage se sont rapidement imposées dans le domaine des sciences biologiques pour réaliser des tests d'hypothèses et calculer des intervalles de confiance de paramètres estimés : Les essais en conditions contrôlées de terrain et de laboratoire [12], l'Ecologie Quantitative [10], la Génétique Quantitative [1], [2] dont la réponse à la sélection [4], les statistiques spatiales [5] et la Génétique des Populations [6], [7], [13]. Cette méthodologie, étudiée en détail dans [8] et [11], offre en effet le grand intérêt de proposer des algorithmes généraux et robustes pour le calcul des variances d'échantillonnage, indépendantes des formules d'estimation des paramètres. Revers de la médaille, la nécessité de réitérer les estimations conduit souvent à des temps de traitement importants se chiffrant en heures voire en jours, même avec des serveurs ou micro-ordinateurs puissants. C'est pourquoi, destinant en particulier le logiciel d'Amélioration des Plantes DIOGENE à des pays du Sud, nous avons cherché à rendre cette technique accessible à partir de micro-ordinateurs bas de gamme. A cet effet, nous avons réduit autant que possible la génération et la lecture de fichiers pour faire appel essentiellement au calcul en mémoire vive à partir d'un fichier de données unique. Une ancienne version du logiciel est décrite dans [3]. Une notice complète concernant la version actuelle qui suit la norme GLM (ce qui permet de traiter simultanément des variables quantitatives et qualitatives) sera disponible fin 2006.

2. Rééchantillonnage : le jackknife et le bootstrap

Nous ne donnons pas une justification détaillée des deux méthodes qui sont maintenant classiques et que l'on trouvera dans [11]. Nous indiquons simplement les principes utiles pour la compréhension des aspects informatiques développés plus loin.

- La méthode du jackknife

Cette méthode procède par sous-échantillonnage exhaustif . Sur une expérimentation concernant N individus, on définit k sous-groupes de taille $(k-1)u$, où $u=\text{int}(N/k)$: 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 est réalisé en éliminant tour à tour les individus

de rangs 1 à *u*, *u*+1 à 2*u*,… (*k*-1)*u*+1 à *ku*. On peut éliminer un seul individu par sous-échantillon : *k*=*N*, *u*=1. Si u>1, les sous-échantillons doivent être représentatifs de l'ensemble de la population (c'est-à-dire de tous les niveaux de facteurs). Ceci peut être réalisé par permutation aléatoire de l'ordre de succession initial des individus. Chaque individu est caractérisé par *n* variables : y_1, y_2 ...y_n et l'on calcule sur la population un paramètre quelconque, $F(y_1,y_2,...y_n)$.
Cette fonction des observations est recalculée sur chaque sous-échantillon. L'autocorrélation positive entre les sous-échantillons, qui possèdent (*k*-2)*u* individus en commun, fait que la variance des valeurs du paramètre sous-estimerait la variance d'erreur. L'estimateur non biaisé de cette variance d'erreur (estimateur de Quenouille-Tukey) est donné par :

$$\hat{S}^2 = \frac{1}{k(k-1)}\left(\sum_{i=1}^{k} F_i^2 - \frac{\sum_{i=1}^{k} F_i^2}{k} \right)$$

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*-1)+1 à *ui* ;
$\hat{F}$ est la valeur calculée sur l'échantillon total (*ku* individus). Ces pseudo-valeurs sont des variables indépendantes et la statistique : $\frac{\hat{F} - E(F)}{\hat{S}}$ suit la distribution du **t** de Student à *k*-1 degrés de liberté.

- La méthode du bootstrap

Il s'agit d'un rééchantillonnage avec remise, qui génére des échantillons de taille *N* et inclut donc la possibilité d'avoir les mêmes données dans des échantillons différents ou dans le même échantillon. Cette méthode s'applique lorsque l'autocorrélation entre les échantilllons aléatoires générés est réduite et donc la proportion de données communes faible. Ces échantillons peuvent être considérés comme indépendants. La variance entre estimations du paramètre est alors une estimation de sa variance d'échantillonnage. Cette méthode est très utilisée en

génétique des populations car celle-ci met en œuvre une structuration simple et robuste (en général, il s'agit d'une population unique ou de hiérarchies à un ou deux niveaux). Elle est plus délicate à utiliser dans le cas de plans expérimentaux en classification croisée ou mixte (croisée et hiérarchique) pour lesquels certaines séquences de tirages avec remise peuvent générer des niveaux de facteurs déconnectés. Mais la méthode présente un avantage important : Le nombre E d'échantillons aléatoires différents possibles à partir de N individus est pratiquement infini dès que N est de quelques dizaines : $E = N^N$.
Les estimations des paramètres étant indépendantes, l'étude de leur distribution sur plusieurs milliers de séquences permet de déterminer leurs intervalles de confiance sans faire l'hypothèse d'une distribution normale.

Variantes des deux méthodes appliquées à la génétique.

Nous avons exposé le principe des deux méthodes utilisées au niveau individuel . Cette modalité suppose que l'échantillon expérimental soit représentatif de la population étudiée (ou, ce qui revient au même, que les valeurs des paramètres ne concernent que cet échantillon). Or, surtout en Génétique des Populations, il se peut que l'échantillon ne puisse être considéré comme représentatif. Le rééchantillonnage se fait alors au niveau des unités génétiques (UG) : clones, familles, populations. Voir [13] pour une discussion.Tous les individus de la même UG sont éliminés à chaque réitération. Les aspects algorithmiques sont exposés ci-dessous.

3. Le fichier de données, génération de variables dérivées

Le système de fichiers de données du logiciel est original et c'est lui qui autorise le mécanisme du rééchantillonnage mis en œuvre. Il est binaire et chaque donnée (indicatif ou observation) est représentée en simple précision sur 4 octets. Un **fichier paramètre,** suffixé par **'.p'** lui est associé. Il comporte toutes les informations utiles au traitement biométrique. La figure 1 explicite la structure de l'enregistrement et la génération de nouvelles variables.

Vecteur *X*

Indicatif 1	...Indic. k	*x*(1,1)	...*x* (1,q)	*x*(p,1)	...*x*(p,q)	..*x*(z,q)

Vecteur *Y*

Indicatif 1	...Indic. k	*y*(1,1)	...*y*(1,q')	*y*(p,1)	...*y*(p,q')	**..***y*(z,q')

Chaque enregistrement **(vecteur *X*)**, stocké en mémoire au moment de son traitement, est défini par trois paramètres : nombre d'indicatifs (*k*), nombre maximum d'individus (*z*) et nombre de variables observées par individu (*q*). Les observations (*x*) sont repérées par leur position intra-individu. L'analyseur syntaxique génère un enregistrement virtuel de même structure (**vecteur *Y***) où les ***q*** observations sont remplacées par ***q'*** fonctions *y* d'un nombre quelconque de variables *x* et/ou de valeurs de *y* déjà définies (récursivité). Les *y* sont définis sous la forme :
y(j)= F[*x*(1), *x*(2)...*y*(i), ctes]. Ainsi, le log de l'accroissement en volume d'un cône :

$$\log(\Delta V) = \log\left(\pi \frac{r_2^2 h_1 - r_1^2 h_2}{3}\right)$$

s'écrira , si r_1, h_1 (rayon et hauteur initiaux) et r_2, h_2 (rayon et hauteur finaux) sont, dans l'ordre, les quatre premiers variables : log((x3**2*x4-x1**2*x2)*pi/3).
Les données manquantes sont codées par '-9' ou '-5' selon que l'individu est mort ou simplement non observable. Tout individu dont l'une des variables *x* définissant au moins un y prend une de ces valeurs est exclu du traitement. Enfin, *n* étant le nombre d'individus de l'enregistrement, si *n* < *z*, un signal de fin logique est codé par '9999'.

Figure 1. ***Structure des enregistrements du fichier de données.***

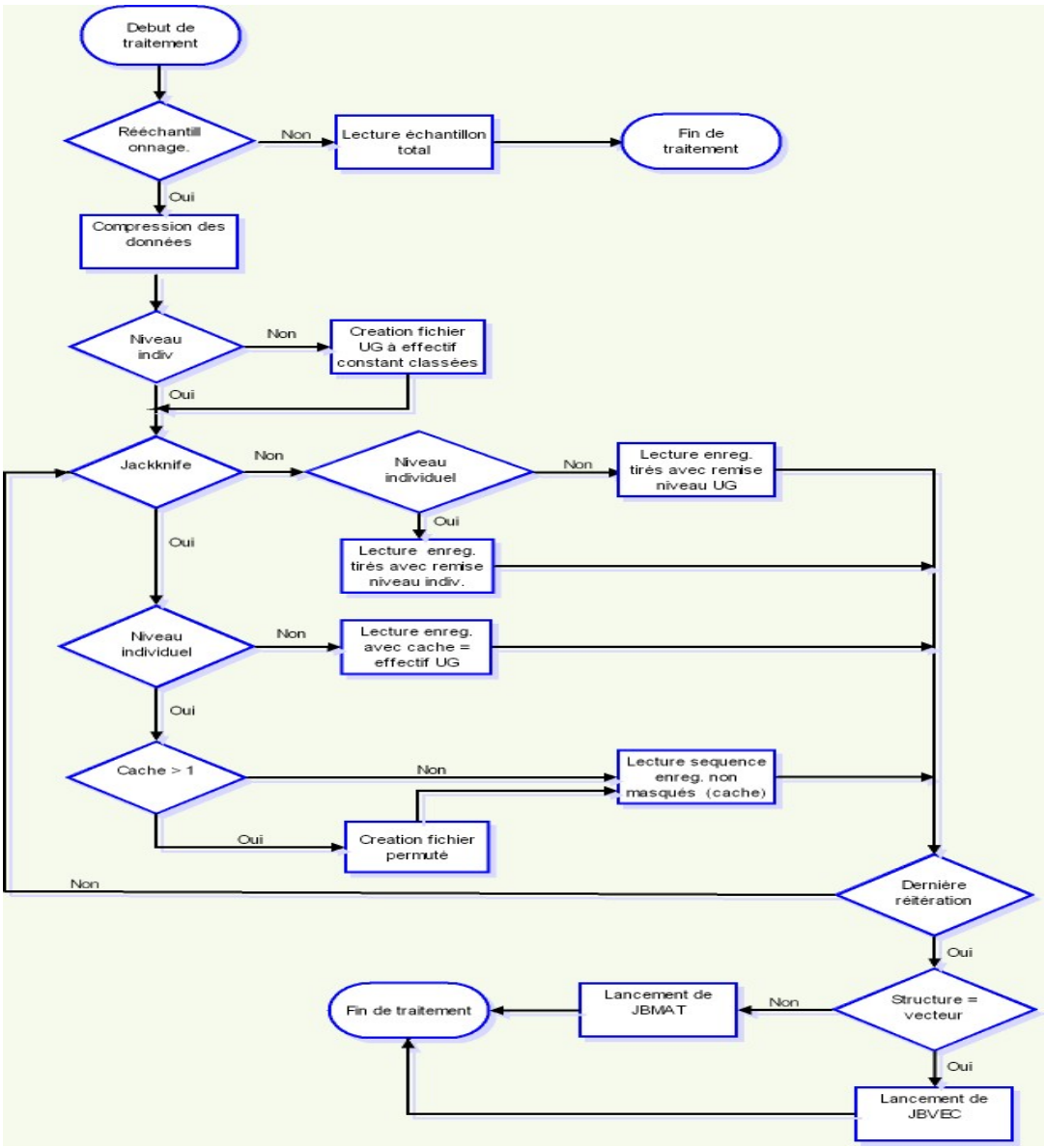

Figure 2. ***Organigramme simplifié schématisant l'implémentation du rééchantillonnage dans le logiciel DIOGENE.***

4. Aspects algorithmiques

Le système de fichier de données à accès direct permet d'éviter toute création de fichiers intermédiaires (sous-fichiers tronqués dans le cas du **jackknife**, fichiers d'individus tirés au sort avec remise dans le cas du **bootstrap**). La figure 2 donne la partie de l'organigramme correspondant au rééchantillonnage (lecture et traitement des données). En cas de rééchantillonnage, le fichier de données est compressé avant traitement, pour que toutes les variables utiles soient présentes dans chaque enregistrement : tout enregistrement incomplet est éliminé. Dans le cas du **jackknife** au niveau individuel, un cache logique masque les enregistrements retirés de l'échantillon total pour constituer chaque sous-échantillon tronqué. Pour le traitement de l'échantillon total, ce cache est rétracté et ne filtre aucun enregistrement. Sa position est actualisée à chaque réitération. Si sa valeur est supérieure à 1, un fichier permuté est substitué au fichier d'origine. Dans le cas du **bootstrap**, le tirage au sort des individus adresse directement les enregistrements correspondants.

Le rééchantillonnage au niveau des unités génétiques (UG) peut être réalisé dans le cadre du **bootstrap** comme dans celui du **jackknife**. Deux utilitaires créent alors un fichier à effectif/UG constant classé par code d'UG croissant. Le nombre d'individus/UG, **n**, permet de calculer le rang du premier individu de la série d'enregistrements correspondant à une UG. Tout tirage pointe sur cet individu. Les enregistrements des **n-1** individus suivants sont lus en séquence. La fin du traitement dépend du type de structure considéré (matrices triangulaires-basses ou vecteurs).

5. Conclusion

L'implémentation du rééchantillonnage qui vient d'être esquissée conduit à une accélération spectaculaire de la vitesse de traitement par rapport aux méthodes traditionnelles qui font appel à une génération de fichiers.

Un PC de bureau HP pavilion w5157.fr avec processeur Intel Dual Core à 2,8 GHZ peut réaliser 2729 réitérations en 1' 30'' sur une analyse diallèle multivariable non orthogonale avec 12 parents et 6 variables (jackknife au niveau individuel avec cache de 1). Le même traitement utilisant une génération de fichiers dure plusieurs heures avec le même matériel.

Références

1] Agwanda C. O., Baradat P., Eskes A. B., Cilas C. and Charrier A.. Selection for bean and liquor qualities within related hybrids of Arabica coffee in multilocal field trials, *Euphytica*, 131: 1-14, 2003.

[2] Baradat P. et Desprez-Loustau M. L.. Analyse diallèle et intégration dans le programme d'amélioration du pin maritime de la sensibilité à la rouille courbeuse, Ann Forest Sci, 54: 83-106., 1997.

[3] Baradat P. et Labbé T., OPEP. Un logiciel intégré pour l'amélioration des plantes. *Traitement statistique des essais de sélection* : 303-330, CIRAD Ed., Montpellier, 1995.

[4] Baradat P., Labbé T. et Bouvet J. M., Conception d'index pour la sélection réciproque récurrente. Aspects génétiques, statistiques et informatiques. *Traitement statistique des essais de sélection*.: 101-150, CIRAD Ed., Montpellier, 1995.

[5] Baradat P., Perrier T. et Raffin A., Papadakis++. Un ajustement multi-dimensionnel du Phénotype à l'Environnement. Document UMR AMAP, Montpellier, 2004.

[6] Besnard G., Baradat P. and Bervillé A.. Genetic relationships in the olive (*Oleuropaea* L.) reflect multilocal selection of cultivars, Theor Appl Genet, 251: 258. 83-106, 2001.

[7] Besnard G., Baradat P., Breton C., Khadari B., Bervillé A., Olive domestication from structure of oleasters and cultivars using nuclear RAPDs and mitochondrial RFLPs, Genet. Select. Evol**.**, 33 : 251-268, 2001.

[8] Efron B., The Jackknife, the bootstrap and other resampling plans. Society for industrial and applied. mathematics, Philadelphie, 1982.

[9] Efron B., and Gong G.. A leisurely look at the bootstrap, the jackknife and cross-validation, Amer. Stat., 37: 36-48, 1983.

[10] Legendre P. and Legendre L.., *Numerical Ecology*, 2rd Edition, Elsevier, Amsterdam , 1998.

[11] Shao J. and Tu D., *The Jackknife and Bootstrap*, Springer, 1995.

[12] Sokal R. R. and Rohlf F. J., *Biometry*, 3rd Edition, Freeman, New York, 1995.

[13] Weir B. S., *Genetic Data Analysis*, Sinauer, Sunderland, 1990.