Méthodologie mathématique des études des liaisons station-production

C. Millier

Ann. Sci. forest., 1973, **30** (3), 351-366

MÉTHODOLOGIE MATHÉMATIQUE DES ÉTUDES DES LIAISONS STATION-PRODUCTION

C. MILLIER

Station de Biométrie,
Centre national de Recherches forestières, I. N. R. A.,
Champenoux 54370 Einville

RÉSUMÉ

On esquisse le principe des méthodes statistiques et mathématiques dans une étude des liaisons station-production en insistant sur les problèmes de causalité. C'est par rapport à elle qu'on envisage la signification des principales techniques (régression, composantes principales et analyse factorielle).

En conclusion, on oppose les modèles descriptifs (ou prédictifs) aux modèles explicatifs qui demandent beaucoup plus d'information qualitative au biologiste et qui peut-être fourniront des compléments à l'expérimentation.

PRÉAMBULE

Nous ne nous donnerons pas ici pour but de décrire très précisément et très abstraitement les différentes techniques mathématiques utilisées dans les problèmes de liaisons station-production. Le temps et la place manquent; l'insertion de cet exposé dans cette journée nous oblige à renvoyer les lecteurs intéressés à des textes de base théoriques ou pratiques.

La présentation sera donc générale et nous essaierons de mettre en valeur les éléments qui justement permettront de préciser les qualités, les limites et les insuffisances de ces techniques. La critique sera d'abord relativement théorique, mais nous espérons qu'elle s'enrichira d'un aspect concret dans la discussion suivante (discussion d'exemples concrets).

1. — CARACTÉRISTIQUES DU MATÉRIEL RECUEILLI LORS DES ENQUÊTES STATION-PRODUCTION

a) Les différents spécialistes coopérant dans une *enquête* de liaison station-production définissent d'abord un *échantillon* constitué de *placettes*. La définition de l'ensemble qu'est l'échantillon pose des problèmes importants que nous négligerons volontairement :

α) problèmes de représentativité. De quelle façon l'échantillon représente-t-il la *population* de toutes les placettes implantables dans le *domaine* d'enquête?

β) dimensions de la placette. Les exigences du pédologue, du dendrométricien, du phytosociologue sont différentes. Le problème est éludé ici, car il s'agit essentiellement d'un problème méthodologique inhérent à chaque discipline. Sachons seulement que ces deux questions rejailliront sur la qualité des résultats, d'abord d'un point de vue immédiat (interprétation), ensuite d'un point de vue inductif (extrapolation de l'échantillon à la population tout entière).

b) Une fois définie la placette, le recueil des données commence. Évidemment abondant. Toutefois l'abondance n'est pas l'écueil majeur : les moyens de calcul modernes permettent une grande liberté d'action (ce qui ne veut pas forcément dire qu'il faut totalement l'utiliser!)

α) certaines données sont globales, c'est-à-dire définies au niveau de la placette tout entière : hauteur dominante, note d'abondance-dominance d'une espèce, niveau de la nappe d'eau. D'autres sont recueillies au niveau d'un autre élément que la placette : l'arbre, l'horizon ou la tranche de sol. Le passage de l'arbre à la placette par exemple devra être défini avec précision et l'information subsidiaire recueillie dans ce passage sera interprétée à l'intérieur de la placette : variation, liaisons entre caractères intra-placettes, etc.

Néanmoins, le niveau de nutrition caractéristique de la placette choisi pour expliquer la production sera-t-il le niveau de nutrition maximum, moyen sur les 10 plus gros arbres, moyen sur l'ensemble des arbres de la placette? Pour les questions de sol, le problème est plus compliqué : si l'on peut estimer qu'à part les relations de concurrence et de consanguinité les performances de deux arbres d'une même placette sont indépendantes, l'indépendance entre les mesures réalisées à des profondeurs différentes sur le même caractère est vraisemblablement irrecevable; le profil d'évolution en fonction de la profondeur est d'ailleurs souvent intéressant pour caractériser le sol.

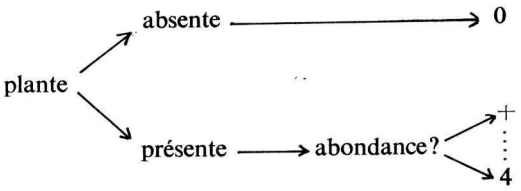
C'est pourquoi très souvent on considère les mesures sur un même élément comme des mesures réalisées sur des caractères différents.

β) la codification des données est vitale et a des conséquences importantes sur les résultats. La qualification d'un caractère déterminera le traitement mathématique. Faut-il se contenter de notes de présence-absence pour caractériser une plante dans une placette? Le niveau de nutrition pour un élément se mesure-t-il par la quantité d'élément contenu dans 100 g de feuilles ou bien est-il mieux traduit par la quantité totale d'élément dans l'arbre?

Ces différents choix dépendent en grande partie de la théorie sous-jacente et sont parfois limités par des questions expérimentales (évaluation de la masse foliaire, mesure de l'abondance-dominance). Ils se traduisent au niveau des calculs mathématiques.

Un exemple : si l'on adopte le cœfficient d'abondance-dominance, on peut être abusé par cette échelle quasi-continue 0 + 1 2 3 4. Devant cet écueil, le phytosociologue a d'ailleurs la précaution de noter + la présence de quelques pieds seulement de l'espèce, ce qui permet de mettre en valeur la différence de nature entre la note 0 et les autres notes (+ 1 2 3 4).

D'un point de vue statistique, on a donc une structure complexe



ce qui la rend particulièrement peu maniable pour les calculs mathématiques (voir ci-après).

2. — LA STRUCTURATION DE BASE

Soit $y^i = (y^i_1, y^i_2, ..., y^i_p)$ l'ensemble des performances réalisées par la placette i pour les caractères 1, 2, ..., p. On peut visualiser y^i comme un vecteur $\overrightarrow{OY}^i$ défini dans un espace dont la base est contituée par les différents *caractères* (ou *variables*) (fig. 1).

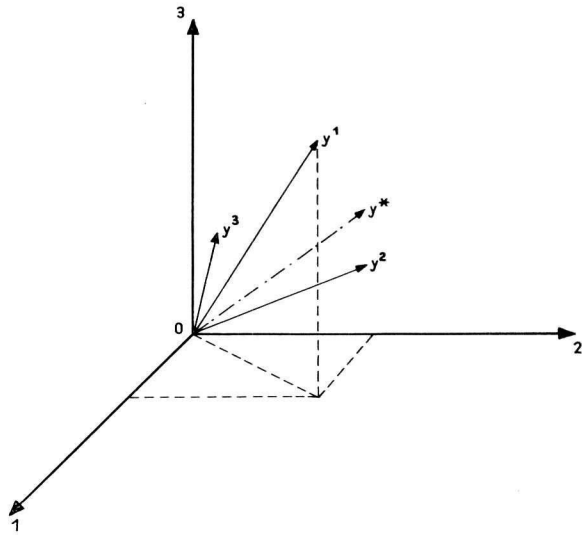


FIG. 1. — *L'espace des placettes*
1 : *hauteur dominante* (m)
2 : *teneur en Ca des aiguilles* (me)
3 : *profondeur de la nappe*
y^1 : *placette n° 1*
y^2 : *placette n° 2*
y^3 : *placette n° 3*
$y^* = a_1 y^1 + a_2 y^2 + a_2 y^3$

FIG. 1. — *Vectorial space of the plots*

Alors $y^* = \Sigma\, a_1 y^1 = a_2 y^2 + a_i y^i + \ldots + a_n y^n$ est un ensemble de performances réalisables par une placette hypothétique appartenant au domaine étudié. Toutefois (et c'est là où on voit l'élément probabiliste) cette placette hypothétique peut avoir la probabilité 0 d'exister.

Par exemple, s'il n'y a qu'un caractère ($p = 1$), la hauteur dominante, si $y^1 = 30$, $y^2 = 32$.

$$\{ a_1, a_2 \} = \{ 1, 1 \} \quad \rightarrow y^* = 62 \text{ m n'existe pas}$$
$$\{ a_1, a_2 \} = \{ 1/2, 1/2 \} \rightarrow y^* = 31 \text{ m peut exister}$$

Soit $y_j = (y_j^1, y_j^2, \ldots, y_j^n)$ l'ensemble des performances réalisées sur le caractère j par les n différentes placettes. On peut visualiser y_j comme un vecteur $\overrightarrow{OY}_j$ défini dans un espace dont la base est constituée par les différentes placettes (fig. 2).

Alors $y^* = \Sigma\, a^j y_j = a^1 y_1 + a^2 y_2 + \ldots + a^p y_p$ est un ensemble de performances réalisables par un caractère hypothétique déductible des caractères initiaux.

Si l'alinéa précédent était recevable (concept de placette hypothétique à partir du continuum du champ d'enquête), celui-ci doit être expliqué. Lorsque, par exemple, le percepteur, à partir de signes extérieurs de richesses (possession d'un château, d'un chien de luxe, etc.) et d'un système de coefficients a^1, a^2, ... a^p calcule un index de signes extérieurs de richesse, il crée un caractère hypothétique qui n'a aucune signification concrète, mais a une valeur opératoire évidente.

La démarche est ici complètement semblable : les caractères hypothétiques créés auront une valeur opératoire mesurée par leur pouvoir explicatif et l'importance relative des a^j permettra d'interpréter la signification de ce caractère hypothétique.

Ces différentes remarques peuvent paraître très secondaires et extérieures à notre dessein. Elles sont en réalité vitales : elles confèrent à l'espace des placettes et à l'espace des caractères une structure d'*espace vectoriel* (E.V.) tout comme l'espace visuel de dimension 3 classique dans lequel nous nous déplaçons (1).

Une fois admise cette structure, le mathématicien l'aménage en définissant des angles et des distances. On pourra dès lors apprécier quantitativement la proximité entre deux placettes $d\,(i, i')$ et la proximité entre deux caractères j et j' c'est-à-dire leur corrélation *(espace euclidiens)*.

Toutes les techniques décrites plus loin supposent l'existence de cette structure, l'une des plus simples structures mathématiques existantes, reflétant ainsi la faiblesse des connaissances théoriques exigibles.

Des bases théoriques plus importantes (en particulier la connaissance des modes d'action de certains caractères sur d'autres) ne détruisent pas forcément la structure d'E.V., mais

1. Voici la définition axiomatique d'un espace vectoriel. Soit x, y, z trois éléments de l'E. V., α, β deux réels, alors 1) $x + y = y + x$ 2) $(x + y) + z = x + (y + z)$ 3) il existe 0 de l'E. V. tel que $x + 0 = x$ pour tout x 4) $\alpha(x + y) = \alpha x + \alpha y$ 5) $(\alpha + \beta)x = \alpha x + \beta x$, 6) $(\alpha\beta)x = \alpha(\beta x)$, 7) $0\ x = 0$, $1\ x = x$.

elles obligent à mesurer la plausibilité de certains y_j. D'autre part, certaines situations très simples ne peuvent pas être rendues par une structure d'E.V.

Reprenons par exemple le coefficient d'abondance-dominance défini en 1 comme la composition de deux caractères élémentaires différents

$$y_1 : \begin{matrix} \text{présence } 1 \\ \text{absence } 0 \end{matrix}$$

$$y_2 = \text{dominance} \begin{cases} + \ \ldots \ 1 \\ 1 \ \ldots \ 2 \\ 2 \ \ldots \ 3 \\ 4 \ \ldots \ 5 \end{cases}$$

la définition « conditionnelle » de y_2 rend impossible la structure d'E.V., puisqu'on ne peut fixer la valeur de y_2 quand $y_1 = 0$.

3. — LES MÉTHODES MATHÉMATIQUES UTILISÉES

a) *Méthodes descriptives*

Reconsidérons les $y^i = (y_1^i, \ldots, y_p^i)$ et les $y_j = (y_j^1, \ldots, y_j^n)$. Ces vecteurs sont plongés respectivement dans des espaces euclidiens de dimension p et n, R^p et R^n. Cette réalité est donc très complexe et est donc hors de portée de la manipulation et de la visualisation (c'est-à-dire estimation des proximités entre placettes et entre caractères).

La tâche du statisticien est uniquement de présenter des documents graphiques utilisables condensant au maximum l'information contenue dans les y^i et les y_j [1].

Pour utiliser des analogies mécaniques, soit $\bar{y} = (\bar{y}_1, \bar{y}_2, \ldots \bar{y}_p)$ le centre de gravité des y^i; remarquons au passage que $\bar{y}$ est une placette hypothétique définie avec $(a_1, a_2, \ldots, a_n) = \left(\frac{1}{n}, \frac{1}{n} \ldots, \frac{1}{n}\right)$ (par exemple).

On peut définir l'inertie de l'ensemble des y^i relativement à chaque direction de l'espace E_p. On peut démontrer qu'elle s'exprime simplement à partir des coefficients qui servent à définir la direction (remarquer qu'ils ne sont autres que les a^j envisagés en 2) et de la matrice d'inertie V.

$$V = \begin{bmatrix} \frac{1}{n} \sum_i (y_1^i - \bar{y}_1)(y_2^i - \bar{y}_2), \frac{1}{n} \Sigma (\bar{y}_2^i - \bar{y})^2 \ldots\ldots\ldots \\ \frac{1}{n} \sum_i (y_1^i - \bar{y}_1)^2, \qquad\qquad , \frac{1}{n} \Sigma (y_1^i - \bar{y}_1)(y_2^i - \bar{y}_2) \ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{bmatrix}$$

Pour définir une inertie, il a fallu :

α) définir des poids relatifs à chaque placette : ici $a_i = \frac{1}{n}$

β) définir une distance dans E^p : ici distance euclidienne classique.

---

1. Nous n'utiliserons pas ici le fait que l'information contenue dans l'espace E^p et celle contenue dans E^n sont équivalentes; simplement elles sont présentées de façon différente. Cette remarque très importante est à la base de très importants développements mathématiques qui ont leur importance pratique. Nous ne pouvons hélas pas nous y attarder.

Cette matrice d'inertie est appelée ici matrice de variance — covariance : on reconnaîtra dans la diagonale principale les variances de chacun des caractères; si on fait $v_{ij} = v_{ij} / \sqrt{v_{ii}\, v_{ij}}$ on reconnaît dans v_{ij} le coefficient de corrélation.

Il existe une direction de l'espace $(a^{1*}_{(1)}, a^{2*}_{(1)}, \ldots, a^{p*}_{(1)})$ qui possède une inertie maximale, c'est-à-dire, pour revenir à notre langage, qui absorbe le maximum de variabilité des placettes.

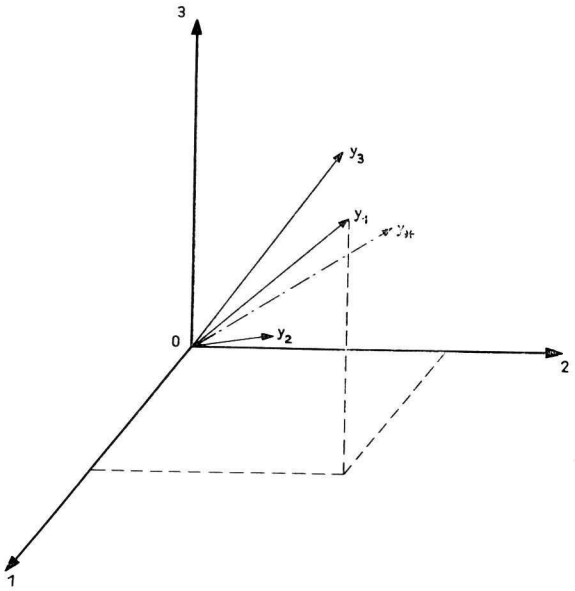


FIG. 2. — *L'espace des caractères*
1 : *placette n° 1*
2 : *placette n° 2*
3 : *placette n° 3*
y_1 : *caractère hauteur (n° 1)*
y_2 : *caractère Ca (n° 2)*
y_3 : *caractère profondeur (n° 3)*
$y_* = a^1y_1 + a^2y_2 + a^3y_3$

FIG. 2. — *Vectorial space of the variables*

On peut également définir une direction de l'espace sans corrélation avec la première qui condense ensuite le maximum de variabilité et ainsi de suite. On s'arrête dès que la part de variabilité absorbée par le sous-espace découvert est jugée suffisante. Nous avons réalisé une *analyse en composantes principales;* les directions trouvées seront appelées *facteurs.*

Au niveau que nous avons atteint, quel est l'acquit? *La simplification de la représentation.* C'est un peu le problème d'un aveugle qui aurait parfaitement reconnu le caractère plan d'un obstacle qui se présente à lui, mais ne saurait pas si c'est un plancher ou un mur.

Son sens de l'équilibre lui permet automatiquement *d'identifier* ce plan comme un mur ou comme un plancher. Nous devons donc maintenant, après avoir simplifié la représentation, identifier cette représentation simplifiée. Aussi comparerons-nous nos directions ($a^*(1)$, $a^*(2)$, $a^*(3)$) à notre connaissance sous-jacente.

Les proximités (corrélations) des différents caractères avec le caractère hypothétique permettront de dégager sa signification.

Les classements des placettes sur ce caractère hypothétique souvent préciseront la signification dans la mesure où on n'a pas intégré dans les p caractères toutes les connaissances que nous avons (on n'a pas par exemple introduit le caractère « type de sol » qualitatif et difficilement manipulable).

b) *Méthodes prédictives*

Soit $y_{p+1} = (y^1_{p+1}, ..., y^n_{p+1})$ un caractère supplémentaire que l'on veut prédire à partir des p autres caractères.

En abrégeant (car le raisonnement de 3*a* est plus complet et plus intéressant), il s'agit de trouver la direction $a^* = (a^{1*}, a^{2*}, ..., a^{p*})$ telle que l'angle entre cette direction et y_{p+1} soit le plus faible possible (1).

$$(\overrightarrow{OY}^*, \overrightarrow{OY}_{p+1}) = \text{minimum.}$$

Cet angle minimum quantifie la liaison qui existe entre y_{p+1} et les autres y. Son cosinus est le coefficient de corrélation multiple (en valeur absolue). Nous avons réalisé une analyse de régression multiple.

[En attribuant à y^i_1, y^i_2, y^i_p un statut de caractère observé explicatif fixe (c'est-à-dire mesuré sans erreur) et sur y^i_{p+1} un statut aléatoire, on peut à partir des résultats calculés sur l'échantillon, tester les différents coefficients a^{j*}, calculer l'erreur avec laquelle on peut prédire pour n'importe quelle placette la valeur y^i_{p+1} connaissant les y^i_j].

4. — PROBLÈMES DE CAUSALITÉ

Pour présenter la discussion des techniques mathématiques décrites en 3, il faut d'abord que nous nous interrogions sur ce que le biologiste attend de telles analyses.

4.1. — *Postulat de départ : statistique et causalité*

Les calculs compliqués ou simples prennent en compte des nombres quantifiant les liaisons entre les caractères : coefficient de corrélation, ou la proximité entre des placettes : distances euclidiennes, coefficients de Jaccard ou Rulczinski (phytosociologie).

La mesure de cette liaison est uniquement « statistique » et ne préjuge en aucun cas des relations causales existant entre caractères. De fait, les diverses publications qui intègrent ces calculs sont souvent décevantes : pour tenir un pari difficile d'objectivité, les auteurs interprètent du bout des lèvres les résultats.

Par exemple, il est dit : « la deuxième composante principale révèle « l'opposition » entre X1 et X5. Cette proposition n'est pas utilisable, car elle n'est pas l'énoncé d'un fait et, pour être utilisée par la suite, ou pour enrichir la connaissance, *cette proposition doit être traduite en termes de causalité.*

A ce niveau, le plus strict positivisme ne peut conduire qu'à une impasse conceptuelle et qu'à considérer les mathématiques comme un appoint à la mode.

4.2. — *Conséquences de l'introduction de la causalité*

La causalité se traduit par des relations complexes entre les différents caractères : relations de hiérarchie, un caractère est déterminé par un autre ; relations de feedback

entre caractères (très cohérentes si l'on imagine que l'unité biologique arbre ou sol cherche à tamponner les effets climatiques et de milieu) ; effets d'interaction.

A côté de cela, la structure d'espace vectoriel paraît anormalement simple. Elle est linéaire, interdit en particulier les variables à seuil, l'introduction aisée d'interactions. Elle met sur pied d'égalité toutes les variables sans tenir compte de leur place dans la chaîne de relations de l'écosystème.

Cette structure est pauvre et passive, objective néanmoins. Sa généralité permet, on le verra, une utilisation paradoxale dans la sélection de structures causales.

4.3. — *Position générale de la causalité dans les liaisons station-production*

Nous avons visualisé (fig. 3) la place relative des différentes unités dans l'écosystème considéré. Le schéma est grossier et certains critiqueraient l'absence de flèches entre certains des composants. Il permet de situer l'isolement « topographique » de l'objet de la phytosociologie.

A l'intérieur de chaque composant, les relations entre les différents caractères peuvent être très variables : relations allométriques de l'arbre (masse foliaire, hauteur, etc.) dérivant de règles morphologiques de développement, relations causales dans le sol et simples relations statistiques au niveau des plantes (en ignorant les problèmes de concurrence et de compétition entre plantes différentes).

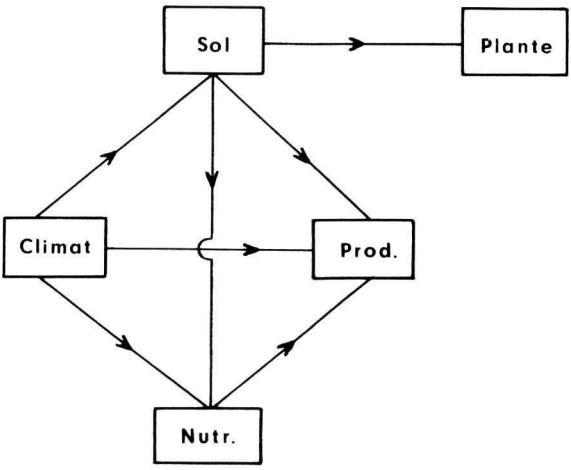


FIG. 3. — *Diagramme causal des différents constituants du système*
FIG. 3. — *Causal diagram between the different components of the biological system*

4.4. — *Causes lointaines et causes proches*

Suivant le niveau de reconnaissance choisi, nous pouvons donc expliquer un phénomène par différentes causes très différentes les unes des autres.

Nous distinguerons causes lointaines et proches, très intuitivement par leur éloignement topographique de la variable qu'on cherche à appréhender.

Prenons un exemple : le diagnostic foliaire peut servir à caractériser la production et la croissance d'une placette. Si, en tant que pédologues, nous cherchons à étendre le problème aux variables de sol, nous faisons intervenir des variables qui sont des causes lointaines car intervenant sur la production par le biais de la nutrition. L'alimentation en eau de l'arbre, plus difficile à mesurer sur l'arbre, sera caractérisée par une mesure de sol appropriée, qui sera alors considérée comme une cause proche de la production.

Ainsi, à moins de contrôler ou de mesurer toutes les causes lointaines ou proches influençant une variable, *il sera impossible de séparer, donc d'identifier les différentes actions.*

Nous considérons la génétique tout à la fois comme une cause lointaine et proche : en particulier, dans le cas d'une forêt naturelle, proche en tant que bruit de fond des analyses effectuées, lointaine parce que les populations d'arbres se sont adaptées progressivement au milieu.

On pourrait considérer de même certaines pratiques sylvicoles dont l'action s'est étendue sur plusieurs siècles.

5. — SIGNIFICATION CAUSALE DES TECHNIQUES MATHÉMATIQUES EMPLOYÉES

5.1. — *Méthodes prédictives* (fig. 4)

Voyons la signification causale de l'analyse de régression multiple. Les variables explicatives peuvent être corrélées entre elles, soit par le hasard de l'échantillonnage, soit parce qu'elles dérivent de caractères placés en amont causalement. En aucun cas, elles ne doivent avoir entre elles de relations de cause à effet.

Pour la première partie de la proposition, remarquons que les plans d'analyse de variance orthogonaux permettent très rigoureusement de réaliser un coefficient de corrélation nul entre deux variables données.

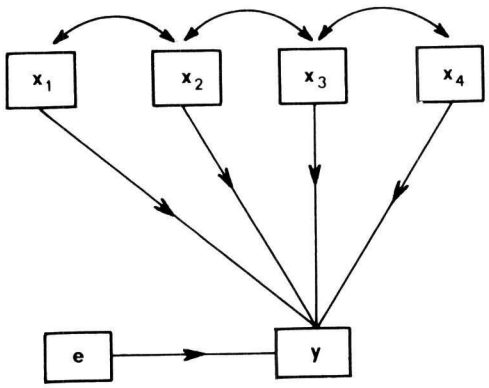
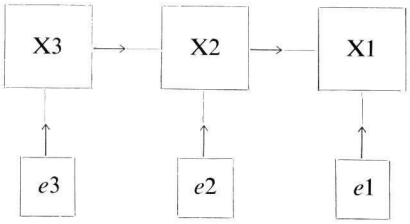


FIG. 4. — *Diagramme causal de la régression multiple*
FIG. 4. — *Multiple regression causal diagram*

Dans ce cadre, on introduit la notion de corrélation partielle : on calcule le coefficient de corrélation entre deux caractères en ayant éliminé l'influence d'un tiers, appelé coefficient de corrélation partielle. Par les termes même employés, on voit que cette mesure n'a de valeur que dans ce schéma causal; il est inconvenant et irrégulier de calculer le coefficient de corrélation partielle entre X1 et X3 dans le schéma ci-dessous (X, variable; *e*, élément aléatoire).



Ces schémas sont irréalistes et incorrects dans la plupart des situations biologiques. Si on cherche à se placer dans cette situation, il faut sélectionner des variables dont on peut supposer *a priori* qu'elles sont sur le même plan.

Est-ce à dire que la technique de régression n'apporte rien? Elle permet de *prédire* correctement une variable expliquée, mais non d'expliquer un comportement. Elle n'est donc rien qu'une *boîte noire* dont on connaît les entrées et la sortie et dont on veut ignorer le fonctionnement interne.

Dans des analyses typologiques, elle permet rapidement d'arriver à des résultats intéressants.

5.2. — *Problèmes particuliers des méthodes prédictives : effet d'atténuation*

Nous n'avons pas insisté jusqu'à présent sur la nature des variables. La variable expliquée reçoit un statut aléatoire qui tient compte de tous les facteurs non mesurés pudiquement nommés erreurs, des erreurs d'adéquation de modèle (on a lissé une forme linéaire alors que la relation est multiplicative ou parabolique) et qui permet d'exécuter des tests statistiques.

Les variables explicatives sont supposées fixes, c'est-à-dire mesurées sans erreur. Cette hypothèse est irrecevable dans les études de variabilité régionale : on ne peut imaginer que la nutrition variable au niveau de la feuille soit appréhendée correctement (= sans erreur) par la quantité d'éléments contenue dans x grammes de matière sèche.

Cet élément, non pris en compte dans les études, se traduit par un effet d'atténuation des liaisons, donc d'affaiblissement des résultats. Si on considère les résultats au niveau de l'arbre (et non de la placette), on peut avoir les informations permettant de régler le problème.

5.3. — *Méthodes descriptives* (fig. 5)

La structure causale est là encore simple. Les caractères sont supposés être les *manifestations* causales de facteurs (= caractères hypothétiques). Dans le cas de l'analyse en composantes principales, ces facteurs sont sans corrélation mutuelle.

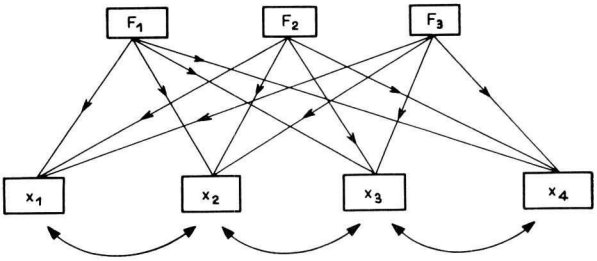


FIG. 5. — *Diagramme causal de l'analyse des composantes principales*
Notez qu'il n'y a pas de liaisons entre F_1, F_2, F_3

FIG. 5. — *Component analysis causal diagram*

Nous examinerons en 5.5 un cas typique d'utilisation correcte d'une telle analyse, mais dans la plupart des cas, les structures sont plus compliquées. La mythique factorielle s'est développée d'abord en psychologie et sociologie, et est conditionnée par l'existence de structures sous-jacentes non accessibles qu'on trouve commode de typer et de séparer : un test psychométrique est la manifestation de différentes qualités fictives mais à valeur opératoire évidente : intelligence, habileté manuelle.

5.4. — *Utilisation causale des résultats des analyses descriptives*

Comme en 5.1 se pose le problème de l'utilité de telles techniques.

Comme on l'a vu plus haut, ces techniques sont d'abord destinées à visualiser au mieux une réalité complexe et comme ce graphique ne peut être interprété que causalement, il faut savoir comment fonctionne l'interprétation et jusqu'à quel point elle peut être poussée.

L'analyse des composantes principales, confrontée à l'expérience et aux connaissances exogènes du biologiste, doit suggérer des structures causales complexes qui enrichiront la connaissance. On courra évidemment le risque de ne pas choisir *exactement* la bonne structure, mais le problème n'est pas différent de celui d'un test statistique.

La problématique de l'interprétation est donc réelle. Les applications de l'analyse des composantes principales reste de ce côté très en retrait des possibilités théoriques. L'attitude du biologiste, l'insuffisance actuelle de la validation des méthodes (simulation de modèles théoriques, extension systématique de la méthode) freinent sensiblement les progrès.

Trop souvent le biologiste choisit entre deux solutions aussi peu réalistes l'une que l'autre :

1) ne retenir que les facteurs qui viennent confirmer une hypothèse sous-jacente du biologiste. Rejeter les autres facteurs même s'ils paraissent expliquer une grande partie de la variabilité;

2) conserver tous les facteurs importants et leur donner une interprétation même claudicante, ce qui revient à accorder une foi totale dans la structure d'espace vectoriel.

La solution optimale consiste à faire progresser concomitamment le modèle causal et les calculs multivariables.

On peut en particulier étudier pas à pas chaque niveau d'explication (études par horizon, études nutritionnelles) de façon à éviter le mélange déjà cité de causes proches et lointaines. On doit en particulier considérer avec critique les caractères synthétiques, d'une part normaux (ph, compacité, etc.), d'autre part calculés (l'activité biologique, rapports d'éléments). Certains ne sont rien d'autre que des estimations *a priori* de composantes principales particulières et risquent de masquer les rapports réels entre les variables, et de préjuger des résultats.

5.5. — *Cas particulier des études phytosociologiques*

On pourra se reporter au cadre très précis qu'implique l'analyse des composantes principales. Les plantes, pures manifestations de phénomènes exogènes, révèlent l'existence en une placette donnée de ces différentes variations.

On doit donc s'attendre à des résultats particulièrement clairs. Le caractère très fruste des mesures effectuées risque néanmoins d'affaiblir la qualité des résultats. En effet, les pourcentages de variation absorbés par les premiers facteurs sont faibles si on les compare aux analyses exécutées sur les données de nutrition par exemple, mais leur interprétation est relativement aisée. On ne pourrait en dire autant de certaines analyses de sol où les composantes principales s'adaptent parfois très mal aux structures complexes des phénomènes pédologiques.

De même, si on cherche à expliquer la production par les données phytosociologiques, on se place également dans le cas d'une structure simple de régression. Les relations n'ont pas d'autre valeur que statistique (il n'y a pas de flèches entre ces deux constituants sur le graphique 3), et la boîte noire peut fonctionner. On obtient un index floristique.

Ce n'est évidemment pas en utilisant de façon explicative cet index qu'on améliorera la production; mais si la tentation est plus grande dans le cas des données de sol, elle est tout aussi inutile, car c'est alors la structure qui est inadaptée.

6. — PROBLÈMES PARTICULIERS LIÉS A LA NON-LINÉARITÉ

Nous avons exclusivement employé des modèles linéaires. Ces modèles, seuls employés en statistiques jusqu'à une époque récente, étaient adaptés au but premier de la statistique : dégager au-delà de la variabilité expérimentale l'influence de *petites* modifications de certains facteurs.

Au contraire, dans les enquêtes station-production, le parti pris est différent; le spectre des conditions de station est très étendu (au contraire des expérimentations en zone très localisée) et l'hypothèse de linéarité des liaisons est très irréaliste : les liaisons sontnon-linéaires, ou plus exactement dépendent de la position de la placette dans le domaine d'enquête.

Dans les techniques décrites ci-dessus, les problèmes de non-linéarité sont *a priori* ignorés; toutefois, en de multiples occasions, on a pu mettre en évidence cette non-linéarité :

par exemple, dans une analyse en composantes principales, les deuxième et troisième facteurs correctement et soigneusement interprétés traduisent la non-linéarité de certaines liaisons qui peuvent être ainsi *identifiées*.

Citons deux non-linéarités importantes qui n'ont pas été encore suffisamment prises en compte :

1) concept de variable limitante : équation de Mitscherlich.

Le modèle de base est multiplicatif et exponentiel de façon à annuler le caractère expliqué si l'un des caractères est en dessous de son seuil ou à le plafonner s'il n'a plus d'influence réelle.

$$y = kx_1^b (1 - e^c(x_2 - x_{20}))...$$

2) plantes à optima (fig. 6). Étant donné un facteur, une plante quelconque a une répartition de type normal autour d'un optimum. La liaison entre deux plantes ayant des optima différents sera très évidemment non-linéaire. L'analyse des distorsions induites sur les composantes principales n'a pas été faite.

7. — CONCLUSIONS PROVISOIRES

7.1. — *Intérêt actuel de ces études*

L'aspect descriptif a été très accusé et, en effet, la première qualité est de fournir des documents synthétiques en quelque sorte optimaux. L'impact visuel de ces techniques est un gain sérieux sur les présentations de tableaux couramment employés.

Pour l'interprétation, les résultats paraissent en général triviaux; dans certains cas, un gain sérieux de connaissance a été enregistré. Toutefois, la recherche scientifique ne consiste-t-elle pas à choisir, entre de multiples possibilités qui sont *a priori* autant d'évidences, celle qui est contenue dans les données?

Dès lors, toutes les applications se sont révélées fructueuses et ont permis d'analyser correctement le champ de variation du domaine d'enquête.

7.2. — *Dépassement de ces études : prédiction et action*

Les études de liaison station-production sont des analyses comparatives conduisant à des prédictions et des typologies. Les méthodes mathématiques employées impliquent des modèles causaux très simples donc irréalistes qui empêchent l'action.

Par définition, la boîte noire représente la négation de la prise en compte de relations réelles et marque bien la limite de telles techniques.

Pour passer de la prédiction à l'action, il faut donc analyser « physiologiquement » l'écosystème et tenir compte des relations causales qui le composent. De nouvelles techniques mathématiques plus compliquées, mais plus riches et plus engagées, permettront d'avancer dans cette voie.

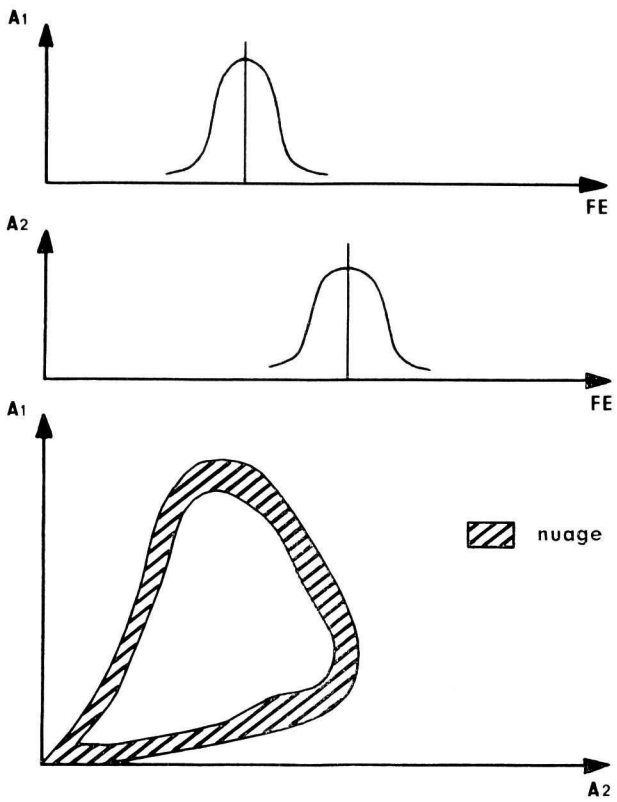


FIG. 6. — *Non linéarité dans la liaison entre deux plantes à optimum sur le facteur écologique Fe*
A_1 : *abondance de la plante 1*
A_2 : *abondance de la plante 2*

FIG. 6. — *Non linearity on the correlation between two species with optimum on the ecological factor Fe*

7.3. — *Conceptions futures de l'expérimentation*

Bien sûr, le traitement même très sophistiqué des données d'une enquête ne se suffira pas entièrement à lui-même pour entamer des actions valables. Il y a une phase de validation qui passe obligatoirement par le recueil de nouvelles informations (par exemple sur un plan d'expérience).

Toutefois, l'expérimentation n'est pas la panacée universelle classiquement décrite et demande parfois à être utilisée aussi précautionneusement qu'une analyse de régression.

Une expérimentation valable doit rassembler les qualités suivantes :

1. — les facteurs contrôlés doivent constituer avec le caractère expliqué une structure causale simple (fig. 7 cas 2). Les interactions introduites permettront d'expliquer certaines complexités;

2. — les facteurs non contrôlés doivent être randomisés (= répartis au hasard) dans l'expérience.

Dans une expérience comparative de provenances, les conditions 1 et 2 sont aisément remplies. Nous pensons que dans les expériences de fertilisation, elles sont plus difficilement vérifiées (cas n° 1, fig. 7).

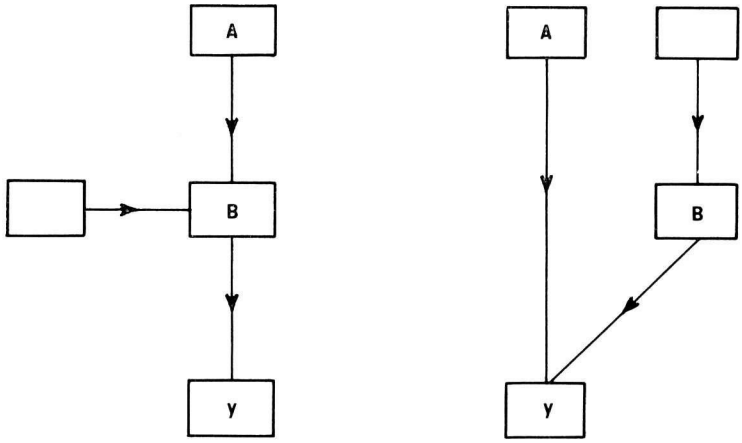


FIG. 7. — *Diagramme causal et expérimentation*
Cas 1 *: l'expérimentation sur* A *et* B *est sujette à caution*
Cas 2 *: l'expérimentation sur* A *et* B *est adéquate*

FIG. 7. — *Causal diagram and experimentation*

Comme les enquêtes, les expériences demandent une sérieuse analyse causale qui doit être intégrée dans l'analyse et qui peut clarifier les résultats.

Bien mieux, les données d'enquête, mieux traitées, pourraient conduire au choix d'expériences « optimales ».

SUMMARY

MATHEMATICAL METHODOLOGY IN THE STUDY OF STAND-PRODUCTION RELATIONSHIPS

Mathematical and statistical methods in the study of stand-production relationships are outlined with a special emphasis on causality. The signification of classical techniques (regression, principal components and factor analysis) is investigated relatively to causality.

In the conclusion we oppose descriptive or predictive models to explicative ones which require much more qualitative information from the biologist and which perhaps will furnish complements to experimentation.

DISCUSSIONS ET INTERVENTIONS

M. GODRON. — Les méthodes d'interprétation ont souvent un intérêt supplémentaire qu'il serait dommage d'oublier : elles permettent souvent de préparer un meilleur échantillonnage, pour la suite des travaux. Ces échantillonnages « progressifs » se sont montrés presque toujours utiles pour les études phyto-écologiques de longue haleine.

Pas de réponse de M. MILLIER.

M. GARBAYE. — Dans le cas où une série de caractères est mesurée plusieurs fois par placette (sur plusieurs arbres par exemple), comment peut-on interpréter ou utiliser les coefficients de corrélation inter et intra-placettes entre ces caractères?

Réponse de M. MILLIER. — Les mesures caractérisant la variabilité intra-placettes quantifient l'action de facteurs très locaux : génétique, microtopographie, erreurs d'échantillonnage, concurrence, etc. Quand on travaille au niveau inter-placettes, ces effets sont « moyennés » et au contraire on relève la variabilité régionale.

M. GOUNOT. — La notion de boîte-noire ne peut pas être séparée de celle d'échelle. Descendre à une échelle plus fine complique parfois les problèmes au lieu de les simplifier. C'est le cas par exemple de la notion d'évapotranspiration qui au niveau de la station est à peu près élucidée par la théorie de PENMAN, alors qu'elle est très complexe au niveau de la feuille de la plante.

M. DECOURT. — On n'échappe pas à la « boîte-noire ». Les systèmes causaux sont des liaisons entre boîtes-noires. Le tournant de notre recherche est peut-être un changement de niveau dans la connaissance de la réalité.

Au lieu de se contenter d'étudier le fonctionnement du peuplement, on « rentre » dans cette première boîte pour étudier des liaisons entre les arbres (2e série de boîtes, etc.).

Ces niveaux correspondent à des buts souvent différents, et un problème important est de bien choisir le découpage de la réalité.

Réponse de M. MILLIER. — Effectivement, les points soulevés par MM. GOUNOT et DECOURT sont très importants. Et il a été montré qu'on peut travailler de manière très efficace à un certain niveau de boîte noire sans savoir ce qu'il y a dedans.

M. BONNEAU. — Les méthodes mathématiques utilisées actuellement sont des méthodes comparatives : elles ne peuvent pas rendre compte de l'existence d'un facteur limitant qui serait constant dans l'ensemble des « placettes » examinées. Existe-t-il des méthodes mathématiques qui pourraient examiner non seulement l'action des facteurs variables mais la « distance » entre tous les facteurs et un « optimum » de chacun introduit dans les données.

Réponse de M. MILLIER. — Cette action ne peut être révélée que dans le cadre d'une enquête plus vaste.