P++ models (*DIOGENE* software) for adjustment to environmental effects

Applications in Genetics. Interest outside of Genetics?

Ph. Baradat (old retired researcher)

INTRODUCTION

- The basic Nearest Neighbor model was initially designed to adjust data at the level of experimental plots (Papadakis 1984, Dagnélie 1987 & 1989, Pichot 1993).
- This model belongs to the '**ARMA**' category (AutoRegressive Moving Average).
- It can be considered as a generalization of an adjustment using control plots (Dagnélie 1987)
- **Reiteration** of adjustment uses a symmetrical processing of neighbor plots (Bartlett 1978, Besag 1983, Azais *et al.* 1990, Goumari 1990).
- Use of competition between adjacent plots was proposed (Besag & Kempton 1986).
- Kempton and Howes (1981) proposed a model using both regression on the nearest neighbors and a block effect, an approach that we also consider.
- Other **reiterated** methods, such as kriging, were applied to control common environment effects for more accurate heritability estimation (Zas 2006).
- At the individual level (Pichot 1993), the method uses as covariate in a simple linear regression the mean of residuals of neighbors for the same variable (one-way ANOVA where the considered factor is the genetic entry).

First version of the multivariate model described below was used by Bertrand (2002) on *Coffea arabica* genetic trials. It resulted in an increase of heritability and reduced confidence intervals for production and quality traits.

The model was further improved and extended up to the present version. It can be applied to **multilocal trials** for simultaneously adjusting several traits across sites.

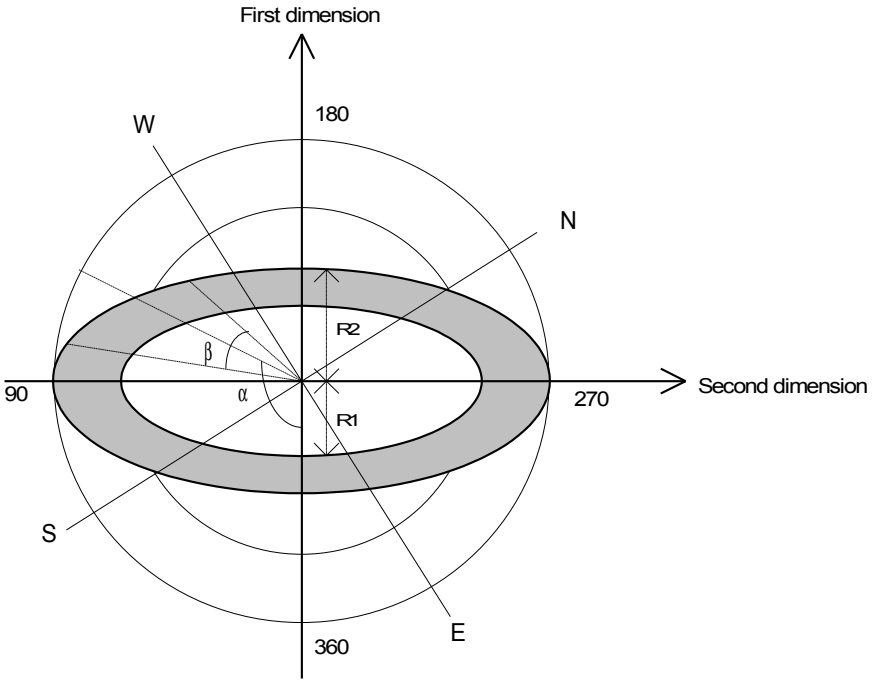
Models description

Single-site model

The basic model is a multiple regression of an observed value for a given individual ("pivot"), located by (x, y) coordinates on the mean residuals for p variables of the surrounding neighbors within a structure, ψ_r defined below.

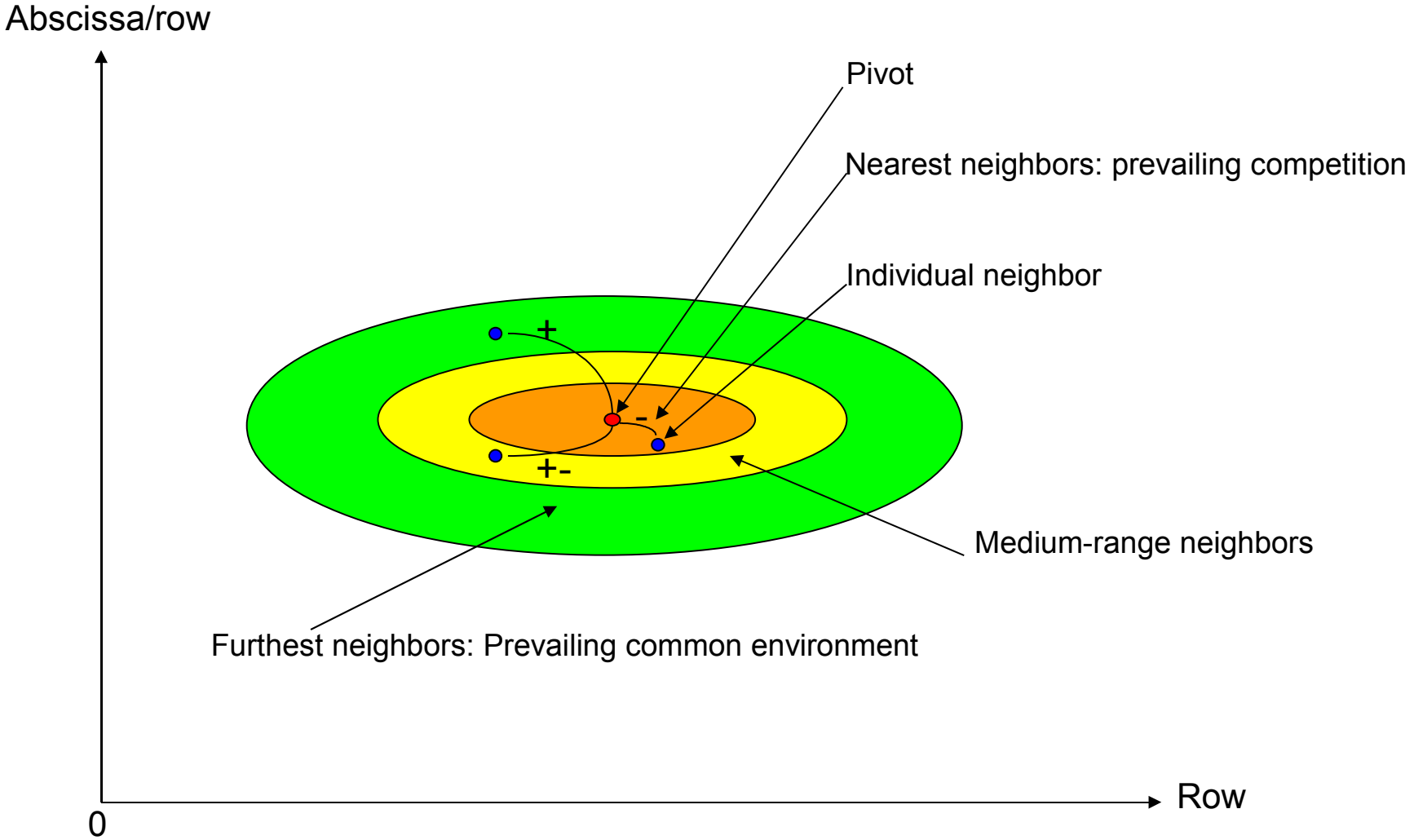
$$Y_{ij(xy)} = \mu + G_i + b_1\overline{E}^1(\psi_r) + b_2\overline{E}^2(\psi_r) + \ldots + b_p\overline{E}^p(\psi_r) + E_{ij(xy)} \quad (1)$$

where G_i if the effect of the genetic unit (clone, family etc.), $\overline{E}^u(\psi_r)$ the mean residual for trait *u* in the relative neighborhood configuration, and $E_{ij(xy)}$ the residual of the pivot.

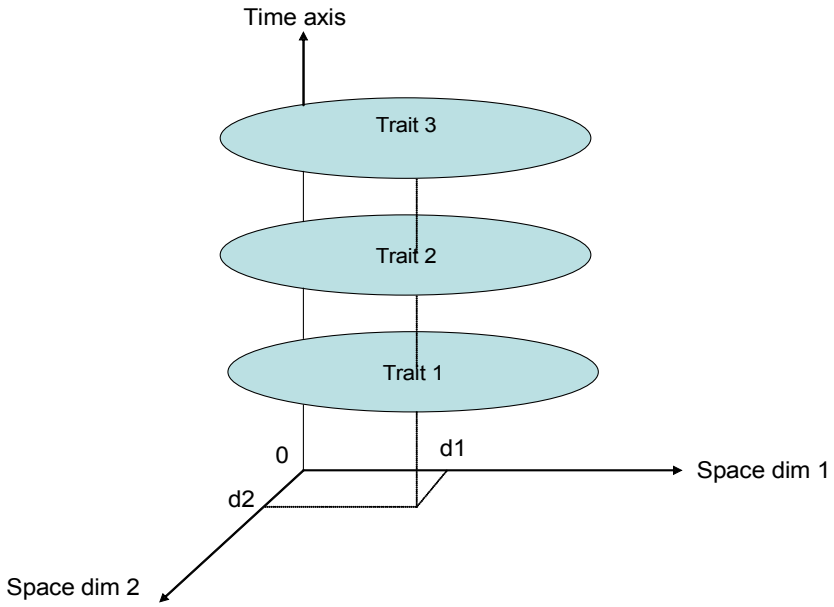


Determination of neighborhood structures (or stitches) surrounding each individual.

A group of neighbors is located at the intersection of angle β with the shaded area which (crown of ellipse). The ellipse center represents the pivot individual.
- γ is the flatness coefficient of the ellipse
- R1 is its minimum radius following the first dimension (plantation rows)
- R2 is its maximum radius in the same direction
- α is the orientation of the bisecting line of the crown sector relatively to the base of the plantation rows and β is the opening angle of the ellipse crown sector.

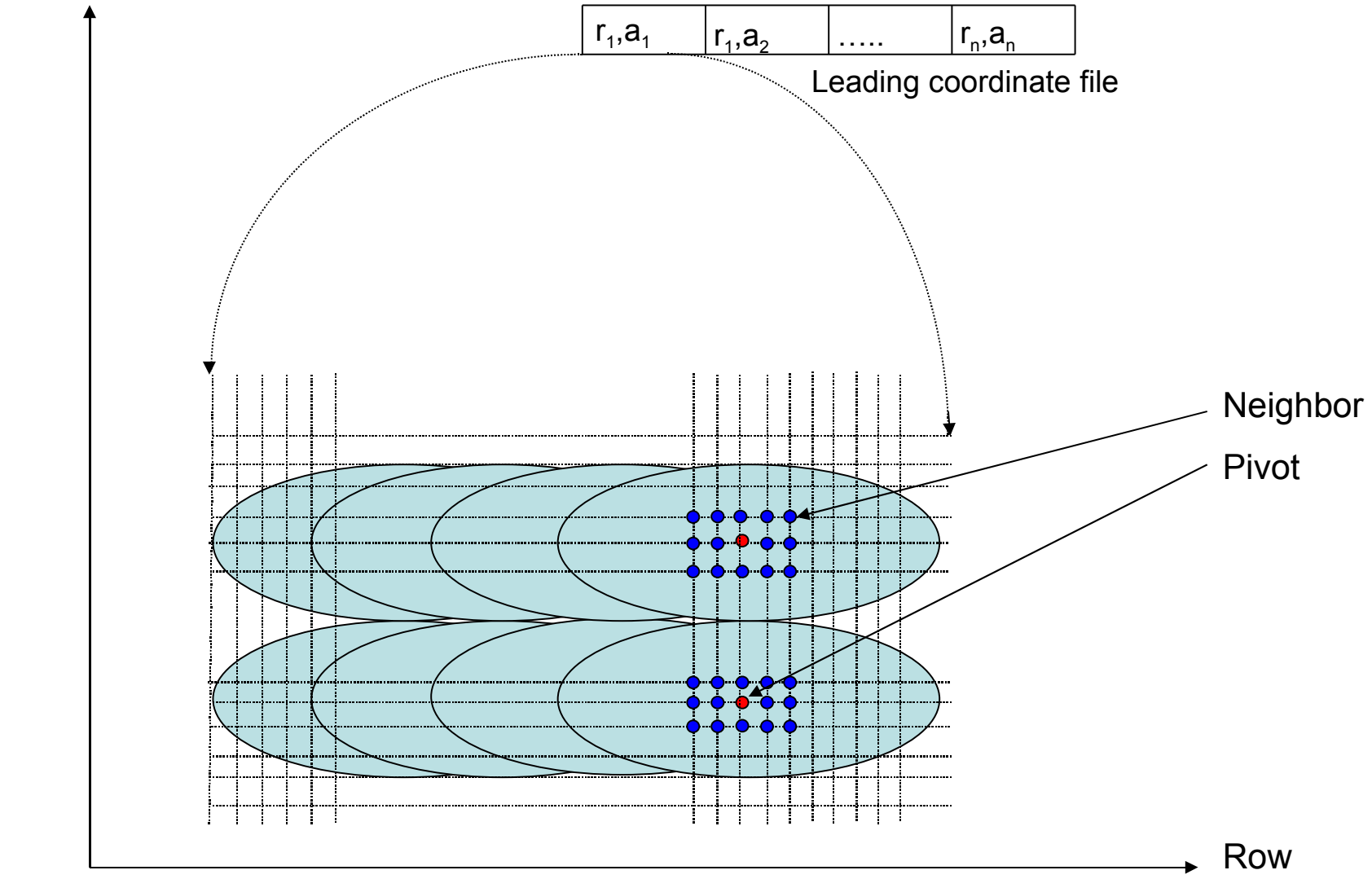


Biological meaning of distance between pivot and surrounding neighbors

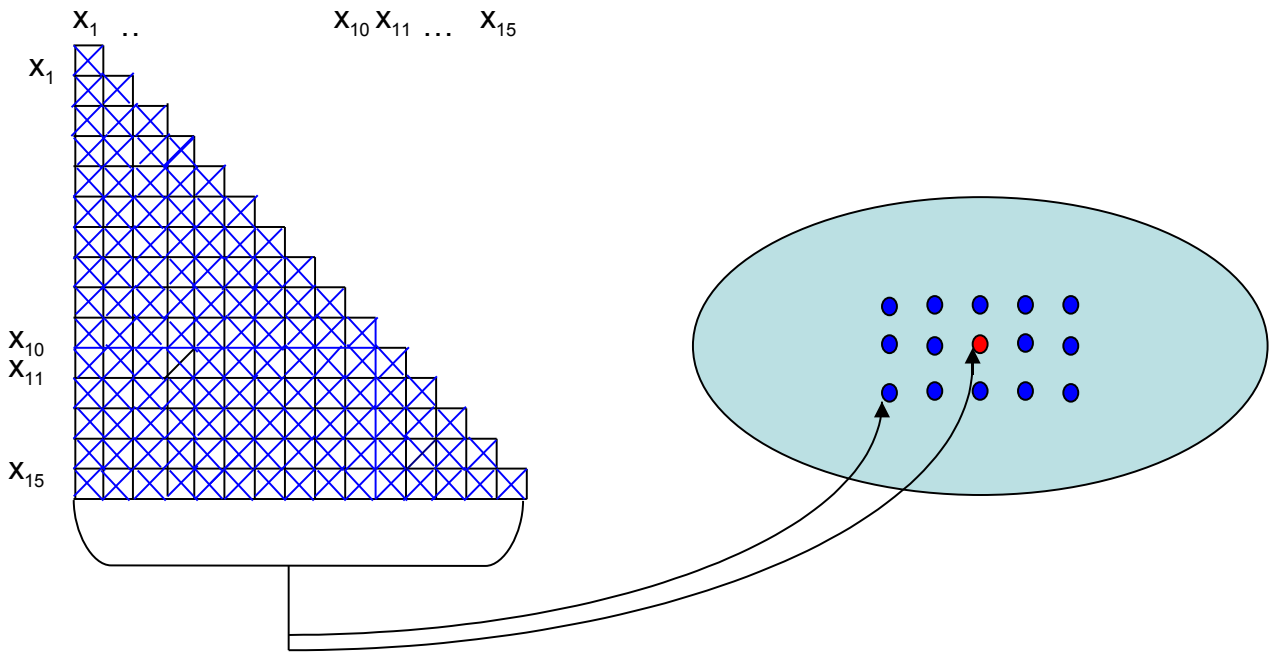


The role played by time in the autocorrelations that the P++ models can deal with.

The models implicitly take into account autocorrelations due to time. For instance, between annual shoots or rings. Time is a discrete coordinate corresponding to the year where the trait has been observed.



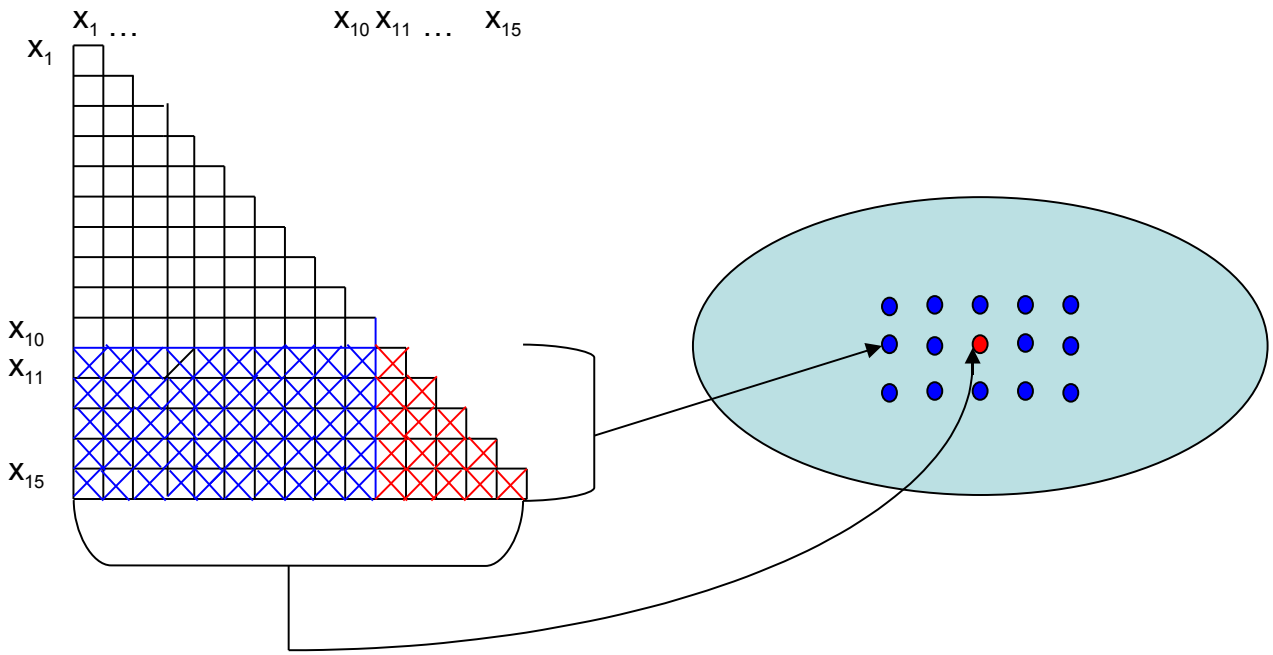
Generation of the successive pivot/neighbors associations using a leading coordinate file



Symmetrical processing of pivots and neighbors

The pivots selected by the leading coordinate file have been observed for the same traits than the surrounding neighbors (x_1-x_{15}).

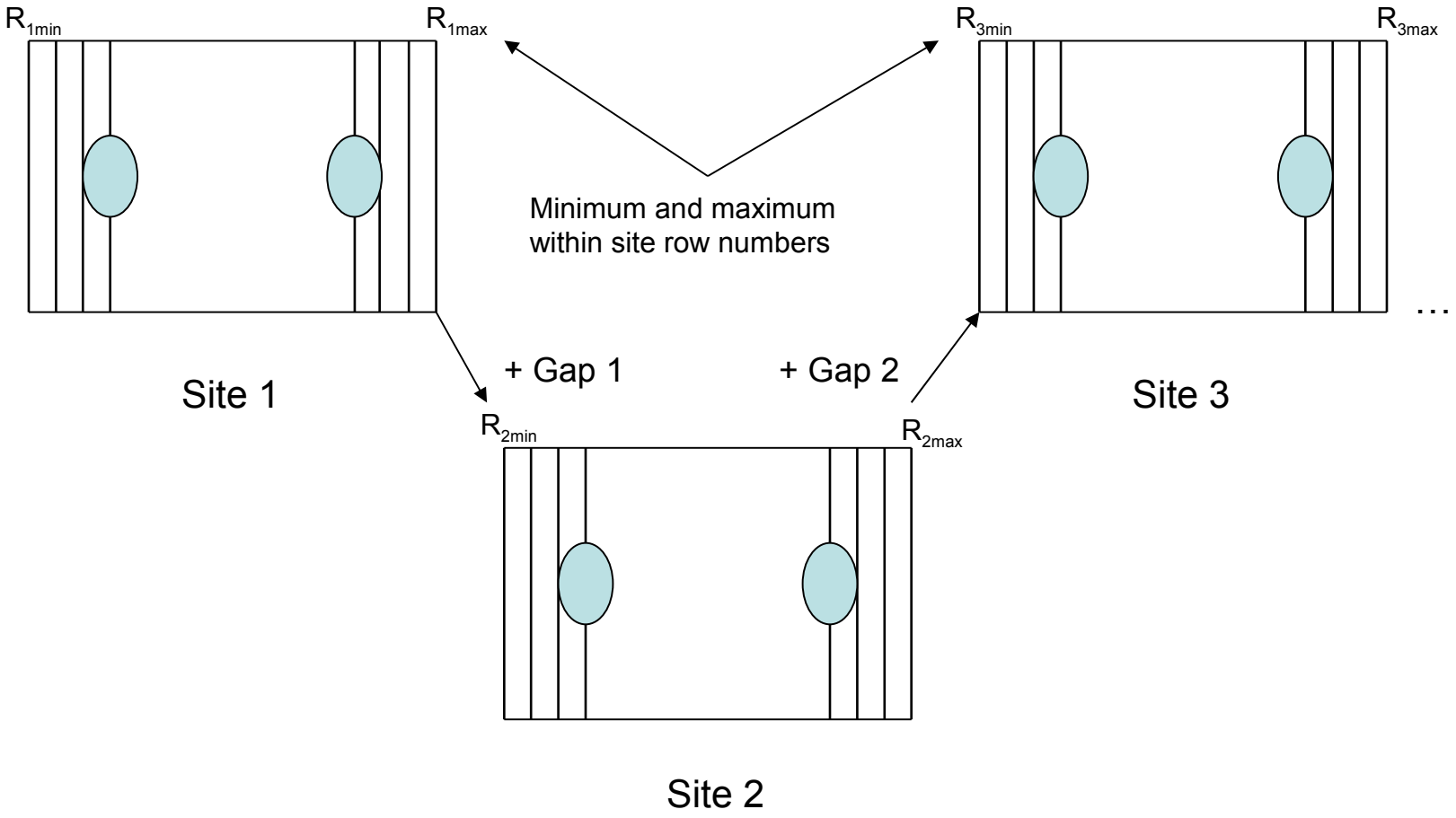
The adjustment process is therefore perfectly symmetric.



Asymmetrical processing of pivots and neighbors

The pivots selected by the leading coordinate file may have values for additional traits
(for instance, traits which are expensive to measure, here x_{11} to x_{15}). All individuals of the general population are observed for 'routine' traits, here x_1 to x_{10}.

The x_{11}-x_{15} traits of pivots are adjusted using their environmental correlations with mean residuals of traits x_1-x_{10}.



Practical way of obtaining disconnected coordinates for groups of trees from different sites

A '**gap**' is added to the row numbers of the sites **2** to ***n*** (***i*** indice) in order that the minimum row number
in site ***i*** is greater than the maximum value in site ***i*-1** by a 'reasonable' amount (50 for instance).

The p variables usually include the variable observed on the pivot individual. After the first run of this model, the mean residuals are re-computed from the adjusted values of all the variables and the multiple regression is run again. The process is reiterated, until the residual variance of each variable (σ^2_E) reaches a plateau.
A second step in the generalization is the choice of the sets of neighbors which allow the best adjustment, using the p covariates of model (1) by a multiple regression with c x p covariates:

$$Y_{ij(xy)} = \mu + G_i + b_{11}\bar{E}^1_1(\psi_r) + b_{21}\bar{E}^1_2(\psi_r) + \ldots + b_{p1}\bar{E}^1_p(\psi_r) + \ldots + b_{pc}\bar{E}^p_c(\psi_r) + E_{ij(xy)} \quad (2)$$

Combination of models (1) and (2), by simultaneously adjusting the pivot observations by a block effect and by a multiple regression on environmental variables gives the model:

$$Y_{ijh(xy)} = \mu + G_i + \beta_h + b_{11}\bar{E}^1_1(\psi_r) + b_{21}\bar{E}^1_2(\psi_r) + \ldots + b_{p1}\bar{E}^1_p(\psi_r) + \ldots + b_{pc}\bar{E}^p_c(\psi_r) + E_{ijh(xy)} \quad (3)$$

In models (1), (2) and (3), a stepwise downward multiple regression with p or c x p explicative variables at the first stage and only one at the last stage allows the identification of the most efficient variables or the configuration x variable combinations.

Computations are reiterated with the adequate set of covariates. The process is stopped when the relative reduction between two runs falls under a predefined value for all the variables.

Generalization of the single-site model for processing multilocal trials

The model may be extended to multilocal trials. The elements below may be added:

Integration of a site effect

If σ_S is the fixed effect of site s, model (3) can be rewritten:

$$Y_{isj(xy)} = \mu + G_i + \sigma_s + b_{11}\bar{E}^1_1(\psi_r) + b_{21}\bar{E}^1_2(\psi_r) + \ldots + b_{p1}\bar{E}^1_p(\psi_r) + \ldots + b_{pc}\bar{E}^p_c(\psi_r) + E_{isj(xy)} \quad (4)$$

We call the "**Multisite P++ method II** " this model, which uses an ANOVA adjustment for site effect and an adjustment by multiple regression.The major modification from model (3) is the change of spatial scale which may cause a Genotype x Environment interaction.

Integration of site and block|site effects

If σ_S is again the effect of site s and β_{sh} the effect of block h within the site s, we call the "**Multisite P++ method III** " the model below:

$$Y_{ishj} = \mu + G_i + \sigma_s + \beta_{sh} + b_{11}\overline{E}^1_1(\psi_r) + b_{21}\overline{E}^1_2(\psi_r) + \ldots + b_{p1}\overline{E}^1_p(\psi_r) + \ldots + b_{pc}\overline{E}^p_c(\psi_r) + E_{ishj} \quad (5)$$

which combines an adjustment to site and block|site effects with a multiple regression.
We call the "**Multisite P++ method I"** the basic model: model (3) which only uses multiple regression of the observed values on mean residuals across all sites.

Software implementation

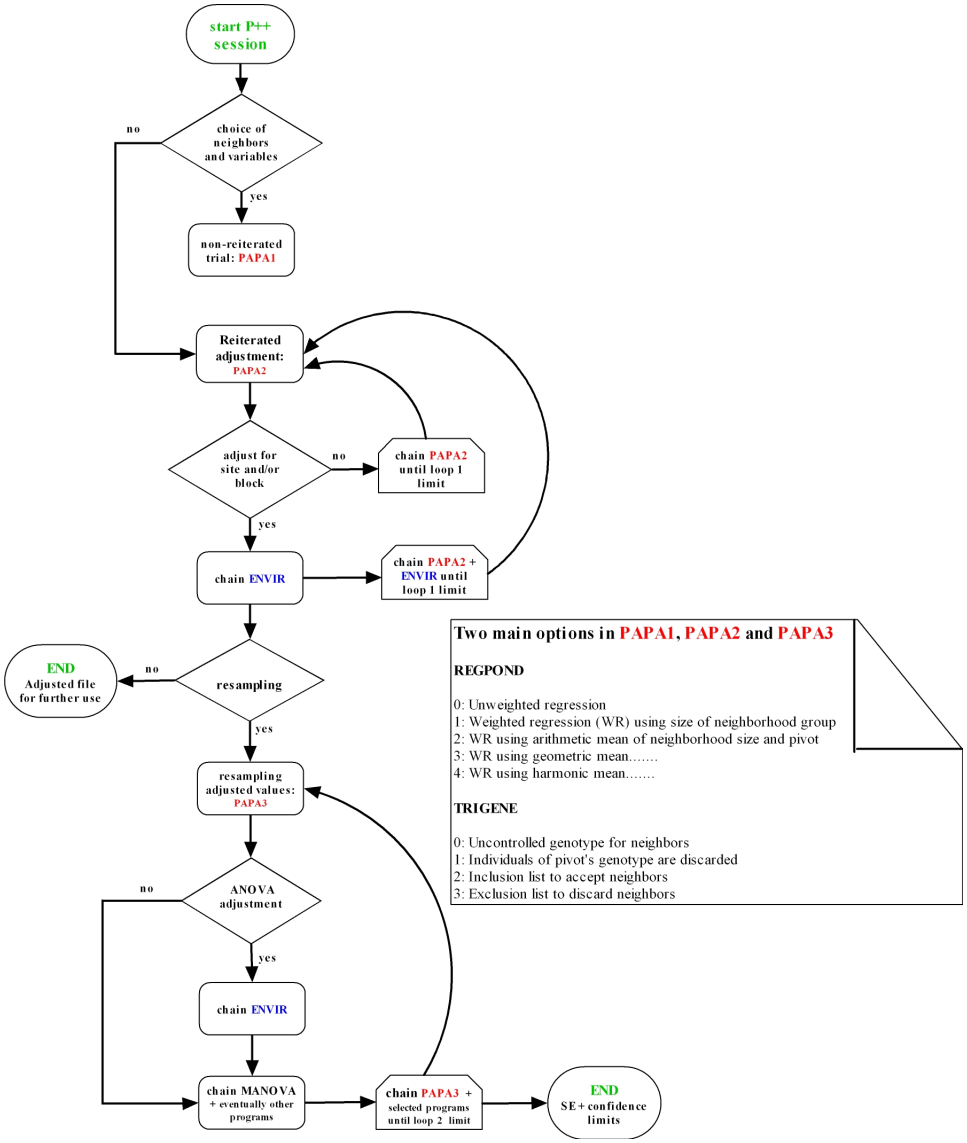
The software is organized into three modules and uses elementary utilities as well as general computation algorithms. The reiterated sequences are controlled by the general reiteration system also used for resampling (JBSTAR). These modules are:

- **PAPA1:** Computation of average residuals according to defined neighbor structures, merging with individual data and computation of a downward multiple regression to determine the appropriate combinations of structures and covariates for the adjustment.

- **PAPA2:** Reiterated adjustment of individual data with the possibility of combining multiple regression with adjustment of block and site effects (general purpose ENVIR program) to fit all the models described above.

- **PAPA3:** Once the adjusted data file is obtained, this module performs resampling (jackknife or bootstrap) to obtain standard errors and confidence intervals on a variety of genetic parameters such as heritability or genetic correlations using appropriate MANOVA programs (which may be followed by other programs, like those required for computation of selection indices including expected genetic gains).

The general flowchart involving these three modules is shown below. REGPOND and TRIGENE parameters modify the multiple regression and filter members of neighborhood groups. Other options concern the geometry and size of neighboring structures.



General flowchart showing the integration of the three modules, PAPA1, PAPA2 and PAPA3, for data processing according to the different P++ sub-models.
See the text for additional legends.

Efficiency of the three multisite P++ methods on three multilocal trials

Three multi-site comparative trials (3 sites per trial) concerning important forest species in Europe and Canada were analysed to compare estimates of narrow and broad sense heritabilities.
The Maritime Pine trial, established with a density of 1250 plants/ha (spacing of 4 m between rows, and of 2 m on the row), comprised 100 half-sib families from controlled pollination with two different pollen mixes (trees selected within first generation progeny trials). The field design comprised 35 one tree plot complete randomized blocks in each site. The three sites were close from each other and similar in respect of fertility.
The Poplar trial, with a density of 6667 plants/ha (spacing of 2 m between rows and 0.75 m on the row), compared on three remote sites quite different for climatic conditions (France, Italy and England) 330 clones from the same full-sib family. The field design was made of six complete randomized blocks with one tree per plot. The cross was realized for QTL location (F2 from *P. deltoides* x *P. trichocarpa* hybrid). Each heritability is therefore a broad sense one corresponding to a narrow genetic basis.
The White Spruce trial comprised 250 maternal progenies from 50 provenances (five progenies per provenance). It was established in three Quebec sites, at a density of 3472 plants/ha (spacing of 2.4 m between rows and 1.2 m/row). The field design comprised six complete randomized blocks in each site with linear plots of five trees.

For Maritime Pine and White Spruce trials, the statistical model is:

$$Y_{ijk} = \mu + P_i + M_{ij} + E_{ijk}$$

where P_i is the effect of pollen mix i in the first case and the effect of provenance i in the second one. M_{ij} is the mother effect nested within level 1 and E_{ijk} the individual deviation (of environmental and genetic origins). The formulae to estimate "within level 1" heritability (half-sib families with an assumed 0 inbreeding coefficient) are the same:

$$\hat{h}^2_{ss} = \frac{4\hat{\sigma}^2_{M|P}}{\hat{\sigma}^2_{M|P} + \hat{\sigma}^2_E}$$

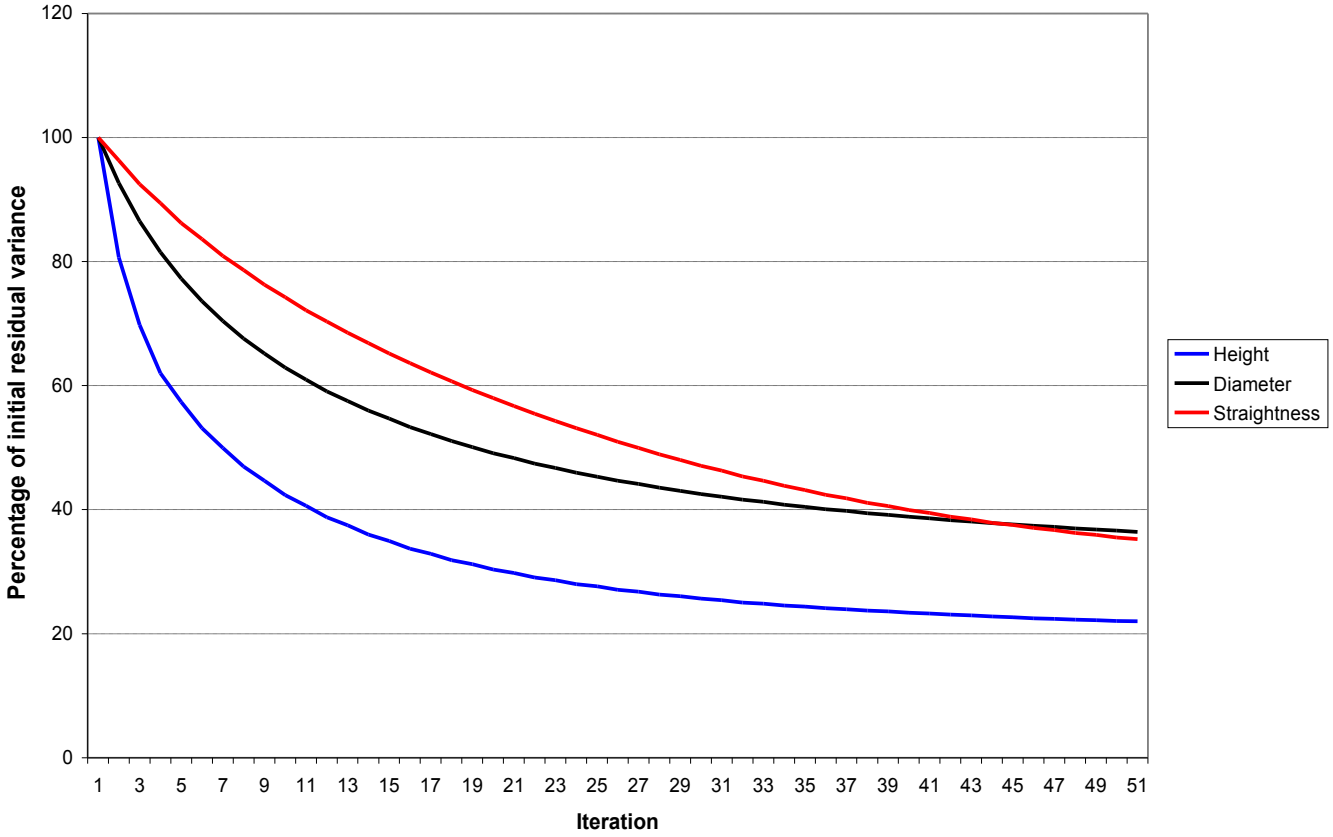
The Poplar clonal trial is relevant from a one-way statistical model:

$$Y_{ij} = \mu + C_i + E_{ij}$$

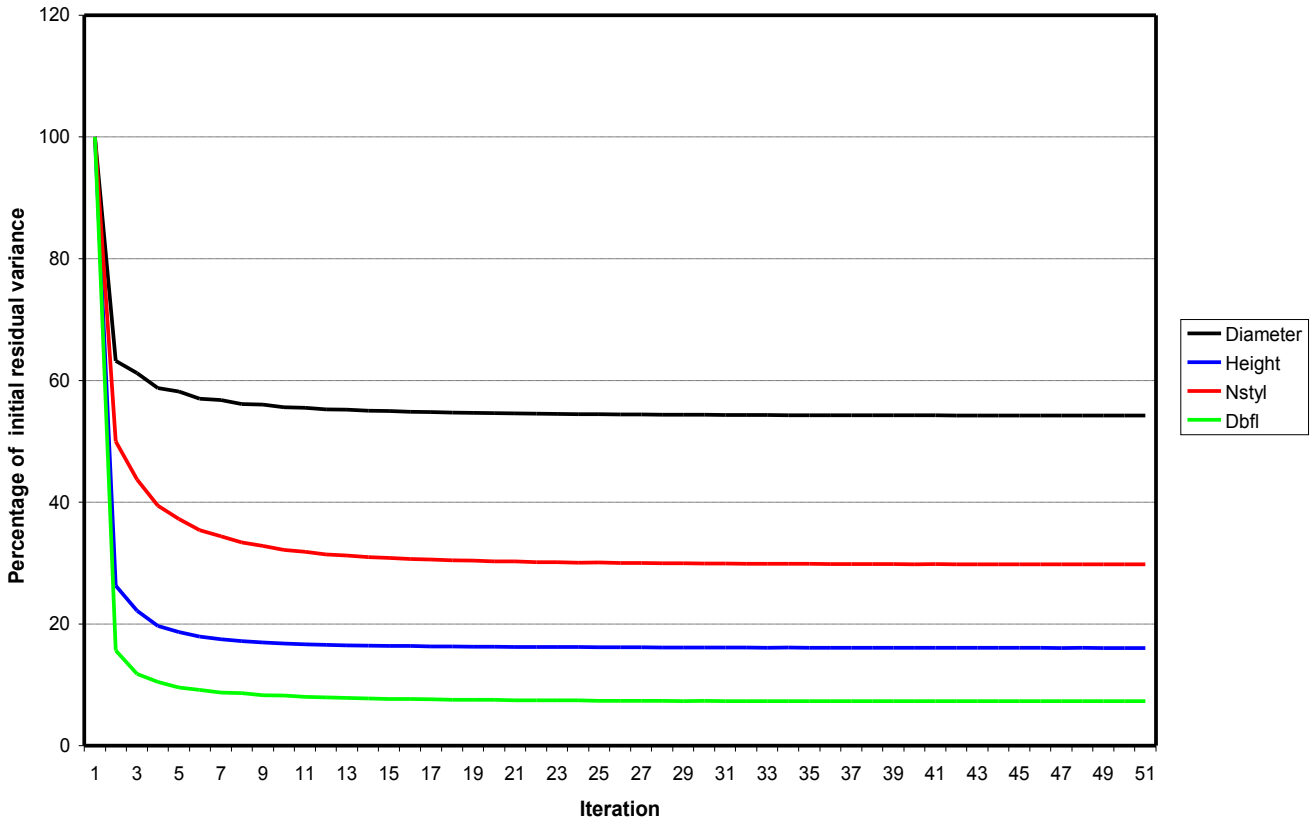
Where C_i is the clone effect and E_{ij} the ramet within clone deviation (completely due to environment) and the corresponding estimation of broad sense heritability is:

$$\hat{h}^2_{sl} = \frac{\hat{\sigma}^2_C}{\hat{\sigma}^2_C + \hat{\sigma}^2_E}$$

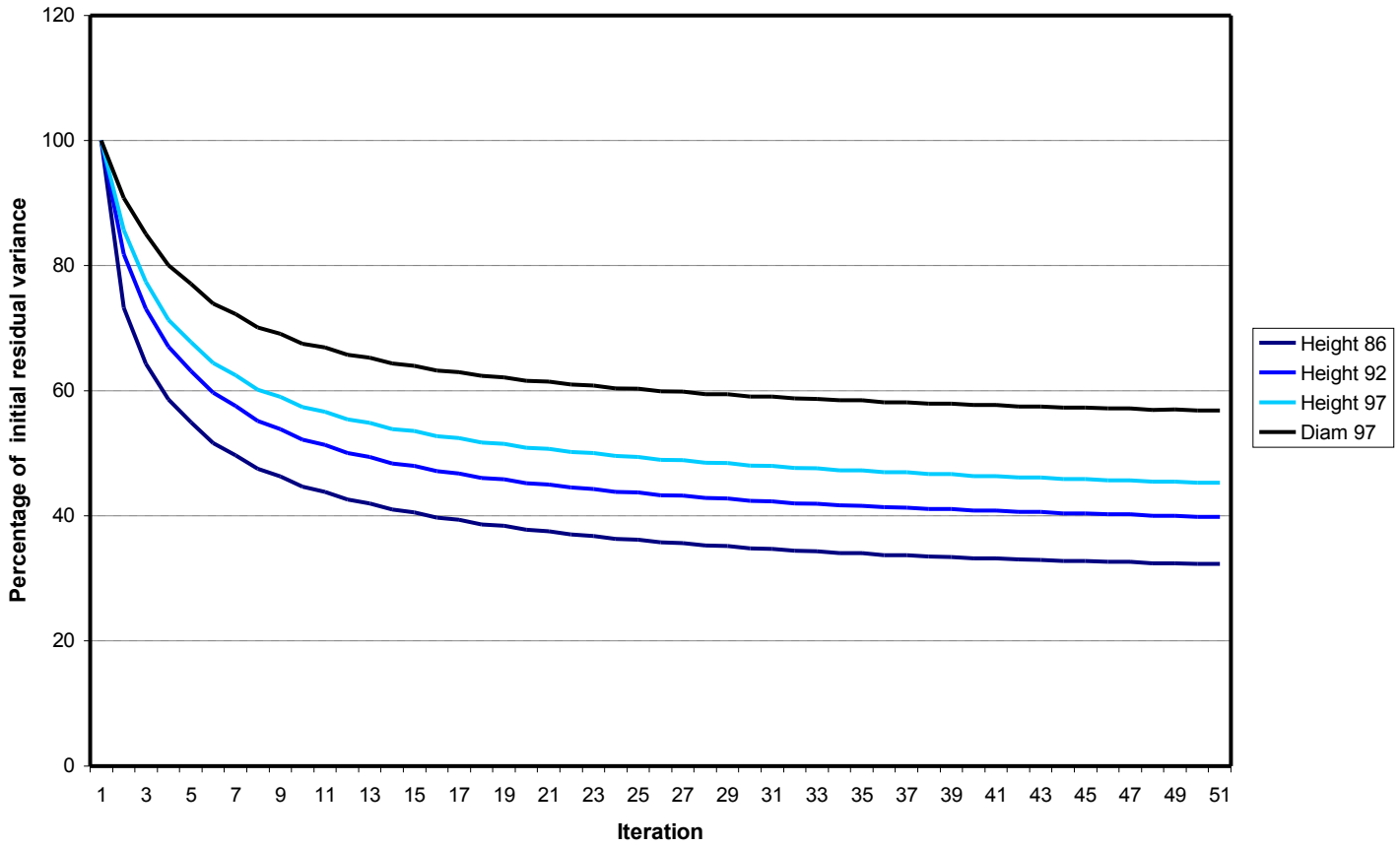
Three neighborhood configurations: nested circles with 2, 3 and 4 within-row spacing units. For each at least one of the covariates was significantly correlated to a pivot's variable.
Fig. 3, 4 and 5 show the reduction of the residual variance for three trials as a function of iteration number for **P++ method I**. The speed and amount of relative decrease is different according to the considered trait. For the Maritime Pine trial, height displayed the maximum response. At the opposite side, the plateau was very quickly reached for the four traits of the Poplar trial.



Reduction of residual variance for three traits according to the number of iterations (P++ method I applied to a multilocal test of half-sib families of Pinus pinaster).



Reduction of residual variance for four traits according to the number of iterations (P++ method I applied to a multilocal test of poplar clones).



Reduction of residual variance for four traits according to the number of iterations (P++ method I applied to a multilocal test of provenances/half-sib families of Picea glauca).

Compared efficiencies of the usual adjustment method (site and block|site), of the three P++ weighted multisite methods (I, II and III, REGPOND option = 1) and of unweighted method I to increase heritability for three multilocal progeny or clonal trials.

Method \ Species / Trait	*Pinus pinaster*			F2 poplar clones				*Picea glauca*			
	H02	C02	V02	C04	H04	Nbs	Dbf	H86	H92	H97	D97
Usual method	0.200	0.109	0.183	0.095	0.099	0.275	0.488	0.147	0.163	0.166	0.090
Unweighted method I	0.692	0.288	0.630	0.139	0.236	0.418	0.658	0.239	0.248	0.252	0.149
Weighted method I	0.869	0.372	0.660	0.133	0.239	0.435	0.676	0.359	0.358	0.346	0.191
Weighted method II	0.872	0.372	0.653	0.134	0.239	0.439	0.676	0.373	0.378	0.362	0.204
Weighted method III	0.752	0.307	0.458	0.138	0.238	0.438	0.677	0.373	0.378	0.361	0.203

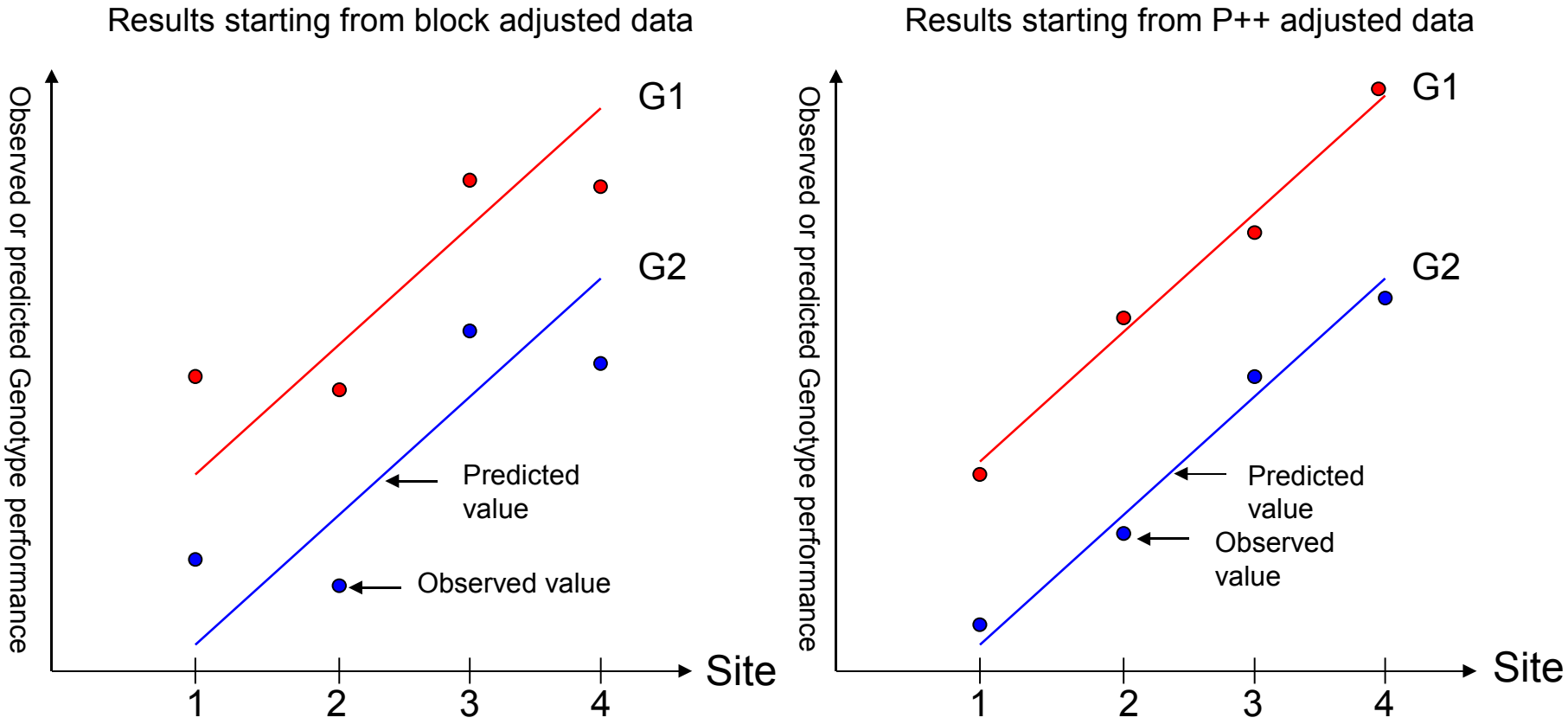
Legend: *H(year): Height at a given year*
C(year): Circumference at a given year
V02: Departure of the stem from verticality in 2002
D97: Diameter in 1997
Nbs: Number of sylleptic ramifications (2003)
Dbf: Number of days before flushing (2004)

We note three important features:

- P++, whatever is the method (I, II or III) always allows substantially greater heritability than a traditional ANOVA adjustment to site and block|site;
- Weighted multiple regression results into greater heritability values;
- At the opposite side, compared with method I, methods II and III never result into a noticeable heritability increase and give similar results (moreover, they can give a lower value in the case of Maritime Pine).

All P++ adjustments drastically reduced the standard errors of heritability estimates.

The favourable influence of weighting by the size of the most internal neighborhood configuration may be due to limitation of incidence of points with higher mortalities or of trial borders that truncate the theoretical geometry of these groups. The lower efficiency of P++ for poplar trial may be related to its young age which did not allow a strong between-tree competition and, more generally, reduced the local environmental influence on phenotypic variability. A greater site homogeneity can explain the abrupt slope of the reduction of residual variances corresponding to the first reiteration.



Graphical interpretation of the role of multisite adjustment to reduce unpredictable G x E interaction

P++ modelisation shows threfore a global efficiency for environment control over several sites compared to an ANOVA adjustment. This property may be be explained by reducing the within-site bias on estimate of genetic effects (reduction of false Genotype x Block interaction).

CONCLUSIONS

- The P++ models are powerful tools for adjusting genetic data and is by far more efficient than ANOVA adjustment, both at the single-site and at the multi-site levels;

- The experimental data demonstrate that their efficiency can be increased by taking into account the sizes of the neighbor groups (weighted multiple regression);

- On multilocal trials, efficiency of the model was not increased by an ANOVA adjustment for site or site + block|site effects;

- Further applications to diversified sets of genetic trials are required to evaluate the practical interest of genetic restrictions on individuals within the neighborhood groups;

- This method can also be used to process data from natural populations where the genetic identification of individuals would result from biochemical, botanical or molecular markers;

- Use of the P++ model in a plant breeding software (with a simple contextual introduction of parameters) makes it routinely usable, for more accurate estimates of genetic parameters, QTL identification, and to achieve greater genetic gains by selection.

References

Azais J-M, Denis J.-B., Dhorne T and Kobilinsky A (1990). Neighbour analysis of plot experiments: a review of the different approaches. Biométrie-Praximétrie, **30**, 15-39.

Baradat P and Labbé T. (2006). Optimisation du rééchantillonnage dans un logiciel d'Amélioration des Plantes. Proc. CARI 06, 435-442.

Baradat P, Raffin A, Bastien C and Beaulieu J (2006). Papadakis++, Un ajustement multi-dimensionnel du Phénotype à l'Environnement. Doc. AMAP DT.1-2006, Montpellier, 157 pp.

Bartlett M S (1978). Nearest Neighbour Models in the Analysis of Field Experiments. J R Statist. Soc. **2**, 147-174.

Bertrand B (2002). L'Amélioration génétique de *Coffea arabica* L. en Amérique Centrale par la voie hybride F1. Thèse ENSAM, Ec. Doct. « Biologie Intégrative », Montp., 275 pp.

Besag J (1983). Contribution to Discussion of Wilkinson *et al*. Journal of the Royal Statistical Society. Series B **45**, 180-183.

Besag J and Kempton R (1986). Statistical Analysis of Field Experiments Using Neighbouring Plots. Biometrics, **42**, 231-251.

Dagnélie P (1987) - La méthode de Papadakis en expérimentation agronomique : considérations historiques et bibliographiques. Biométrie-Praximétrie, **27**, 49-64.

Gleeson A C and Cullis B R (1987). Residual Maximum Likelihood (REML) Estimation of a Neighbour Model for Field Experiments. Biometrics, **43**, 277-288.

Goumari A (1990). Analyse comparative des resultats d'essais en champ selon les techniques des blocs aléatoires complets, des lattices et des plus proches voisins. Biométrie-Praximétrie, **30** (3-4), 91-105.

Kempton R A and Howes C W. (1981). The Use of Neighbouring Plot Values in the Analysis of Variety Trials. Applied Statistics **30** (1), 59-70.

Papadakis J. (1984). Advances in the analysis of field experiments. Comm. Acad. Athènes, **59**, 326-342.

Pichot C (1993). Variabilité au stade adulte de *Populus trichocarpa* et prédiction juvénile-adulte chez *P. trichocarpa* et *P. deltoides*. Thèse de doct. INA-PG, Paris, 235 p. + annexes.

Ryan T P (1997). Modern regression methods. Wiley, NY, 515 pp.

Sébastien B (1993). Modèles linéaires avec résidus spatialement autocorrélés, application à l'expérimentation agricole. Thèse doct. INA-PG, Paris, 205 pp.

Zas R (2006). Iterative kriging for removing spatial autocorrelation in analysis of forest genetic trials. Tree Genetics & Genomes, DOI 10.1007/s11295-006-0042-4.

Interest outside of Genetics?

- No constraint about the nature (random or fixed) of adjusted effets
- In *DIOGENE*'s applications, genetic models are added independently
- Therefore, the P++ models can be used in fields concerned by 'environmental' controlled effects:

  - Fertilization trials
  - Sylvicultural practices (pruning…)
  - Mixed multispecies stands

……

Try P++ and see if it may be useful in your research field.

Actualités

>>>>>> Voir toutes les actualités

tidal cycle to regional coastal changes. Nature Geoscience, 1 (3): 169-172. ...[En savoir plus]

25/02/2008

Vient de paraître :

Slope stability and erosion control: Ecotechnological solutions

Norris J.E., Stokes A., Mickovski S.B., Cammeraat E., van Beek R., Nicoll B.C., Achim A. 2008. Slope stability and erosion control: Ecotechnological solutions. Dordrecht: Springer. 290 p. [En savoir plus]

Unité Mixte de Recherche AMAP

botAnique et bioInforMatique de l'Architecture des Plantes

Directeur : **Daniel BARTHELEMY**
Directeur adjoint : **Daniel AUCLAIR**.

UMR AMAP
TA A51 / PS2
34398 Montpellier cedex 5

Umr Cirad 51 - Umr Inra 931 - Umr Cnrs 5120 - Umr Ird 123

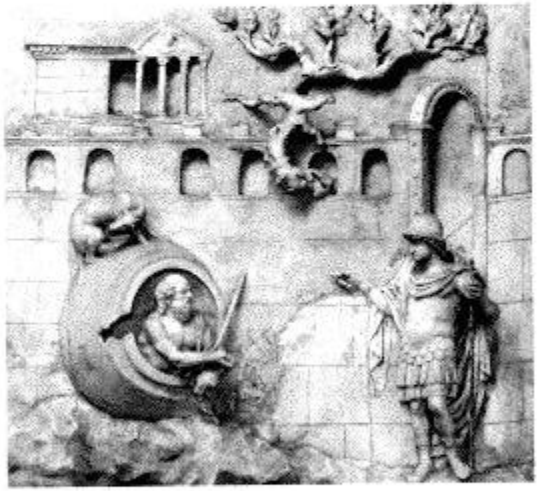
Production logicielle

Liste des logiciels développés au sein de l'unité

AMAPSim	AMAPsim propose un noyau de simulation du développement architectural d'une plante interconnectable avec des applications externes.
Capsis	Capsis est une plate-forme de développement de modèles de croissance et de dynamique forestière. Il permet de construire et d'évaluer des scénarios sylvicoles en s'appuyant sur un modèle pour une espèce et une région donnée (fertilité stationnelle, densité initiale, intensité, type et nature des éclaircies, élagage...). Il est utilisé par des chercheurs pour évaluer leurs modèles, des gestionnaires forestiers pour aider à l'élaboration de guides sylvicoles et pour l'enseignement.
Diogene	Interprétation d'observations sur plantes annuelles et pérennes en plantations comparatives ou populations naturelles. Traite également les analyses de marqueurs moléculaires.
	MHM, for Multiscale Heterogeneity Map, is a convenient tool aiming at

Logiciels AMAP

DIOGENE



Logiciel libre (licence GPL)

Systèmes d'exploitation :

Unix (Solaris) et **Linux** (toutes distributions)

Traitement d'observations sur plantes annuelles ou pérennes en plantations comparatives ou populations naturelles. Observations de terrain et marqueurs moléculaires.
Processing data on annual or perennial plants in comparative trials or natural populations. Field

***DIOGENE*, version 2008/1.2**

(Disponible par téléchargement ou sur CD)