

Comparaison de quelques techniques statistiques descriptives et inductives à propos d'un plan expérimental multivarié

J. PAGES (1)

INTRODUCTION

Dans la plupart des expérimentations, les biologistes sont actuellement tentés de prendre en compte un grand nombre de variables; il n'est pas utile ici d'insister sur les avantages qu'il y a à considérer des observations multivariées, c'est-à-dire à connaître les individus auxquels on s'intéresse, non par une mesure seulement, mais par plusieurs.

Si cette conception de l'expérimentation enrichit considérablement les interprétations que l'on peut tirer de l'étude d'un phénomène biologique, elle élargit aussi l'éventail des techniques statistiques utilisables pour renforcer les conclusions. En effet, pour le traitement d'une expérimentation univariée, on peut utiliser directement les techniques inductives (analyse de variance notamment), et les outils descriptifs s'avèrent souvent trop frustes; par contre, lorsque plusieurs mesures interviennent, les premières s'alourdissent rapidement et, parallèlement les secondes sont alors à même de fournir des résultats très intéressants.

Les plans expérimentaux multivariés constituent par excellence le terrain sur lequel les deux optiques statistiques auxquelles il vient d'être fait allusion s'appliquent avec profit; la forte structure du recueil des données (il s'agit d'expérimentation et non d'observation) rend possible l'emploi d'outils inductifs, et la pluralité des mesures confère un grand intérêt à une

approche descriptive. Face à cette situation, le biologiste est souvent en droit d'être embarrassé quant au choix de la technique à employer : l'objectif de cette note est précisément de proposer, à partir de l'étude d'une expérimentation, quelques indications susceptibles d'éclairer l'utilisateur dans ce choix.

I – LE PROBLEME BIOLOGIQUE

Pour illustrer notre propos, nous utiliserons certaines données issues du très important protocole expérimental mis en place par BARLOY et CHEROUVRIER (Chaire de Phytotechnie, E.N.S.A.R.) en 1975. De ce protocole, dont le thème général est l'étude de la physiologie du maïs, nous retiendrons, dans cet exposé, uniquement les analyses de teneur en matière sèche, au moment de la récolte, des différents organes de la plante. Ainsi, **pour chaque pied, nous disposerons de cinq mesures**, se référant à la matière sèche de la tige, des feuilles, des spathes, de la rafle et des grains. En outre, à la demande de MM. BARLOY et CHEROUVRIER, nous adjoindrons à ce tableau la mesure de hauteur au moment de la récolte. Cet ajout se justifie à la fois par l'intérêt scientifique qu'il y a dans la mise en relation des mesures de matière sèche et de hauteur, et par l'avantage pratique qu'il y aurait à remplacer des mesures de matière sèche (difficiles) par des mesures de hauteur (faciles) si cela était possible, dans l'étude de tel ou tel aspect de la croissance.

(1) Chaire de Statistique et Informatique E.N.S.A., 65, rue de Saint-Brieuc 35042 RENNES CÉDEX

Le protocole expérimental proprement dit est décrit en détail dans CHEROUVRIER (1976) et nous n'en rappellerons que les traits principaux (fig. 1, Communication suivante). L'expérience s'étend sur trois grandes parcelles, appartenant à un même champ, qui constituent le contrôle d'hétérogénéité appelé communément «facteur bloc». A l'intérieur de chaque bloc, on cultive sur une même bande de terrain, cinq variétés de maïs (1). Dans chaque bande, quatre emplacements fixes d'observation sont choisis au hasard, qui comportent chacun 40 pieds successifs.

Nous disposons ainsi de 3 (blocs) x 5 (variétés) x 4 (emplacements) = 60 emplacements et de 60 x 40 = 2 400 pieds de maïs. Cependant, dans le cadre de cette étude, nous n'utiliserons pas les mesures sur les 2 400 pieds mais seulement leur moyenne par emplacement.

En ce qui concerne les teneurs en matière sèche, le problème essentiel de cette expérimentation est de proposer une typologie des répartitions de masses au moment de la récolte. Cette typologie aura pour but de préciser les liaisons existant entre les différents organes; par ailleurs, elle sera mise en relation avec l'appartenance à un bloc ou à une variété donnée.

Dans cette étude, deux sortes de données seront prises en compte : les pourcentages de matière sèche (par rapport au total de la plante) qui conduisent à une approche strictement physiologique, et les teneurs brutes qui, résulant chaque type de répartition en fonction du résultat global, sont d'intérêt plus directement agronomique. Ces deux aspects seront abordés de façon complémentaire.

II — APPROCHE MÉTHODOLOGIQUE

1. — Étude de la variabilité des emplacements

Face à un assez grand nombre d'observations, la première préoccupation est de tenter d'apprécier la variabilité des différentes mesures sur l'ensemble de l'échantillon. A ce niveau, il est bon de ne pas faire intervenir la structure du plan expérimental mais simplement de reconnaître, parmi les emplacements, plusieurs tendances susceptibles d'être caractérisées par la présence commune de certains niveaux de plusieurs variables.

Pour cela, on soumettra l'ensemble des données (tableau croisant 60 emplacements et six mesures)

à une A.C.P. (2). Ces données seront au préalable centrées et réduites afin d'éviter tout biais dû à l'hétérogénéité des mesures.

Par ailleurs, on peut s'interroger sur l'aptitude d'autres méthodes factorielles, essentiellement de l'analyse des correspondances, à traiter ces données. En ce qui concerne le tableau initial (60 emplacements x 6 mesures) un tel traitement serait soumis à des artefacts indésirables : en particulier, la pondération par la somme des lignes, qui intervient lors du calcul de la distance du χ^2 , n'a pas grand sens puisqu'elle met en jeu des mesures différentes.

Le traitement du tableau limité aux matières sèches exprimées en pourcentage par A.F.C. est, lui, parfaitement approprié et présente deux spécificités par rapport à une A.C.P. (portant elle aussi sur le tableau des cinq pourcentages) :

- dans le calcul des distances, intervient une pondération par la somme des colonnes qui accroît l'importance des différences observées sur des petites valeurs;

- à profil égal, chaque mesure joue un rôle proportionnel à son niveau moyen.

Ce deuxième point est le plus important et relie chaque analyse à une optique de recherche distincte. En analyse en composantes principales, les mesures jouent, *a priori*, des rôles équivalents : ce qui correspond à une étude purement physiologique. Par contre, en affectant à chaque pourcentage de matière sèche une importance proportionnelle à la place qu'il occupe dans la plante, on se situe sur un plan plus «agronomique» où chaque résultat est pondéré par le niveau de ses conséquences pratiques. Aucun de ces deux points de vue n'est supérieur à l'autre : ils sont simplement différents.

Enfin, un autre traitement possible consiste à subdiviser chaque mesure en classes et à soumettre à l'analyse des correspondances le tableau booléen obtenu, composé de classes disjonctives complètes. Pour la constitution de ces classes nous adoptons la procédure usuelle, justifiée par NAKACHE (1973), qui vise à obtenir des effectifs égaux.

Les traitements proposés sont au nombre de quatre : deux A.C.P., l'une sur les données brutes, l'autre sur les pourcentages; deux A.F.C., l'une sur les données brutes découpées en classes, l'autre sur les pourcentages.

(1) F 7 x E P 1, L G 11, Primeur 170, INRA 258, ANJOU 256.

(2) A.C.P. analyse en composantes principales; A.F.C. analyse factorielle des correspondances.

En principe, ces analyses visent simplement à étudier la variabilité globale de l'échantillon sans tenir compte du plan expérimental; toutefois, ce dernier est quand même pris en compte lors de l'interprétation au cours de laquelle on recherche systématiquement si les emplacements correspondant à une même modalité d'un facteur se regroupent. Cette démarche revient à essayer d'identifier les directions principales de variabilité générale aux facteurs du plan expérimental.

2. Étude de la variabilité des variétés

Cette partie vise à présenter différentes modalités d'introduction de la structure du plan expérimental dans les techniques statistiques. Ce faisant, nous ne chercherons plus à appréhender les données d'un point de vue général, sans *a priori* mais tenterons de répondre à des questions précises. Ces questions seront introduites successivement : tout d'abord, nous nous intéresserons uniquement au facteur essentiel du plan qu'est l'appartenance des parcelles à une variété donnée. Ainsi la première préoccupation sera de chercher à discriminer, non plus les emplacements comme précédemment, mais les variétés. Ensuite seulement, nous essaierons de faire intervenir la structure complète de l'expérience.

a) Analyse en composantes principales des variétés

La formulation du problème en tant que recherche d'une discrimination entre variétés conduit à examiner les points moyens tous blocs confondus, des variétés. L'étude de ce nouveau tableau croisant cinq individus (les variétés) avec les six variables peut être faite par simple consultation, mais aussi à l'aide d'une analyse en composantes principales. Une telle analyse n'est pas à négliger : elle répond parfaitement à l'objectif qui est de séparer les variétés et, à ce titre, complète très bien l'analyse des données brutes. En outre, elle est très facile à mettre en oeuvre. Son défaut majeur est de ne pas tenir compte de la variabilité à l'intérieur de chaque variété : il limite, semble-t-il, ce procédé, à un rôle de simple indicateur. Il est possible cependant d'atténuer cet inconvénient en projetant les emplacements sur les plans factoriels issus de l'analyse des centres de gravité. On obtient ainsi une représentation du nuage initial des 60 emplacements dans un espace de dimension restreinte dont la caractéristique est de séparer au mieux les centres de gravité des groupes qui nous intéressent. La qualité du regroupement des emplacements par variété sur ce plan sera un excellent indicateur de la signification de ce dernier et, par suite,

de l'importance de la perte d'information subie lors de la prise en compte des seuls centres de gravité.

b) Analyse factorielle discriminante

Pour tenir compte de la variabilité à l'intérieur de chaque variété, nous formulerons l'objectif légèrement différemment. Étant donné un échantillon d'individus (60 emplacements) repérés par une série de mesures (teneurs en matière sèche, hauteur finale) et une partition *a priori* sur cet échantillon (l'appartenance à l'une des cinq variétés), est-il possible de retrouver la classe d'un individu en considérant seulement les mesures dont il a été l'objet ? Cette question revient toujours à rechercher les spécificités des classes; pour y répondre on peut utiliser des programmes d'analyse factorielle discriminante qui fournissent, lorsque le nombre des groupes est supérieur ou égal à trois, des représentations graphiques des individus étudiés.

Les axes factoriels que l'on recherche ici maximisent le critère variance inter-classe/variance intra-classe ou ce qui revient au même (ROMEDER, 1973) le rapport variance totale/variance intra-classe. Les variances auxquelles il est fait allusion sont celles des projections (sur des axes), des points représentant les emplacements dans l'espace de départ (R^6 puisque six mesures). Une fois ces axes obtenus, on peut construire des représentations graphiques planes à l'instar des autres méthodes factorielles. Cette utilisation de l'analyse discriminante répond à un objectif essentiellement descriptif : fournir des directions de l'espace initial R^6 qui séparent bien les emplacements appartenant à des variétés différentes et séparent peu ceux d'une même variété.

En outre, on peut montrer (ROMEDER, 1973) que l'analyse factorielle ainsi proposée équivaut à la recherche des axes principaux d'inertie du nuage de points comprenant uniquement les centres de gravité mais en utilisant la métrique W^{-1} (ou T^{-1}) (1). Cette remarque permet de mieux situer la procédure sommaire du paragraphe II, 2, a : au choix de la métrique près, il s'agit d'une analyse discriminante usuelle. Or, le choix de la métrique W^{-1} (ou T^{-1}) repose implicitement (2) sur l'hypothèse selon laquelle la matrice variance intra-classe est la même d'une classe à l'autre. Dans notre exemple, le biologiste a des raisons de penser *a priori* que cela n'est pas vérifié; en effet, parmi les cinq variétés se trouvent des hybrides simples dont tous les individus sont génétiquement identiques, et des hybrides trois voies et doubles dont les individus sont génétiquement

(1) W , T matrice de covariance intra-classe et totale

(2) et même très explicitement si l'on recherche les fonctions discriminantes de façon classique (TOMASSONE, 1974)

différents. Dans une telle situation, l'emploi de la métrique W^{-1} (ou T^{-1}) n'est peut-être pas justifié; auquel cas, l'analyse discriminante n'apporte pas grand chose à une A.C.P. sur centres de gravité.

c) Approche inductive

Pour prendre en compte complètement la structure du plan expérimental, il est nécessaire de se tourner vers des techniques inductives. L'objectif global de l'étude ne change pas : mettre en évidence des types de répartition de matière sèche caractéristiques des différentes variétés. Cependant, nous emploierons ici un intermédiaire : le classement des variétés selon chacune des mesures effectuées. Les techniques inductives permettront justement d'obtenir ces classements avec le plus de précision possible.

Bien que le plan expérimental contienne plusieurs mesures, nous n'utiliserons pas de procédure multivariée proprement dite. Dans leur emploi courant, ces dernières visent à tester des hypothèses très globales, comme par exemple : toutes les populations sont semblables vis-à-vis de toutes les mesures. Une telle hypothèse ne concerne que rarement le biologiste, et même dans ce cas, le gain de puissance obtenu par une procédure multivariée de préférence à plusieurs procédures univariées est douteux (RAO, 1972).

Comparaisons multiples

Pour obtenir un classement, nous effectuerons des tests de comparaisons multiples de moyennes en utilisant conjointement les trois procédures classiques : p.p.d.s., Duncan, Newman et Keuls (1). Depuis longtemps, la p.p.d.s. est critiquée, pour son risque global très élevé. Toutefois, dans l'esprit de FISHER qui l'a proposée, elle ne doit être employée qu'à l'issue d'une analyse de variance ayant mis en évidence un facteur significatif. Cette condition (qui n'est pas requise pour les deux autres procédures) rend, à nos yeux, cette technique acceptable.

Quoi qu'il en soit, ces trois procédures nécessitent une estimation de la variance supposée commune (ce qu'il convient de vérifier) des populations composées. Cette estimation sera obtenue à partir d'une analyse de variance qui devient, de ce fait, un préalable aux comparaisons multiples. Cependant, ce sont ces dernières qui correspondent à notre objectif et c'est pourquoi nous les avons traitées en premier, même si dans l'application cet ordre est inversé.

Enfin, nous appliquerons ces tests indépendamment sur les cinq mesures de matière sèche. Chaque classement obtenu présentant un intérêt propre pour le biologiste, nous ne nous sommes pas préoccupé d'un risque afférent à l'ensemble des mesures.

Analyse de variance

Le plan expérimental est décrit dans la présentation du problème et des données. On est en présence d'un plan complet et équilibré à deux facteurs (variété, bloc) avec répétitions (quatre emplacements par couple de modalités des deux facteurs). Si l'on appelle :

i l'indice des blocs $i \in (1, I) \quad I = 3$

j l'indice des variétés $j \in (1, J) \quad J = 5$

k l'indice des répétitions $k \in (1, K) \quad K = 4$

le modèle peut s'écrire :

$$(1) Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk} \quad \text{avec}$$

γ moyenne considérée

μ moyenne générale

α facteur bloc

β facteur variété

$\alpha\beta$ interaction

e résidu aléatoire

(2) On suppose ici les facteurs fixes. En ce qui concerne les variétés, la population dont sont issues les cinq choisies peut être considérée comme normale et infinie et le facteur variété ne peut être pris en compte comme aléatoire, même si le biologiste désire, en fin de compte, généraliser ses résultats.

(3) A l'intérieur d'un bloc, les quatre répétitions d'une même variété sont toujours placées sur une même bande de terrain. Cette disposition est celle d'un split-plot. Tous les résidus ne sont pas indépendants et, en particulier $Cov. (e_{ijk}, e_{ijk'})$ différent de 0. Le modèle du split-plot revient ici à considérer la corrélation entre e_{ijk} et $e_{ijk'}$ constante quels que soient k et k' , et quel que soit le couple (i, j) concerné. Par contre, d'un bloc ou d'une variété à l'autre, ces résidus sont indépendants, ce qui, il est vrai, n'est pas en accord parfait avec la disposition sur le terrain. Pour chaque mesure, nous

(1) La méthode des contrastes n'est pas adaptée ici où aucune partition *a priori* des variétés ne semble être en rapport étroit avec les critères mesurés.

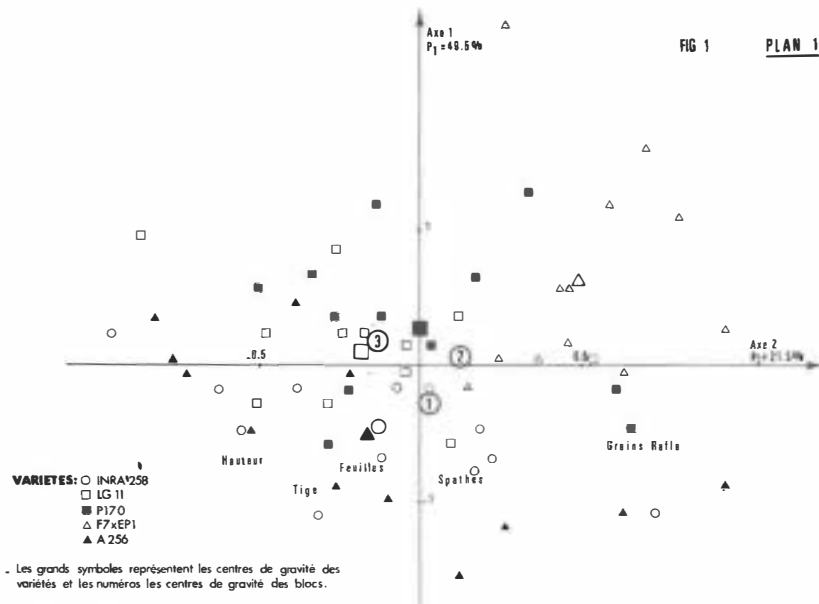


Figure 1 — Plan 1 — PLAN FACTORIEL ISSU DE L'ANALYSE EN COMPOSANTES PRINCIPALES (A.C.P.) DES MATIERES SECHES ET HAUTEURS FINALES. 60 emplacements répartis en cinq variétés. Chaque mesure est une moyenne sur 40 plantes (60 emplacements, 6 mesures).

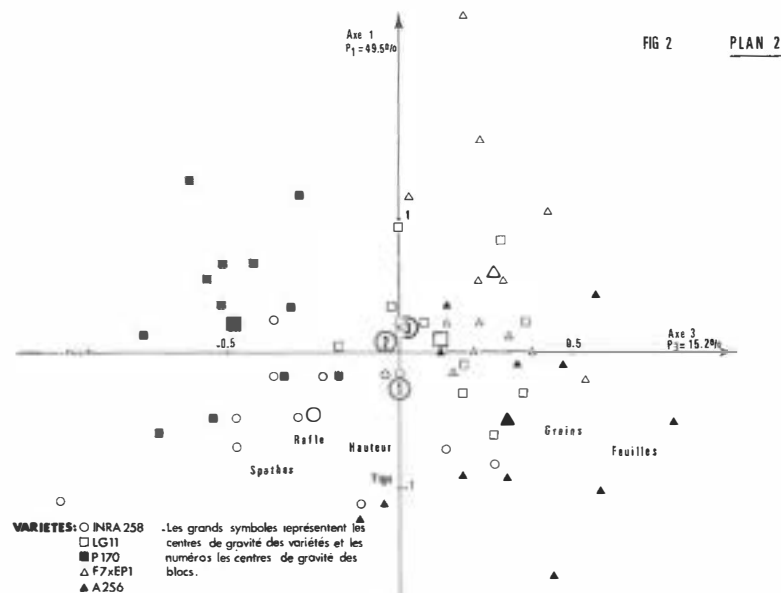


Figure 2 — Plan 2 — PLAN FACTORIEL ISSU DE L'ANALYSE EN COMPOSANTES PRINCIPALES (A.C.P.) DES MATIERES SECHES ET HAUTEURS FINALES. 60 emplacements répartis en cinq variétés. Chaque mesure est une moyenne sur 40 plantes (60 emplacements, 6 mesures).

effectuerons le test global du facteur variété, se référant aux hypothèses suivantes :

$$H_0 \quad \beta_j = 0 \quad \forall j$$

$$H_1 \quad \exists \text{ au moins } 1 \beta_j \neq 0$$

III — APPLICATIONS

1. Analyses factorielles sur emplacement.

(fig. 1 et 2) (1)

(1) Les représentations graphiques des A.C.P. sont du type de celles décrites dans LEBART-FENELON (1975). Les coordonnées d'un point variable sont les coefficients de corrélation entre cette variable et les différents facteurs.

Le premier axe exprime un effet taille : globalement toutes les variables sont corrélées. Le deuxième oppose la matière sèche de l'épi (grains, rafle) à celle du vert (feuilles, tiges). Le troisième oppose feuilles-grains, à rafle-spathes.

Ces trois axes fournissent les trois critères essentiels de classification des pieds de maïs. Parmi eux, seuls

le premier et le troisième correspondent à des spécificités variétales : utilisés conjointement, ils caractérisent bien les cinq variétés comme le montre le plan 2 repéré par les axes 1 et 3. Ce plan suggère deux classements des variétés (tabl. 1).

- TABLEAU 1 -

CLASSIFICATION DES VARIÉTÉS SELON LES FACTEURS DE L'A.C.P. SUR EMBLEMENTS

Hiveau	faible	moyen	fort
Critère			
taille générale (axe 1)	P 7 < A P 1	L G 11 P 170	A 256 I 258
matière sèche feuilles et grains VS. rafle et spathe (axe 3)	P 170	L G 11 I 258	A 256 P 7 < E P 1

Une A.C.P. et une A.F.C. ont été appliquées sur les pourcentages; ces deux analyses, pratiquement identiques, ont fourni un plan factoriel (1, 2) semblable au plan (2, 3) de l'A.C.P. sur données brutes. (Ceci est naturel : le passage aux pourcentages supprime tout effet taille puisque la somme des mesures est constante pour tous les individus. Les facteurs 2 et 3 de l'analyse sur données brutes deviennent alors les deux premiers).

2. — A.C.P. des points moyens des variétés d'analyse factorielle discriminante

Ces deux traitements conduisent à un premier plan factoriel semblable, du point de vue de la position relative des variétés, au plan 2 (axes 1 et 3) de l'analyse précédente. Toutefois, les emplacements appartenant à une même variété y sont mieux regroupés surtout par l'analyse discriminante. Ce fait peut être quantifié en calculant le pouvoir discriminant (rapport variance inter-classes/variance totale) relatif à chacun des axes mis en cause (tabl. 2)

- TABLEAU 2 -

POUVOIR DISCRIMINANT DE FACTEURS ISSUS D'ANALYSES DIFFÉRENTES

Facteur	taille générale axe 1	grains V.S. rafle: axe 2 ou 3
Analyses		
A.C.P. sur p. réelles	0,38	0,66
A.C.P. sur points moyens	0,47	0,67
A. discriminante	0,74	0,65

Le pouvoir discriminant du premier axe s'accroît légèrement d'une A.C.P. à l'autre et devient très forte avec l'analyse discriminante. Ce résultat, très naturel en soi, n'est pas contradictoire avec le fait que, sur chaque plan, la position relative des variétés (schématisée par leur centre de gravité) est la même. Nous sommes dans une situation analogue à celle de la figure 3 où ne sont représentées que trois variétés (chacune par des ellipsoïdes représentant la dispersion des emplacements autour du point moyen) et deux mesures; les

projections des trois centres de gravité apparaissent toujours dans le même ordre mais celles des emplacements se regroupent par variété de façon plus ou moins nette selon les axes.



Figure 3 - ANALYSES FACTORIELLES D'UN MEME NUAGE. Chaque ellipsoïde représente la dispersion des emplacements d'une même variété autour de son centre de gravité.

3. — Approche inductive

Les résultats des analyses de variances peuvent être résumés par le tableau des risques rendant significatives les valeurs des tests F observées (tabl. 3).

- TABLEAU 3 -

ANALYSES DE VARIANCES : RISQUES RENDANT SIGNIFICATIVES
LES VALEURS DE F OBSERVÉES

données	brutes	pourcentages
variables		
tige	0,0009	0,0012
feuilles	0,0019	0,3586
spathe	0,0186	0,0091
rafle	0,2640	0,0015
grains	0,2543	0,0533

Ceux des comparaisons multiples, effectuées uniquement lorsque le facteur est significatif à un risque inférieur à 5 p. 100, sont donnés figure 4. Signalons tout de suite que les différents tests, p.p.d.s., Duncan et Newman-Keuls ont donné les mêmes résultats. La raison en est, en particulier, le faible nombre de moyennes comparées.

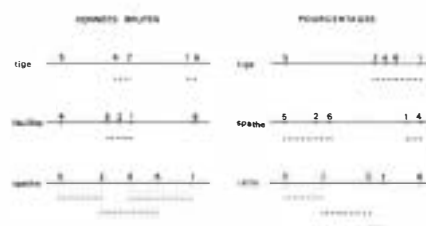


Figure 4 — RÉSULTATS DU TEST DE NEWMAN-KEULS. Les lignes pointillées mettent en évidence les groupes de moyennes non significativement différentes.

Ces résultats sont tout à fait cohérents avec les précédents. Sans entrer dans les détails, qui se trouvent dans CHEROUVRIER (1976), on remarque que pour les données brutes les classements proposés par l'analyse de variance sont toujours très voisins de celui fourni par le critère «taille générale»; à quelques exceptions près (par ex. P. 170 et ses faibles valeurs pour les feuilles) les

différences de structure sont d'un ordre inférieur à celui du facteur taille.

IV – CONCLUSION

Les méthodes qui viennent d'être employées conduisent à des résultats cohérents entre eux. Cependant, cette cohérence, qui traduit le fait qu'une forte structure des données peut être mise en évidence par la plupart des techniques, ne doit pas faire oublier les différents points de vue qui ont été adoptés. Nous ne ferons que mentionner la recherche des facteurs principaux de variabilité : cet objectif est servi sans ambi-

guïté par l'A.C.P. sur tableau de données initial. Par contre, on notera trois approches pour mettre en évidence les particularités des variétés :

- séparer les variétés au moyen des variables initiales (analyses de variance);

- séparer les variétés au moyen des principaux facteurs de variabilité générale (A.C.P. sur tableau initial);

- recherche de la «meilleure» fonction linéaire des variables initiales du point de vue de la séparation des variétés (analyse discriminante ou A.C.P. sur centres de gravité).

SUMMARY

Comparison of some descriptive and inferential statistical methods applied to a multivariate experimental design.

In a statistical study, inferential methods are preferably used. When experimental conditions do not allow it (e. g. hypothesis of normality or homosce-

daslicity not verified), the biologist commonly turns to descriptive methods: This attitude is however dangerous because each of these methods possesses its own aim and does not provide the same type of result. In this paper, several analysis (analysis of variance, principal components analysis, discriminant analysis) are simultaneously applied to one experimental design, and their aims are detailed.