UNE ESTIMATION DYNAMIQUE DES RENDEMENTS

OLIVIER BOVEY
U.N.C.A.A.
83, 85 Avenue de la Grande Armée
75782 PARIS CEDEX 16

Résumé

OPTICOOP est un réseau national d'enquêtes culturales dont l'objectif est de déterminer les techniques qui jouent le plus sur le niveau de rendement obtenu, concernant plusieurs cultures. Ces résultats sont ensuite réexploités sous la forme d'un simulateur agronomique développé sur micro ordinateur. Ce logiciel doit aider l'agriculteur dans son cheminement technique vers le meilleur résultat. L'estimation du rendement s'appuie sur l'analyse factorielle. La notion d'itinéraire technique conduit à rechercher une méthode d'estimation dynamique du rendement.

Mots-clefs

Enquête culturale - Itinéraire technique - Rendement - Analyse factorielle - Régression par boule - Analyse par sous-tableau - Relation de récurrence.

Introduction.

OPTICOOP est un réseau national d'enquêtes culturales dont l'objectif est d'expliquer les variations du rendement, pour une culture donnée, en fonction des techniques mises en oeuvre.

L'Union Nationale des Coopératives Agricoles d'Approvisionnement, (U.N.C.A.A.), dispose ainsi d'un vaste observatoire des pratiques culturales qu'elle met à la disposition de ses coopératives adhérentes en vue de définir la meilleure stratégie de conseil technique destinée aux agriculteurs.

Chaque enquête est réalisée à l'intérieur d'une petite région naturelle dans laquelle plusieurs agriculteurs sont interrogés à propos des interventions qu'ils ont effectuées sur une parcelle de leur choix recevant la culture étudiée (blé, pois, colza, mais, etc.). Les pratiques culturales ainsi observées sont ensuite hiérarchisées en fonction du niveau de rendement qu'elles ont permis d'atteindre.

Depuis la création de ce système, il y a une dizaine d'années, environ cent mille parcelles ont déjà été analysées. L'idée est très vite apparue d'exploiter ces résultats en vue de les intégrer dans un simulateur agronomique permettant d'estimer un niveau de rendement en fonction d'un itinéraire technique. Ce logiciel conçu pour tourner sur micro-ordinateur doit guider l'agriculteur dans le choix de la meilleure technique en terme de potentiel de rendement.

Eu égard à la nature des données, la procédure d'estimation mise en oeuvre fait appel à l'analyse factorielle. En effet, la technique du point supplémentaire est partout présente : régressions typologique et par boule (P. CAZES).

L'analyse par sous-tableau permet, en outre, de simplifier un certain nombre de calculs lorsque l'estimation du rendement devient dynamique, c'est à dire lorsqu'elle est autorisée à n'importe quelle étape de réalisation de l'itinéraire.

1.Règles d'estimation.

Le principal objectif assigné aux enquêtes OPTICOOP est de déterminer les techniques qui sont habituellement associées aux meilleurs rendements et d'extrapoler ces résultats pour des parcelles n'ayant pas participé à l'étude. On suppose donc qu'il existe une relation fonctionnelle entre le rendement et les techniques mises en place :

$$R_i = F(T_i^{(1)}, T_i^{(2)}, \ldots, T_i^{(p)}) + e_i,$$

où R_i représente le rendement observé sur la i-ème parcelle,
$T_i^{(1)}, T_i^{(2)}, \ldots, T_i^{(p)}$, les techniques adoptées,
e_i, un terme aléatoire,
F, une fonction de production dont la forme est inconnue.

Dans la mesure où la fonction F est difficilement identifiable, l'estimation va s'appuyer sur un tableau de régression qui est en quelque sorte un tableau de contingence multiple croisant des niveaux de rendement avec des modalités techniques. Cette méthode présente aussi l'avantage d'intégrer facilement le mélange de caractères discrets et continus, ces derniers étant préalablement découpés en classes. On ajoute en dessous de ce tableau toutes les parcelles sous forme disjonctive complète conservées comme individus supplémentaires. A partir des facteurs sous-jacents à ce tableau, deux solutions sont alors envisageables.

Rendement X 1ère technique	(...) Tableaux de contingence	Rendement X p-ème technique
0 0 0 1 0 0 0	(...)	0 0 1
	Itinéraires techniques des parcelles sous forme recodée	

D'une part, la formule de transition permet de calculer les coordonnées factorielles d'un itinéraire technique non observé dans l'enquête et de l'affecter à la classe de rendement de laquelle il se rapproche le plus selon la métrique euclidienne. Cette méthode permet d'obtenir une estimation en niveau du rendement de la parcelle (régression typologique).

$F_{\alpha,isup} = (1/\sqrt{\lambda_\alpha})\Sigma_j(\delta_{isup,j} * G_{\alpha,j})/CardQ$,
$F_{\alpha,isup}$ est la coordonnée de l'individu supplémentaire sur l'axe α,
λ_α est la valeur propre de l'axe α,
$\delta_{isup,j}$ vaut 1 si l'individu a adopté la modalité technique j, 0 sinon,
CardQ est le nombre total de techniques retenues.

L'affectation se fait suivant la règle :
R^{i0} est tel que $D_{isup,i0} = Min(D_{isup,i}, i = 1,\ldots,CardI)$,
R^{i0} est le niveau de rendement estimé,
$D_{isup,i}$ est la distance euclidienne entre l'individu supplémentaire et la classe de rendement i,
CardI est le nombre de classes de rendement.

D'autre part, toujours à partir de la formule de transition, les coordonnées de l'individu supplémentaire sont rapportées non plus aux différentes classes de rendement, mais aux parcelles observées appartenant à son voisinage sur les axes factoriels. La moyenne ou la médiane des rendements observés sur ces parcelles constitue l'estimation recherchée (régression par boule).

La forte variabilité des résultats obtenus avec la régression par boule conduit à s'interroger sur le bien fondé de la méthode. Toutefois, les observations agronomiques montrent qu'un même itinéraire technique conduit à des rendements très différents. En règle générale, cette variabilité trouve son explication dans le fait que tous les facteurs déterminants n'interviennent pas dans le modèle par absence de mesure. Il en résulte qu'on ne peut apprécier les performances d'un itinéraire à partir d'un seul indicateur tel que la moyenne. Le potentiel, mesuré par le maximum des rendements enregistrés dans les conditions décrites par l'itinéraire, et le minimum, reflétant un niveau plancher, constituent pour l'agriculteur des éléments de décision incontestables. Il en va de même des paramètres exprimant la variabilité des résultats (écart-type, coefficient de variation) qui peuvent s'interpréter comme des mesures de risque inhérent à la stratégie adoptée. Enfin, deux itinéraires techniques indifférents du point de vue du rendement peuvent se départager grâce à leur coût financier.

2.Estimation dynamique.

Jusqu'à présent on supposait que l'estimation était possible une fois que tous les éléments de l'itinéraire étaient connus. L'efficacité d'une combinaison technique était alors privilégié plutôt que le choix d'une technique à un instant donné relativement aux décisions déjà prises. Ce dernier point, plus proche du raisonnement d'un agriculteur, est réalisable grâce à une estimation dynamique des rendements.

Un itinéraire technique peut être perçu comme une suite d'interventions échelonnées dans le temps, mises à part les conditions d'implantation qui sont les données initiales dans le processus d'élaboration du rendement. Choisir le meilleur itinéraire, c'est choisir la meilleure intervention à chaque étape de réalisation de la culture compte tenu de tout ce qui a déjà été fait. La variable de contrôle étant le rendement réalisé sur la parcelle, il doit être possible de le calculer en fonction des seules techniques déjà mises en place. Cela suppose que les facteurs décrivant les états intermédiaires soient disponibles, ce qui peut conduire à stocker un volume de données considérable, à savoir tous les jeux de facteurs associés aux différents tableaux définissant les itinéraires uniquement composés des interventions réalisées jusqu'à l'époque considérée. Heureusement, le déroulement temporel d'un itinéraire technique limite le nombre de tableaux à prendre en compte. Par ailleurs, son caractère récurrent qui fait que tout ce qui peut être entrepris à un instant donné dépend des interventions passées, invite à rechercher une éventuelle relation mathématique entre deux jeux de facteurs consécutifs. Or, une telle relation existe, elle découle directement des résultats issus de l'analyse par sous-tableau (B. ESCOFIER). En effet, il est possible de calculer les nouveaux facteurs d'un itinéraire quand une intervention vient l'enrichir, en fonction des

facteurs associés à cette opération seule et de ceux relatifs à toutes celles qui l'ont précédée. On s'aperçoit aussitôt que cette méthode réduit notablement le volume d'informations à stocker puisque seuls les facteurs associés aux techniques, et non plus aux itinéraires, sont absolument nécessaires, les autres pouvant se déduire par calculs. Il devient également envisageable de mettre au point une règle de décision qui consisterait à conseiller la modalité technique assurant le meilleur potentiel compte tenu des étapes déjà franchies.

3.Formalisation du problème.

Pour simplifier, un itinéraire technique est caractérisé par une suite de p interventions chacune entièrement décrites à l'aide d'un ensemble fini de modalités.

3.1.Les tableaux simples.

A partir des observations réalisées sur N parcelles, on construit une suite de tableaux de contingence $T^{(1)},T^{(2)},...,T^{(p)}$ croisant les modalités des p interventions avec les classes de rendement.

Chaque tableau $T^{(k)}$, $k=1,...,p$, décrit ainsi les N parcelles à partir de r classes de rendement et de s_k modalités techniques. Plusieurs indices de liaison peuvent être définis, tous fonction de la statistique du *chi-deux* :

$\Phi^2_k = X_k^2/N$, *phi-2* ou inertie totale dans l'analyse des correspondances du tableau $T^{(k)}$,

$V_k = \Phi_k/\sqrt{\min(r-1,s_k-1)}$, coefficient de CRAMER.

En vertu de la formule de reconstitution du tableau des données, $T^{(k)}$ est entièrement défini par ses marges et ses facteurs issus de l'analyse des correspondances : $F^{(k)}$ sur les rendements et $G^{(k)}$ sur les modalités techniques. $F^{(k)}$ est une matrice ayant r lignes et $\min(r-1,s_k-1)$ colonnes. De même, $G^{(k)}$ est une matrice ayant s_k lignes et $\min(r-1,s_k-1)$ colonnes. Enfin, les valeurs propres issues de l'analyse des correspondances de $T^{(k)}$, $\lambda_\alpha^{(k)}$, $\alpha=1,...,\min(r-1,s_k-1)$, décrivent la distorsion de l'information, en terme d'inertie, observée sur les axes factoriels, due aux liaisons entre les rendements et les techniques.

3.2.Les tableaux composés.

On considère maintenant la suite des tableaux $R^{(k)}$ $k=1,...,p$ obtenus en concaténant les k premiers $T^{(k)}$. On obtient de la sorte un tableau de contingence multiple décrivant les associations entre les modalités techniques des k premières interventions et les classes de rendement. Il comporte donc r lignes et $\Sigma_i^k s_i$ colonnes. Ce tableau est également caractérisé par un *Chi-deux* ($X_k^{2'}$) et en conséquence par tous les indicateurs de liaison déjà définis sur les $T^{(k)}$. On a par ailleurs la relation suivante : $X_k^{2'} = X_1^2 + X_2^2 + ... + X_k^2$. Il existe également une relation de récurrence entre les $R^{(k)}$. En effet, on a : $R^{(k)} = R^{(k-1)} \oplus T^{(k)}$, le signe $\oplus$ désigne l'opération de concaténation de deux

tableaux de contingence ayant le même ensemble de lignes. Pour terminer, les tableaux $R^{(k)}$ sont de même entièrement déterminés par leurs marges et leurs facteurs issus de l'analyse des correspondances : $F'^{(k)}$ en ligne, $G'^{(k)}$ en colonne. $F'^{(k)}$ est une matrice ayant r lignes et $\min(r-1,\Sigma^k s_i-k)$ colonnes; $G'^{(k)}$ une matrice ayant $\Sigma^k s_i$ lignes et $\min(r-1,\Sigma^k s_i - k)$ colonnes. Les valeurs propres issues de l'analyse des correspondances de $R^{(k)}$, s'écrivent : $\lambda'^{(k)}_\alpha$, $\alpha=1,\ldots,A_k$ où $A_k = \min(r-1,\Sigma^k s_i - k)$ à savoir le nombre d'axes factoriels associés au tableau $R^{(k)}$; en général $A_k=r-1$ à partir d'un certain $k=k_0$.

3.3.Calculs des facteurs.

L'analyse par sous-tableau a été conçue au départ dans le but de traiter par l'analyse factorielle les tableaux de grande dimension qui débordaient des capacités de calcul disponibles. La solution était simple, il suffisait de réduire les dimensions du problème en décomposant le tableau initial en plusieurs sous-tableaux et de procéder à l'analyse des correspondances sur chacun d'eux. Les facteurs extraits de la sorte permettaient ensuite de calculer une approximation des facteurs associés au tableau initial. Lorsque la décomposition résulte naturellement d'un critère donné, par exemple une variable de stratification, cette méthode offre un intérêt supplémentaire puisqu'il est alors possible d'entreprendre une étude intra et inter sur les groupes définis par cette variable, ce qui permet de peaufiner l'interprétation des données. Dans le cas présent, le clivage s'opère selon les techniques. On obtient ainsi une suite de tableaux dont les marges sont proportionnelles de sorte que la méthode devient tout à fait applicable.

L'analyse par sous-tableau énonce qu'il est possible de calculer $F'^{(k)}$ et $G'^{(k)}$ à partir des $F^{(i)}$ et $G^{(i)}$, $i=1,\ldots,k$. Si d'aventure, il est possible de mettre à jour une relation de récurrence sur les $F'^{(k)}$ et $G'^{(k)}$, on peut espérer de la sorte réduire dans des proportions importantes le volume de calculs à réaliser.

Une première remarque confirme qu'une telle relation existe bien. En effet, le nombre de parcelles ayant obtenu tel niveau de rendement avec telle modalité technique peut-être déterminé au moyen de n'importe quel ensemble de facteurs inhérents à tout tableau dans lequel figure la modalité prise en compte. On en déduit une relation fondamentale entre facteurs issus de tableaux différents qui s'exprime sous forme de système d'équations non-linéaires. En effet, on a :

$$N_{ij} = (k-1)N_{i.}N_{.j}\left(1 + \Sigma_\alpha\left(F'^{(k-1)}_{\alpha i}G'^{(k-1)}_{\alpha j}\right)/\sqrt{\lambda'^{(k-1)}_\alpha}\right) =$$

$$kN_{i.}N_{.j}\left(1 + \Sigma_\alpha\left(F'^{k}_{\alpha i}G'^{k}_{\alpha j}\right)/\sqrt{\lambda'^{k}_\alpha}\right), \text{ d'où}$$

$$(k-1)\left(1 + \Sigma_\alpha\left(F'^{(k-1)}_{\alpha i}G'^{(k-1)}_{\alpha j}\right)/\sqrt{\lambda'^{(k-1)}_\alpha}\right) =$$

$$k\left(1 + \Sigma_\alpha\left(F'^{k}_{\alpha i}G'^{k}_{\alpha j}\right)/\sqrt{\lambda'^{k}_\alpha}\right),$$

N_{ij} représente le nombre de parcelles ayant atteint le niveau de rendement i avec la modalité technique j.

$N_{i.}$ est égal au nombre de parcelles ayant obtenu le niveau de rendement i.
$N_{.j}$ représente le nombre de parcelles ayant adopté la modalité technique j.

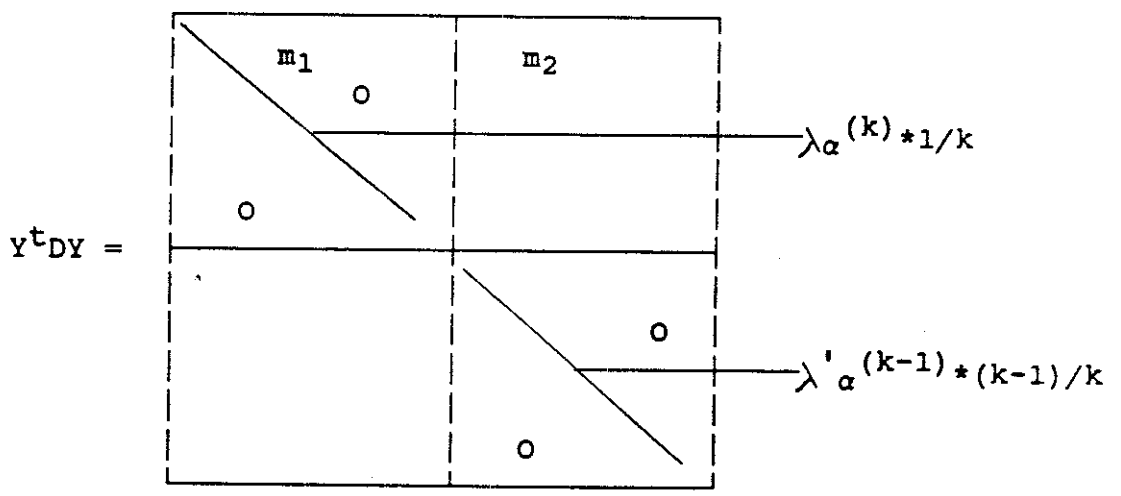
Cependant, cette approche ne permet pas de manière simple d'établir une règle de passage entre les différents ensembles de facteurs.

Plusieurs caractéristiques communes aux différents tableaux renseignent par ailleurs sur la nature de cette relation. On constate, tout d'abord, que leurs centres de gravité dans l'espace des modalités techniques sont identiques. Qui plus est, les distances entre modalités techniques sont égales, et cela quel que soit le tableau à partir duquel elles sont calculées. En fait, seules les coordonnées factorielles changent, ce qui laisse à penser que le passage d'une représentation à l'autre ne correspond qu'à une rotation des axes factoriels.

Or, en suivant strictement la démarche adoptée dans l'analyse par sous- tableau, pour obtenir $F'^{(k)}$ et $G'^{(k)}$, il suffit de retenir $F'^{(k-1)}$ et $F^{(k)}$, en les pondérant respectivement par $\sqrt{(k-1)/k}$ et $\sqrt{1/k}$, et de construire la matrice Y de la façon suivante:

$$Y=[\sqrt{1/k}*F^{(k)},\sqrt{(k-1)/k}*F'^{(k-1)}]$$

Ensuite, on calcule Y^tDY où D est la matrice diagonale des poids des r classes de rendement. Les valeurs propres de Y^tDY ne sont rien d'autres que les valeurs propres de $R^{(k)}$ recherchées, c'est à dire $\lambda'^{(k)}_\alpha$, $\alpha=1,\ldots,A_k$. En posant $m_1=\min(r-1,s_k-1)$ et $m_2=A_{k-1}$, Y^tDY est donc une matrice carrée symétrique de format m_1+m_2.



Les vecteurs propres de Y^tDY permettent de calculer $F'^{(k)}$. En effet, si C_α désigne le vecteur propre associé à $\lambda'^{(k)}_\alpha$ alors: $F'^{(k)}_\alpha = Y\,C_\alpha$. Autrement dit, le facteur d'ordre α sur les classes de rendement relatif au tableau $R^{(k)}$ s'écrit comme une combinaison linéaire des facteurs $F^{(k)}$ et $F'^{(k-1)}$, ce qui définit la relation de récurrence recherchée.

Pour ce qui concerne les $G'^{(k)}$, il est plus intéressant de raisonner sur les modalités techniques et leurs coordonnées factorielles. Ainsi, $G'^{t(k-1)}_j$ $j=1,\ldots,A_{k-1}$, représente le

vecteur des coordonnées factorielles de la modalité technique j; c'est donc la j-ème ligne transposée de la matrice $G^{'(k-1)}$. De même, $G_j^{t(k)}$ $j=1,...,s_k$, donne les coordonnées factorielles de la modalité j relatif à $T^{(k)}$. Pour calculer $G^{'(k)}$, deux cas de figure sont à considérer selon qu'il s'agit d'une modalité technique de $R^{(k-1)}$ ou de $T^{(k)}$:

Si j se rapporte à une modalité technique appartenant à $T^{(k)}$, alors $G'^{t(k)}_j = \Gamma^{(k)} G_j^{t(k)}$ où $\Gamma^{(k)}$ est une matrice carrée de format m_1 et de terme général :

$$g_{\alpha\beta} = \sqrt{\{k*(\lambda'^{(k)}_\beta / \lambda^{(k)}_\alpha)\}} * c_{\alpha\beta},$$

$c_{\alpha\beta}$ est la β-ème composante du vecteur propre C_α, $\beta = 1,...,m_1$.

Si j se rapporte à une modalité technique appartenant à $R^{(k-1)}$, alors $G'^{t(k)}_j = \Gamma'^{(k)} G'^{t(k-1)}_j$ où $\Gamma'^{(k)}$ est une matrice carrée de format m_2 et de terme général:

$$g'_{\alpha\beta'} = \sqrt{\{(k/k-1)*(\lambda'^{(k)}_\beta / \lambda'^{(k-1)}_\alpha)\}} * c_{\alpha\beta'},$$

$c_{\alpha\beta'}$ est la β'-ème composante du vecteur propre C_α, $\beta' = m_1+1,...,m_1+m_2$, $\beta = \beta' - m_1$.

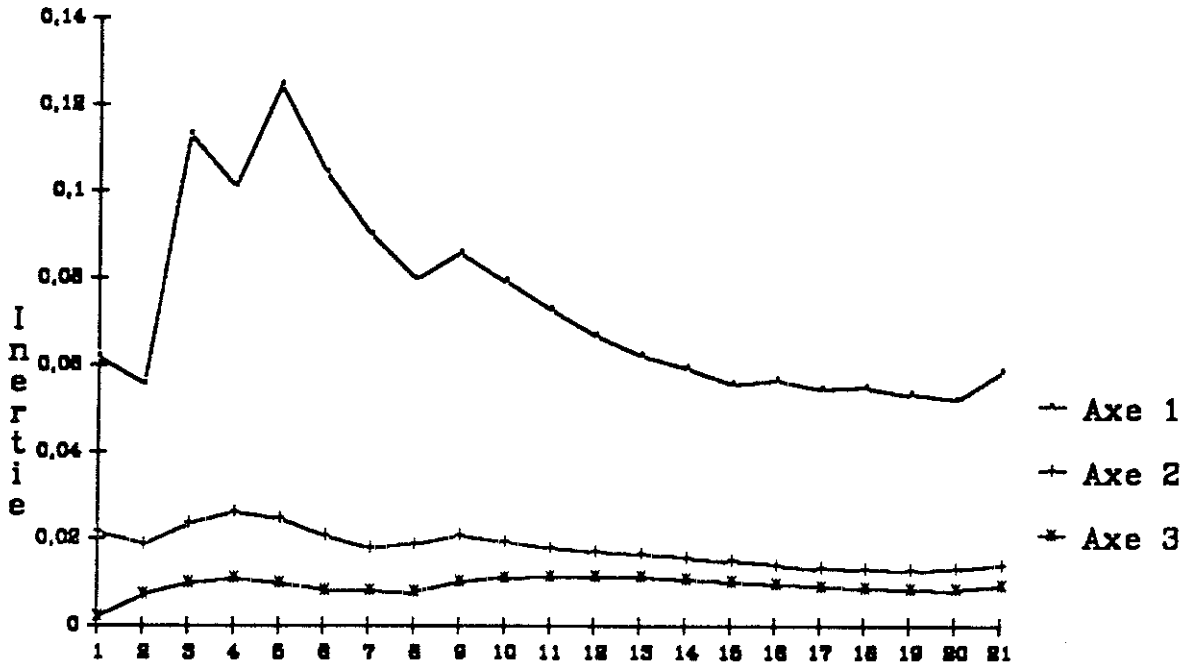
L'intérêt de cette écriture est de montrer clairement que les nouvelles coordonnées factorielles d'une modalité technique s'exprime en fonction des anciennes par le biais d'une application linéaire qui correspond en fait à une rotation. Ce résultat est directement perceptible quand l'espace factoriel se réduit à un plan. Dans ce cas, $\theta_1 = \arccos(g_{11})$ est une mesure de l'angle associé à la rotation effectuée par les modalités techniques de $T^{(k)}$. De même, $\theta_2 = \arccos(g_{31})$ correspond à l'angle associé à la rotation appliquée aux modalités de $R^{(k-1)}$.

4.Conclusion.

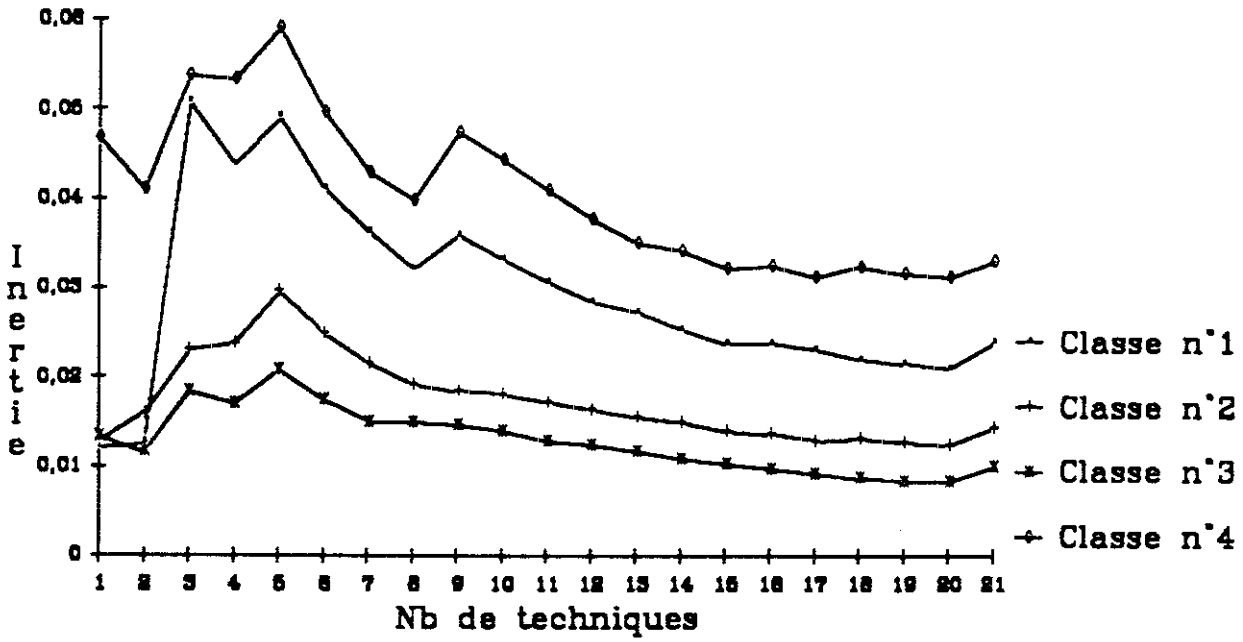
L'estimation dynamique du rendement conduit à une utilisation particulière de l'analyse par sous-tableau. En effet, toute cette méthode a été conçue au départ pour traiter les grands tableaux. Il s'agissait de ventiler les données en plusieurs sous-tableaux, d'extraire les facteurs associés à ces derniers, moins longs à calculer, de ne retenir que les plus significatifs pour, finalement, en déduire une approximation des facteurs relatifs au tableau complet.

En outre, cette méthode répond correctement aux exigences, posées au départ, de stockage des données nécessaires et de rapidité des calculs. En vertu de la relation de récurrence ainsi mise en évidence, il est facile de se rendre compte que seuls les $F^{(k)}$ et les $G^{(k)}$, sont nécessaires pour effectuer tous les calculs. En revanche, à chaque étape, il faut diagonaliser la matrice Y^tDY, opération relativement rapide, en raison de sa faible dimension, et donc encore supportable. La formule de transition permet, quant à elle, de calculer les coordonnées factorielles des itinéraires techniques observés ou non dans l'enquête, indispensables dans la régression par boule.

Valeurs Propres des R(k) (Fig. 1)



Inertie par classe de rendement des R(k) (Fig. 2)



Enfin, il peut être intéressant de suivre, par exemple, sur un graphique l'évolution des valeurs propres des $T^{(k)}$, ou bien des $R^{(k)}$ (fig. 1), afin d'apprécier le poids de chaque intervention au moment de son introduction dans l'itinéraire. On constate, néanmoins, une convergence rapide des valeurs propres autour d'un point d'équilibre. La répartition de l'inertie par classe de rendement (fig. 2) renseigne, quant à elle, sur la distorsion de l'information relative à la dernière technique appliquée. On dispose ainsi de plusieurs indicateurs qui peuvent s'avérer très utile dans le choix des interventions à retenir dans l'itinéraire technique.

Bibliographie

P. CAZES, Régression par boule et par l'analyse des correspondances, Laboratoire de statistique, Université P. et M. Curie, 1976.

P. CAZES, Analyse des correspondances multiples, Bulletin de l'ADDAD n°12, 1983.

B. ESCOFIER, Analyse des correspondances par sous-tableau, Session d'analyse des données perfectionnement, Tome I, INSEE, ADDAD, ISUP, 1981-1982.

B. ESCOFIER J. PAGES, Analyses factorielles simples et multiples, Dunod, 1988.

J. P. FENELON, Statistiques normatives et génie statistique, A propos des prévisions en agro-industries, Actes des journées Agro-Industrie et méthodes statistiques, 1990.