

Statistical analysis of forest genetic experiments. Some key points.

Catherine Bastien

▶ To cite this version:

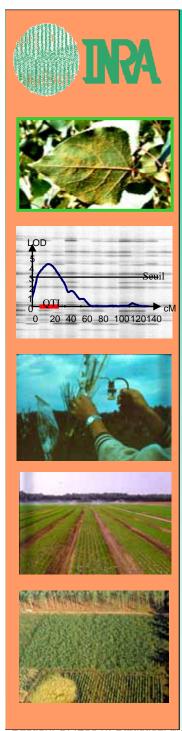
Catherine Bastien. Statistical analysis of forest genetic experiments. Some key points.. PROFOREST Workshop on "New approaches in forest tree genetics", Aug 2004, Varsovie, Poland. hal-02759762

HAL Id: hal-02759762 https://hal.inrae.fr/hal-02759762v1

Submitted on 4 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Statistical analysis of forest genetic experiments Some key points

Catherine Bastien
INRA, UAGPF Orléans

PROFOREST Workshop, Warsaw, 24-27 August 2004



Objectives of forest genetic field experiments

- Comparison of different populations of a given species for quantitative and qualitative traits expressed in forest conditions: provenance tests
- Genetic evaluation in forest conditions of phenotypical selections:
 progeny tests (« + » trees open-pollinated progenies, polymix
 progenies, controlled crosses), clonal tests, multisite experiments
- Backward selection in clonal seed orchard on multitrait evaluation in forest conditions of phenotypical selections: progeny tests (« + » trees open-pollinated progenies, polymix progenies, controlled crosses)
- Forward selection on multitrait evaluation in forest conditions for long-term breeding strategies: progeny tests (« + » trees openpollinated progenies, polymix progenies, controlled crosses)
- Evaluation of genetic variability of natural and artificial populations for *quantitative* and *qualitative* traits expressed in forest conditions : progeny tests



forest genetic field experiments

Genotype: provenance, progeny-family, clone

A basic common model: Fisher (1918)









Interaction between genotype and environment

Fixed situation

 Precise estimation of genotypic values and genotype stability over a given set of environmental conditions

Random situation Precise estimation of genetic and GxE variances in a multitrait context



forest genetic field experiments

Prediction of G_i values in a given experiment

$$\mathbf{P_{ij}} = \mathbf{G_{i}} + \mathbf{B_{lock}} + \mathbf{R_{ij}}$$

Controlled experimental variation



To maximize for a better control of environmental variation

Residual Uncontrolled variation

To minimize for maximum precision (experimental designs)

* Complete or incomplete block design with single or multitree plots



forest genetic field experiments

- 1 Test and adjustment for local environmental effects:
 - Efficiency of block designs
 - Correction with spatial analysis: Papadakis iterative method





forest genetic field experiments



Prediction of Breeding Values Ai before genetic thinning in clonal seed orchards

Fisher's key insights: Each individual pass to its offspring a fraction of its genetic value which at a minimum is equal to ½ genetic additive value A

Evaluation criteria

Breeding objective

Own performance P₁

Performance of offspring P₂

Correlated Traits P_n

Molecular markers M_n

Breeding value A_i



Multiple linear regression

$$A = b_1P_1 + b_2P_2 + ... + b_nP_n + ... + c_nM_m$$



forest genetic field experiments

- 1 Test and adjustment for local environmental effects:
 - Efficiency of block designs
 - Correction with spatial analysis: Papadakis iterative method
- 2 <u>Estimation of breeding values</u>
 - BLUP's
 - variance components estimation

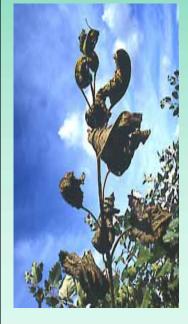


forest genetic field experiments

Multitrait selection and economic weights of the different selection objectives

Adaptation (biotic & abiotic factors)

Volume production







Stem quality

Wood quality



Selection Index : $I = a_1G_1 + a_2G_2 + ... + a_nG_n$



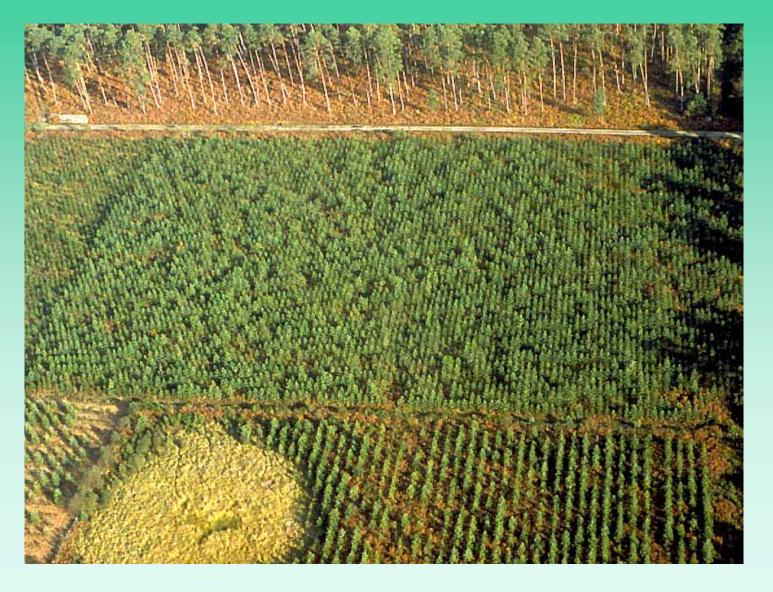
forest genetic field experiments

- 1 Test and adjustment for local environmental effects:
 - Efficiency of block designs
 - Correction with spatial analysis: Papadakis iterative method
- 2 <u>Estimation of breeding values</u>
 - BLUP's
 - variance components estimation
- 3 Multi-trait selection
 - Prediction of response to selection
 - Independent Culling vs. Index
 - Economic vs technical weights in selection index



forest genetic field experiments

Control of environmental variation





forest genetic field experiments

Control of environmental variation by block effects

Example : analysis of total height of a clonal test in a 6 complete block design

ANOVA Table 2003 Total Height

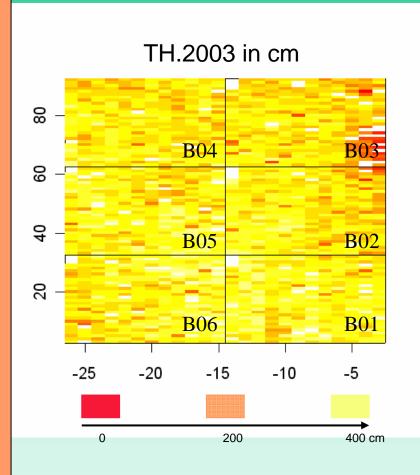
	Df	SSE	MS	F-test	P-value
Bloc	5	603119	120624	72.36	0.000
Genotype	354	1988304	5617	3.3692	0.000
Residuals	1718	2864033	1667		

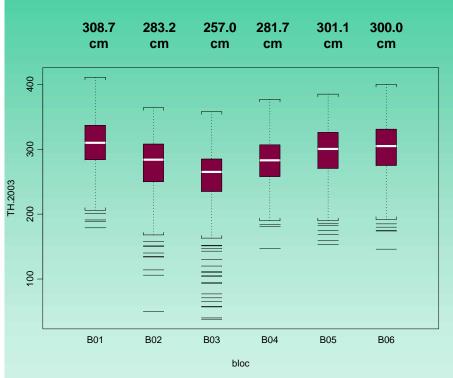
Strong block effects! $CV_r = 14.1\%$



Control of environmental variation by block effects

Example : analysis of total height of a clonal test in a 6 complete block design

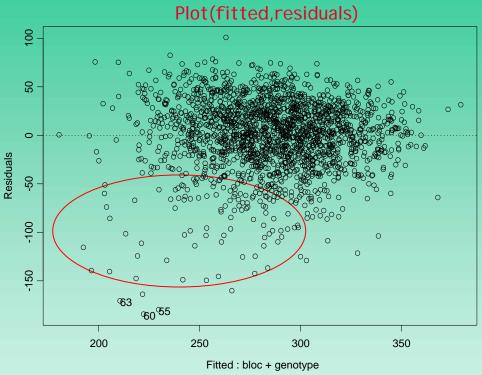




block effects will control part of environmental variation. What does remain?



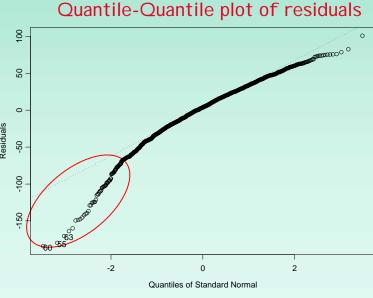
Control of environmental variation by block effects Example: analysis of total height of a clonal test in a 6 complete block design



Analysis of residual variation

o A lot of plants with relative low height

transplantation effect? local environmental effects?

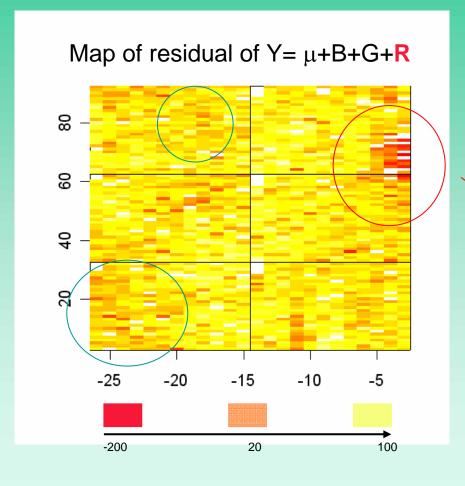


Analysis of microsatellite sequences in Scots pine (p. 145-158). Presented at PROFOREST Workshop on "New approaches in forest tree genetics", Varsovie, POL (2004-08-24 - 2004-08-27). Varsovie, POL: Forest Research Institute.



Control of environmental variation by block effects Example: analysis of total height of a clonal test in a 6 complete block design

spatial distribution of residual variation



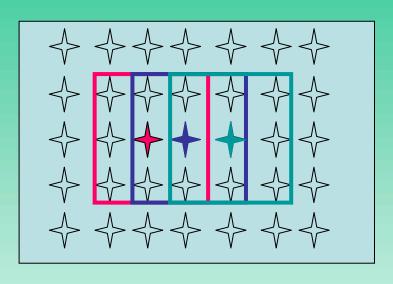
Environmental variation still exist within block!





Control of environmental variation by spatial analysis Example: analysis of total height of a clonal test in a 6 complete block design

Papadakis iterative method



Environmental variation is measured by the neighborhood residual information (Ψr)

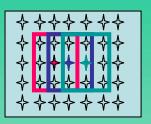
$$\begin{split} \mathbf{P_{ij}} &= \mu + \mathbf{G_i} + \mathbf{b} \; \mathbf{E}(\Psi \mathbf{r}) + \mathbf{R'_{ij}} \\ \mathbf{E}(\Psi \mathbf{r}) &= \sum_{i'j'} \; \mathbf{R'_{i'j'}} \; / \; \mathbf{n_{(r)}} \\ \mathbf{P'_{ij}} &= \mathbf{P_{ij}} - \mathbf{b} \; \mathbf{E}(\Psi \mathbf{r}) \end{split}$$



I terative procedure



Control of environmental variation by spatial analysis Example: analysis of total height of a clonal test in a 6 complete block design



Papadakis iterative method

Neighborhood: 5 trees x 9 trees

ANOVA Table 2003 on Total Height corrected by Papadakis

Df	SSE	MS	F-test	P-value
Bloc 5	7023	1405	1.0711	0.3745
Genotype 354	1861064	5257	4.0089	<0.0001
Residuals 1718	2252986	1311		

Reduced residual variation

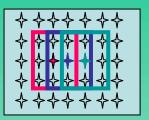
 $CV_r = 12.8\%$



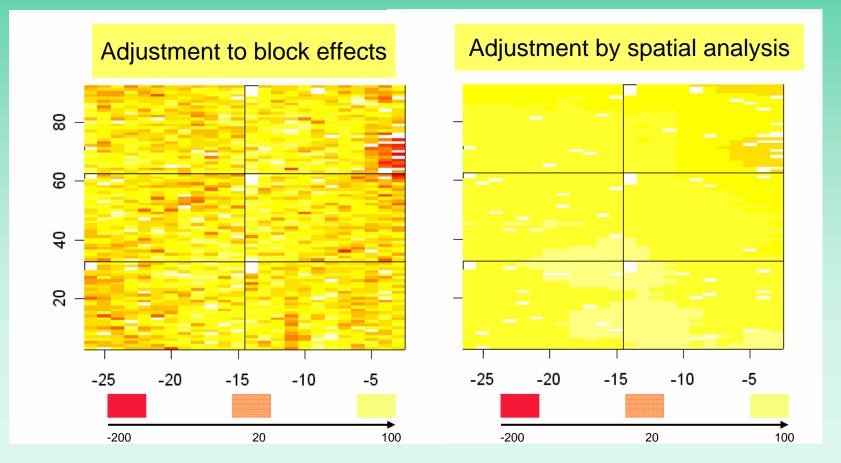
for local environmental effects

Adjustment

Control of environmental variation by spatial analysis Example: analysis of total height of a clonal test in a 6 complete block design



Papadakis iterative method



Final choice: adjustment by spatial analysis and elimination of five rows in block 03

Analysis of microsatellite sequences in Scots pine (p. 145-158). Presented at PROFOREST Workshop on "New approaches in forest tree genetics", Varsovie, POL (2004-08-24 - 2004-08-27). Varsovie, POL: Forest Research Institute.



Control of environmental variation by spatial analysis

Papadakis iterative method

Kempton RA and Howes CW 1981. The use of neighbouring plot values in the analysis of variety trials. Applied Statistics 30 (1), 59-70

Dagnélie P. 1989. The method of Papadakis in Agricultural Experimentations. An overview Bulletyn Oceny Odmian, 21-22, 111-122.

Besag J and Kempton R. 1986. Statistical Analysis of Field Experiments Using neighbouring plots. Biometrics 42, 231-251.

Bartlett MS. 1978. Nearest neighbour models in the Analysis of Field Experiments.J.R. Statist. Soc. 2, 147-174.



forest genetic field experiments

Estimation of breeding values and phenotypic variance components















forest genetic field experiments



Prediction of Breeding Values A_i before genetic thinning in clonal seed orchards

Finding the optimal regression coefficients b_n

$$A = b1P1 + b2P2 + ... + bnPn$$

$$Y = f(X) = b X$$

$$b = \frac{cov(X,Y)}{var(X)}$$

BLUP = **B**est linear unbiased **p**rediction

Evaluation criteria

Own performance P₁

$$\mathbf{b_1} = \text{Cov}(P_1, A) / \text{var}(P_1)$$

$$b_1 = Cov(A + D + A + A) / var(P_1)$$

$$\mathbf{b_1} = \mathbf{Cov}(\mathbf{A}, \mathbf{A}) / \mathbf{var}(\mathbf{P_1})$$

$$b_i = h^2$$

Analysis of microsatellite sequences in Scots pine (p. 145-158). Presented at PROFOREST Workshop on "New approaches in forest tree genetics", Varsovie, POL (2004-08-24 - 2004-08-27).



Prediction of Breeding Values A_i before genetic thinning in clonal seed orchards

Evaluation criteria

Open pollinated progeny performance P₂

$$A = b_2 P_2$$

n = nb. ind in progeny

Var(mean) = common variance + specific / n

$$\mathbf{b_2} = \text{Cov}(P_2, A) / \text{var}(P_2)$$

$$b_2 = \frac{1}{2} V_A / var(P_2)$$

$$b_2 = \frac{1}{2} V_A / (V_{Fam} + V_{resid} / n)$$

Heritability of progeny test

$$\mathbf{b_2} = \mathbf{h^2}_{Fam} = \frac{2n}{n + 4 - h^2}$$

b₂ depends on the number of progeny and on the heritability



Prediction of Breeding Values A, before genetic thinning in clonal seed orchards

Evaluation criteria

Open pollinated progeny performance P₂

k traits measured

$$A = b_2^{1} P_2^{1} b_2^{2} P_2^{2} + ... + b_2^{k} P_2^{k}$$

 M_P = matrix of phenotypic variances-covariances

 M_A = matrix of additive genetic variances-covariances

$$b_2 = \frac{1}{2} M_A M_P^{-1}$$

b₂ for A_{Total height age 15}

Total height age 15

0.562

Total height age 15
Total height age 10
Girth age 15
Branch angle age 10

0.714

From Bastien 1999, unpublished data



Prediction of Breeding Values A_i before genetic thinning in clonal seed orchards

- Efficiency of BLUP estimation proved in many animal and plant breeding programs
- BLUP estimation is always superior to phenotypical selection on progeny means
- Measuring correlated traits could increase significantly precision of breeding values estimation
- BLUP could be easily calculated with all softwares including linear model predictions [SAS, ASREML, Splus,....]
- BLUP needs only accurate estimation of M_A (heritabilities and additive genetic correlations)



Estimation of variance components

- Two key statistical ANOVA identities
 - Total variance = between-group variance (V_{Fam}) + within-group variance (V_W)
 - Variance(between groups) = covariance (within groups)
- One key genetic property of Fisher model (Kempthorne 1957)

X and Y, two individuals

$$Cov(X, Y) = 2 r_{XY} V_A + u_{XY} V_D$$

In practice

Open-pollinated progenies collected randomly in most Scots pine stands could be considered as a random sample of half-sib progenies

$$V_{Fam} = Cov (HS)$$

 $V_{Fam} = V_A / 4$



4 V_{Fam} gives an estimation of V_A



Estimation of variance components according to the experimental design



Estimation of variance components

Two methods

Expected means squares of
Analysis of Variance
(ANOVA)
Henderson III

Restricted maximum likelihood estimation (REML)

- Independent estimation of fixed and random effects
- Biaised estimation in case of non-orthogonal (unbalanced) designs
- difficulty to analyze jointly variety of relatives

- •Simultaneous estimation of fixed and random effects
- no demand on design or balance of data
- no demand on design or balance of data
- now available in most statistical softwares



Estimation of variance components

Example : analysis of total height and branch angle in a Scots pine progeny test

Expected means squares of Analysis of Variance (ANOVA)

ANOVA Table on Total Height adjusted to block effects Y'= Y - Block

		Df	SSE	MS	E(MS)
Fixed	Bloc	41	100978	2463	$V_R + k\phi_{bloc}$ $V_R + nV_{Fam}$
Random	Genotype	64	1039806	16247	$V_R + nV_{Fam}$
Random	Residuals '	1935	7260365	3752	V_R

Average n = 31.4 trees per progeny

$$\hat{V}_R = 3752$$
 $\hat{V}_{Fam} = (16247 - 3752)/31.4 = 397.9$
 $\hat{V}_A = 4*\hat{V}_F = 1591.7$

$$\hat{\mathbf{h}}^2 = 1591.7 / (397.9 + 3752) = 0.383$$

Analysis of microsatellite sequences in Scots pine (p. 145-158). Presented at PROFOREST Workshop on "New approaches in forest tree genetics", Varsovie, POL (2004-08-24 - 2004-08-27). Varsovie, POL : Forest Research Institute.



Estimation of variance components

Example : analysis of total height and branch angle in a Scots pine progeny test

Expected means squares of Analysis of Variance (ANOVA)

MANOVA Total Height , Branch angle adjusted to block effects Y'= Y - Block

		Df	SCPE	MCP	E(MCP)
Fixed	Bloc	41	-232.27	-5.66	$Cov_R + k\phi_{bloc}$ $Cov_R + nCov_{Fam}$
Random	Genotype	64	722.97	11.30	Cov _R + nCov _{Fam}
Random	Residuals '	1911	2423.75	1.27	Cov _R

Average n = 31.2 trees per progeny

$$Cov_R = 1.27$$
 $Cov_{Fam} = (11.30-1.27)/31.2 = 0.32$
 $Cov_A = 4*Cov_F = 1.28$
 $r_A^{\circ} = 1.28 / \sqrt{(1591.7 *0.45)} = 0.047$

Analysis of microsatellite sequences in Scots pine (p. 145-158). Presented at PROFOREST Workshop on "New approaches in forest tree genetics", Varsovie, POL (2004-08-24 - 2004-08-27).



Estimation of variance components

Example : analysis of total height and branch angle in a Scots pine progeny test

Restricted maximum likelihood estimation (REML)

Model: $Y_{adj} = \mu + B_{lock} + F_{amily} + R_{esidual}$

Fixed Random Random

	Total Height	Branch angle
V _{Fam}	411.86	0.116
Sd(V _{Fam})	95.21	0.024
V_R	3753	0.609
Sd(V _R)	29.33	0.020
h ²	0.395	0.640
Sd(h²)	0.085	0.051

Precision of variance component estimations depends on nb. of progenies



Estimation of variance components

Example : analysis of total height and branch angle in a Scots pine progeny test

Restricted maximum likelihood estimation (REML)

Model:
$$Y_{adj} = \mu + B_{lock} + F_{amily} + R_{esidual}$$

Fixed Random Random

Estimation of Covariance components

$$Cov(X+Y) = V_X + V_Y + 2 Cov(X,Y)$$

$$Cov(X,Y) = \frac{1}{2} (Cov(X+Y) - (V_X + V_Y))$$

Total Height-Branch angle

Cov_{Fam} -0.120

 Cov_R 3753

r_A -0.017



Multi-trait selection

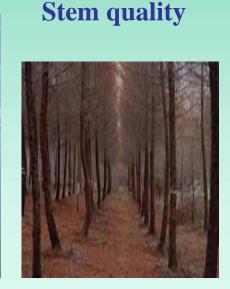
3- Multitrait selection and economic weights of the different selection objectives

Adaptation (biotic & abiotic factors)

Volume production







Wood quality



Selection Index : $I = a_1G_1 + a_2G_2 + ... + a_nG_n$



forest genetic field experiments Prediction of response to selection



(natural stands, provenance tests)





Seed collection



Multisite evaluation progeny testing



Clonal Collection for recombination







Forward selection Genetic thinning in seed orchard



Realized genetic gain





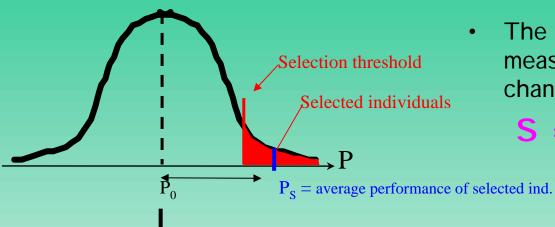
selection

to

Response

forest genetic field experiments

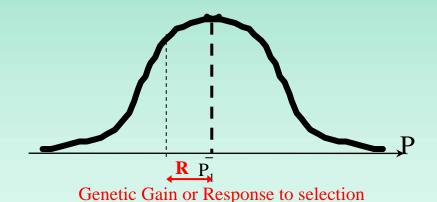
Prediction of response to selection



The selection differential S measures the within-generation change in the mean

$$S = P_s - P_0$$

Recombination of selected individuals

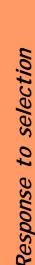


The response R is the betweengeneration change in the mean

$$R = P_1 - P_0$$

Analysis of microsatellite sequences in Scots pine (p. 145-158). Presented at PROFOREST Workshop on "New approaches in forest tree genetics", Varsovie, POL (2004-08-24 - 2004-08-27).

Varsovie, POL: Forest Research Institute.





forest genetic field experiments

Prediction of response to selection

 Selection can change the distribution of phenotypes. We typically measure this by changes in mean.

This is a within-generation change measured by $S = P_s - P_0$

 Selection can also change the distribution of breeding values (changes in allele frequencies).

This is a the response to selection, the change in the trait in the next generation (between-generation change) measured by

$$R = P_1 - P_0$$



selection

to

Response

Prediction of response to selection The Breeder's Equation

 $R = h^2 S$

- Note that no matter how strong S, if h² is small, the response is small
- S is a measure of selection, R the <u>actual</u> response. One can get lots of selection but no response

Applications

- In agriculture and forestry breeding
- Construction of divergent pedigree for QTL mapping and gene expression (microarray) analysis: inferences about nb. Of loci, effects and frequencies
- Evolutionary inferences: correlated charcaters, effects on fitness, long-term response



selection

to

Response

Prediction of response to selection

The Selection Intensity, i

Populations with the same selection differential (S) may experience very different amounts of selection. The selection intensity i provided a suitable measure for comparisons between populations,

$$i = \frac{s}{\sigma_p}$$

$$R = h^2 S = i h^2 \sigma_p = i h \sigma_A$$

Since h = correlation between phenotypic and breeding values

Response = Intensity * Accuracy * Spread in V_A



Response to selection

Prediction of response to selection

The correlated response

Selection on Trait 1, predicting response of Trait 2

$$R_2 = i_1 r_{A_{1,2}} h_1 h_2 \sigma_{p2}$$



to selection

Response

Prediction of response to selection

A general formulation X = trait selected

Y= trait measured

$$R = i \rho \sigma_{Ax}$$

$$\rho = 2. \text{ r. } r_{Ax,y}. h'_{Y}.\sqrt{(n/[1+(n-1)t])}$$

Ollivier,2002

r = coancestry coefficient between candidate and ind. measured (OP progeny \rightarrow parent-offspring \rightarrow r=1/4)

 $\Gamma_{AX,Y}$ = genetic correlation between X and Y if different

 \mathbf{h}'_{v} = heritability of the selection criterion (ind. Values, progeny means)

n = nb. of measures on the candidate (nb. offspring per parent)

t = correlation between observations on the same candidate (OP progeny \rightarrow h² / 4)



Response to selection

Response to selection with progeny testing

Forward selection

$$\rho = 2. \text{ r. } r_{Ax,y}. h'_{Y}.\sqrt{(n/[1+(n-1)t])}$$

Ollivier,2002

	Response X	Response Y			
r	1⁄4	1/4			
$r_{Ax,y}$	1	r _{AX,Y}			
h' _Y	h _X	h _X			
n = nb. of measures on the candidate (nb. offspring per parent)					
t	h ² _X /4	$h_{x}^{2}/4$			





Response to selection with progeny testing

$$R_X = 0.5 i h_X . \sqrt{(n/[1+(n-1)h_x^2/4])} \sigma_{Ax}$$

$$R_{Y/X} = 0.5 i r_{AX,Y} h_{X} . \sqrt{(n/[1+(n-1)h_{X}^{2}/4])} \sigma_{AY}$$

Response X Total height

Response Y Branch angle

h²

0.395

0.640

$$\mathbf{O}_{A}^{2}=4^{*}\mathbf{O}_{Fam}^{2}$$
 1647.44

0.464

Forward selection on X

Phenotypic selection on X

$$R_{X = 62.4 \text{ cm}}$$

 $R_{Y/X = -0.43}$

$$R_{X = 44.8 \text{ cm}}$$

 $R_{Y/X = -0.16}$



Response to selection with progeny testing

Forward selection

$$R = 2 h_f^2 S_f = 2 i h_f^2 \sigma_{Pf}$$

Selection on 2 parents (male and female)

$$h_f^2 = \frac{\sigma_f^2}{\sigma_{f+}^2 \sigma_R^2/n}$$

$$\sigma^2_{Pf} = \sigma^2_{f} + \sigma^2_{R}/n$$

Response X
Total beight

Total height

 h^2 0.395

 $\sigma_{\rm Pf}^2 = 537$

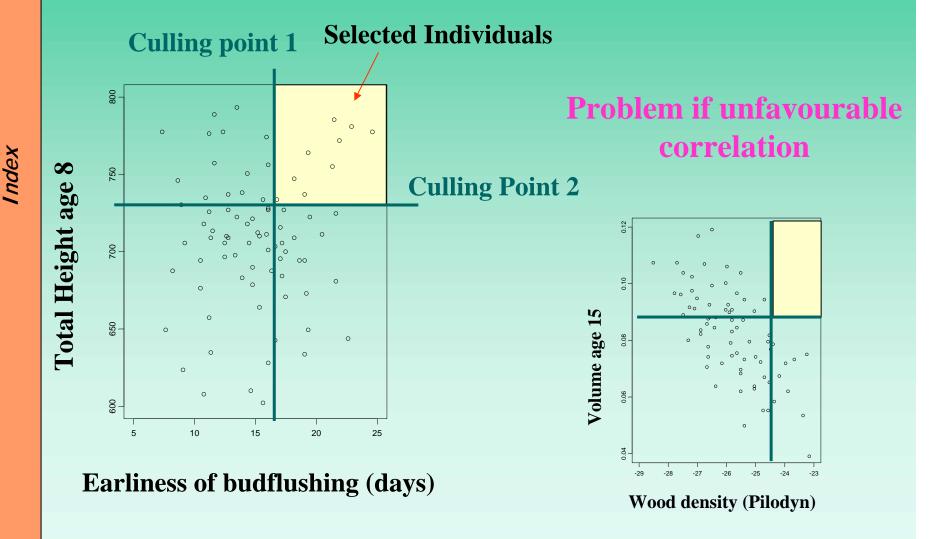
Rx = 62.4

Response to selection



Multi-trait selection

Independent Culling



Analysis of microsatellite sequences in Scots pine (p. 145-158). Presented at PROFOREST Workshop on "New approaches in forest tree genetics", Varsovie, POL (2004-08-24 - 2004-08-27).

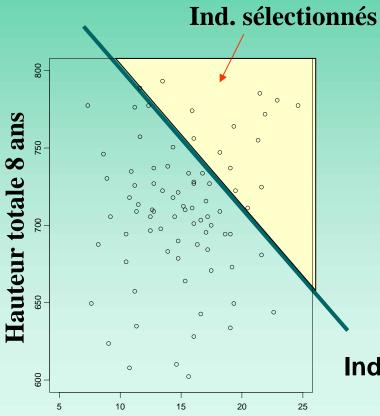
Varsovie, POL: Forest Research Institute.



Multi-trait selection Index Selection

$$I = w_1 A_1 + w_2 A_2 + \dots w_q A_q$$





Disadvantage in one trait off set by advantage in the other

Index line

Débourrement (jrs)



Index Selection vs. Independent Culling

Theoretical comparisons

Practical considerations

If same total of nb. of individuals measured on all traits:

genetic gain

Index S > Independent Culling > Tandem selection

•Index selection:

-must keep all individuals until all traits measured

-cull in one stage

•Traits differ greatly in costs to measure

- Traits differ greatly in age of evaluation
- •Selection intensity may be greater for multistage (culling) selection

Analysis of microsatellite sequences in Scots pine (p. 145-158). Presented at PROFOREST Workshop on "New approaches in forest tree genetics", Varsovie, POL (2004-08-24 - 2004-08-27).

Varsovie, POL: Forest Research Institute.



Index Selection

$$I = w_1 A_1 + w_2 A_2 + \dots w_q A_q = [w'A]$$

w = vector of technical or economical weights

$$I = b_1 P_1 + b_2 P_2 + \dots b_q P_q = [b'P]$$

b = vector of weights for phenotypic predictors

BLUP properties : $A = M_P^{-1} M_A Z_{centered}$

$$b = [M_P^{-1} M_A w]$$

Example $σ_P$ h^2 r_A w b Wood density 0.4 0.3 0.5 5 → 0.53 Volume 0.2 0.5 0.5 -1 -0.31

In general, weights on phenotypic information sources are not esay to « recognize »



Response to Index Selection

$$I = w_1 A_1 + w_2 A_2 + \dots w_q A_q = [w'A]$$

w = vector of technical or economical weights

$$R = R_2 = \frac{i \ W' \ M_A}{\sqrt{W' \ M_P \ W}}$$

$$R_k$$



Response to Index Selection

Example: HUMPTULIPS Population

$$I = w_1$$
 BudFlush $+ w_2$ TH $+ w_3$ Ang $+ w_4$ Br $+ w_5$ Def

Estimation of maximum relative genetic expected gains

W	BudFlus	sh TH	Ang	Br	Def
(-1,0,0,0,0)	-55%	8.9%	1.8%	12%	-19%
(0,1,0,0,0)	-24.4%	20%	-0.5%	-0.4%	9.9%
(0,0,1,0,0)	-4.4%	-0.4%	22.5%	11.0%	-3.9%
(0,0,0,10)	-32.2%	-0.4%	12.1%	20.5%	-16.3%
(0,0,0,0,-1)	-34%	-6.4%	2.8%	10.8%	-31%

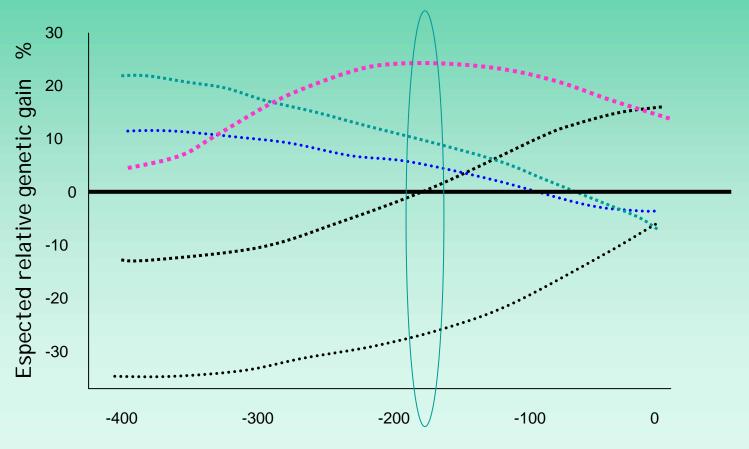


Response to Index Selection

Simulation of expected genetic gains with varying w

Example: HUMPTULIPS Population

$$I = w_1$$
 BudFlush $+ w_2$ TH $+ w_3$ Ang $+ w_4$ Br $+ w_5$ Def



Coefficient w5 with w1=0 w2=5 w3=2 w4=2



$$I = w_1 A_1 + w_2 A_2 + \dots w_q A_q = [w'A]$$

w = vector of technical or economical weights

$$A = \frac{1}{2} M_A M_{Pf}^{-1} P$$

$$I = \frac{1}{2} w' M_A M_{Pf}^{-1} P$$

$$\sqrt{W' M_{PFam} W}$$

$$\sigma^{2}_{PFam} =$$

$$\sigma^2_{PFam} = \sigma^2_{Fam} + \sigma^2_{R}/n$$

Total height

Branch angle

$$M_{Fam} = 411.86 -1.451 -1.451 0.116$$

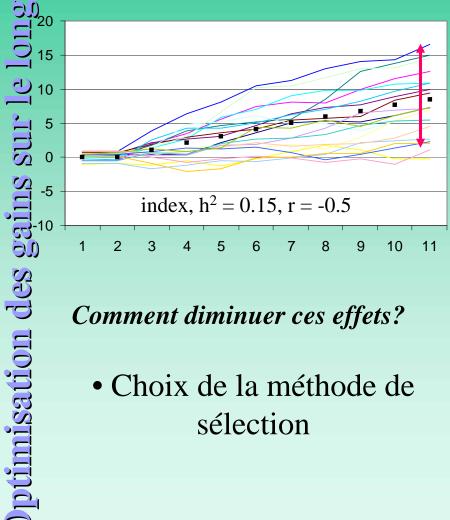
Total height

Branch angle

$$M_{PFam} = 537 -1.528$$
 $-1.528 0.136$

Sélection multi-caractères et Liaisons génétiques défavorables

Quels sont les effets?



- •Augmentation de la variation du progrès génétique (imprévisibilité)
- Perte du mérite général

Comment diminuer ces effets?

• Choix de la méthode de sélection

