MÉTHODOLOGIE DESTINÉE AUX ESSAIS DE SÉLECTION DU CACAOYER

R. LOTODÉ *

Ph. LACHENAUD **

Les difficultés rencontrées pour obtenir une efficacité acceptable des dispositifs expérimentaux sont la conséquence :

— des techniques d'installation de l'arbuste ;
— du comportement de l'arbuste (interaction avec les facteurs du milieu) ;
— de la durée de l'expérience.

FACTEURS DE PERTURBATION DES ESSAIS

Hétérogénéité du sol

La mise en place des cacaoyers est effectuée après défrichement de la forêt ou de la jachère ancienne. Les diverses essences, depuis longtemps *in situ*, ont, par leurs exigences très diverses (retombées foliaires, explorations du sous-sol), induit une hétérogénéité importante dans la couche superficielle du sol. Elle est encore accentuée parfois par les andainages interlignes, après abattage, constitués d'essences et surtout de quantités de matières végétales variées. Aucun travail, partiellement homogénéisant, du sol n'est effectué pour éviter toute détérioration rapide de l'horizon superficiel alors mis à nu.

Les coefficients de variation pour les principaux éléments du sol sont de l'ordre de :

— 20 à 25 % pour le phosphore total et l'azote total ;
— 25 à 35 % pour le carbone organique et le magnésium échangeable ;
— 40 à 50 % pour le potassium et le calcium échangeable.

Cela signifie que pour K et Ca, par exemple, il faut analyser un échantillon composite constitué à partir d'une centaine de micro-prises pour avoir une valeur moyenne entourée d'un intervalle de confiance de ± 10 % ! L'intervalle de confiance est $\pm 2\,CV/\sqrt{n}$ (n étant le nombre de micro-prises dans la parcelle à l'étude).

Une courbe donnant la valeur de la précision en fonction de n peut être tracée, pour chaque élément.

Hétérogénéité génétique

Lorsqu'un clone ou un mélange clonal est utilisé, dans certains essais, on est assuré de l'homogénéité des plants au point de vue potentiel de production ou réactions aux diverses agressions. Ce n'est pas le cas lorsque des plants issus de semis sont utilisés. L'hétérogénéité est impor-

* Ancien chef du Service de Biométrie de l'IRCC, BP 5035, 34032 Montpellier Cedex.

** Agronome, IRCC, Station de Divo (Côte d'Ivoire). Adresse actuelle : IRCC/CIRAD, BP 701, 97387 Kourou, Cedex.

tante à tous points de vue dans la descendance de croisements intervariétaux, surtout si la recherche de l'hétérosis conduit à adopter des parents éloignés génétiquement. Toutefois, on note, en ce qui concerne les productions, que la variabilité des données individuelles pour un clone est presque du même ordre de grandeur que celle des descendances, confirmant l'importance de la composante « environnementale » de la variance totale.

Hétérogénéité d'origines diverses

En pépinière, malgré toutes les précautions qui peuvent être prises, les développements sont variés, à cause des micro-environnements parfois hétérogènes, des dates de semis ou de bouturage insuffisamment groupées, mais on s'efforcera de prélever, pour les essais, un ensemble de plants le plus homogène possible.

En champ, les micro-environnements (microtopographie, développement variable des arbres), les accidents de croissance dus aux attaques d'insectes (sur bourgeons terminaux notamment, entraînant un retard dans la croissance), aux blessures (matchette, etc.) sont, de même, facteurs d'hétérogénéité. Il faut également noter les perturbations provoquées par la perte accidentelle d'arbustes, inévitable dans les essais de longue durée. Les remplacements effectués avec des plants du même âge au cours de l'année suivant la mise en place pourront être intégrés à l'essai ; au-delà, ce n'est plus possible, même si l'arbuste croît normalement, ce qui n'est pas toujours le cas.

Cet ensemble de facteurs d'hétérogénéité est à l'origine d'un coefficient de variation de 40 à plus de 50 % pour la production annuelle individuelle. En cumulant les récoltes de quelques années, le CV s'abaisse, mais pas de manière spectaculaire, car les arbres haut ou bas producteurs le restent le plus souvent chaque année.

Quand les parcelles élémentaires sont constituées de placeaux de vingt plants (4 × 5) ou de lignes de dix plants, techniques autrefois utilisées, le coefficient de variation des moyennes de productions parcellaires reste de l'ordre de 25 à 35 %, même après cumul des récoltes sur quelques années, et cinq à six répétitions ne permettent de différencier significativement que les premiers des derniers des divers classements obtenus. La faible efficacité de ces dispositifs classiques les a remis en question et une étude fondamentale sur la taille optimale des parcelles élémentaires a été alors entreprise (1, 2). Les données ont été reprises récemment et l'utilisation d'une technique d'ajustement différente a permis de mieux cerner le problème.

TAILLE DES PARCELLES ÉLÉMENTAIRES ET ÉVOLUTION DE LA PRÉCISION

Le premier point étudié a été celui du cumul des productions individuelles après dix ans dans une plantation de cacaoyers Trinitario au Cameroun, installée sous ombrage léger constitué d'essences choisies dans la forêt initiale. Des parcelles fictives aussi compactes que possible ont été constituées avec un effectif d'arbres variant de 1 à 96 (1, 4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 64, 96), sans ligne de bordure inter-parcellaire, mais avec une ligne de bordure générale. Le problème des bordures sera abordé plus loin.

Les résultats ont été ceux présentés dans le tableau I.

Le graphique 1 montre l'évolution de la variance de la moyenne parcellaire V_n, en fonction de l'effectif n dans la parcelle (x).

Si la répartition des arbres, pour ce qui concerne la donnée de production, était réalisée strictement au hasard, c'est-à-dire si l'hypothèse de l'indépendance de la production d'arbres voisins était vérifiée, la variance V_n suivrait la loi générale théorique :

$$V_n = V_1/n .$$

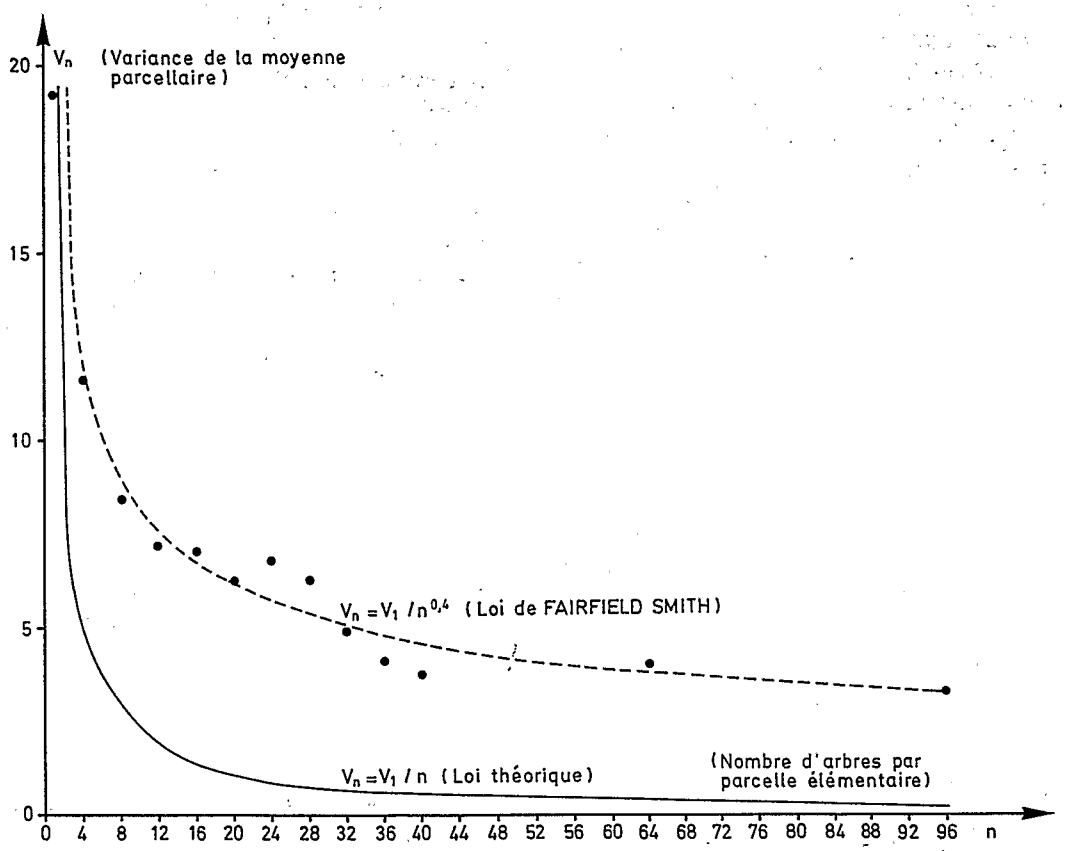
C'est une branche d'hyperbole ayant pour asymptote 0 quand n croît (graphique 1 : trait plein). Il n'en est rien, V_n observé est constamment supérieur à la valeur V_n théorique.

Fairfield Smith (An empirical law describing heterogeneity in the yields of agricultural crops, 1938) a proposé, pour l'ajustement, l'adoption d'une formule empirique $V_n = V_1/n^b$ où b est un coefficient variant de 0 à 1, mesurant la dépendance entre productions d'arbres voisins.

TABLEAU I

Evolution des variances des moyennes parcellaires en fonction de l'effectif d'arbres par parcelle

Effectif d'arbres par parcelle (n)	Nombre de parcelles (r)	Variance de la moyenne parcellaire (V_n) (kg de fèves fraîches)	Variance ajustée ($V_n = V_1/n^{0,4}$)
1	2 008	19,16	19,83
4	485	11,54	11,39
8	241	8,47	8,63
12	161	7,18	7,34
16	119	6,99	6,54
20	94	6,19	5,98
24	79	6,79	5,56
28	67	6,29	5,23
32	59	4,83	4,96
36	52	4,04	4,73
40	46	3,71	4,53
64	28	4,00	3,76
96	19	3,23	3,19



Graphique 1. — Evolution de la variance de la moyenne parcellaire (V_n) en fonction de l'effectif d'arbres dans la parcelle (n)

Si b = 1, $V_n = V_1/n$, il y a indépendance.

Si b = 0, $V_n = V_1$, il y a dépendance totale et la variance reste constante.

On trouvera bien sûr toujours une valeur comprise entre 0 et 1 : plus b sera faible, plus la dépendance sera forte.

En partant de $V_n = V_1/n^b$ et en passant aux logarithmes, il vient :

$$\text{Log } V_n = \text{Log } V_1 - b \text{ Log } n .$$

Un ajustement linéaire Log V_n en fonction de Log n, effectué sur les données du tableau I aboutit à l'équation :

$$\text{Log } V_n = 2{,}9872 - 0{,}397 \text{ Log } n \quad (b = 0{,}397) .$$

Le coefficient de corrélation linéaire est très hautement significatif : r = 0,974 (r 0,001 = 0,80 pour 11 degrés de liberté).

En pondérant chaque point par les nombres de

degrés de liberté des variances correspondantes, on trouve $b = 0{,}382$. Adoptons $b = 0{,}4$.

On a donc une formule empirique :

$$V_n = V_1/n^{0,4}$$

V_1 étant la valeur ajustée, soit, en remplaçant n par 1 dans l'équation de régression :

$$\text{Log } V_1 = 2{,}9872 \quad \text{ou} \quad V_1 = e^{2{,}9872} = 19{,}83 .$$

Les valeurs ajustées pour V_n sont données dans le tableau I en dernière colonne et sont figurées sur le graphique 1 (trait pointillé).

L'avantage de cette formule est d'être plus manipulable mathématiquement que l'ajustement hyperbolique $V_n = V_1/n + b$ ou $V_n = (V_1 - b)/n + b$ qui avait été adopté dans les premières études.

Cette formule étant retenue, il faut maintenant répondre à la question suivante : étant donné un champ d'essai de surface donnée qui peut être caractérisé par N arbres, quelle est la taille de la parcelle élémentaire qui donnera la plus grande précision dans la comparaison des « traitements » ?

Après analyse de variance, la différence significative entre deux traitements est donnée par la formule habituelle :

$$d = q \times \sqrt{\text{variance de la moyenne d'un traitement}}$$

(q de Keuls ou de Duncan).

La variance de la moyenne d'un traitement étant le rapport du carré moyen résiduel de l'analyse de variance sur le nombre de répétitions, la question précédente devient donc : quelle est la taille de la parcelle élémentaire qui donne la plus petite variance de la moyenne parcellaire, sachant qu'on a la relation $V_n = V_1/n^{0,4}$.

Cas où une ligne de bordure interparcellaire n'est pas nécessaire

Ce cas est celui d'essais clonaux, d'hybrides intervariétaux, de cultivars, etc., que nous regrouperons sous le terme général de familles.

Soit :

N : le nombre d'arbres utiles dans un champ d'essai (en éliminant donc une ligne de bordure périphérique).

n : le nombre de plants par parcelle élémentaire,

r : le nombre de répétitions,

k : le nombre de clones, hybrides, traitements en comparaison ($n \cdot k \cdot r = N$),

$\bar{x}$: la moyenne générale pour la variable étudiée des N arbres. Cette moyenne est constante quel que soit l'effectif n par parcelle élémentaire.

V_1 = variance des données individuelles (S_1 = écart type),

$CV_1 = S_1/\bar{x}$, le coefficient de variation des données individuelles,

$V_n = V_1/n^{0,4}$, la variance de la moyenne parcellaire, la différence significative entre deux traitements est :

$$d = q \cdot \sqrt{V_n/r} \quad (r = N/n \cdot k)$$

soit

$$d = q \cdot \sqrt{\frac{V_1}{n^{0,4}} \Big/ \frac{N}{n \cdot k}}$$

$$d = \frac{q}{\sqrt{N}} \cdot \sqrt{k} \cdot S_1 \sqrt{n/n^{0,4}}$$

et comme $\sqrt{n/n^{0,4}} = \sqrt{n^{1-0,4}} = \sqrt{n^{0,6}} = n^{0,3}$.

Il vient :

$$d = q \cdot \sqrt{\frac{k}{N}} \cdot S_1 \cdot n^{0,3} .$$

Si on appelle D la différence significative en pour-cent de la moyenne générale,

$$D = 100\, d/\bar{x}$$

soit :

$$D = 100\, q \cdot \sqrt{\frac{k}{N}} \cdot \frac{S_1}{\bar{x}} \cdot n^{0,3}$$

$$D = 100\, q \cdot \sqrt{\frac{k}{N}} \cdot CV_1 \cdot n^{0,3} .$$

CV_1 est connu ou calculable pour une variété, un environnement donné ; N et k sont donnés ; appelons C la quantité $100 \cdot \sqrt{\frac{k}{N}} \cdot CV_1$, il vient :

$$D = C \,.\, q \,.\, n^{0,3} .$$

On peut alors établir le tableau II.

On constate que la différence significative D croît relativement rapidement dès que n est supérieur à 1. Pour une parcelle élémentaire de seize arbres, par exemple, D est multiplié par 2,3 par rapport à D calculé pour une parcelle élémentaire d'un arbre. Ce coefficient est même sous-estimé, car la valeur de q croît également puisque le nombre de degrés de liberté affecté à l'erreur décroît quand n croît. La progression de la précision dépend bien entendu de la valeur de la pente b de la régression Log V_n en fonction de Log n. Cette progression sera d'autant plus grande que b est petit. Ce coefficient est aisément calculable dans un environnement donné, si on possède les données individuelles dans une plantation relativement grande pour l'estimer avec une précision correcte.

TABLEAU II

Evolution de la différence significative entre deux moyennes de traitement en fonction de l'effectif d'arbres dans la parcelle élémentaire

Effectif d'arbres dans la parcelle élémentaire (n)	Différence significative entre deux moyennes de traitement (D en % de la moyenne générale)
1	$C.q.2^{0,3}$ = C.q
2	C.q x 1,23
4	$C.q.4^{0,3}$ = C.q x 1,52
8	C.q x 1,87
12	C.q x 2,11
16	C.q x 2,30
25	C.q x 2,63
36	C.q x 2,93
40	C.q x 3,02
64	C.q x 3,48

Incontestablement, dans un champ de surface donnée, le dispositif expérimental qui donne la plus grande précision est la randomisation totale arbre par arbre, quand on admet que les lignes de bordure interparcelles ne sont pas nécessaires.

Les considérations suivantes expliquent ou renforcent ce résultat :

— aucune donnée n'est à estimer s'il y a perte d'arbres ; seuls les survivants sont pris en compte et l'analyse est valide même si les moyennes par famille sont calculées avec des effectifs variables ; par contre, dans un plan en blocs, les moyennes parcellaires, seules prises en compte, seront calculées sur des effectifs différents, après un certain temps, et n'auront donc pas, dans l'ensemble, la même signification (on est toutefois obligé de passer outre dans l'analyse). La présence d'un manquant perturbe les productions des arbres voisins : cette perturbation est difficile à mesurer, aussi a-t-on intérêt à ce qu'elle intervienne sur des arbres appartenant à plusieurs familles afin de la répartir et donc de l'atténuer, alors que dans un plan avec parcelles élémentaires, elle interviendra le plus souvent sur des arbres de la même famille ;

— des taches de fertilité médiocre apparaissent inéluctablement dans un champ d'essai : toutes les familles ont des chances d'y être représentées par un ou plusieurs individus et tous seront donc affectés, alors que dans un plan avec parcelles élémentaires, une ou deux familles peuvent être pénalisées par hasard, ceci entraînant de plus un accroissement accidentel important du carré moyen résiduel de l'analyse de variance, diminuant donc la précision ;

— chaque arbuste d'une famille donnée (hybride, clone, cultivar, etc.) est en contact avec, au mieux, huit familles différentes, qui varieront dans l'ensemble du champ. Les compétitions seront donc très variées et la réaction globale de la famille à l'environnement immédiat interviendra dans la précocité, la récolte. C'est un critère de sélection à ne pas négliger, les familles devant être distribuées le plus souvent en mélange aux planteurs ; ceci est la réponse à la critique souvent entendue concernant la perturbation provoquée par les compétitions entre arbustes voisins ;

— les conditions de pollinisation pouvant varier dans le périmètre de l'essai, la dispersion des arbustes d'une même famille est la garantie d'une certaine homogénéisation de ces conditions pour l'ensemble des familles ;

— aucune information n'est perdue, toutes les données individuelles sont prises en compte et le nombre de degrés de liberté affecté au carré moyen résiduel est bien supérieur à celui obtenu dans tout autre dispositif.

Estimation de la précision d'un dispositif en randomisation totale

Si on connaît approximativement le coefficient moyen de variation des productions individuelles intrafamille en un lieu donné, on peut, à l'aide de formules simples de statistique, établir une liaison entre le nombre d'arbres à adopter par famille et les différences entre moyennes que l'on désire significatives.

Pour étudier ce problème, on ne va d'abord prendre en considération que le risque de première espèce α : c'est-à-dire la probabilité de conclure à une différence alors qu'en réalité il n'y en a pas. Les calculs seront effectués dans le cas le plus simple : comparaison de deux familles (ou traitements).

Illustrons graphiquement ce risque de première espèce. Supposons que la différence vraie Δ entre les deux familles A et B soit nulle ($\Delta = 0$). Si nous expérimentons ces deux familles un grand nombre de fois, les différences d_1, d_2, ...,

$d_i, \ldots, d_m$ observées dans chaque essai entre A et B vont se distribuer selon une loi normale de moyenne nulle. Si on appelle S_d l'écart type de ces différences, la quantité $e = d/S_d$ suivra une loi normale réduite (moyenne nulle, écart type égal à 1).

L'expérimentateur fixe α. On répartit α aux deux extrémités de la courbe, le test étant bilatéral (on ne sait *a priori* si la différence éventuelle entre A et B est dans un sens ou dans l'autre) (fig. 1).

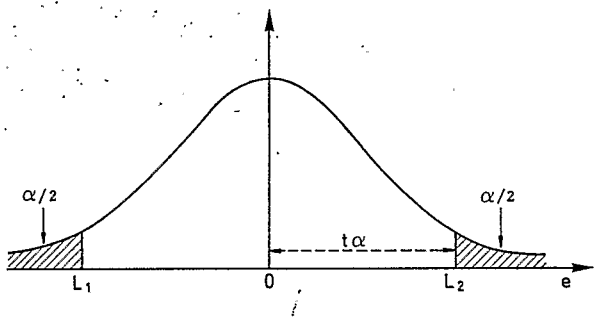


Fig. 1. — Distribution de $e = d/S_d$ (loi normale centrée réduite), représentation du risque de première espèce

On conclura donc à une différence significative dès que $|d/S_d| = t\alpha$; or, $S_d = \sqrt{\frac{S^2 A}{r} + \frac{S^2 B}{r}}$, $S^2 A$ et $S^2 B$ étant les variances observées dans les traitements A et B. En supposant que $S^2 A$ et $S^2 B$ sont homogènes et estimés par $S^2 e$, on a $S_d = \sqrt{\frac{2 S^2 e}{r}} = Se \sqrt{2/r}$ et l'égalité plus haut devient :

$$\frac{d}{Se \sqrt{2}} \cdot \sqrt{r} = t\alpha$$

(r étant le nombre de répétitions pour les familles A et B)

$$d = t\alpha \cdot \sqrt{2} \cdot Se/\sqrt{r}\,.$$

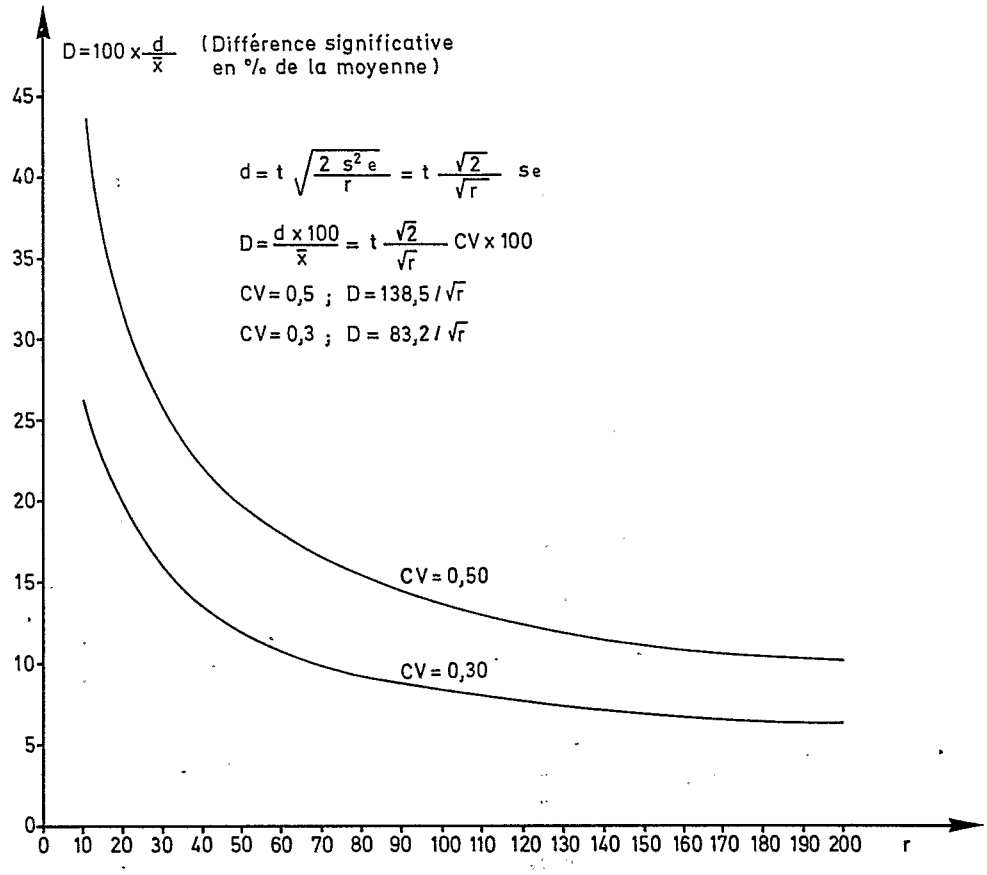
En divisant par $\bar{x}$ (moyenne générale de l'essai), on a :

$$d/\bar{x} = t\alpha \cdot \sqrt{2} \cdot \frac{Se}{\bar{x}} \cdot \frac{1}{\sqrt{r}}$$

donc, en pour-cent de la moyenne, on a :

$$D = t\alpha \cdot \sqrt{2} \cdot CV/\sqrt{r}$$

ou encore : $$r = 2\left(\frac{CV}{D}\right)^2 t\alpha^2\,.$$



Graphique 2. — Précision d'un dispositif en randomisation totale

C'est le nombre de répétitions nécessaire pour qu'une différence D (en pour-cent de la moyenne générale) soit significative quand on s'attend, dans l'essai, à un coefficient de variation CV, avec un test effectué au seuil α (risque de première espèce).

Le nombre de degrés de liberté affecté à l'erreur étant élevé dans ce type de dispositif, on peut « approximer » $t\alpha$ par 1,96 (seuil 0,05) et on a :

$$D = 100 \times 2{,}77 \times CV/\sqrt{r}$$
$$D = 277 \times CV/\sqrt{r}$$

(en pour-cent de la moyenne générale).

On peut tracer la courbe donnant D en fonction de r pour un CV donné. Le graphique 2 montre la liaison entre D et r pour deux coefficients de variation.

$$50\,\% \rightarrow D = 138{,}5/\sqrt{r}$$
$$30\,\% \rightarrow D = 83{,}2/\sqrt{r}\,.$$

On constate que dans le premier cas, il faut adopter r voisin de 80 pour avoir D situé près de 15 % de la moyenne générale. Au-delà, l'augmentation de la précision est relativement faible en fonction de r. Dans le deuxième cas, pour une précision semblable, il faut $r \neq 30$ et il semble inutile pour la même raison d'aller au-delà de 60.

Prise en compte du risque β de deuxième espèce

Le risque β de deuxième espèce doit être pris en compte.

Supposons que la différence vraie entre les deux familles soit $\Delta = K$ et non plus 0. La quantité $e = d/S_d$ se distribuera selon une loi normale de moyenne $M = K/S_d$ et d'écart type unité.

La figure 2 indique la distribution de e dans le cas où $\Delta = K$ et dans celui où $\Delta = 0$.

La partie hachurée sur la gauche de la courbe 2 correspond au risque β de deuxième espèce, qui est la probabilité de ne pas déceler une différence qui, en réalité, existe. En effet, à chaque fois que e sera inférieur à L2, alors que la véritable différence est $\Delta = K$, on conclura que $\Delta = 0$, alors qu'en réalité $\Delta = K$.

Il est clair que β diminue (donc que la quantité $(1 - \beta)$, appelée puissance du test, augmente) quand α augmente. Il y a antagonisme entre les deux risques.

Le risque β ne se situe qu'à une extrémité de la courbe (2) : il est unilatéral.

La distance OL2 est égale à $t\alpha$ (test bilatéral correspondant au risque α). La distance ML2 est égale à $t2\beta$ (test unilatéral correspondant au risque β).

$$OM = t\alpha + t2\beta$$

or :
$$OM = K/S_d = \frac{K}{\sqrt{\frac{2\,S^2\,e}{r}}}$$

On a donc :

$$\frac{K}{\sqrt{\frac{2\,S^2\,e}{r}}} = t\alpha + t2\beta\,.$$

Après résolution, on obtient :

$$r = 2\left(\frac{Se}{K}\right)^2 \cdot (t\alpha + t2\beta)$$

ou, en divisant Se et K par $\bar{x}$ (moyenne générale de l'essai) et en multipliant par 100 :

$$CV = 100\,.\,Se/\bar{x} \quad \text{et} \quad D = 100\,.\,K/\bar{x}$$

(différence entre les moyennes des deux traitements en % de la moyenne générale).

$$D = (t\alpha + t2\beta) \cdot \sqrt{2} \cdot CV/\sqrt{r}$$

et
$$r = 2\left(\frac{CV}{D}\right)^2 (t\alpha + t2\beta)^2\,.$$

r est le nombre de répétitions qu'il est nécessaire de faire dans une expérience où on attend

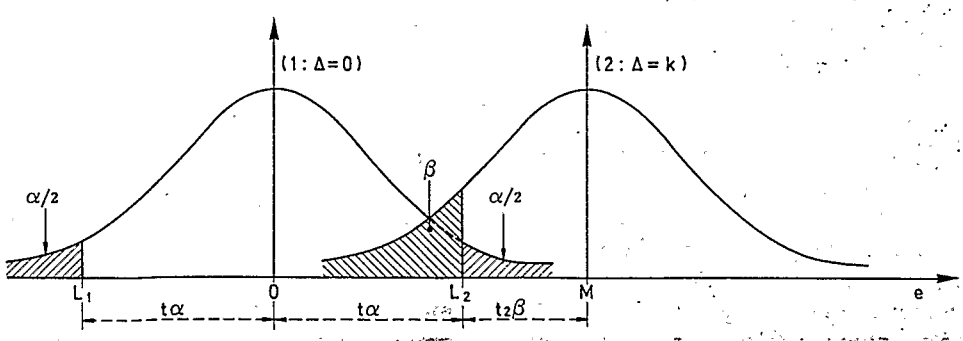


Fig. 2. — Représentation du risque de deuxième espèce

un coefficient de variation CV, pour avoir moins de β chances de ne pas déceler une différence qui serait supérieure D % avec un test fait au seuil α (risque de première espèce).

Par rapport à la formule établie dans le cas où on ne considère que le risque de première espèce, il y a simplement intervention de $t2\beta$. Le graphique 2 établi sans cette intervention est donc optimiste.

On peut construire des abaques correspondant à différentes valeurs de α, β, CV. Un compromis entre α et β doit être trouvé.

Ces calculs ont été réalisés pour le cas de deux familles. Dans le cas de la comparaison globale de plusieurs familles, ils sont nettement plus laborieux, mais avec les formules données on a une approximation suffisante.

Exemples

— **Randomisation totale arbre par arbre**

CV attendu : 40 % (optimiste)

D que l'on veut significatif en pour-cent de la moyenne générale : 20 %.

$\alpha = 0{,}05$ $\qquad t\alpha \neq 2$ (n élevé)

$\beta = 0{,}25$ $\qquad t2\beta = 0{,}674$

$$r = 2\left(\frac{40}{20}\right)^2 (2 + 0{,}674)^2 = 57 .$$

Si on ne tenait pas compte du risque de deuxième espèce, r serait égal à 32.

— **Randomisation totale parcelle par parcelle**

CV attendu : 20 % (optimiste)

$D = 20\ \%$

$\alpha = 0{,}05$

$\beta = 0{,}25 \quad (t2\beta = 0{,}674)$

$$r = 2\left(\frac{20}{20}\right)^2 (2 + 0{,}674)^2 \neq 14 .$$

Si on ne tenait pas compte du risque de deuxième espèce, r serait égal à 8.

On constate que, du fait des valeurs relativement élevées des CV, r atteint rapidement des valeurs qu'il est difficile parfois d'envisager.

Comment améliorer la précision ?

— Il faut abaisser le coefficient de variation par la recherche d'un champ d'essai le plus homogène possible, quitte à abandonner dans ce champ des zones à fertilité manifestement médiocre (où des arbres hors essais seront installés). La randomisation arbre par arbre n'impose aucune contrainte dans la forme du champ.

— Il faut tenter d'associer à la variable étudiée une covariable indépendante permettant une mesure de certains facteurs de perturbation. Une covariable semble intéressante : la moyenne des données mesurées sur les huit arbres contigus à chaque arbre de l'essai (croissance ou production). Cette moyenne, $\frac{1}{8}\sum_{i=1}^{8} V_i$ (V_i représentant la valeur de la variable sur l'arbre contigu i), représente une sorte d'indice de fertilité ou de valeur de l'environnement pour la petite zone entourant l'arbre.

On contrôlera ainsi, pas à pas, les conditions micro-locales de croissance et de production. On peut tenter une analyse de covariance en prenant pour chaque arbre, comme covariable, la valeur de cet indice et, s'il y a effectivement une liaison variable/covariable dans l'ensemble de l'essai, l'introduire dans l'analyse, ajuster les moyennes par hybride, éliminant donc les variations observées des indices arbre par arbre pour l'ensemble de l'essai. Cette technique a été utilisée avec succès dans un essai d'hybrides en Côte d'Ivoire sur des données de croissance (coefficient de corrélation entre variable et covariable : 0,5, CV passant de 51 à 44 %).

On peut améliorer les valeurs de cet indice en calculant :

$$\frac{1}{8}\sum_{i=1}^{8} (V_i - \overline{VH_i})$$

$\overline{VH_i}$ représentant la moyenne des valeurs de la variable sur l'ensemble des arbres (dans l'essai) de la famille situé en i. On tiendra compte alors de l'identité des familles contiguës à chaque arbuste, alors qu'on ne le fait pas dans le premier indice.

$$\begin{array}{ccc} V_{5}\,\circ & V_{6}\,\circ & \circ\, V_7 \\ V_4\,\circ & (Y) & \circ\, V_8 \\ V_3\,\circ & V_2\,\circ & \circ\, V_1 \leftarrow V_i \end{array}$$

Y = performance de l'arbre (variable à étudier),

V_i performance du voisin i (i variant de 1 à 8),

$\overline{VH_i}$ = performance moyenne dans le champ d'essai de la famille située en i,

X = covariable associée à Y

$$= \left(\sum_{i=1}^{8} (V_i - \overline{VH_i})\right)/8.$$

On peut calculer l'indice :

- sur les huit arbres voisins,
- sur quatre arbres (deux situés dans la même ligne, deux situés au même niveau dans les lignes voisines),
- sur les deux arbres situés dans la même ligne, etc., et on adopte l'indice conduisant à la plus grande diminution du carré moyen résiduel de l'analyse de variance.

C'est la méthode dite des « plus proches voisins ».

Si un gradient de fertilité apparaît nettement dans le champ d'essai, on peut adopter la randomisation totale arbre par arbre à l'intérieur des différents blocs constitués sur le terrain. Des disparitions d'arbres survenant au cours de l'essai, l'analyse de variance nécessite préalablement un tirage au sort pour avoir des effectifs égaux d'arbres par famille, par bloc, pour tous les blocs.

On pourrait tenter aussi une analyse de covariance sur le nombre de manquants contigus à chaque arbre en affectant éventuellement à chacun d'eux un coefficient de pondération proportionnel au temps écoulé depuis la mort de l'arbre. Nous sommes ici sceptiques et nous pensons qu'il vaut mieux procéder aux remplacements continus, quitte à les considérer hors essai, s'ils sont effectués après un an.

Les productions individuelles suivent une distribution log-normale donc unimodale. On sait que, dans ce cas, une moyenne, dès qu'elle est calculée sur un effectif suffisant, suit une loi quasi-normale : 5 pour une distribution peu dissymétrique, 10 pour une distribution très dissymétrique. On a donc peu à s'occuper de la normalité dans un dispositif en randomisation totale par plant où les effectifs sont suffisamment élevés. Par contre, l'homogénéité des variances intra-groupes est une condition importante à vérifier pour que l'analyse de variance soit valide et éviter une perte d'information avec, pour conséquence, une diminution importante de l'efficacité de l'analyse. Si les variances intra-groupes sont liées aux moyennes suivant une loi simple, il existe des transformations appropriées qui permettent d'homogénéiser des variances (logarithme, racine carrée, etc.).

Mais le plus souvent une loi n'apparaît pas clairement et on peut être amené alors à supprimer une ou plusieurs familles anormales, le plus souvent à faible moyenne, donc inintéressantes, afin de se placer, pour celles qui sont retenues, dans des conditions acceptables d'homogénéité de variance.

L'installation de ces dispositifs et le contrôle sont plus laborieux que pour un plan avec parcelles élémentaires, mais le gain important en précision plaide très nettement en sa faveur.

Cas où une ligne de bordure interparcellaire est nécessaire

L'évolution de la précision des essais dans un champ de N arbres en fonction de la taille des parcelles élémentaires utiles sera étudiée.

Le problème à résoudre est, comme précédemment : quelle est la taille de la parcelle élémentaire qui donne la plus petite variance de la moyenne parcellaire dans ce champ de N arbres sachant que l'on a toujours la relation $V_n = V_1/n^b$ (V_1 étant la variance des données individuelles, b étant un coefficient à estimer dans une étude préalable comme on l'a vu précédemment).

Soit une parcelle élémentaire utile rectangulaire de dimensions l et l' (l et l' représentant les nombres de cacaoyers sur les longueur et largeur du rectangle, l' pouvant être égal à l). Le nombre de cacaoyers par parcelle élémentaire utile est donc $n = l \times l'$. Supposons qu'on adopte une ligne de bordure. Le nombre de cacaoyers par parcelle élémentaire totale est donc : $(l+2) \times (l'+2)$.

La variance de la moyenne parcellaire calculée sur $l \times l'$ arbres est, d'après la loi de Fairfield Smith : $V_1/(l \times l')^b$. Le nombre de parcelles élémentaires que l'on peut mettre dans un champ de N arbres est $N/[(l+2).(l'+2)]$. La variance de la moyenne des moyennes parcellaires est :

$$V_n = [V_1/(l \times l')^b]/[N/[(l+2) \times (l'+2)]]$$

$$V_n = \frac{V_1}{N} \cdot \frac{(l+2)(l'+2)}{(l \times l')^b} .$$

Il faut calculer l et l' minimisant la valeur V_n, en annulant les dérivées $\partial V_n/\partial l$ et $\partial V_n/\partial l'$.

$$\frac{\partial V_n}{\partial l} = \frac{V_1}{N} \cdot \frac{(l'+2)(l.l')^b - b.(l.l')^{b-1}.l.(l+2).(l'+2)}{(l.l')^{2b}} .$$

$\partial V_n/\partial l = 0$ quand :

$$(l'+2).(l.l')^b = b.(l.l')^{b-1}.l.(l+2).(l'+2) .$$

La résolution de cette équation aboutit à :

$$l = 2b/(1-b) .$$

L'expression V_n étant symétrique en l et l', la résolution de $\partial V_n/\partial l' = 0$ conduit également à :

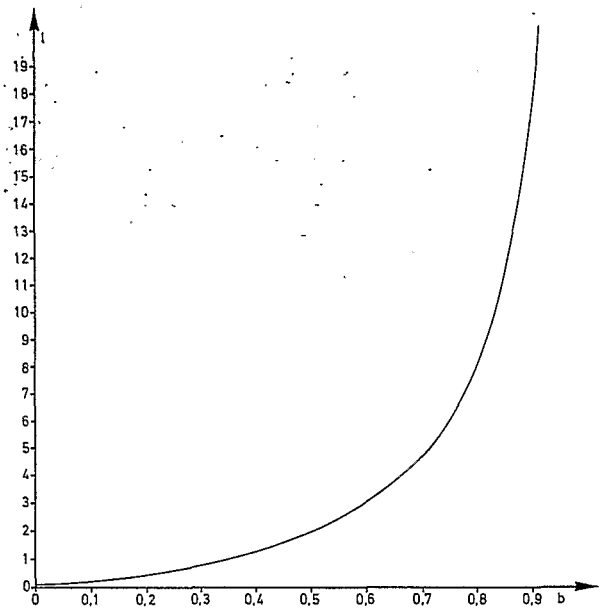
$$l' = 2b/(1-b) ,$$

soit à une parcelle utile carrée de dimension :

$$l = 2b/(1-b) .$$

Voici donc une liaison permettant de déterminer l'optimum concernant la longueur du côté de la parcelle utile carrée (l en nombre de plants) en fonction de la valeur du coefficient b de la loi de Fairfield Smith, calculé dans une écologie donnée.

Le graphique 3, qui montre l'évolution de l en fonction de b variant de 0 à 1, permet de prendre une décision. On peut prendre comme stratégie d'adopter pour valeur de l le nombre entier immédiatement supérieur à la valeur lue sur la courbe. Dans notre cas par exemple, pour b ≠ 0,4, on prendra l = 2, soit un effectif de 4 par parcelle utile et de 16 par parcelle totale.



Graphique 3. — Evolution de la taille de la parcelle élémentaire utile (représentée par la longueur du côté de la parcelle carrée, en nombre de plants = l) en fonction de la valeur de b, coefficient de Fairfield Smith

Notons la faible dimension optimale obtenue du fait de la faible valeur de b, conséquence des conditions de l'étude : ombrage définitif conservé, constitué d'essences choisies et variées de la forêt antérieure. Après déforestation totale au départ et conduite ultérieure sans ombrage définitif, il est logique de penser que la dimension optimale doit être plus grande, puisque certains facteurs d'hétérogénéité ont disparu, avec pour conséquence une corrélation moins importante entre performances d'arbres voisins.

Ce calcul donne la dimension optimale de la parcelle quand la taille du champ d'essai est donnée, mais d'autres considérations sont à prendre en compte :

- il faut prévoir un nombre d'arbres plus important que l'optimum au départ, étant donné les pertes ultérieures ;
- les productions individuelles suivant une loi log-normale, les moyennes parcellaires suivront une loi normale, si elles sont calculées avec un effectif suffisant (supérieur à 10). La transformation logarithmique des données a ici un intérêt moins évident ;
- il est parfois nécessaire, dans certains essais, d'obtenir un effet de masse : modification d'un micro-climat par un ombrage particulier (temporaire ou définitif).

Douze arbres semblent être en conséquence une limite inférieure à adopter pour la taille de la parcelle élémentaire utile entourée d'une ligne de bordure, même si le calcul aboutit à une taille optimale inférieure comme cela a été obtenu avec les données utilisées.

Quelle doit être la limite supérieure ? Elle est donnée par le fait que le plan d'expérience qui sera installé sur le champ doit conduire à une analyse de variance comportant un nombre de degrés de liberté à l'erreur supérieur à 10, 10 étant le strict minimum admissible (15 est plutôt recommandé).

En conclusion, dans un champ de surface donnée, si le calcul aboutit à une taille inférieure à douze, on adopte douze ; si le calcul aboutit à une taille supérieure à douze, étant donné le protocole à mettre en place, on détermine le nombre de degrés de liberté affectés au résidu aléatoire : s'il est supérieur à 10, on adopte la taille calculée, s'il est inférieur à 10, on diminue la taille de la parcelle (ayant pour conséquence l'augmentation du nombre de répétitions) jusqu'à ce que ce nombre soit supérieur à 10, diminuant ainsi la précision de l'essai, mais se plaçant dans les conditions de validité de l'analyse de variance.

Si on n'est pas limité par la surface du champ d'essai, la limite inférieure doit toujours, pour les mêmes raisons, être fixée à 12, mais quelle limite supérieure adopter ? L'observation du graphique 1 montre qu'à partir de vingt-cinq arbres utiles environ par parcelle élémentaire utile, l'augmentation de la taille de celle-ci n'amène qu'une diminution négligeable de la variance de la moyenne parcellaire. On n'a aucun intérêt à aller au-delà, à moins que la variable étudiée nécessite la constitution d'une ambiance qui ne peut être obtenue qu'à partir d'une surface plus grande, alors à estimer agronomiquement (essai d'ombrage définitif par exemple).

L'équation liant la précision et le nombre de répétitions en fonction des seuils α et β et du CV attendu est inchangée.

L'intérêt de l'utilisation de la méthode des plus proches voisins n'a jamais été démontré avec des parcelles élémentaires quelle que soit d'ailleurs la culture.

Nota

Les calculs ont été effectués sur des données concernant des cacaoyers, mais les résultats obtenus sont tout aussi valables pour le caféier étant donné la similitude dans les conditions d'installation (sol), les variations interarbustes dans les performances, etc. La démarche à suivre pour prendre une décision est strictement la même.

CONCLUSIONS ET RECOMMANDATIONS

La méthodologie à adopter pour les essais de sélection est basée sur les résultats de l'étude fondamentale précédente, pour ce qui concerne les dispositifs à utiliser lorsqu'on peut supprimer sans inconvénient les lignes de bordure autour des parcelles élémentaires utiles.

Dispositif expérimental

• La randomisation totale arbre par arbre doit être retenue quelles que soient les caractéristiques du champ d'essai. Le coefficient de Fairfield Smith sera, dans les conditions de culture du cacaoyer, toujours inférieur à 1. Une estimation de ce coefficient b, dans des conditions écologiques données, permettra toutefois de quantifier l'intérêt de ce dispositif par rapport à celui en parcelles de n arbres. Les multiples avantages du dispositif concernent :

— l'inégalité des effectifs, inéluctable après un certain temps ;

— la représentativité de la donnée analysée ;

— la possibilité d'abandon de certaines zones infertiles *a priori* et même *a posteriori* (termitières, affleurements rocheux, etc.) ;

— l'adoption d'une forme quelconque du champ d'essai ;

— la perturbation occasionnée sur les voisins par la disparition d'un arbre ;

— la pollinisation ;

— la compétition interarbre prise en compte en tant que critère supplémentaire de sélection ;

— l'intérêt d'un nombre élevé de degrés de liberté affecté au carré moyen résiduel de l'analyse de variance.

• La méthode des plus proches voisins (Papadakis) sera utilisée, car elle permet de prendre en compte, par une analyse de covariance, l'hétérogénéité du milieu. Le calcul de la covariable peut varier suivant notamment le dispositif de plantation (on peut prendre la moyenne des performances relatives des huit, des quatre, ou éventuellement des deux voisins). Les arbres en bordure n'auront que cinq voisins et ceux en coin trois, car on ne peut prendre en compte les données des arbres de la bordure générale, qui sont placés dans des conditions trop particulières, qui influencent de façon trop importante leurs performances.

• Si un gradient de fertilité est manifeste (essai installé sur terrain pentu, en région accidentée), la randomisation totale doit être effectuée dans des blocs allongés, la plus grande dimension des blocs étant perpendiculaire au gradient. Dans ce cas, l'hétérogénéité étant prise en compte par la stratification en blocs, la méthode des plus proches voisins n'a plus autant sa raison d'être et son utilisation soulèverait de toute façon beaucoup de difficultés. Une contrainte intervient ici : pour que le plan soit analysable, il doit être équilibré et les effectifs par famille, par bloc (pour tous les blocs) doivent être identiques, entraînant un tirage au sort des plants pour s'y conformer (l'effectif commun retenu étant le plus petit observé). Ceci entraînera une perte d'informations.

Variables à analyser

Critères végétatifs

En pépinière

Les plants d'un futur essai sont tout d'abord élevés en pépinière. Les effectifs semés par famille sont différents et, à la plantation, on mettra en place des échantillons d'effectifs identiques par famille, constitués par les plants les plus beaux. Les données suivantes seront relevées :

• taux de sélection (rapport du nombre de plants prélevés sur le nombre total de plants obtenus et sur le nombre de graines semées),

• diamètre des plants retenus (pris sous les cotylédons),

• taux de plants chétifs (calculé sur le nombre total de plants obtenus).

Elles pourront servir comme critère supplémentaire de choix dans le cas d'égalité de performances en champ ultérieurement.

En champ

La croissance des plants en champ, mesurée à partir des diamètres à 20 cm au-dessus du sol, est un critère à prendre en considération pour des analyses ultérieures. Il faut que cette croissance soit mesurée après que le « précédent pépinière » puisse être considéré comme négligeable.

La croissance mesurée entre un an et deux ans de plantation semble être une donnée répondant au mieux à cette contrainte (3). La croissance de la surface de la section ($\Delta s = (\pi/4) . (D_2^2 - D_1^2)$, D_2 et D_1 étant les diamètres mesurés en années 2 et 1, semble plus représentative de la vigueur intrinsèque d'un plant que l'accroissement du diamètre ($\Delta D = D_2 - D_1$). En fait, cela conduit à amplifier les écarts intra et interfamilles. Quoi qu'il en soit, les études peuvent être menées avec ces deux variantes du critère vigueur.

Notons que la section du tronc ne représente pas toujours une circonférence parfaite ; dans le cas où l'irrégularité est nette, on doit prendre pour diamètre la moyenne de deux mesures, l'une effectuée suivant le plus grand diamètre et l'autre perpendiculairement à la précédente.

Le taux de mortalité au champ est un critère supplémentaire à relever.

Critères de production

La deuxième ou la troisième année suivant la plantation, les premières cabosses apparaissent. Combien de récoltes ensuite doit-on cumuler pour avoir un classement fiable des familles en comparaison ? Un consensus apparaît pour quatre récoltes, soit un arrêt des contrôles à la fin de la septième année suivant la plantation. On a en effet toujours constaté que l'adjonction de récoltes supplémentaires dans le cumul n'amenait pas de modifications autres que des modifications de détail dans le classement arrêté à cette date. Le calendrier des contrôles doit donc être le suivant :

Année 0 : mesures en pépinière juste avant la plantation,

Année 1 : mesure du diamètre à 20 cm au-dessus du sol. Remplacements pouvant être utilisés ultérieurement malgré une petite perturbation.

Année 2 : mesure du diamètre à 20 cm au-dessus du sol. Remplacements hors essais.

Année 3 : première récolte. Evaluation du nombre et du poids de cabosses.

Années 4, 5, 6, 7 : quatre récoltes. Evaluation du nombre et du poids de cabosses.

Les modalités de récolte seront celles proposées par Ph. Lachenaud (annexe 1, p. 289) avec dénombrement et pesée des cabosses par arbre et utilisation d'un coefficient calculé sur un échantillon de deux à trois cents cabosses pris sur la totalité des cabosses de l'essai (poids de fèves fraîches/poids de cabosses) pour estimer le poids de fèves fraîches par arbre et par passage (4).

Cette méthode permet de réduire nettement le temps nécessaire au contrôle des récoltes individuelles effectuées auparavant avec écabossage arbre par arbre. Ce coefficient α doit être estimé à chaque passage puisqu'il évolue au cours de la récolte. Une modélisation de cette évolution dans le temps est à l'étude (Ph. Lachenaud) et permettra de réduire à nouveau le temps de contrôle.

La régression positive significative production/croissance, qui a été mise en évidence, peut permettre éventuellement le repérage des familles les plus intéressantes au niveau phytotechnique, c'est-à-dire productives et de faible vigueur (annexe 2, p. 289).

Stratégie à suivre

Etant donné ce qui précède, deux étapes sont envisageables : la mise en place d'un grand nombre de familles avec tri rapide sur le critère de croissance ; un essai de confirmation des meilleures familles.

Mise en place d'un grand nombre de familles avec tri rapide sur le critère de croissance

Pour éviter de mobiliser une surface trop importante, puisque l'essai ne doit durer que deux ans, on peut adopter une forte densité, qui alors ne perturbera pas la croissance à observer sur ce court laps de temps. On peut aller jusqu'à un point tel que la compétition interarbres soit encore inexistante deux ans après la plantation

(par exemple, 2 500 arbres à l'hectare, soit un espacement de 2 m × 2 m et peut-être plus). L'arrachage après deux ans permet la mise en place sur le même terrain d'essais du même type. Un grand nombre de familles peut ainsi être mis à l'épreuve sans investissement important en surface.

Essai de confirmation des meilleures familles

Une installation plurilocale est recommandée, car une interaction lieu × famille est possible. Deux densités peuvent être expérimentées, si les moyens le permettent : la densité classique recommandée dans la région et une densité calculée en fonction des observations de croissance de l'essai préalable. Cette densité (DE) peut être estimée de la façon suivante. Elle doit être inversement proportionnelle à la vigueur (V). La loi la plus simple est : $DE = K/V$ (V étant la moyenne des croissances individuelles des sections ΔS et K un coefficient). On peut estimer K d'après les valeurs DET (densité retenue pour le témoin, obtenue dans des essais de densités antérieurs) et VT (moyenne des croissances individuelles chez le témoin) : $K = DET \times VT$. Pour les familles repérées, on peut donc estimer la densité optimale par la relation $DE = K/V$, V étant la « vigueur » calculée pour celles-ci.

La surface des champs d'essais de confirmation doit être réduite pour obtenir le maximum d'homogénéité, avec des lignes de dimension réduite également pour faciliter les opérations de récolte et la saisie des données : vingt familles par exemple avec un nombre d'arbres par famille qui ne devra pas être inférieur à soixante pour que les différences significatives, fonction du coefficient de variation observé, ne soient pas trop importantes.

Deux témoins, issus de pollinisations manuelles, seront introduits dans chaque essai de façon à permettre un regroupement des classements : graphique à deux dimensions pour deux classements, l'un étant porté sur l'abscisse, l'autre sur l'ordonnée et les positions relatives des familles étant obtenues par projection des points figurant les familles (sur les deux axes) sur la droite obtenue en joignant les deux points correspondant aux deux témoins ; pour trois classements et au-delà, la droite utilisée pour la projection des divers classements est celle joignant les points/témoins dans un espace à 3, 4, ... dimensions (méthode multidimensionnelle). Le choix des témoins pourra évoluer ultérieurement avec l'accroissement des performances des familles mises en essai.

Pratiques d'homogénéisation

Afin de réduire au maximum les coefficients de variation, toutes les techniques d'homogénéisation possibles devront être utilisées.

— Recherche d'un champ d'essai le plus homogène possible (la forme peut être quelconque ; des zones à fertilité manifestement médiocre *a priori*, et même *a posteriori*, peuvent être abandonnées).

— Utilisation de matériel végétal de vigueur semblable à la sortie de la pépinière. Pour cela, il est nécessaire de semer les différentes familles qui seront mises en comparaison dans une fourchette de dates la plus étroite possible, et en nombre suffisant pour permettre le choix.

— Traitements avec insecticides endothérapiques mensuels, en pépinière, puis pendant environ dix-huit mois au champ.

— Désherbages totaux.

— Pas d'ombrage définitif.

— Paillage.

— Fumure minérale d'après le « diagnostic sol » dès la plantation.

— Traitements phytosanitaires recommandés par les spécialistes.

Problèmes à résoudre au moment de l'analyse

Normalité des distributions

Les distributions individuelles par famille suivent une loi log-normale (quasi normale pour celles à productions élevées). Une condition de validité de l'analyse de variance est la normalité des distributions. Mais les statisticiens sont d'accord pour conclure que l'analyse de variance est « peu sensible à la non-normalité des populations considérées tant en ce qui concerne le niveau de signification que la puissance du test » (5). Pratiquement, il suffit d'éviter d'inclure dans l'analyse des distributions trop dissymétriques : distributions en i dans notre cas qui concernent les familles à faible production. Ces familles seront à éliminer de l'analyse finale (elles auraient été de toute façon éliminées du choix final).

Homogénéité des variances intrafamilles

Ici les avis sont plus partagés. Dagnelie déclare que l'hypothèse d'égalité des variances est d'importance relativement secondaire, quand les

effectifs des échantillons sont égaux ; par contre, s'ils sont variables, le risque de première espèce peut être influencé considérablement (5). Lison (6) pense que cette condition est plus importante que la première. C'est certain, car il est évident qu'une variance anormalement élevée dans une seule famille va causer une augmentation anormale du carré moyen résiduel et, par conséquent, tous les tests effectués au moyen de cette valeur vont perdre en sensibilité.

En conclusion, l'analyse de variance est toutefois « robuste » pour ce qui concerne l'homogénéité des variances. On peut se situer un peu au-dessus des seuils donnés par les divers tests d'homogénéité sans pour cela rejeter l'analyse de variance.

Mais on constate le plus souvent une hétérogénéité importante, telle qu'on ne peut passer outre. Dans ce cas, il existe, suivant les circonstances, plusieurs façons de procéder :

- Si les variances (ou les écarts types) sont liées aux moyennes suivant une loi simple, il existe des transformations appropriées qui permettent de les homogénéiser (annexe 3, p. 290).

- Si entre les variances (ou les écarts types) et les moyennes, aucune loi n'apparaît nettement, et c'est malheureusement le cas le plus fréquent, il faut observer de plus près les données ; une variance anormalement élevée provient le plus souvent d'une ou de plusieurs données anormalement basses : il est possible qu'on ait pris en compte des plants qui ne méritaient plus de l'être pour des causes diverses : coups de matchettes intempestifs, retard anormal dans le développement dû à des attaques de borers, des bris de tiges, etc. Il faudrait, après un contrôle sévère sur le terrain, placer hors essais tous ces arbustes. Cette élimination s'ajoutant à la mortalité, il est nécessaire de mettre en place un effectif un peu plus élevé que celui estimé par les graphiques de l'étude précédente.

- Si, malgré tous ces efforts, l'homogénéisation des variances est encore loin d'être obtenue, on peut en dernier ressort utiliser le test de Keuls (ou celui de Duncan) généralisé, qui est expliqué en annexe 4, p. 291. Les calculs sont plus laborieux, car on utilise les variances calculées pour chaque famille et non plus une variance estimée par le carré moyen résiduel de l'analyse globale.

Les familles étant mises en place pour obtenir tout d'abord un classement suivant le caractère productivité, il est nécessaire de les protéger le plus parfaitement possible contre les divers aléas. On ne pourra donc, dans un premier temps, utiliser l'essai pour établir un classement suivant les caractères de résistance (ou sensibilité, ou attractivité) divers en champ. On ne pourra le faire, en champ, que lorsque le classement suivant le potentiel de production sera arrêté ; le dispositif en randomisation totale arbre par arbre convient très bien pour ce genre d'étude, car c'est une garantie de l'homogénéité des conditions d'infestations diverses pour l'ensemble des familles. Par contre, on peut envisager parallèlement l'utilisation de tests au laboratoire ou de tests précoces en pépinière, mais ceci est un autre problème.

BIBLIOGRAPHIE

1. LOTODÉ (R.). — Possibilités d'amélioration de l'expérimentation sur cacaoyers. *Café Cacao Thé* (Paris), vol. XV, n° 2, avril-juin 1971, p. 91-104.
2. LOTODÉ (R.), MULLER (R. A.). — Problems of experimentation with cocoa trees. *In* : « *Phytophthora* disease of cocoa », édité par P. H. Gregory, Longman (Londres), 1974, p. 23-50.
3. NYA NGATCHOU (J.), LOTODÉ (R.). — Variabilité de la précocité chez les premiers hybrides obtenus au Cameroun et recherche d'une corrélation entre le diamètre des troncs à un âge donné et la précocité des hybrides. 2e Conférence Internationale sur les Recherches Cacaoyères, Salvador et Itabuna, 19-26 novembre 1967, p. 98-101.
4. LACHENAUD (Ph.). — Une méthode d'évaluation de la production de fèves fraîches applicable aux essais entièrement « randomisés ». *Café Cacao Thé* (Paris), vol. XXVIII, n° 2, avril-juin 1984, p. 83-88.
5. DAGNELIE (P.). — Théorie et méthodes statistiques. Ed. J. Duculot (Gembloux), vol. 2, 1970, p. 124.
6. LISON (L.). — Statistique appliquée à la biologie expérimentale. Gauthier-Villars (Paris), 1958, p. 199.
7. KRAMER (C. Y.). — Extension of multiple range tests to group means with unequal numbers of replications. *Biometrics* (New York), septembre 1956, p. 307-309.
8. SIEGEL (S.). — Non parametric statistics for the behavioral sciences. McGraw-Hill (Londres), 1956, p. 184-194.
9. DUNNETT (C. W.). — Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association* (Washington), vol. 75, 1980, p. 796-800.

Annexe n° 1

CONTRÔLES À EFFECTUER DANS LES ESSAIS EN RANDOMISATION TOTALE

Année 0 : Année de plantation.

Année 1 : — Remplacements des arbres morts.
— Repérage des arbres porteurs de chérelles et/ou de cabosses, qui seront mis à l'essai pour leur probable auto-compatibilité.
— Eventuellement première récolte individuelle en nombre de cabosses (saines, rongées, pourries).

Année 2 : — Derniers remplacements (hors essais).
— Récolte individuelle en nombre et poids de cabosses.

Année 3 : — Récolte individuelle en nombre et poids de cabosses.

Année 4 : — Récolte individuelle en nombre et poids de cabosses selon la technique suivante, qui permet à deux observateurs assistés de deux manœuvres de traiter plus de trois mille cabosses par jour.

Les deux manœuvres récoltent les cabosses sur chaque arbre individuellement en avançant par ligne et en déposant les cabosses au pied de l'arbre. Les deux observateurs passent ensuite : l'un d'eux dénombre les cabosses et les pèse à l'aide d'un peson et d'un seau, l'autre note les résultats.

A la fin de chaque passage de récolte, le coefficient poids de fèves fraîches/poids de cabosses sera déterminé pour un tas de deux cents à trois cents cabosses mélangées ; (deux cents à trois cents cabosses de la récolte seront, le jour même, globalement pesées et cassées ; les rachis seront enlevés et les fèves fraîches pesées).

Toutes ces données permettront d'évaluer la production de fèves fraîches, par arbre, pour chaque famille.

Année 5 et suivantes : Idem à l'année 4.

Sur les registres, les divers événements de la vie de l'arbre (mort, remplacement, recépage, etc.) seront indiqués en toutes lettres dans l'emplacement réservé à cet effet.

Sur les plans et relevés d'existence, on utilisera les symboles suivants :

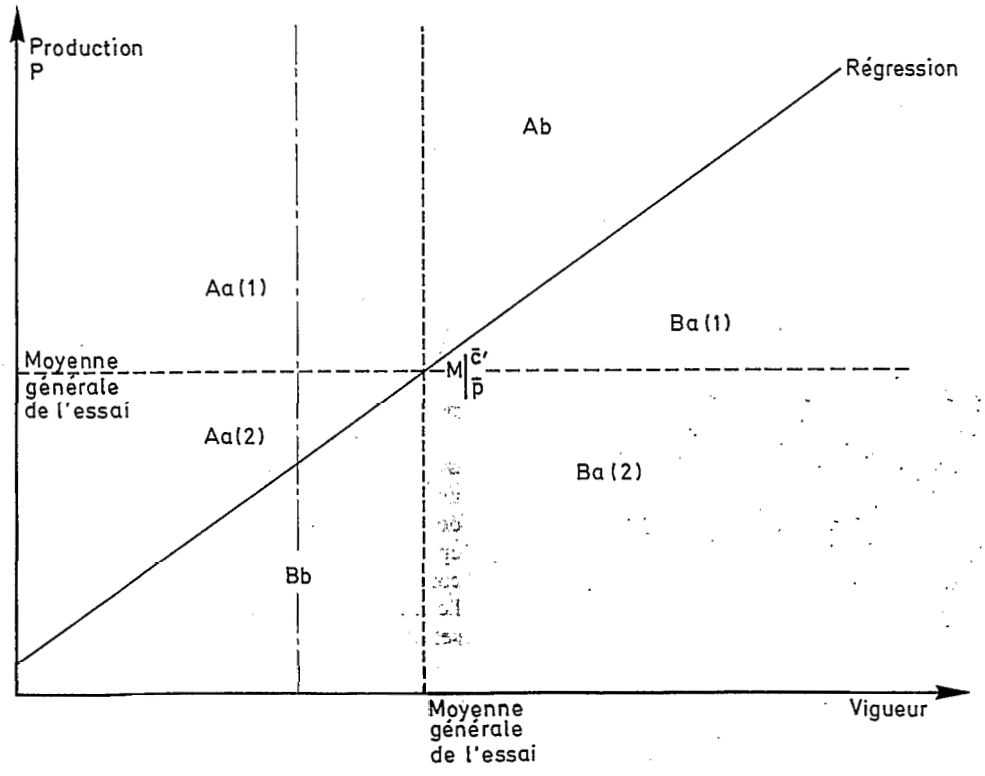
0 : arbre mort,
□ : obstacle à la plantation,
F : fruits (chérelles et/ou cabosses).

Pour se repérer dans les parcelles on peut suggérer :

— **Pancartes de lignes.**
— Etiquettes de localisation en rhodoïd tous les cinq arbres et toutes les deux lignes.

Annexe n° 2

RELATION PRODUCTION — VIGUEUR



Seules sont intéressantes les familles se situant dans les secteurs :

Ab : pour l'établissement de vergers à faible densité.
Aa(1) : **pour l'établissement de haies fruitières à forte densité.**

Annexe n° 3

LES TRANSFORMATIONS DE VARIABLES

Les analyses de variances, conduisant à une comparaison de moyennes de traitements à partir d'une estimation de la variance moyenne intra-traitement, ne sont valides que si on est en droit de calculer cette variance moyenne. On ne peut le faire que si les variances intra-traitements ne sont pas significativement différentes à un seuil donné. Il existe plusieurs tests approximatifs de comparaison de variances :

- test de Bartlett, à formule un peu lourde à manier, conduisant à une valeur distribuée suivant une loi de χ^2, si les variances sont homogènes. Le nombre de variances est quelconque et les nombres de d.d.l. affectés à chacune d'elles également.

- test de Hartley, qui a tabulé le rapport s^2 max./s^2 min. donnant un seuil à ne pas dépasser, seuil variant avec le nombre de variances et le nombre de degrés de liberté affecté à chaque variance (il doit être constant ou quasiment constant). La table est limitée à douze variances.

- test de Cochran, qui a tabulé le rapport $s^2 \text{ max.} / \sum_i s_i^2$ donnant également un seuil à ne pas dépasser, variant avec le nombre de variances et le nombre de d.d.l. affecté à chacune d'elles (constant ou quasiment constant). La table est limitée à vingt variances.

Les deux derniers tests sont plus approximatifs que le premier, surtout celui de Hartley.

Notons que l'analyse de variance est robuste et qu'on peut être en limite des seuils sans inconvénient. Si on s'en écarte nettement, il faut songer à une transformation homogénéisante.

Notons aussi que ces tests sont très sensibles à la non-normalité des distributions, donc peu robustes. Deux transformations simples sont le plus souvent à envisager :

- Si les écarts types des distributions intra-traitements sont approximativement fonction linéaire des moyennes, la transformation logarithmique : $\log x$, $\log(x+1)$, $\log(x+k)$ homogénéise les variances. C'est en général le cas pour les mesures de longueur, de surface, de volume, de production.

- Si les variances sont fonction linéaire des moyennes, la transformation racine carrée est à utiliser ($\sqrt{x}$ ou $\sqrt{x+0{,}5}$). C'est le cas, le plus souvent, pour des données énumératrices (distribuées suivant une loi de Poisson).

Ces deux transformations sont des cas particuliers de la formule générale suivante :

Supposons que les variances des distributions sont fonction de leurs moyennes suivant une loi : $\sigma_x^2 = f(m)$.

(x = variable aléatoire, m = moyenne des x pour une distribution donnée).

Avec une fonction de transformation quelconque $x \to g(x)$, on a, approximativement :

$$\sigma_g^2 = \left(\frac{dg}{dm}\right)^2 \cdot f(m).$$

Si on veut $\sigma_g^2 = \text{Cte} = C^2$, il vient

$$\frac{dg}{dm} = \frac{C}{\sqrt{f(m)}}$$

et

$$g(m) = \int \frac{C\,dm}{\sqrt{f(m)}}$$

Exemple :

- $s^2 x = km^2$ (ce qui revient à $sx = k' m$)

$$g(m) = \int \frac{C\,dm}{k' m} = \lambda \log m .$$

C'est bien la transformation logarithmique proposée plus haut.

- $s^2 x = km$

$$g(m) = \int \frac{C\,dm}{\sqrt{km}} = \lambda m^{1/2} .$$

C'est la transformation racine carrée.

- $s^2 x = km^b$ $\quad (\log s^2 = k' + b \log m)$

On a :

$$g(m) = \int \frac{C\,dm}{\sqrt{km^b}} = \lambda \int \frac{dm}{m^{b/2}} = \lambda m^{1-1/2\,b}$$

transformation utilisée dans les distributions agrégatives (insectes).

Les deux précédentes transformations sont des cas particuliers de celle-ci : $b = 2$, $b = 1$.

Une transformation particulière peut être utilisée dans les problèmes relatifs aux proportions : la transformation angulaire $\arcsin \sqrt{p/100}$. Ceci est justifié pour des variables binomiales ou pour des rapports compris entre 0 et 1, même quand il s'agit de variables continues. Elle ne peut être adoptée que lorsque l'effectif ou le dénominateur du rapport est constant ou sensiblement constant.

Annexe n° 4

LA COMPARAISON MULTIPLE DES MOYENNES

— Lorsque les moyennes par traitement sont calculées à partir d'effectifs égaux et lorsque les variances intra-traitements sont homogènes, les tests de comparaison multiples sont bien connus : Newman-Keuls, Duncan, Dunnett (si on compare tous les traitements à un témoin). Ce sont des tests approximatifs. Par rapport à la méthode de Keuls, la méthode de Duncan est caractérisée par un risque de première espèce supérieur et un risque de deuxième espèce inférieur. Il peut donc arriver, avec la méthode de Duncan, d'obtenir un F non significatif (test exact) et des distances parfois significatives (à la limite). Ce n'est pas grave, car quelle importance peut-on donner au fait que la signification soit atteinte au seuil 5,2 % au lieu de 5 % ?

— Si les variances sont homogènes, mais si les effectifs sont inégaux, on peut utiliser les formules suivantes généralisant le test de Keuls ou celui de Duncan (7) :

$$D_h = \frac{q_h}{\sqrt{2}} \sqrt{\frac{CMR}{n_i} + \frac{CMR}{n_j}}.$$

D_h étant la distance significative entre deux moyennes espacées, après classement, de h moyennes (y compris celles en comparaison) ;

q_h étant lu dans les tables de Keuls ou de Duncan pour ν degrés de liberté (nombre de d.d.l. du carré moyen résiduel obtenu dans l'analyse de variance) ;

CMR étant le carré moyen résiduel ; n_i et n_j étant les effectifs des deux traitements en comparaison.

C'est un test approximatif conservatif, c'est-à-dire un peu sévère.

— Si les variances ne sont pas homogènes, on n'est plus en droit d'effectuer une analyse de variance, car le carré moyen résiduel n'est plus une variance moyenne intra-traitement qu'on peut utiliser pour effectuer les tests de comparaisons multiples de moyennes. Si aucune transformation ne parvient à homogénéiser les variances, on a alors recours à un test non paramétrique.

Supposons un essai en randomisation totale arbre par arbre avec k hybrides et n_J arbres pour l'hybride J, $N = \sum_J n_J$ arbres en tout.

On range l'ensemble des N données de la plus petite à la plus grande et on remplace chaque donnée par son rang.

Soit R_J la somme des rangs concernant les individus de l'hybride J.

On calcule l'expression

$$H = \frac{12}{N(N+1)} \sum_{J=1}^{k} \frac{R_J^2}{n_J} - 3(N+1).$$

H suit approximativement une loi de χ^2 à $(k-1)$ degrés de liberté, si les k populations ne sont pas différentes.

C'est le test de Kruskal-Wallis (8).

Il est toutefois nécessaire que les n_j soient suffisamment grands ($n_J > 5$), mais c'est le cas dans ce type d'essai. Si $n_J \leq 5$, Kruskal-Wallis a tabulé les valeurs de H pour $k = 3$ seulement.

Si le test conduit à accepter l'hypothèse d'égalité, dès le départ, le travail s'arrête. Sinon, il faut recommencer le test en enlevant l'hybride pour lequel R_J est minimum, et ainsi de suite jusqu'à ce qu'on trouve un groupe d'hybrides homogènes pour placer le premier crochet ; on recommence en enlevant l'hybride pour lequel R_J est maximum, etc. C'est très laborieux !

Aussi vaut-il mieux faire tout ce qui est possible pour homogénéiser les variances et adopter une variance commune, même si on est en limite avec les tests de comparaison de variances ! On peut, également, éliminer un ou deux hybrides perturbateurs pour se placer dans les limites de validité.

Toutefois, il existe un test approximatif facile d'emploi recommandé par Dunnett (9) ; c'est un test « conservatif » comme celui de Kramer.

Soit à comparer deux moyennes $\bar{x}_i$ et $\bar{x}_j$ calculées sur des effectifs n_i et n_j, les variances des deux séries de données étant s_i^2 et s_j^2 hétérogènes. Supposons que $\bar{x}_i$ et $\bar{x}_j$ sont séparées par h moyennes après classement (y compris ces deux en comparaison). Soit $q_{i,h}$ la valeur de q lue dans la table de Keuls ou celle de Duncan pour $(n_i - 1)$ degrés de liberté et colonne h, $q_{j,h}$ pour $(n_j - 1)$ d.d.l., et colonne h. On calcule :

$$\tilde{q} = \frac{(q_{i,h} \cdot s_i^2/n_i) + (q_{i,h} \cdot s_J^2/n_J)}{s_i^2/n_i + s_J^2/n_J}.$$

C'est la moyenne de $q_{i,h}$ et $q_{J,h}$ pondérée par les variances des moyennes $\bar{x}_1$ et $\bar{x}_J$. *

On calcule ensuite la différence significative :

$$D = \frac{\tilde{q}}{\sqrt{2}} \sqrt{\frac{s_i^2}{n_i} + \frac{s_J^2}{n_J}}.$$

C'est encore une généralisation du test habituel.

Un cas particulier est celui où les effectifs n_i et n_J sont égaux. Dans ce cas $q_{i,h} = q_{j,h} = q_h$ et on obtient :

$$D = \frac{q_h}{\sqrt{2}} \sqrt{\frac{s_i^2}{n_i} + \frac{s_J^2}{n_J}}.$$

* Dans un essai en randomisation totale pour lequel les nombres de degrés de liberté des variances par famille sont élevés, $q_{i,h} \neq q_{J,h}$ et il est inutile de calculer $\tilde{q}$.

LOTODÉ (R.), LACHENAUD (Ph.). — **Méthodologie destinée aux essais de sélection du cacaoyer.** *Café Cacao Thé* (Paris), vol. XXXII, n° 4, oct.-déc. 1988, p. 275-292, 2 fig., 3 graph., 2 tabl., 9 réf., 4 annexes.

L'efficacité des dispositifs expérimentaux est fonction des techniques d'installation de l'arbuste, de l'hétérogénéité du milieu choisi, de la durée de l'expérience. Après avoir présenté les facteurs de perturbations des essais, les auteurs étudient la taille qu'il faut donner aux parcelles élémentaires pour avoir la plus grande précision possible, premièrement dans le cas où une ligne de bordure interparcellaire n'est pas nécessaire, deuxièmement dans celui où une ligne de bordure interparcellaire est nécessaire. La précision des essais est étudiée et des possibilités d'amélioration de celle-ci sont présentées.

LOTODÉ (R.), LACHENAUD (Ph.). — **Methodology for cocoa selection trials.** *Café Cacao Thé* (Paris), vol. XXXII, n° 4, oct.-déc. 1988, p. 275-292, 2 fig., 3 graph., 2 tabl., 9 réf., 4 annexes.

The effectiveness of the experimental designs depends upon the techniques used for planting the trees, the heterogeneity of the environment and the length of the experiment. The authors describe the factors which disturb the trials and then study how big the plots need to be to make the trials as accurate as possible, firstly for cases where a border row between the plots is not necessary and secondly for cases where it is. The accuracy of the trials is studied together with the scope for improvement.

LOTODÉ (R.), LACHENAUD (Ph.). — **Methodik für Kakaobaum-Zuchtversuche.** *Café Cacao Thé* (Paris), vol. XXXII, n° 4, oct.-déc. 1988, p. 275-292, 2 fig., 3 graph., 2 tabl., 9 réf., 4 annexes.

Die Zuverlässigkeit einer Versuchsanordnung richtet sich nach dem Verfahren der Strauchansiedlung, der ungleichmässigen Beschaffenheit des Substrats und der Versuchsdauer. Die Autoren schildern zunächst verschiedene, die experimentelle Arbeit beeinflussende Störfaktoren und befassen sich im Blick auf einen möglichst hohen Präzisionsgrad mit der optimalen Grösse der Elementarparzelle, je nachdem ob eine Parzellenabgrenzung notwendig ist oder nicht. Die Versuche werden auf ihre Genauigkeit geprüft und es ergehen Anregungen zu ihrer Verbesserung.

LOTODÉ (R.), LACHENAUD (Ph.). — **Metodología destinada a las pruebas de selección del árbol de cacao.** *Café Cacao Thé* (Paris), vol. XXXII, n° 4, oct.-déc. 1988, p. 275-292, 2 fig., 3 graph., 2 tabl., 9 réf., 4 annexes.

La eficacia de los dispositivos experimentales depende de las técnicas de instalación del arbusto, de la heterogeneidad del medio ambiente escogido y de la duración del experimento. Tras presentar los factores que pueden perturbar las pruebas, los autores analizan el tamaño que es preciso dar a las parcelas elementales para obtener la mayor precisión posible, en primer lugar en caso de no ser necesaria una línea de delimitación interparcelaria y, en segundo lugar, de ser necesaria una línea de delimitación interparcelaria. Por último, se estudia la precisión de las pruebas y se presentan diversas posibilidades para lograr su mejoramiento.