JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

Communication. Statistique et recherche agronomique. Comptes rendus de la demi-journée d'étude du 15 décembre 1981

Journal de la société statistique de Paris, tome 123, nº 2 (1982), p. 93-143
<http://www.numdam.org/item?id=JSFS_1982__123_2_93_0>

STATISTIQUE ET RECHERCHE AGRONOMIQUE

Comptes rendus de la demi-journée d'étude du 15 décembre 1981

RÔLE DU DÉPARTEMENT DE BIOMÉTRIE DE L'INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE

Depuis le début du siècle, les mathématiques appliquées ont joué un rôle important dans la recherche agronomique et, réciproquement, celle-ci a été une source d'inspiration constante, pour le développement des statistiques en particulier. Les travaux de Sir Ronald FISHER en sont la plus illustre, mais non l'unique preuve. Et c'est certainement de cette imbrication que la notion actuelle de biométrie tire ses origines.

Le constat qui vient d'être énoncé ne doit étonner en rien. Si l'on considère les sciences de la vie en général, force est de reconnaître que la formalisation mathématique peut, et souvent doit, intervenir en plusieurs étapes du travail de recherche, qu'il s'agisse de l'observation d'un phénomène, de sa description, de sa représentation ou des décisions prises dans le but d'agir sur lui. Cette formalisation est d'ailleurs tellement intégrée à la démarche scientifique dans certains domaines qu'elle passe inaperçue. Les plans d'expérience, les histogrammes, la régression, l'analyse de variance font partie du vocabulaire et de la pratique quotidienne de nombre de chercheurs agronomes. De la même manière, lorsqu'un enzymologiste utilise l'équation de Michaelis et Menten, se souvient-il que ce sont les mathématiques qui ont permis de comprendre correctement les cinétiques enzymatiques et d'en donner une représentation opératoire? Dans de nombreux domaines, la représentation mathématique fait ainsi partie du processus de connaissance du phénomène étudié et les chercheurs concernés ne peuvent pas l'éviter. De plus, c'est la plupart du temps à travers cette représentation qu'ils pourront déterminer leur intervention de manière optimale, pour planifier une expérience, pour réaliser un sondage, pour contrôler le niveau d'une population d'insectes nuisibles, pour quantifier un effort de pêche acceptable... La biométrie s'intègre donc de façon naturelle dans la démarche propre des biologistes. L'importance de son apport est seulement fonction du degré de formalisation du phénomène étudié. Une démonstration par l'absurde de cette affirmation est la déception qu'éprouvent généralement nos interlocuteurs lorsqu'ils déposent sur nos bureaux des lots de données, attendant que le biométricien fasse pour eux des « découvertes ».

D'autre part, c'est une évidence, les phénomènes étudiés sont de plus en plus complexes. Les soucis actuels de développement d'une agriculture performante, évitant les gaspillages, exigent une intégration des connaissances du niveau moléculaire au niveau écologique et des possibilités de gestion des systèmes agraires, tâches pour lesquelles l'apport de la biométrie est essentiel. Les phénomènes considérés ne sont pas immédiatement formalisés de façon opératoire, et l'exploitation de ces représentations pose souvent des problèmes de méthode importants. Pensons par exemple à la gestion du patrimoine génétique d'une population animale ou végétale, à la gestion des ressources d'une exploitation agricole, à la protection des cultures. L'utilisation routinière des méthodes classiques, ne suffit donc plus et l'activité des biométriciens doit s'exercer à tous les niveaux de l'échelle qui va du calcul statistique élémentaire à la recherche en mathématique appliquée.

Tenant compte de ces deux faits : intégration sans rupture de la biométrie dans les travaux de recherche agronomique et diversité des niveaux de besoins, plusieurs possibilités peuvent être envisagées pour satisfaire les exigences d'un institut de recherche tel que le nôtre.

La première consiste à faire appel à des équipes de mathématiciens, n'appartenant pas nécessairement à l'INRA, spécialisés dans des domaines particuliers des mathématiques appliquées. Si cette solution peut être très efficace pour des problèmes difficiles précis, elle ne permet guère de résoudre l'ensemble des problèmes courants et risque d'aboutir à l'adaptation coûte que coûte du problème biologique à la méthode mathématique.

La seconde consiste à disséminer des biométriciens dans diverses équipes de recherches. La continuité affirmée plus haut entre biométrie et biologie peut alors être réalisée pour un temps dans de bonnes conditions... sur un problème limité. Mais cette pratique, si elle est exclusive, ne peut en aucun cas être source de progrès pour la biométrie qui évolue grâce à la confrontation de problèmes issus de divers domaines biologiques et à leur mise en forme par des méthodes mathématiques générales. De plus, l'isolement de ces biométriciens entraînerait à moyen ou à long terme une sclérose certaine, sans parler du risque d'asservissement à des tâches routinières imposées par leur compétence particulière.

La troisième solution est celle adoptée par l'I.N.R.A., qui s'est doté d'un département de biométrie constitué d'équipes réparties dans plusieurs centres de recherches. Elle n'exclut pas d'éventuels recours aux deux précédentes, mais elle est la seule qui puisse assurer efficacité et stabilité à la démarche biométrique au sein de notre Institut.

Pour remplir son rôle, le département de biométrie de l'I.N.R.A. agit de plusieurs manières.

Il constitue *une force de proposition;* ses équipes collaborent avec celles des autres départements, participent à leurs recherches et sont à même d'apprécier leur besoins. Parallèlement, elles poursuivent des recherches en mathématiques appliquées et se tiennent informées des progrès accomplis dans ce domaine. Ceci leur permet d'introduire ces progrès là où ils sont nécessaires pour la progression de la recherche agronomique et d'orienter leurs propres efforts scientifiques en fonction des besoins de cette dernière. Ainsi, le département de biométrie a-t-il pris une part importante dans l'introduction de l'analyse des données et, plus récemment, de la modélisation puis de la gestion de banques de données à l'I.N.R.A. Dans ces divers domaines, comme dans d'autres, il joue ce rôle de force de proposition et de relais, non seulement au niveau de la recherche mais aussi à celui du développement.

La formalisation mathématique est, dans certains domaines des sciences de la vie, un élément nécessaire de la démarche scientifique — tout comme les connaissances biochimiques sont nécessaires à de nombreux biologistes — aussi est-il indispensable que les chercheurs concernés acquièrent, au moins jusqu'à un certain point, des connaissances en biométrie. Pour atteindre cet objectif, le département de biométrie propose des actions de formations, dont les motivations et les modalités sont exposées en détail plus loin. D'autre part, les contacts quotidiens des biométriciens avec leurs collègues d'autres disciplines, permettent, grâce à des échanges répétés, de faire en sorte que la tâche des biométriciens ne se limite plus à des interventions ponctuelles, mais qu'elle s'intègre harmonieusement dans le travail de recherche. Cela permet en outre que l'autonomie des biologistes vis-à-vis de l'utilisation des méthodes biométriques se développe. Pour mener à bien cette double mission de recherche et de développement, la répartition des équipes de biométrie sur plusieurs centres est particulièrement favorable. Mais les effectifs encore faibles — 23 biométriciens pour quelques 1 200 chercheurs que compte l'I.N.R.A. — doivent être développés afin de permettre à toutes ces équipes d'atteindre la taille critique nécessaire.

Le département de biométrie est un *lieu de synthèse.* Le fait de ne pas être lié à d'autres départements de recherche permet à ses scientifiques de prendre en charge des problèmes d'intérêt général et d'introduire dans une discipline les méthodes et les concepts développés pour une autre. S'il est vrai que nombre de recherches menées au sein du département sont issues d'une application précise, il est aussi vrai que les méthodes développées dépassent souvent largement le cadre de cette application. Les exemples abondent : l'étude de l'interaction, commencée pour l'analyse de l'interaction entre génotype et milieu, a été largement formalisée; l'étude sur les composantes de la variance, développée à propos de problèmes de génétique appliquée a non seulement été étendue, mais elle peut servir à d'autres domaines d'application; la dynamique de population, commencée à propos de contrôle de l'aleurodes, a largement débordé de ce domaine... Là encore, ce qui est vrai pour la recherche l'est aussi pour le développement. Ainsi, la structure du département de biométrie a permis la mise en commun des moyens

LES GRANDS THÈMES DE RECHERCHE ET DE DÉVELOPPEMENT DU DÉPARTEMENT DE BIOMÉTRIE

1. RECUEIL DES DONNEES
   - 1.1 Échantillonnage
   - 1.2. Planification expérimentale
   - 1.3. Saisie automatique des données
2. ANALYSE DES DONNÉES
   - 2.1. Représentations factorielles
   - 2.2. Classification automatique
   - 2.3. Comparaison de tableaux
   - 2.4. Méthodes graphiques
3. MODÈLES LINÉAIRE ET NON LINÉAIRE
   - 3.1. Modèle aléatoire et mixte
   - 3.2. Interaction
   - 3.3. Tables de contingence
   - 3.4. Modèle non linéaire
   - 3.5 Robustesse
4. SYSTÈMES SPATIAUX ET/OU TEMPORELS
   - 4.1. Processus ponctuels
   - 4.2. Séries chronologiques
   - 4.3. Modélisation
5. LOGICIELS
   - 5.1. Logiciels sur mini-ordinateur
   - 5.2. Langages de programmation spécialisés
   - 5.3. Gestion de banques de données

informatiques de traitement, en particulier celle de logiciels scientifiques.

La structure du département de biométrie lui permet d'assurer l'intégration des méthodes mathématiques adaptées dans la démarche de la recherche agronomique, en veillant à ce que l'ensemble des besoins, du traitement élémentaire à la recherche, soient satisfaits. Cinq thèmes de recherche ont été définis. Ils correspondent en partie aux étapes de la démarche scientifique, déjà mentionnées plus haut, — de l'observation à la prise de décision en passant par la description et la représentation —, sachant qu'à chacun de ces stades, le traitement informatique doit intervenir. Cette correspondance ne peut être strictement respectée, afin de tenir compte de la diversité des objets mathématiques mis en jeu. Les cinq thèmes actuellement développés sont indiqués dans le tableau ci-joint. Ce tableau représente une vision très analytique de notre activité, vision où l'aspect méthodologique est primordial.

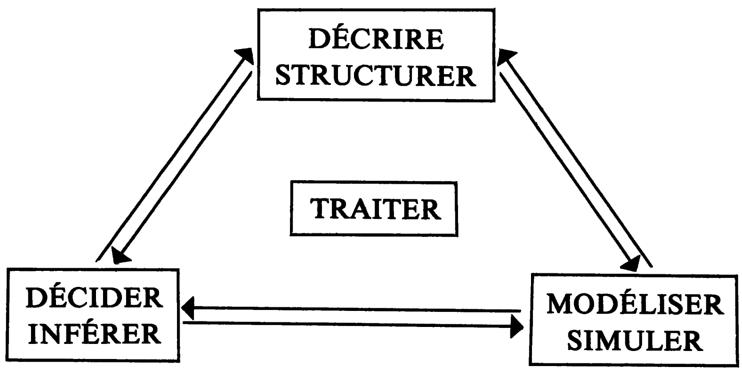
De ces *cinq thèmes* on peut dégager trois types d'approche des problèmes biologiques; ces types d'approche correspondent davantage à la vision que le biologiste a de l'outil mathématique. Très schématiquement, il peut dire que l'outil mathématique peut lui permettre :

i) de faire une description des phénomènes biologiques en dégageant certaines structures cachées;

ii) de prendre certaines décisions à partir de raisonnements inférentiels, ou

iii) de modéliser des phénomènes pour mieux les comprendre et surtout pour mieux imaginer des situations qu'il serait impossible d'approcher expérimentalement de manière systématique.

Nous pouvons donc aussi accepter le schéma simpliste suivant :



Il y a bien *souvent confusion* entre la description et l'analyse des données, la décision et la statistique, la modélisation et l'étude de systèmes. Mais dans la pratique, il n'y a *pas d'analogie complète;* il est même très facile de trouver de nombreux exemples où la statistique est un outil de description, et la modélisation un outil de décision.

Pour mener à bien ces recherches le département de biométrie dispose de cinq laboratoires propres (Avignon, Jouy-en-Josas, Nancy, Toulouse et Versailles) et de quatre laboratoires rattachés à des Établissements supérieurs d'enseignement agronomique (Montpellier, Paris-Grignon, Rennes) ou à une Université (Paris-Sud Orsay).

Les textes qui suivent s'intègrent aux thèmes présentés ci-dessus; ils correspondent à des travaux en cours de jeunes chercheurs.

Pour tout renseignement s'adresser à Département de Biométrie I.N.R.A. 78350 Jouy-en-Josas.

REGARD STATISTIQUE SUR LA SÉLECTION DU BLÉ

par JEAN-MARC AZAIS

Laboratoire de biométrie de l'I.N.R.A. — Versailles

Cet article décrit une étude statistique des expérimentations en parcelles dans un programme de sélection du blé. Après avoir présenté le problème de la sélection d'une bonne variété parmi un grand nombre, nous décrivons comment ce problème est abordé en pratique. Nous présentons ensuite les données sur lesquelles nous avons travaillé. Nous exposons enfin les résultats obtenus :

— un plan en blocs complets aurait donné des résultats équivalents à ceux du plan lattice utilisé.

— à coûts fixes, on améliore la précision en augmentant le nombre de lieux et en diminuant le nombre de répétitions.

— des méthodes graphiques peuvent constituer une bonne aide à la sélection

— le problème de la sélection en plusieurs étapes peut être abordé par simulation.

This paper describes a statistical study of experimentations in corn selection program. After presenting the problem of selecting a good variety among various ones, we describe how this problem is practically solved. The basic data of our works are then presented. The results we found are finally exposed :

- *complete block design should have given results similar to those of the used lattice*
- *for a given cost, accuracy is improved by increasing the number of locations and reducing the number of replicates*
- *graphic methods can be a good help for selection*
- *the problem of multistage selection can be solved by simulation*

Dieser Artikel beschreibt eine statistische Studie der Experimentation oder Parzellen in einem Programm der Selektion von Getreide. Nachdem man zuerst das Problem der Selektion einer guten Getreideart unter vielen beschrieben hat, beschreiben wir wie dieses Problem in der Praxis angegriffen wird. Wir geben dann die Tatsachen, auf denen wir gearbeitet haben. Dann geben wir die Resultate :

— Ein Plan in completen Blocks hätte die gleichen Resultate gegeben wie der Plan lattice, den wir verwendet haben.

— Bei gleichbleibenden Preisen verbessert man die Genauigkeit, indem man die Anzahl der Plätze erhöht und die Anzahl der Wiederholungen vermindert.

— Die graphischen Methoden können von guter Hilfe für die Selektion sein.

— Das Problem der Selektion in mehreren Etappen kann mit einer Simulation begonnen werden.

Journal de la Société de statistique de Paris, tome 123, n° 2, 1982.

(I) DESCRIPTION D'UNE MÉTHODE DE SÉLECTION : SÉLECTION PÉDIGRÉE APRÈS HYBRIDATION.

Le blé est une plante autogame; les variétés commerciales sont des lignées pures : c'est-à-dire composées d'individus identiques génétiquement et stables par reproduction. La sélection cherche de nouvelles lignées pures, meilleures que les précédentes, pour les commercialiser ensuite sous forme de semences.

1. *Généralités*

A partir de lignées ou de populations existantes, on réalise un certain nombre de croisements artificiels :

Exemple $(A \times B)$ ou $(A \times B) \times (C \times D)$ qui constituent la génération F_1. On reproduit ensuite ces épis par autofécondation naturelle pour obtenir successivement les générations F_2, F_3, F_4... Les hybrides F_1 ne sont pas stables par reproduction : ils peuvent donner naissance à des individus très différents entre eux; leurs descendants F_2, F_3 ont par contre tendance à devenir stables (homozygotes) au cours des générations. On ne reproduit pas toutes les plantes : celles qui représentent des défauts apparents sont éliminées. A partir de la génération F_4 on considère les choix visuels comme trop imprécis : on va donc faire des essais quantitatifs. Chaque année le sélectionneur réalise une vingtaine de croisements F_1 et obtient 500 F_4 qui seront testés quantitativement. Ces tests quantitatifs vont faire l'objet de notre étude.

2. *Objectif de la sélection quantitative.*

Qu'est-ce qu'une bonne variété?

Le sélectionneur cherche une variété qui ait un bon rendement dans la zone géographique qu'il s'est défini : c'est le premier critère. Mais il cherche aussi une garantie de régularité de ce rendement : sinon l'agriculteur risque d'être déçu une année et il n'achètera plus la semence. La valeur économique d'une variété sera donc définie à partir de la moyenne et de la variance du rendement dans la zone géographique considérée.

Parmi le lot de 500 F_4 on veut choisir en quelques années une variété (et une seule) que l'on espère performante pour la présenter à l'inscription comme variété commerciale. Il n'est pas forcément nécessaire de prendre la meilleure du lot : Si l'on en choisit une légèrement plus faible mais qui reste au-dessus du niveau des variétés commerciales déjà existantes, on sera quand même satisfait.

Contraintes de la sélection quantitative.

Elles ont deux origines : les quantités de semences disponibles et le budjet total du programme de recherche.

— Pour ce qui est des semences : la première année (F_4) on a des graines pour ensemencer deux parcelles. Ce n'est qu'au stade F_5 que l'on pourra en pratique semer par variété autant de parcelles que l'on voudra.

— Pour ce qui est du budjet : chaque parcelle a un prix de revient P_1 donné qui correspond à la location du terrain, aux frais de semis, de culture et de ramassage. L'installation d'un lieu d'expérimentation représente par ailleurs une dépense supplémentaire P_2 indépendante du nombre de parcelles qui y sont semées : ce sont les frais de transport du personnel et du matériel.

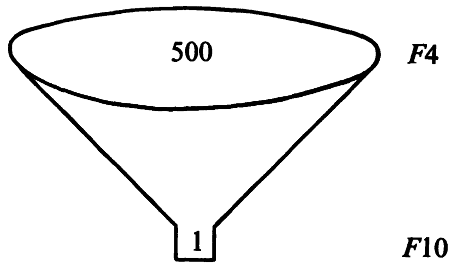
Le budget total vaut donc $B = \sum_{i=1}^{n} (np_i . P_1 + nl_i . P_2)$

n est le nombre d'années d'expérimentation.
np_i (resp. nl_i) est le nombre de parcelles (resp. lieux) de l'année i. Toute modification éventuelle des paramètres doit laisser B constant.

3. *Réalisation pratique de la sélection quantitative.*

La première année l'ensemble des 500 F_4 est partitionné en groupes de 22, chaque groupe de 22 est comparé à trois variétés témoins (toujours les mêmes) dans un plan d'expérience lattice 5 × 5 à deux répétitions. Les parcelles sont d'environ 10 m^2. Sur chacune d'elles on mesure le rendement par hectare (de l'ordre de 50 qx), la précocité, la résistance aux maladies et on note l'aspect de la plante. L'analyse statistique de ces résultats permet d'obtenir des estimateurs corrigés des effets du terrain.

Après examen des résultats de chacune des variétés, le sélectionneur conserve la moitié environ des variétés pour les expérimenter l'année suivante de la même façon, sauf qu'ayant plus de semence il expérimentera chaque variété en trois lieux. A la fin de l'année, il opère une nouvelle sélection et ainsi de suite jusqu'à la sixième année et dernière année où il choisit la variété qui lui paraît la meilleure. Le dispositif se présente comme un entonnoir.



L'individu choisi est présenté à l'inscription du catalogue des variétés commerciales. La sélection des variétés que l'on conserve est faite de manière totalement empirique : le sélectionneur regarde les résultats de la variété et suivant son intuition il prend sa décision. Ce n'est pas une mauvaise méthode en soi : il est difficile de tenir compte de tous les paramètres dans une méthode de sélection parfaitement déterminée. Mais dans le cas précis la sélection « à la main » est extrêmement lourde : un paramètre mesure sur 500 F_4 cela représente 5 000 nombres or on doit sélectionner simultanément les F_4, F_5... F_9 correspondants aux différents programmes démarrés successivement : cela représente environ 70 000 nombres à examiner avec soin : c'est fastidieux; il y a des risques de fluctuation des barèmes pendant les quelques semaines que dure ce travail.

(II) DONNÉES DISPONIBLES.

Nous avons travaillé sur des résultats de deux années d'expérimentation d'une entreprise de production de semences : France Canada Semences. Chaque année, il y avait du matériel de tous âges de F_4 à F_9 correspondant à des programmes démarrés en des années différentes. Cela représente pour chacune des deux années 1 000 variétés : 500 F_4, 250 F_5, etc.

Pour chaque F_5 il y a 9 parcelles : 3 par lieux et sur chaque parcelle 11 paramètres sont mesurés : au total cela fait un volume de 140 000 nombres. Un grand travail informatique a été nécessaire pour extraire de ces fichiers les données correspondant à un problème précis. De plus beaucoup de données manquaient : par exemple une variété qui aurait dû se trouver en 3 lieux ne se trouvait qu'en 2 car la semence manquait, certaines notes n'avaient pas été mises car les sélectionneurs n'avaient pas eu le temps de le faire. Le cas échéant il a fallu réduire le nombre de cas étudiés ou estimer les données manquantes.

Tous les paramètres mesurés comprennent une part d'erreur de mesure ou de ramassage et sont fortement influencés par l'environnement : climat, sol, méthode de culture. Par contre les risques d'interversion de sacs de graines ou de résultats expérimentaux sont très réduits car les sélectionneurs y font très attention.

(III) Éléments d'optimisation de la sélection.

Nous n'avons utilisé que les données de rendement. L'aboutissement de notre étude est constitué par des simulations de différentes stratégies de sélection : par ce mot nous entendons la répartition des nombres de répétitions, de lieux d'expérimentation et des taux de sélection au cours des six années (ce dernier point détermine la forme de l'entonnoir de la figure 1). Avant d'aborder cela nous nous sommes attachés à obtenir des estimateurs aussi précis que possible.

1. *Choix du type de plan d'expérience.*

Le type de plan utilisé était un lattice 5×5. Les mêmes données peuvent s'analyser comme un plan en blocs complets pour 25 traitements : il suffit de supposer que les effets des blocs incomplets du lattice sont nuls.

Nous avons cherché le type d'analyse donnant des estimateurs des contrastes de comparaison aux témoins (rendement de la variété — moyennes des rendements des témoins) les plus précis. La variance moyenne de ces estimateurs se calcule à partir de la variance résiduelle σ^2 de l'analyse de variance : il suffit de diviser σ^2 par l'efficacité du plan d'expérience : 1 pour l'analyse « type blocs complets », 5/6 pour le lattice. Si c'est l'analyse « type blocs complets » qui donne les meilleurs résultats c'est que la finesse du lattice ne sert à rien il conviendrait à l'avenir d'utiliser un plan en blocs complets.

Nous avons fait notre comparaison en différents lieux et années, les résultats ont été contradictoires : dans certains cas l'analyse lattice donnait de meilleurs estimateurs, dans d'autres cas c'était l'analyse « type blocs complets ». L'équipe des expérimentateurs étant habituée au plan lattice il est probable que celui-ci soit maintenu. Mais le résultat le plus important c'est qu'on ne peut définir un plan optimal en agronomie : d'abord pour des raisons d'assolement on ne peut pas refaire les essais au même endroit d'une année sur l'autre et de toutes façons on ne retrouvera jamais le même climat. Or ces facteurs ont une importance prédominante.

2. *Modélisation globale à l'échelle d'une année.*

Si l'on expérimente une variété en plusieurs lieux on prendra comme estimateur du contraste de comparaison aux témoins la moyenne des estimateurs dans les différents lieux. Mais comme nous l'avons déjà vu la valeur en rendement d'une variété est influencée par l'environnement : il y a une forte interaction variété-milieu. En un lieu donné on estime en fait la valeur moyenne d'une variété augmentée de l'interaction de la variété avec le milieu cultural. Nous considérons cette interaction comme une source d'erreur qui masque la vraie valeur de la variété. Dans le cas d'un plan en blocs complets cela nous conduit à écrire le modèle linéaire suivant :

$$Y_{ijk} = \mu + v_i + l_j + r_{jk} + I_{ij} + E_{ijk}$$

i (resp. j, resp. k) est l'indice de la variété (resp. lieu, resp. bloc complet) Y étant le rendement parcellaire.

Les deux derniers termes étant les deux effets aléatoires indépendants

$$I_{ij} \sim N(O, \sigma_V^2)$$

les I_{ij} sont indépendants entre eux.

$$E_{ijk} \sim N(O, \sigma_e^2)$$

les E_{ijk} sont indépendants entre eux.

Dans ce cas les estimateurs des contrastes sont des moyennes :
(moyenne des résultats de la variété — moyenne des résultats des témoins).

La variance de la moyenne des estimateurs des contrastes dans les différents lieux vaut :

$$\frac{\sigma_V^2}{n_l} + \frac{\sigma_e^2}{n_l . n_b}$$

n_l nombre de lieux
n_b nombre de blocs par lieux

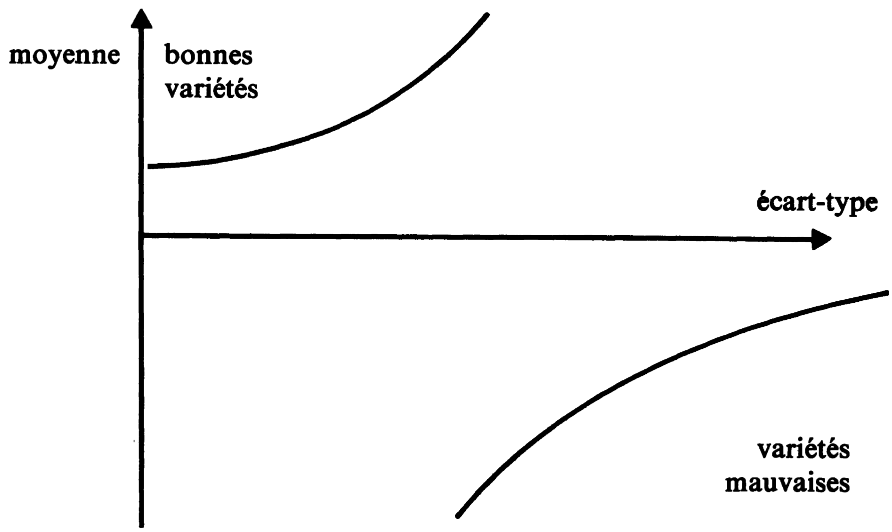
dans le modèle (1) la somme des carrés des termes d'interaction de l'analyse de variance à effets fixes a pour espérance :
$(n_b \sigma_V^2 + \sigma_e^2)d$ d étant le nombre de degrés de liberté associé. Cela permet d'estimer σ_V^2. Nous avons trouvé

$$\sigma_V^2 = 13{,}36 \text{ (qx/ha)}^2$$
$$\sigma_e^2 = 13{,}77 \text{ (qx/ha)}^2$$

avec ces valeurs plutôt que de semer 3 répétitions en 3 lieux il aurait mieux valu :
— soit 2 répétitions en 4 lieux : variance 5,06 $(\text{qx/ha})^2$
— soit 1 répétition en 5 lieux : variance 5,42 $(\text{qx/ha})^2$
au lieu de 5,98 $(\text{qx/ha})^2$ avec la répartition actuelle

3. *Aide graphique à la sélection*

Nous avons proposé de représenter sur un même diagramme (moyenne, écart-type) toutes les variétés de même âge :



l'écart-type est l'écart-type de la variété en fonction des différents milieux. Moyenne et écart-type sont calculés à partir des différents contrastes en chacun des lieux et chacune des années. Il y a donc un graphique représentant les F_4, un graphique représentant les F_5, etc. Ce graphique permet au sélectionneur de visualiser tous les résultats en un seul coup et de sélectionner rapidement toutes les variétés qui sont franchement bonnes ou franchement mauvaises. Il n'a pas été possible par contre sur ce diagramme de définir une frontière précise entre les variétés à conserver et celles à éliminer.

4. *Recherche d'une meilleure stratégie par simulation.*

La stratégie utilisée par France-Canada-Semences se définit ainsi :
500 entrées, la première année 2 répétitions, en un seul lieu, la dernière année 3 répétitions en 10 lieux, les autres années 3 répétitions en 3 lieux avec tout le long un taux de sélection de 45 %. Pour comparer d'autres stratégies à celle-là et éventuellement en trouver une meilleure nous avons fait des simulations par la méthode du « bootstrap »; l'idée générale est d'approcher la loi d'une variable aléatoire par la loi empirique sur un échantillon.

Dans le cas d'un plan en blocs complets les estimateurs du rendement d'une variété sont de simples moyennes : on peut donc estimer les contrastes de comparaison aux témoins à l'intérieur de chaque bloc : on prend le rendement de la variété moins la moyenne de rendement des témoins dans le bloc.

Pour ces contrastes on peut écrire le modèle suivant :

$$Y_{ijk} = V_i + I_{ij} + E_{ijk} \quad (2)$$

(En réalité nous avons utilisé un modèle plus complexe) les indices ont le même sens que dans le modèle (1)

V_i est la valeur de la variété i

I_{ij} est l'interaction de la variété i avec le lieu j

E_{ijk} l'erreur expérimentale

Nous supposerons tous les termes du modèle (2) comme aléatoires.

Sur l'échantillon de toutes les F_5 nous avons estimé les paramètres par les estimateurs classiques d'un modèle linéaire non orthogonal.

Pour générer une variété nouvelle artificielle on tire au hasard un indice i de variété. Pour avoir la valeur moyenne d'une variété i dans un milieu donné on tire au hasard un terme d'interaction parmi I_{il} ... I_{in} et on ajoute V_i. Pour générer un terme d'erreur expérimentale on tire un terme au hasard parmi tous les E_{ijk}. Cette méthode permet de ne faire aucune hypothèse sur la forme des lois. Elle donne des valeurs qui sont construites à partir des données expérimentales. A partir de cela on peut simuler : on génère 500 variétés, on les observe en un certain nombre de lieux, on peut les sélectionner si on s'est donné une méthode précise pour choisir les variétés à conserver en expérimentation.

Résultats

L'analyse de la stratégie acctuelle pour une valeur économique définie par : $E(m, \sigma) = m - 0{,}2\,\sigma$ montre une perte moyenne de 1 qx/ha (sur 60) entre la variété choisie et la meilleure des 500. Cette perte est due surtout à un mauvais choix la dernière année. Nous sommes en train de simuler d'autres stratégies où on est plus sévère en début de sélection et ou par contre on augmente le nombre de lieux la dernière année.

CONSTITUTION ET UTILISATION DES STOCKS DE FOIN DANS LES ÉLEVAGES PYRÉNÉENS, CONSTRUCTION ET RÉSULTATS D'UN MODÈLE

J.-L. CHARPENTEAU, *Laboratoire de biométrie,*
et M. DURU, *Station d'agronomie,*
I.N.R.A., Centre de Toulouse (1)

En Pyrénées centrales, les conditions d'hivernage des troupeaux sont le révélateur de l'ensemble de la contrainte climatique qui pèse sur l'éleveur. Nous présentons ici la conception, la construction et quelques résultats d'un modèle de simulation de cette contrainte. Deux possibilités d'en diminuer les effets sont envisagées : d'une part l'amélioration des structures d'exploitation, d'autre part la mise en œuvre d'actions correctrices conjoncturelles. Les résultats des simulations effectuées montrent la plus grande efficacité de la deuxième solution dans la réduction de la variabilité interannuelle des conditions d'hivernage.

In central Pyrennées, cattle wintering conditions reveal the climatic constraint for the breeder. We present here the conception, the construction and some results of a simulation model of this constraint. Two possibilities of reducing the effects are studied : on the one hand, the improvement in running structures, on the other hand, the implementation of correcting actions linked to the situation. The results of the used simulations show the greater efficiency of the second solution for reducing the variability of wintering conditions over years.

In den Zentralpyrenäen die Bedingungen der Ueberwinterung der Tierherden bringen den Zwang zum Ausdruck, dem der Herdenbesitzer unterworfen ist. Wir geben hier die Conception, die Construction und einige Resultate eines Modells der Simulation dieses klimatischen Zwanges. Zwei Möglichkeiten diese Wirkungen zu vermindern sind ins Auge gefasst : Einerseits eine Verbesserung der gesamten Wirtschaft, andererseits die Anwendung von correctiven conjunkturellen Massnahmen. Die Ergebnisse der angewendeten Simulationen zeigen die grössere Wirksamkeit der zweiten Lösung, die Reduktion der interannuellen Variabilité der Ueberwinterung.

INTRODUCTION

Dans les Pyrénées centrales les conditions dans lesquelles s'effectue l'hivernage des troupeaux, constituent le point de convergence des différents aléas climatiques qui peuvent intervenir tout au long de l'année. En effet :

— la longueur de l'hiver dépend à la fois des premières neiges, des gelées automnales et du démarrage de la végétation au printemps suivant;

(1) BP 12, 31320 CASTANET-TOLOSAN

— à longueur d'hiver égale, la quantité de fourrage nécessaire dépend des possibilités de pâturage hivernal c'est-à-dire de l'importance de l'enneigement;

— la quantité de foin récoltée en vue de l'hiver dépend du climat du printemps précédent;

— enfin la qualité de ce foin dépend de la pluviométrie estivale.

Le grand nombre d'interventions possibles du climat d'une part, et la variabilité observée de chaque élément climatique d'autre part nous ont conduit à en envisager l'étude à travers la construction d'un modèle (M. DURU et J.L. CHARPENTEAU, 1981). C'est la conception, la construction et quelques résultats de ce modèle que nous présentons ici.

I — Conception du modèle

Les objectifs assignés à ce modèle sont de trois ordres.

— Il doit représenter le mieux possible les mécanismes d'intervention des différentes contraintes climatiques, leur hiérarchie et leurs interactions sur le résultat final de l'hivernage.

— Il doit nous permettre de déceler quels paramètres de structure de l'exploitation seront les déterminants de la sensibilité de cette exploitation et éventuellement quelle gamme de variation de ces paramètres est admissible.

— Enfin, nous espérons pouvoir après l'analyse du fonctionnement de cette partie du système d'exploitation qu'est la constitution des stocks de fourrage et leur consommation, être capables de déceler et d'analyser des possibilités de « régulations tactiques ».

En fait, il s'agit pour nous de construire un outil prenant en compte simultanément toutes les idées à priori que l'on a sur le phénomène étudié.

Nous n'avons pas conçu cet outil comme un producteur potentiel de références. En effet, dans cette étude c'est plus la comparaison entre deux simulations que la valeur du résultat de chacune d'elle qui nous intéresse. C'est pourquoi, nous ne pouvons qualifier notre modèle de « biologique » et c'est pourquoi les deux principaux phénomènes biologiques mis en œuvre sont modélisés de façon assez sommaire (Bibliographie pour l'alimentation des animaux et relation linéaire pour la pousse de l'herbe). Toujours dans le même esprit d'analyse du fonctionnement d'un processus, notre modèle n'est pas prévisionnel. C'est la raison pour laquelle tous les événements sont pris en compte de façon entièrement déterministe et que les effets partiels du climat sur le fonctionnement de parties du système sont pris en compte à posteriori (ex. : climat du printemps sur la récolte de foin, longueur de l'hiver sur le nombre de rations à distribuer).

Ainsi défini, voyons maintenant la composition de ce modèle. L'ensemble du fonctionnement du système est représenté sur la figure 1. Nous avons choisi pour cette représentation les conventions de FORRESTER (FORRESTER, 1980). C'est-à-dire que les rectangles figurent les variables d'état du système, que les cercles représentent les calculs intermédiaires nécessaires, que les flux de matière sont représentés en traits pleins et les flux d'information en tiretés. Par ailleurs, le symbole de la vanne est employé lorsqu'il s'agit d'un point où l'on peut intervenir sur un flux.

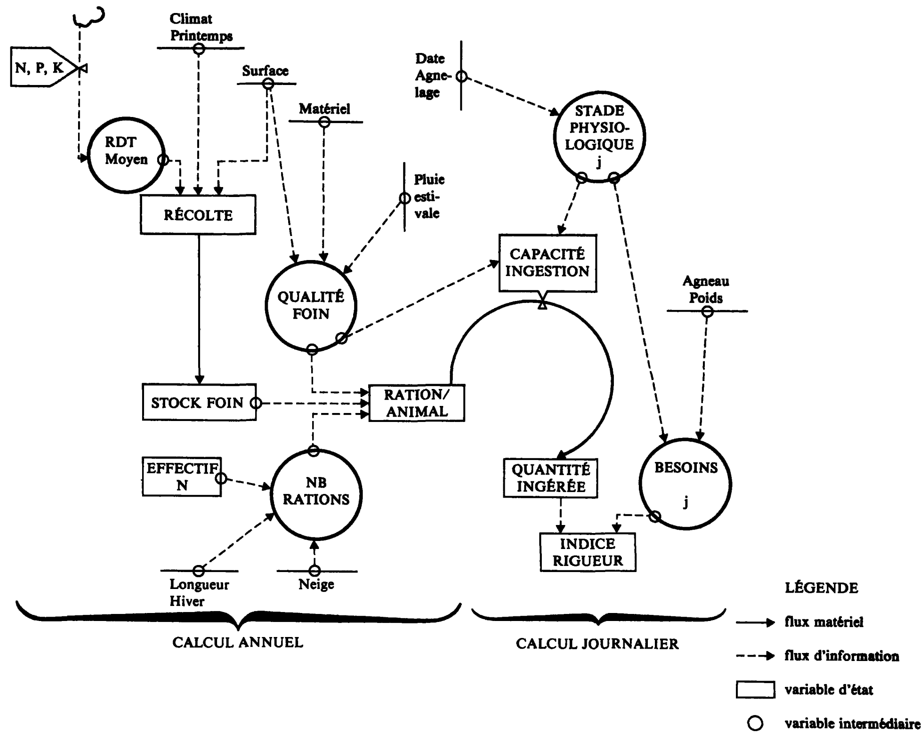


FIGURE 1

Diagramme de FORRESTER du système étudié

D'une façon claire, sur cette figure, notre modèle se décompose en deux parties : une partie dans laquelle les calculs se font une fois et une seule dans l'année, c'est le calcul des paramètres de la récolte de foin (quantité, qualité), une autre partie dans laquelle les calculs sont journaliers ou presque, c'est la consommation de cette récolte par les animaux. Cette façon de procéder montre en outre concrètement pourquoi il est impossible d'utiliser ce modèle pour faire de la prévision. En effet, le poids et les caractéristiques des rations offertes sont déterminés d'après la connaissance exacte de l'hiver.

Les variables de structure d'exploitation que nous avons choisies de faire varier sont :

— la surface fauchée,
— le rendement moyen des prés fauchés,
— le nombre de jours nécessaires à la fenaison (matériel, main-d'œuvre),
— la taille du troupeau,
— la date d'agnelage du troupeau.

Nous appelons ces différentes variables les commandes du système. Leur champ de variation est précisé dans l'annexe 1.

Les entrées sont constituées uniquement des données climatiques relatives aux différentes périodes de l'année, c'est-à-dire :

— l'indice de potentialité climatique de TURC (TURC, 1972) cumulé sur la période 1er mai-5 juillet pour le printemps et la détermination de la biomasse récoltable;

— les séquences de jours avec et sans pluie pour la période estivale pour déterminer les possibilités de récolte du foin et ainsi la qualité moyenne du foin;

— la longueur de l'hiver,

— les séquences de jours où la neige couvre ou non le sol pendant l'hiver pour calculer le nombre de rations à fournir pour l'hivernage.

Enfin, la seule sortie de notre modèle est constituée par le calcul d'une quantité que nous appelons « indice de rigueur de l'hiver ». Cet indice est égal au produit du déficit global en énergie ou en azote par la quantité d'énergie ou d'azote que doit obligatoirement prélever l'animal au pâturage pour équilibrer ses besoins. Ainsi, nous prenons en compte à la fois l'importance du déficit alimentaire et sa répartition au cours de l'hiver.

Dans un deuxième temps nous avons envisagé la possibilité d'intervenir sur le fonctionnement de l'exploitation de façon non systématique. Alors que précédemment, tout changement dans les valeurs d'une ou plusieurs variables de commande était définitif, cette nouvelle optique vise à étudier l'effet de modifications temporaires du système de production. Nous appelons ces modifications « régulations ». Elles visent essentiellement à contrebalancer des effets négatifs du climat et sont définies comme suit :

— possibilité de fertilisation azotée en cas de démarrage tardif de la végétation,

— achat de foin si la récolte est insuffisante,

— achat de foin ou diminution de l'effectif du troupeau si l'hiver est précoce.

Enfin la possibilité d'une gestion des stocks de foin par report d'une année sur l'autre a été envisagée. L'ensemble de ces régulations et leur insertion dans l'ensemble du modèle constituent la figure 2.

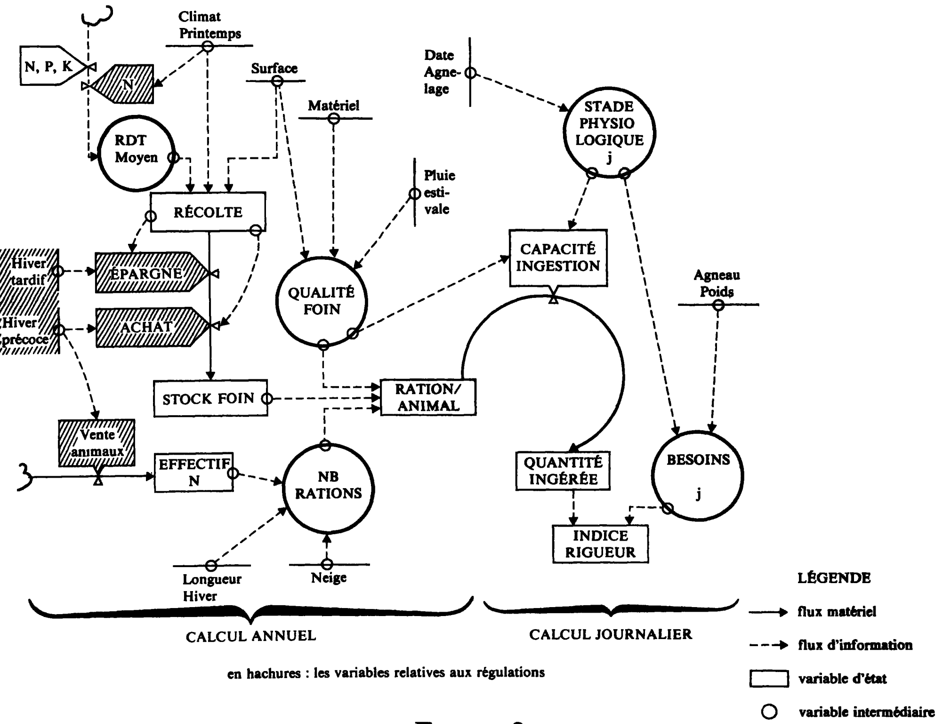


FIGURE 2

Diagramme de FORRESTER du système étudié avec régulations

II — Identification du modèle

Par construction, à chaque flux d'information représenté sur les figures 1 et 2 correspond, soit une fonction algébrique soit un algorithme de calcul ou d'exploration de données. C'est cet ensemble qui constitue la partie « matérielle » de notre modèle, que nous allons décrire rapidement ici. Le détail et les justifications agronomiques de ces relations sont décrits ailleurs (DURU M., CHARPENTEAU J.-L., 1981, et DURU M., CHARPENTEAU J.-L., 1982). L'ensemble des sigles et abréviations est rappelé en annexe 2.

II-1. *Calculs annuels*

• Récolte de foin

$$QMS = \frac{RDT}{50} \times (2 \times ITURC + 29) \times S$$

avec :

QMS : quantité de matière sèche récoltée (t)
RDT : rendement moyen de la parcelle (t/ha)
ITURC : indice de TURC
S : surface fauchée (ha)

• Qualité du foin

$$TENUF = 72.25 - 0.0022\ TREP - 0.0018\ TREP^2$$
$$TEMAD = -0.909\ TREP + 122.3$$

avec

TENUF : teneur en énergie (UF/kg)
TEMAD : teneur en matière azotée digestible (t/kg)
TREP : nombre de jours de repousse du foin.

L'algorithme de calcul de TREP à partir des jours avec et sans pluie est donné figure 3. Dans ce calcul, l'intervention de la structure de l'exploitation se fait à travers le paramètre NECES : nombre de jours nécessaires à la récolte d'une quantité fixée de foin.

Nombre de rations

$$NBRAT = (NJ0 \times 2) + (NJ1 \times 3)$$

avec :

NBRAT : nombre de rations
NJ0 : nombre de jours sans neige sur le sol
NJ1 : nombre de jours avec neige.

• Poids de la ration

Pour tenir compte des pratiques observées chez les éleveurs nous distinguons les cas d'agnelage groupé et dispersé.

En cas d'agnelage dispersé :

$$PRAT = QMS/NBRAT/EFF$$

où

EFF : est l'effectif du troupeau.

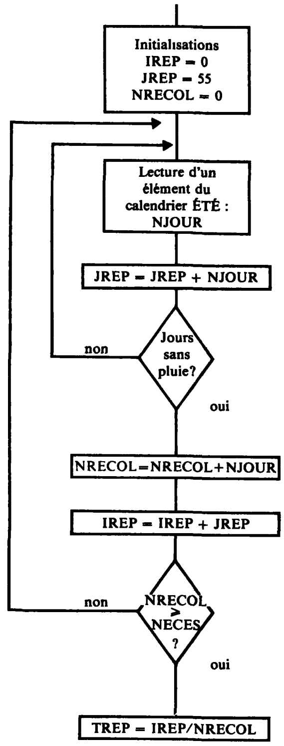


FIGURE 3.

Calcul du nombre de jours de repousse de la récolte (signification des variables : annexe 2)

En cas d'agnelage groupé on calcule

$$X = QMS/(1.325 \times NRAT1 + NRAT2)/EFF$$

et on donne :

$$PRAT = X \times 1.325 \text{ avant le 01 mars}$$

et

$$PRAT = X \text{ après}$$

(NRAT1 et NRAT2 sont respectivement le nombre de rations nécessaires avant et après le 01 mars).

II-2. *Calculs journaliers ou quasi journaliers*

- Besoins des animaux

En cas d'agnelage dispersé, les besoins d'une brebis fictive ont été pris :

$$BESUF = 1.11 \text{ (UF)}$$
$$BESMAD = 111 \text{ (g de MAD).}$$

(I.N.R.A., 1978)

En cas d'agnelage groupé, nous avons les formules suivantes :

$$BESUF = (-0.0052NJ^2 + 0.0307NJ + 142.85)/100 \text{ (UF)}$$
$$BESMAD = -0.006NJ^2 + 0.231NJ + 231.158 \text{ (g de MAD)}$$

pendant les 5 mois qui suivent l'agnelage avec NJ : nombre de jours écoulés depuis l'agnelage.

Si NJ > 150 nous avons :

BESUF = 0.7 (UF)

BESMAD = 54 (g de MAD).

- Ingestion de matière sèche par les animaux

$$MSI = COEF \times (-0.00159\ TREP + 0.3642)$$

où

MSI : matière sèche ingérée en kg

TREP : caractérise comme précédemment la qualité du foin

COEF : vaut 2.65 dans les 5 mois après l'agnelage et 2.0 ailleurs.

L'ensemble des calculs relatifs au poids de la ration et à son ingestion est représenté sur l'algorithme de la figure 4.

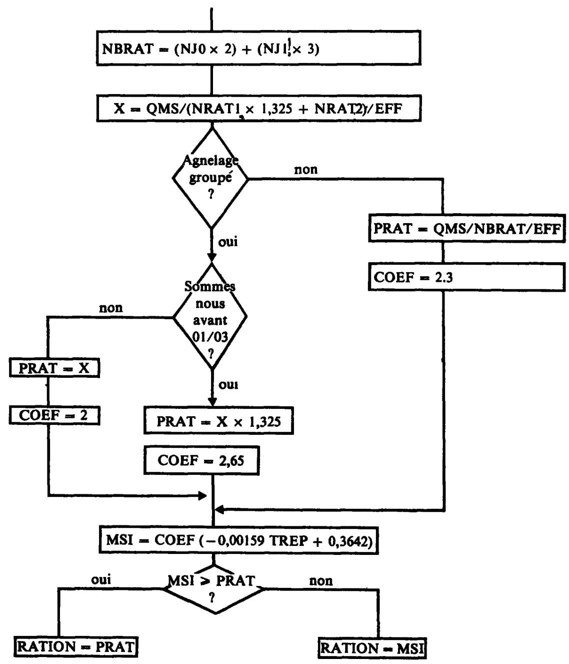


FIGURE 4.

Calcul du poids de la ration ingérée
(signification des variables : annexe 2)

- Calcul de l'indice de rigueur de l'hiver

Le déficit global est calculé de la façon suivante :

$$DEF = (QMS \times TENUF) - \sum_{j} (RATIO_j \times TENUF)$$

où

RATIO est la quantité de matière sèche effectivement ingérée le jour j

On a bien sûr la formule équivalente pour calculer le déficit en azote.

La composante journalière, elle, est calculée de la façon suivante :

Soit $DEFJ_j$ le déficit en énergie ou en azote le jour j, nous calculons pour les jours où ce déficit est positif ou nul la quantité :

$$\text{COMPJ} = \sum_j \text{DEFJ}(j) / \text{NJO}$$

où

COMPJ est la composante journalière de l'indice de rigueur

et NJ0 le nombre de jours sans neige.

L'indice de rigueur de l'hiver est donc égal à :

$$\text{IRH} = \text{DEF} \times \text{COMPJ}$$

Cet indice est calculé pour l'énergie (UF) et pour l'azote (MAD).

II-3. *Les régulations : règles de décision et mise en œuvre*

- Fertilisation azotée au printemps (R1)

On décide d'épandre 50 unités d'azote le 15 mai si l'indice de TURC cumulé du 1er janvier au 15 mai est inférieur ou égal à 2.

Cette opération permet d'augmenter le rendement potentiel de 0,6 T de matière sèche à l'hectare.

- Achat de foin au vu de la récolte (R3)

On décide d'acheter du foin si l'indice de TURC cumulé du 1er mai au 25 juin est inférieur ou égal à 9.

- Interventions au début de l'hiver

Si l'hiver débute avant le 16 novembre on décide soit de diminuer l'effectif du troupeau de 25 % (R2) soit de porter le stock de foin à 120 % de la récolte (R4).

- Régulation combinée (R10).

Le décideur a à sa disposition toutes les possibilités d'intervention (sauf la variation d'effectif du troupeau) décrites ci-dessus. Les décisions sont prises en « temps réel » et l'achat de foin se fait 1 fois au maximum.

- Gestion de stocks, mise en réserve

On décide de mettre 15 % de la récolte en réserve pour l'année suivante si l'hiver débute après le 22 novembre et si la récolte est au moins égale à 120 % de la moyenne.

Le tableau 1 récapitule les fréquences d'interventions des différentes régulations envisagées. L'observation de ce tableau appelle plusieurs remarques. En effet, si chaque

TABLEAU 1

Fréquence d'intervention des régulations

Année	61	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	Tot.
R1								×	×				×	×			×	×	6
R2					×					×		×	×		×	×		×	7
R3							×									×		×	3
R4					×					×		×	×		×	×		×	7
R10					×		×	×	×	×		×	×	×	×	×	×	×	12
Stock	×		×	×		×			×										5

régulation élémentaire n'intervient pas plus d'une année sur deux (le paramètrage des règles de décision a d'ailleurs en partie été choisi de façon à respecter ce critère) la régulation combinée elle, est mise en œuvre deux années sur trois. Enfin, comme on pouvait à priori s'y attendre, il n'y a pas recouvrement entre régulations et report de stock.

Toutes ces remarques sont évidemment conditionnées par le choix que nous avons fait dans le paramétrage de nos règles de décisions. Ces différents choix n'ont d'autres qualités que celles de paraître raisonnables et de pouvoir servir d'exemple.

III — Résultats des simulations

L'exploration d'un champ de variation suffisamment large pour les variables de commande du système ainsi que l'introduction des régulations nous ont conduit à effectuer un nombre considérable de simulations. De cet ensemble de résultats (DURU M. et CHARPENTEAU J.L., 1982) nous avons extrait un échantillon suffisamment restreint mais néanmoins suffisamment éloquent.

Nous avons choisi de représenter les résultats relatifs à 4 hypothèses de structure d'exploitation. Dans ces hypothèses, l'une constitue une référence couramment rencontrée dans les exploitations (H1), les autres représentent des améliorations pour la chaîne de récolte (H3), la fertilisation (H2) ou les deux cumulés (H4) (tableau 2). Les calculs ont été effectués dans tous les cas avec un troupeau de 200 brebis. De plus, nous ne présentons ici que les résultats relatifs à l'énergie (UF) et dans le cas d'un agnelage groupé.

TABLEAU 2

Hypothèses de structure d'exploitation

Hypothèse	N°	Surface fauchée (ha)	Rdt moyen (t/ha)	Stock ingéré (t)	Indicateur chaîne récolte
Situation de départ	H1	6,8	3,5	23,8	3,0
Amélioration récolte	H2	6,8	3,5	23,8	0,7
Amélioration rendement	H3	6,8	4,5	30,6	3,0
Amélioration conjointe	H4	6,8	4,5	30,6	0,7
Récolte rendement					

III-1. *Étude sans régulations*

Pour quelques années caractéristiques nous avons représenté l'indice avec ses composantes (figure 5a). Puis, pour l'ensemble des années nous avons représenté les valeurs et variations de l'indice (figure 5b).

Sur le premier graphique (figure 5a), les déplacements des coordonnées de l'indice pour les hypothèses H2, H3, H4 relativement à l'hypothèse de référence montrent que la variabilité interannuelle est supérieure aux variations induites par les différentes hypothèses. Dans le tableau 3 sont indiquées les moyennes, les écarts types et les coefficients de variations de l'indice correspondant à la figure 5b. D'une manière générale, l'amélioration de *l'efficacité* (moyenne de l'indice) va de pair avec celle de la *sensibilité* (variance de l'indice).

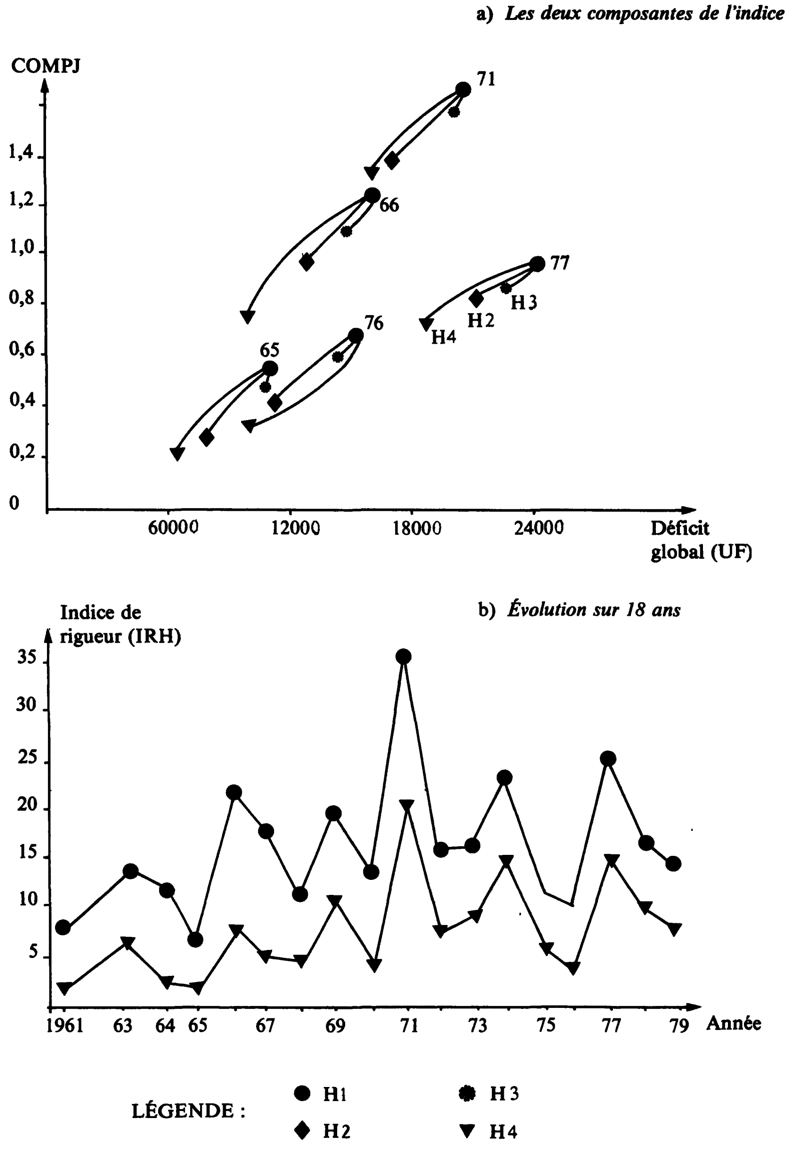


FIGURE 5.

Indice de rigueur de l'hiver (IRH) pour l'énergie, agnelage groupé.

TABLEAU 3

Indice de rigueur de l'hiver pour l'énergie, agnelage groupé : statistiques élémentaires

Hypothèses	UF		
	$\bar{x}$	σ	CV (%)
H1	16,6	6,8	41
H2	9,8	5,5	56
H3	14,8	6,5	44
H4	8,0	5,1	64

A ce niveau, notons pour mémoire que les graphiques relatifs à l'azote (MAD) montreraient des effets différents des améliorations de structure. En particulier, l'hypothèse H3 améliore plus l'indice MAD qur l'indice UF.

Le fait que les courbes de la figure 5b ne se recoupent quasiment jamais, ainsi que la persistance d'une grande variabilité interannuelle nous font penser que l'amélioration des structures (en restant dans le domaine du « raisonnable » économiquement) n'est pas capable de résoudre de façon satisfaisante les problèmes de l'hivernage. En effet, la persistance d'hivernages difficiles laisse entier le problème d'une répercussion d'une année sur l'autre des effets néfastes de cet hivernage (état du troupeau, performances de reproduction, etc.). Il faut donc envisager la possibilité d'intervenir au coup par coup, éventuellement plusieurs fois dans l'année si nécessaire.

III-2. *Étude avec régulations*

Les résultats des simulations (tableau 4) font apparaître que l'application d'azote au printemps (R1) et l'achat de foin au vu des stocks (R3) diminuent très peu les moyennes et les écarts types de l'indice. Par contre, l'achat de foin (R4) ou la réduction de l'effectif du troupeau (R2) en cas d'hiver précoce réduisent nettement les moyennes des indices UF et MAD et très fortement les écarts types.

TABLEAU 4

Indice de rigueur de l'hiver pour l'énergie, agnelage groupé, avec régulations : statistiques élémentaires

	UF		
	$\bar{x}$	σ	CV (%)
H1	16,6	6,8	41
H2	9,8	5,5	56
H1R1	15,2	7,0	46
H1R2	12,9	4,5	35
H1R3	16,1	6,8	42
H1R4	14,2	4,9	34
H1R10	13,6	5,0	36

La figure 6 montre bien que les 4 années les plus difficiles, les régulations R4 et R2 sont mises en œuvre mais ces années restent encore difficiles. On peut penser qu'un achat de foin plus important réduirait de manière encore plus significative la variabilité interannuelle. La régulation combinée (R10) apporte peu d'amélioration par rapport aux précédentes.

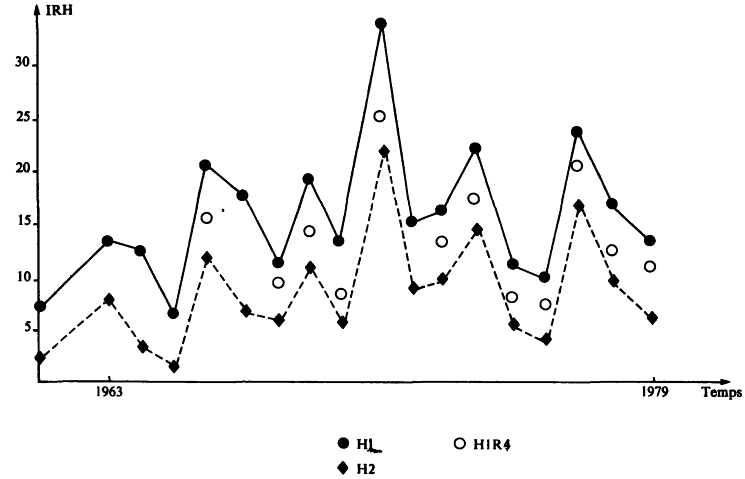


FIGURE 6.

Indice de rigueur de l'hiver pour l'énergie, agnelage groupé avec régulation

Enfin, la mise en réserve de 15 % du stock les années où la récolte est bonne et l'hiver tardif diminue sensiblement les écarts types et les coefficients de variation pour les hypothèses H2 et H4 (figure 7, tableau 5). Toutefois, une analyse plus détaillée des résultats montre que l'indice n'est abaissé que pour 1 ou 2 hivers rigoureux. Dans les autres cas, des mises en réserve se suivent 2 années consécutives ou concernent de bonnes années et n'ont donc pas d'effet sur la diminution de la variabilité interannuelle.

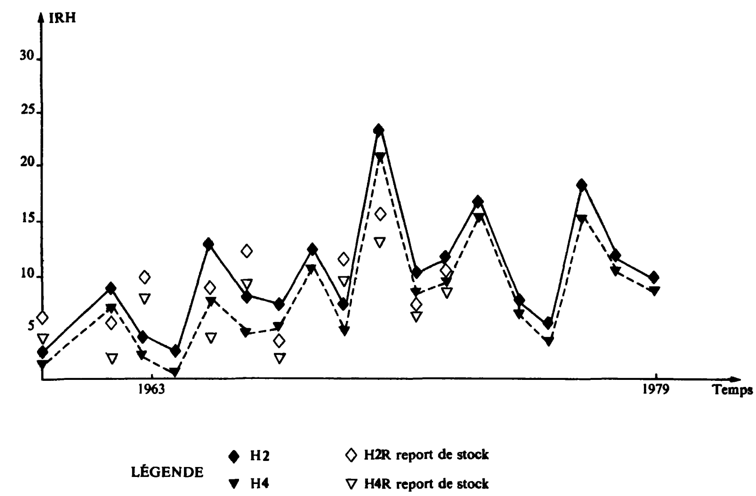


FIGURE 7.

Indice de rigueur de l'hiver pour l'énergie, agnelage groupé — Régulation report de stock

TABLEAU 5

Indice de rigueur de l'hiver pour l'énergie, agnelage groupé, avec report de stock : statistiques élémentaires

Hypothèses	UF sans report de stock		UF avec report de stock	
	$\bar{x}$	σ	$\bar{x}$	σ
H1	16,6	6,8	16,1	6,7
H4	8,0	5,1	7,6	5,0

III-3. *En guise de validation*

Les objectifs assignés au modèle ainsi que les méthodes employées pour le construire, ne permettent pas d'effectuer une validation classique. Il importe néanmoins de s'assurer du caractère vraisemblable des indications fournies par ce modèle. Nous utiliserons pour cela deux résultats :

— d'une part, par enquête nous avons constaté qu'il y a accord entre les éleveurs et notre modèle sur le classement en rigueur des hivernages 1976, 1977 et 1978.

— Par ailleurs, on peut considérer que le fonctionnement du système étudié, une année donnée, peut être caractérisé par les variables suivantes :

- structures d'exploitations,
- croissance moyenne des agneaux sur l'hiver,
- variation du poids des brebis,
- prélèvement moyen au pâturage.

Ces variables, selon que l'on se situe à l'intérieur du modèle ou en observateur d'une exploitation particulière ont des statuts différents. En effet, la croissance des agneaux et les variations de poids des brebis sont les résultats de l'exploitation mais sont des données du modèle. Inversement, le prélèvement au pâturage est une donnée de l'observation en exploitation mais est le résultat du modèle.

Ces différences de statut permettent néanmoins d'affirmer que c'est à travers la comparaison de telles séries de résultats que pourra se faire la confrontation de notre modèle à la réalité. A cet effet, nous avons regroupé dans le tableau 6 les résultats observés par A. GIBON (GIBON, 1981) dans une exploitation voisine de notre structure H1, l'hiver 1979/1980, les observations de G. BALENT (BALENT, GIBON, 1980) le même hiver sur les parcours et les paramètres et résultats d'une de nos simulations pour la même année également. L'observation de ce tableau montre que s'il n'y a pas parfaite concordance entre l'observaton et la simulation, l'écart entre les deux reste du domaine de l'acceptable. L'observation de la figure 5 nous montre que cet hiver 1979/1980 se situe dans les meilleurs.

Par son caractère ponctuel, cette confrontation des résultats de notre modèle aux données réellement observées ne peut évidemment pas servir de validation. Néanmoins, le cadrage ainsi réalisé permet de sortir un peu du champ initial de nos préoccupations (recherche des moyens pour réduire la variabilité interannuelle) et porter a posteriori sur certaines situations un diagnostic de faisabilité ou non de l'ajustement ressources-besoins.

TABLEAU 6

Comparaisons simulations-observations pour l'hiver 1979/1980

	Structure	Croissance agneaux	Variation réserve corporelles brebis	Prélèvement pâturage	
				UF	MAD (g)
Observation A. GIBON	Voisine de H1	200 g	oui		
Observation G. BALENT				0,6-0,8	40-60
Simulation agnelage groupé	H1	200 g	oui	0,77	65
Simulation agnelage dispersé	H1	200 g	oui	0,87	76

CONCLUSION

Les résultats de cette étude peuvent susciter deux types de conclusions. Des conclusions agronomiques concernant essentiellement le type de mesures à prendre en vue de réduire la variabilité interannuelle de la rigueur de l'hivernage et des conclusions de type méthodologique relatives au type de modèle mis en œuvre et à la pertinence par rapport aux objectifs fixés.

Le bilan agronomique de ces simulations à déjà été fait (DURU M., CHARPENTEAU J.L., 1982). Il constate essentiellement la supériorité d'une stratégie basée sur la mise en œuvre de régulations sur celle qui consiste à améliorer une ou plusieurs caractéristiques de structure, et il pose le problème de la mise en œuvre de telles stratégies.

Sur le plan méthodologique, il faut s'interroger sur la validité des approximations faites dans les entrées biologiques du modèle et dans la prise en compte déterministe de l'environnement. Dans le cadre des objectifs assignés au modèle (étude essentiellement théorique portant sur l'identification de processus plutôt que production de références ou prévision), et dans la mesure où nos simulations ont été faites dans un espace « raisonnable » pour les variables de commande, on peut dire que cette pratique se révèle efficace et fiable. Elle ne peut néanmoins constituer qu'une étape dans une tentative de modélisation du fonctionnement de l'exploitation agricole. Les étapes suivantes viseront à améliorer les approximations faites ici de deux manières :

— par l'inclusion de véritables modèles biologiques,

— par la prise en compte de l'environnement soit comme source d'aléa soit comme source d'informations (mémoire).

REMERCIEMENTS

Ce travail s'est effectué sur le contrat DGRST « Recherches sur l'élevage en montage Pyrénéenne ». Nous remercions ici tous les membres de l'équipe pluridisciplinaire constituée à l'occasion de ce contrat.

BIBLIOGRAPHIE

BALENT G., GIBON A. (1980)
Le pâturage hivernal dans les Pyrénées Centrales.
Recherches sur l'élévage en Pyrénées. Ronéo 95 p. INRA Toulouse

DURU M., CHARPENTEAU J.L. (1981)
The farming system in the Pyrenees : a model of the constitution and utilisation of hay stocks.
Agricultural System 7 137-156.

DURU M., CHARPENTEAU J.L. (1982)
Constitution et utilisation des stocks de foin dans les Pyrénées : résultats de simulations.
Document interne.

FORRESTER J.W. (1980)
Principes des systèmes
Presses Universitaires de Lyon.

GIBON A. (1981)
Pratiques d'éleveurs et résultats d'élevages dans les Pyrénées Centrales. Logique de la conduite des troupeaux et possibilités d'amélioration.
Thèse de docteur ingénieur 105 p.

INRA (1978)
L'alimentation des ruminants
INRA Publications (Route de St-Cyr - 78000 Versailles).

TURC L. (1972)
Indice climatique de potentialité agricole
Science du sol 2 81-101.

ANNEXE 1

Champ de variation des commandes du système

a) Rendement × surface fauchée = stock espéré

Stock / Rdt	21	24	27	30	33	36	39	42
3,5	6,0	6,8	7,7	8,6	9,4	10,2	11,1	12,0
4,5	4,7	5,3	6,0	6,7	7,3	8,0	8,7	9,3
5,5	3,8	4,4	4,9	5,4	6,0	6,5	7,1	7,6

b) **Indicateurs de chaîne de récolte utilisés dans le modèle et leur signification en terme de matériel et de main-d'œuvre**

Indicateur de chaîne de récolte en jour/ha	MATERIEL	Main d'œuvre en unité travailleur
0,7	Ensilage-Séchage en grange	3
2	2 tracteurs — 1 presse	2
3	1 tracteur — 1 presse	2
4	Motofaucheuse — 1 presse	2
5	Motofaucheuse vrac	2

c) Taille du troupeau

EFF = 200 brebis

d) Dates d'agnelage

groupé : toutes les brebis agnèlent le 1er octobre
dispersé : distribution uniforme de l'agnelage sur tout l'hiver.

ANNEXE 2

Signification des sigles et abréviations

BESMAD	Besoin en matière azotée digestible en g/animal/jour
BESUF	Besoin énergétique en unité fourragère/animal/jour
COEF	Coefficient de calcul
COMPJ	Composante journalière de l'indice de rigueur
DEF	Déficit global sur l'hiver
DEFJ	Déficit journalier
EFF	Effectif du troupeau en brebis mères
ITURC	Indice climatique de potentialité agricole cumulé du 01/05 au 05/07
IRH	Indice de rigueur de l'hiver
MAD	Matière azotée digestible
MSI	Quantité de matière sèche ingérable par l'animal pour une ration
NBRAT	Nombre de rations nécessaires pour un animal pendant l'hiver
NECES	Nombre de jours nécessaires à la récolte d'une quantité de foin
NJ	Nombre de jours écoulés depuis l'agnelage
NJ0	Nombre de jours sans neige
NJ1	Nombre de jours avec neige
NRAT1	Nombre de rations nécessaires entre le début de l'hiver et le 01/03
NRAT2	Nombre de rations nécessaires après le 01/03
NRECOL	Nombre de jours de récolte effective
PRAT	Poids possible d'une ration de foin
QMS	Quantité de matière sèche récoltée
RATIO	Poids de la ration effectivement distribuée
RDT	Rendement moyen d'une surface en tonnes de matière sèche/ha

S	Surface fauchée en ha
TEMAD	Teneur d'un fourrage en matière azotée digestible en g/kg
TENUF	Valeur énergétique d'un fourrage en UF/kg
TREP	Nombre de jours de repousse moyen d'une quantité de foin
UF	Unité fourragère (1730 K cal.)

QUE FAIRE DES DONNÉES DE TÉLÉDÉTECTION?

C. DERVIN

Chaire mathématiques et informatique
Institut national agronomique, Paris-Grignon [1]

La télédétection est de plus en plus employée en agriculture et dans les sciences utiles pour l'agriculture : de l'inventaire des forêts et des cultures à l'étude de la propagation de certaines épidémies végétales, sans oublier l'espionnage! Dans cet article, nous essayons de montrer en quoi les méthodes statistiques de l'analyse des données peuvent être utiles pour mieux interpréter les données, dont la caractéristique essentielle est l'important volume.

The remote detection is increasingly used in agriculture and the sciences useful in agriculture : from the inventory of forests and crops to the study of the spreading of some plant epidemics, without forgetting spying! In this paper, we try to show how statistical methods of data analysis can be useful for a best interpretation of data whose main characteristic is the important volume.

Die Télédetection wird immer mehr in der Landwirtschaft und den Wissenschaften, die von Interesse für die Landwirtschaft sind verwendet : von dem Inventar der Wälder und der Getreidefelder bis zum Studium der Ausbreitung gewisser Pflanzenepidemien, ohne dass man die Wirtschaftsspionage vergisst!. In dem vorliegenden Artikel versuchen wir zu zeigen, inwiefern die statistichen Methoden der Analyse der Gegebenheiten nützlich sein können um die gegebenen Tatsachen zu erklären deren wesentliches Charakteristum sein Volumen ist.

Un des moyens de reconnaître des objets est d'étudier leurs propriétés spectrales. Pour cela on étudie les caractéristiques des ondes électromagnétiques *réfléchies* ou *émises* par ces objets. Ces ondes sont reçues par des capteurs qui peuvent être ACTIFS ou PASSIFS.

Les appareils passifs, dénommés SCANNERS, enregistrent :

— les radiations dues au rayonnement solaire qui est réfléchi par les objets;

— les radiations dues à une auto-émission.

Les appareils actifs émettent et reçoivent. Ils envoient, dans une longueur d'onde déterminée, des radiations sur la surface à étudier, qui absorbe, transmet et réfléchit le rayonnement. L'onde réfléchie est reçue par l'appareil récepteur.

1 — *Divers capteurs*

1) *Appareils photographiques :* Ils travaillent dans la bande spectrale allant de 400 à 900 nm (visible et proche infrarouge)

2) *Capteurs non photographiques :* Il faut distinguer les appareils à balayage de ceux qui ne possèdent pas ce système.

Journal de la Société de statistique de Paris, tome 123, n° 2, 1982.

(1) 16, rue Claude Bernard, 75005 Paris.

Balayage :

Un réflecteur rotatif (dont la vitesse est adaptée à la vitesse de déplacement de l'avion ou du satellite...) balaie la zone suivant des bandes parallèles.

Ces bandes sont perpendiculaires à la ligne de vol, et sont avec ou sans recouvrement.

SCANNERS MULTISPECTRAUX :

Ils reçoivent, dans plusieurs bandes spectrales, les radiations émises par la surface survolée. Ils sont munis d'un système de balayage qui dirige le faisceau sur un spectromètre muni de plusieurs détecteurs. Chaque détecteur émet un signal électrique correspondant à l'énergie contenue dans une bande de spectre. Ces signaux sont alors, soit transmis directement soit stockés sur bande magnétique.

(Par exemple le scanner Daedalus a 10 canaux allant de 380 nm à 1100 nm).

SCANNER INFRAROUGES :

Ils ne reçoivent que les radiations infrarouges émises ou réfléchies par la surface survolée.

SPECTRO-RADIOMÈTRES :

Ils émettent et reçoivent des rayonnements. Leur source de rayonnement éclaire la surface survolée qui en réfléchit une partie. Celle ci est reçue par le spectro-radiomètre. Ces radiomètres travaillent dans une portion étroite du spectre : par exemple, les radars opèrent dans le domaine des micro-ondes, c'est-à-dire pour des longueurs d'onde > 6 mm.

LES DONNÉES :

Il faut distinguer 2 branches majeures de la télédétection.

1 — la télédétection orientée vers l'image, s'attache surtout aux aspects picturaux des données et utilise couramment la photointerprétation manuelle comme principale méthode d'analyse. Ce n'est pas à ces méthodes que nous nous intéresserons.

2 — La télédétection orientée vers le numérique, mais il faut savoir que dans ce cas, l'image peut être utile pour visualiser les données ou même illustrer les résultats d'une analyse. Cependant l'image n'est pas le but profond de l'étude.

En quoi ces données sont elles *différentes?*

Elles sont gigantesques. En effet un satellite de télédétection a une résolution au sol qui peut aller

de 2 000 m (SMSI),

à 100 m (LANDSAT),

à 25 m (SEASAT A).

Certains équipements aéroportés peuvent même atteindre moins de 15 m de résolution au sol.

Dans chaque « image », chaque point élémentaire, appelé pixel, peut être caractérisé par un niveau de gris quantifié, allant généralement de 0 à 255. (Représenté en binaire par 8 bits). Une fenêtre de résolution $N \times N$ peut donc prendre 2^{8N2} formes possibles. Pour une fenêtre de 256 lignes et 256 colonnes on obtient 65 536 pixels c'est-à-dire 524 288 bits. Une image Landsat couvre une surface correspondant à un carré de 158 km de côté et contient donc environ 7,5 millions de pixels. Enfin, pour illustrer ces nombres : au centre de Fucino, en Italie, 700 km de bande magnétique ont servi à stocker les données en provenance de Landsat 1 du 31 mars au 20 novembre 1975 (8 mois). Dans le même ordre d'idée, un radar aéroporté à vue oblique peut produire 12 m de film en 6 h en couvrant 75 000 km^2. Pour une résolution au sol de 15×15 m, cela représente 3.3 10^8 pixels par image.

TRAITEMENT DES DONNÉES :

1) *Les transformations*

Avant d'utiliser quelque méthode que ce soit, des transformations des données doivent être faites pour corriger certaines distorsions dues au capteur, à l'atmosphère ou au soleil.

On utilise pricipalement :

— le rapport des signaux dans deux canaux contigus;

— le rapport du signal dans un canal sur la somme de tous les autres;

— le rapport de la somme sur la différence des canaux contigus.

Ces transformations entraînent évidemment la perte d'un canal (bande spectrale) qui peut être importante quand les canaux ne sont pas nombreux. Cependant ces transformations sont souvent insuffisantes pour pallier à des défauts tels que le roulis et le tangage de l'avion transportant le capteur ou la présence d'une masse nuageuse. Nous ne parlerons pas ici des problèmes posés par le recalage des images, nécessaire quand on souhaite faire des traitements multi-temporels.

2) *Méthodes supervisées*

Elles doivent leur nom au fait qu'elles nécessitent une continuelle intervention de l'utilisateur. On les appelle également méthodes de CLASSEMENT (par opposition aux méthodes de CLASSIFICATION).

Elles sont basées sur la théorie de la décision.

Le contexte de chaque pixel est ignoré. On n'utilise que l'information contenue dans la mesure de pixel (ou point élémentaire)

a) *Classement point par point*

On suppose que les mesures spectrales (mesures suivant les différentes longueurs d'onde) correspondent aux composantes d'un vecteur X.

La stratégie de classement consiste à affecter X à la classe C_i ($i = 1,n$) qui minimise la perte moyenne c'est-à-dire qui minimise $L_x(C_i) = \sum_{j=1}^{m} \lambda_{ij}\, p(C_j/x)$

où λ_{ij} = coût résultant du classement dans C_i d'une observation qui est dans C_j.
$p(C_j/X)$ = probabilité (a posteriori) que l'observation X soit dans la classe C_j.
On prend souvent $\lambda_{ij} = 0$ si $i = j$ (pas de perte si le classement est correct)
$\lambda_{ij} = 1$ si $i \neq j$

Il faut donc minimiser.

$L_x(C_i) = 1 - p(C_i/X)$

On retrouve ainsi la méthode du maximum de vraisemblance.

En pratique, la probabilité a posteriori $p(C_j/X)$ peut être difficile à déterminer.

On montre que $p(C_i/X)$ est proportionnel à $p(X/C_i)\,p(C_i)$
où $p(X/C_i)$ = loi de probabilité des observations de la classe C_i
$p(C_i)$ = probabilité a posteriori de la classe C_i.

On obtient alors : $X \in$ classe $C_i \Leftrightarrow p(X/C_i)\,p(C_i) = \max_j p(X/C_j)\,p(C_j)$

On fera souvent la supposition que les lois de probabilité des classes sont des lois gaussiennes multidimensionnelles. Dans ce cas on peut alors explicitement utiliser la méthode du maximum de vraisemblance.

Dans la pratique, il faudra avoir recours à des mesures au sol (d'où le nom de classifications supervisées) afin d'obtenir des estimations non absurdes du vecteurs moyenne et de la matrice des convariances des m classes. Ces estimations sont obtenues à partir d'échantillons, choisis comme représentatifs des classes considérées.

Pourquoi avoir supposé une distribution gaussienne?

Tout d'abord, cette supposition nous permet d'obtenir une règle de classement relativement simple à mettre en œuvre. En effet, les paramètres à estimer pour chaque classe sont en nombre limité. Cela nous permet de limiter le nombre des échantillons nécessaires pour chaque classe (il est évident que ces échantillons sont très coûteux). Mais cette hypothèse est restrictive car beaucoup de classes « naturelles » ont des distributions beaucoup plus complexes.

Appliquer une distribution gaussienne peut amener à des résultats décevants. Il faut pourtant remarquer que des processus qui paraissent sophistiqués peuvent, de façon satisfaisante, être approchés par la réunion d'un petit nombre de processus gaussiens. Cette hypothèse nous donne donc un bon rapport qualité/prix.

Si l'utilisateur de méthodes de classement souhaite obtenir de bons résultats il doit veiller à :

1) adapter le nombre d'échantillons à l'estimation des vecteurs moyenne et de la matrice des covariances de ses classes. Une mauvaise estimation au départ peut être cause de nombreux échecs.

2) des classes multimodales ne peuvent être évidemment pas supposées gaussiennes, donc unimodales; dans ce cas, on peut essayer de découper ces classes en sous-classes unimodales et dont les distributions peuvent être approchées par des lois de densité gaussienne.

b) *Classement par groupes :*

Au lieu de classer point par point tous les éléments d'une « image » on peut souhaiter classer des ensembles de points que l'on sait membres d'une même classe. Ces points peuvent être, par exemple, le résultat d'un algorithme de segmentation destiné à isoler des surfaces relativement homogènes.

Le principal intérêt de cette méthode est, évidemment, d'accélérer le traitement, l'accélération étant d'autant plus forte que la taille moyenne des groupes est grande par rapport au pixel.

On généralise alors la méthode du maximum de vraisemblance en prenant

$$x = (X_1, \dots X_s) \in C_i \iff p(x/C_i) = \max_j p(x/C_j)$$

On peut alors faire comme ci-dessus des suppositions de distribution gaussienne et de transformations monotones.

3) *Méthodes non supervisées ou classification :*

Il existe 2 principales méthodes de classification, les méthodes divisives et les méthodes agglomératives.

Les *méthodes divisives* partent de l'ensemble des données et les partitionnent en un nombre (maximum) de classes.

Les *méthodes agglomératives,* au contraire, regroupent les données l'une après l'autre de façon à former des classes les plus homogènes possible.

Dans les 2 cas, le regroupement ou la division s'appuie sur la définition d'une mesure de similarité (distance euclidienne, distance du X^2...)

Les méthodes agglomératives, du type KMEANS, « nuées dynamiques », « centres mobiles » ou « ISODATA » sont les plus couramment utilisées.

Les méthodes présentent :

— un avantage : elles sont basées sur des algorithmes itératifs et, elles sont donc évolutives;

— un inconvénient : elles nécessitent le choix du nombre de classes que l'on souhaite obtenir (ce choix peut être un peu moins restrictif pour certains algorithmes qui ne demandent qu'un nombre maximum ou imposent un seuil limite de remplissage d'une classe.) Ces algorithmes partent de classes initiales déterminées par leur centre de gravité. Les données sont regroupées, une par une, autour de ces centres, suivant une distance choisie au départ. Ensuite, les points associés à chaque centre initial servent à calculer les nouveaux centres de chaque classe. Ces nouveaux noyaux servent d'entrée à l'itération suivante. Les itérations continuent jusqu'à ce qu'il y ait stabilisation des classes.

Quand on parle de classification, on pense également aux *classifications ascendantes hiérarchiques*. Ces algorithmes sont très peu utilisés en télédétection car ils se heurtent aux problèmes de grands fichiers de données. En dehors des temps de calcul trop longs pour traiter de gros fichiers, il faut obtenir des résultats qui soient rapidement et facilement interprétables.

En effet, quand il s'agit de classer en différentes cultures une image satellite, l'obtention d'une carte des grisés (ou mieux en couleurs) — définissant l'appartenance de chaque point à une classe — est une sortie classique à obtenir à partir des algorithmes agglomératifs. Ce n'est pas le cas des classifications ascendantes hiérarchiques. On pourrait croire que quiconque s'est orienté vers les méthodes divisives refusera les agglomératives ou vice-versa. En effet, si l'on ne s'intéresse pas aux guerres d'école, les « télédétecteurs » utilisent souvent les 2 méthodes séquentiellement.

En particulier :

— on utilise les méthodes divisives pour établir les classes initiales;

— on utilise les méthodes agglomératives à partir de ces classes initiales, afin de les faire évoluer et ainsi, les améliorer.

On remarque immédiatement les dangers que font courir les méthodes de classification. Il est recommandé de ne pas conserver (ou tout du moins de ne pas le faire aveuglement) des classes trop petites qui risquent de n'avoir aucune existence réelle (points aberrants ou inintéressants car trop rares)

4) *Données multitemporelles :*

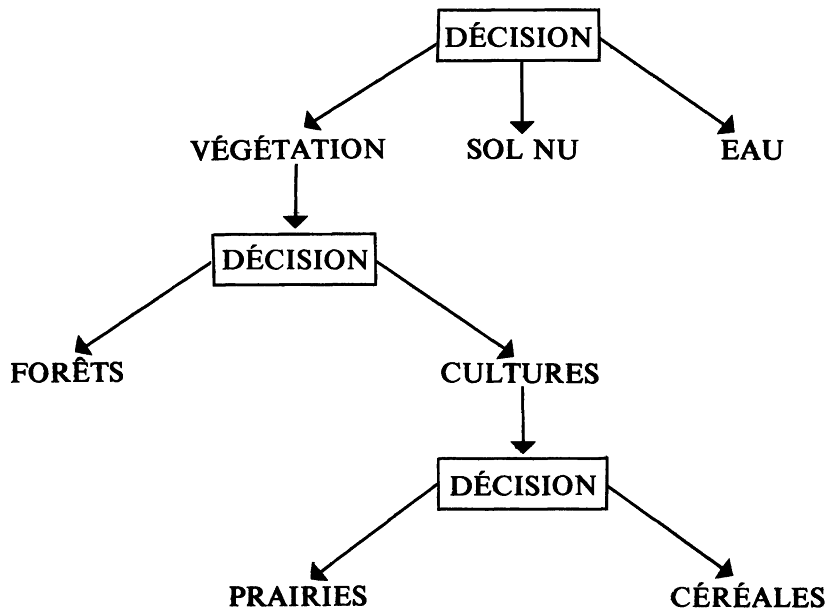
Les satellites Landsat ont amené la notion de répétitivité dans les observations multispectrales. Bien que les satellites n'aient pas des orbites très stables (ce qui pose des problèmes de recalage d'image), on peut cependant utiliser (après traitements numériques) des données recueillies lors de plusieurs passages. L'intérêt « stratégique » de ces scènes multitemporelles est évident (surveillance des cultures tant du point de vue sanitaire que du point de vue prévision, discrimination d'espèces à partir de différence de croissance...)

a) *Concaténation des vecteurs :*

On peut, et c'est une méthode « naturelle », concaténer simplement les vecteurs de données provenant de plusieurs passages et ensuite appliquer une des méthodes décrites ci-dessus. Cette approche directe, qui a déjà fait ses preuves, peut cependant présenter quelques difficultés :

— accroissement considérable des dimensions;
— accroissement des zones tests pour les méthodes de classement.

b) *Arbre de décision :* on peut imaginer qu'une analyse puisse se traduire par un arbre de décision tel que celui qui est dessiné ci-contre. Malheureusement, cet arbre de décision n'est pas facile à programmer



De plus la détermination d'un arbre optimal n'est pas toujours facile à faire surtout dans les cas non banals. Il est en effet impossible d'envisager tous les arbres possibles. Il est souvent nécessaire de restreindre la forme des arbres.

c) *Classement en cascade :*

C'est une extension de la théorie de la décision utilisée pour les méthodes de classement.

Pour 2 temps t_1 et t_2

On note D_k = classes au temps t_1 où $k = 1, m_1$

C_i = classes au temps t_2 où $i = 1, m_2$

X_1 correspond au vecteur X au temps t_1

X_2 correspond au vecteur X au temps t_2

$$X_2 \in C_i \Leftrightarrow \sum_{k=1}^{m_1} p(X_1, X_2/D_k, C_i)\, p(C_i/D_k)\, p(D_k) =$$

$$= \max_j \sum_{k=1}^{m_1} p(X_1/X_2)\, p(D_2/C_j)\, p(C_j/D_k)\, p(D_k)$$

Si on suppose que X_1 et X_2 sont indépendants pour D_k et C_i données alors

$$X_2 = \max_j \sum_{k=1}^{m_1} p(X_1/D_k)\, p(X_2/C_j)\, p(C_j/D_k)\, p(D_k)$$

5) *Réduction des données :*

Il peut être intéressant de réduire le volume des données à transmettre ou à stocker. L'idée de base de la réduction des données est d'éliminer la redondance dans les données sans altérer fondamentalement leur information. Pour les images, des techniques de transformation spectrale sont utilisées. On ne stocke (ou on ne transmet) que les composantes considérées comme intéressantes. Ces transformations peuvent aller jusqu'aux analyses en composantes principales (appelées dans ce cas Harhunen-Loeve) ou aux analyses discriminantes. Elles fournissent des transformations linéaires qui nous permettent d'aboutir à une réduction des dimensions.

Il faut noter que l'analyse en composantes principales en plus du fait qu'elle réduit la redondance linéaire des mesures, peut également éliminer certains bruits aléatoires (en éliminant les composantes de faible variance). L'analyse discriminante permet de tenir compte de l'information sur les classes. Elle maximise la variance interclasse relativement à la variance interclasse. Les méthodes sont couramment utilisées en analyse des données multidimensionnelles.

CONCLUSION

Dans ce texte, nous n'avons pas parlé des problèmes rencontrés par les utilisateurs de la télédétection. En fait les statistiques sont actuellement assez peu employées, en France, pour le traitement des données de télédétection. En effet :

- Les algorithmes ou les méthodes qui ont été mis au point pour le traitement des données ne concernent souvent que de petits fichiers de données. Pour traiter les fichiers très gros de télédétection, les télédétecteurs ont souvent repris des algorithmes ou des programmes déjà existant et les ont adaptés (avec plus ou moins de succès) à leurs problèmes. Certaines transformations ont abouti à des résultats satisfaisants mais d'autres ont fourni des logiciels dont la logique, la performance et même parfois la cohérence sont insuffisants. De plus, dans le souci d'être généraux et universels, les logiciels sont toujours très gros et de ce fait difficiles à utiliser (quand ils ne sont pas utilisés comme de simples « boîtes noires »).

• Les statistiques sont fortement concurrencées par les traitements interactifs avec visualisation sur écran en couleur. Du point de vue traitement, ce sont souvent des combinaisons de couleurs, des seuillages, des lissages... quant à la visualisation, elle est faite sur des écrans-couleur et sert de base à la photointerprétation. Il faut remarquer que les résultats obtenus par ce type de méthodes sont souvent intéressants et permettent de résoudre les problèmes posés.

• Enfin il faut atténuer la dichotomie supervisée — non supervisée. Cette séparation entre 2 types de méthodes, devient de plus en plus artificielle. Quand on améliore un algorithme appartenant au « non supervisé » en utilisant des « vérités-terrains », ne fait-on pas en quelque sorte du « supervisé »?

BIBLIOGRAPHIE

R.M. HARALICK et J.C. SIMON (1980). — *Issues in Digital Image processing Nato advanced study Institute Series* Series E : Applied Science N° 34

J.L. VAN GENDEREN et W.G. COLLINS (1975). — *Remote sensing data processing* (The remote sensing society)

Francis LEVY (1977). — *Le traitement automatisé de l'image. Documentation et recherches Inter-Photothèque* — La documentation française

MODÈLE MIXTE : RISQUE QUADRATIQUE ET SÉLECTION

Bruno GOFFINET

Laboratoire de biométrie, I.N.R.A.-C.R. Toulouse, (1)

On étudie la prédiction du vecteur aléatoire U dans un modèle mixte $Y = X\beta + ZU + E$ vis-à-vis de la fonction de risque quadratique. On montre d'une part les qualités du B.L.U.P. (Best Linear Unbiased Predictor) mais d'autre part sa non admissibilité en exhibant un meilleur prédicteur.

On recherche ensuite les critères optimaux de sélection dans ce même modèle et on montre l'intérêt du classement à partir des estimations B.L.U.P.

We study the forecast of a random vector U in a mixed model : $Y = X\beta + ZU + E$ with respect to the quadratic risk function. We show on the one hand the BLUP (Best Linear Unbiased Predictor) qualities and on the other hand its non admissibility when producing a best predictor.

We find then the optimal selection criteria in this model ans we show the interest of classification from BLUP estimations.

Der Verfasser untersucht die Voraussagen eines zufälligen Vektors U in einem nicht einheitlichen Modell $Y = X\beta + ZU + E$ in Beziehung zu der Funktion des quadratischen Risiko. Man zeigt einerseits die Eigenschaften des B.L.U.P. (Best linear unbiased Predictor), aber andererseits seine Unzulässigkeit, indem man einen besseren Predictor anwendet.

Man schätzt daraufhin die optimalen Kriterien der Selektion in dem gleichen Modell und zeigt das Interesse für eine Klassifikation, indem man von den Schätzungen des B.L.U.P. ausgeht.

Les sélectionneurs animaux et végétaux sont souvent conduits à utiliser des modèles linéaires mixtes du type :

$$Y = X\beta + ZU + E$$

avec X et Z indicatrices fixes et connues relatives au plan d'expérience.

$$U : \begin{pmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_N \end{pmatrix} \sim N(O, D)$$

$$E \sim N(O, \gamma_E I); \ \text{cov} \ (u, E) = 0$$

$$\text{var} \ (Y) = V = ZDZ' + \gamma_E I$$

β fixe inconnu

Nous considérons ici que V est connue.

---

(1) BP 12, 31320 CASTANET TOLOSAN.

Journal de la Société de statistique de Paris, tome 123, n° 2, 1982.

L'objectif de la sélection porte généralement sur le vecteur U; chaque u_i étant une caractéristique (effet génétique additif par exemple) d'un individu i. Le vecteur β est en fait considéré comme perturbateur.

Nous allons étudier deux critères fréquemment utilisés pour faire cette sélection.

I — Le critère quadratique

On cherche le vecteur $\widehat{U} = \begin{pmatrix} \widehat{u}_1 \\ \vdots \\ \widehat{u}_i \\ \vdots \\ \widehat{u}_N \end{pmatrix}$ qui minimise : $\Omega_1 = E\left[\sum_{i=1}^{N} (u_i - \widehat{u}_i)^2\right]$

Si tous les paramètres de la loi conjointe de U et Y sont connus on sait que $\widehat{u}_i = E(u_i/Y)$ minimise Ω_1.

Ici on ne connaît pas l'espérance de Y. Nous sommes donc amenés à limiter la classe des estimateurs dans laquelle nous recherchons le minimum de Ω_1.

a) *Le B.L.U.P.*

— HENDERSON (voir (1)) propose la recherche du « Best Linear Unbiased Predictor » (B.L.U.P.) c'est-à-dire $\widehat{U}^B$ vérifiant

$$\begin{cases} E(\widehat{U}^B) = E(U) = 0 \\ \widehat{U}^B \text{ linéaire en } Y \text{ et} \\ \widehat{U}^B \text{ minimisant } \Omega_1 \end{cases}$$

On obtient : $\widehat{U}^B = DZ'V^{-1}(Y - X\widehat{\beta}_V)$
avec : $\widehat{\beta}_V = (X'V^{-1}X)^- X'V^{-1}Y$ estimateur maximum de vraisemblance de β.

Mais, cet estimateur a d'autres qualités, dans le cas de la multinormalité, qui sont peut-être plus intéressantes pratiquement.

* Dans la classe des estimateurs invariants par rapport à une translation des paramètres fixes $\widehat{u}^B$ minimise Ω_1 (voir (2)).

* Si on note un estimateur $\widehat{u} = f(Y) : \{f_i(Y)\}$, on peut considérer le risque quadratique Ω_1 comme une fonction de β et f.

$$\Omega_1 = R(f, \beta) = E\left(\sum_{i=1}^{N} (u_i - f_i(Y))^2\right)$$

Alors, pour cette fonction de risque R, la fonction $f_B(Y) = \widehat{U}^B$ est minimax

(voir (2)), c'est-à-dire : $\max_\beta R[f_B, \beta] = \min_f \max_\beta R[f, \beta]$

Cette dernière propriété exprime que le B.L.U.P. est « protégé » contre les valeurs « les pires » de β.

b) ***Non admissibilité du B.L.U.P.***

Dans les dispositifs hiérarchisés (classiques en sélection animale où β_j représente l'effet élevage et u_{ij} la valeur génétique additive du mâle i de l'élevage j) du type :

$$Y_{ij} = \beta_j + u_{ij} + E_{ij}$$
$$\beta_j \; j = 1, \ldots, J \text{ effet fixé} \geqslant 4$$
$$U_{ij} \; i = 1, \ldots, n_j \text{ effet aléatoire} \sim N(O, \gamma_U)$$
$$\sum_{j=1}^{J} n_j = N \quad \text{cov}(U_{ij}, U_{i'} j) = 0 \quad i \neq i' \text{ ou } j \neq j'$$
$$E_{ij'} \text{ résiduelle} \sim N(O, \gamma_E) \; \text{cov}(E_{ij}, E_{ij'}) = 0$$
$$\text{pour } i \neq i' \text{ ou } j \neq j'$$

Le B.L.U.P. n'est pas admissible pour la fonction de risque $R[f,\beta] = \Omega_1$

En effet, dans ce cas, $f_B(Y) = \hat{U}^B : \{\hat{u}_{ij}^B\}$ $\hat{u}_{ij}^B = \dfrac{\gamma_u}{\gamma_u + \gamma_E}(Y_{ij} - Y_{.j})$

et soit un nouvel estimateur :

$$\begin{cases} f_S(Y) = \hat{U}^S = \{\hat{u}_{ij}^S\} \text{ où} \\ \hat{u}_{ij}^S = \dfrac{\gamma_u}{\gamma_u + \gamma_E}(Y_{ij} - Y_{.j}) + \dfrac{(J-2)\gamma_u.(Y_{.j})}{\sum_{j=1}^{J}(Y_{.j})^2 n_j} \\ \text{avec } Y_{.j} = \sum_{j=1}^{n_j} Y_{ij}/n_j \end{cases}$$

$fs(Y)$ est tel que (voir (2))

$$R[f_S, \beta] \leqslant R[f_B, \beta] \qquad \forall \beta$$

avec l'inégalité stricte pour certaines valeurs de β.

Ce nouvel estimateur $\hat{U}^S$ conserve évidemment la propriété minimax mais pas celle d'invariance par rapport à une translation des paramètres fixes. Si donc, on s'intéresse au critère quadratique, les arguments en faveur de l'utilisation du B.L.U.P. doivent faire apparaître l'intérêt de l'invariance.

c) *Bayesien empirique*

Dans le cas où $n_j = n \; \forall j$, on peut montrer aussi (voir (2)) que dans le cadre du modèle (2)

$$R[f_{SB}, \beta] \leqslant R[f_B, \beta] \; \forall \beta$$

avec l'inégalité stricte pour certaines valeurs de β.

Avec

$$f_{SB}(Y) = \hat{U}^{SB} : \{\hat{u}_{ij}^{SB}\}$$

$$\hat{u}_{ij}^{SB} = \frac{\gamma_u}{\gamma_u + \gamma_E}(Y_{ij} - Y_{.j}) + \frac{(J-1)\gamma_u.(Y_{.j} - Y_{..})}{n\sum_{j=1}^{J}(Y_{.j} - Y_{..})^2}$$

Considérons maintenant que les effets $\{\beta_j\}$ sont aléatoires $\beta \sim N(\mu, g_\beta I)$ le meilleur estimateur linéaire sans biais devient :

$$\widehat{u}_{ij}^{R} = \frac{\gamma_u}{\gamma_u + \gamma_E}\left((Y_{ij} - Y_{..}) - \frac{\gamma_\beta n\,(Y_{.i} - Y_{..})}{(\gamma_u + \gamma_E) + \gamma_\beta n}\right)$$

Pratiquement on ne connaît pas γ_β; remplaçons dans l'expression de $\widehat{u}_{ij}^{R}$, γ_β par $\widehat{\gamma}_\beta$:

$$\gamma_\beta = \sum_{j=1}^{J} \frac{(Y_{.j} - Y)^2}{J-1} - \frac{\gamma_u}{n} - \frac{\gamma_E}{n}$$

On obtient alors une estimation $\widetilde{\widehat{u}}_{ij}^{R}$ de $\widehat{u}_{ij}^{R}$ et cette estimation coïncide avec $\widehat{u}_{ij}^{SB}$.

Ainsi, pour la fonction de risque Ω_1, l'apport de l'hypothèse « β aléatoire », sans préciser la variance γ_β, n'apporte aucun gain supplémentaire (du moins si on utilise la stratégie classique de remplacement de γ_β par γ_β) dans le cadre du modèle (2).

II — Le critère de sélection

a) *Définition de l'objectif*

On cherche à sélectionner les individus i pour lesquels les valeurs prises par les variables aléatoires u_i seront les plus grandes.

Il y a deux types de formalisation de ce problème qui correspondent à deux types de contrainte.

Nous devons sélectionner les individus i en utilisant les valeurs prises par le vecteur Y.

On définit donc N régions R_i $i = 1,\ldots, N$ telles que si l'observation y du vecteur aléatoire Y appartient à R_i, l'individu i est sélectionné.

Notons $F_i(y)$ les indicatrices qui prennent la valeur 1 si $y \in R_i$, O sinon.

L'objectif est alors de maximiser :

$$\Omega_2 = E\left(\sum_{j-1}^{N} F_i(Y)\, u_i\right)$$

sous l'une des contraintes suivantes :

1) le nombre de retenus est fixé à q quelles que soient les observations y, c'est-à-dire :

$$\sum_{i=1}^{N} F_i(Y) = q \;\forall\, y$$

2) l'espérance du nombre de retenus est fixé (généralisation de la contrainte de COCHRAN (voir (3))

$$E\left[\sum_{i=1}^{N} F_i(Y)\right] = \alpha \qquad (\alpha \text{ n'est pas nécessairement entier}).$$

On obtient alors les stratégies qui maximisent Ω_2.

$$\Omega_2 = E\left(\sum_{i=1}^{N} F_i(Y)\, u_i\right) = E\left[\mathrm{E}\left(\sum_{i=1}^{N} F_i(Y)\; u_i/Y\right)\right]$$

qui conduit à (voir (4)).

Contrainte 1 :

La stratégie optimale consiste à classer les individus selon $E(u_i/Y)$ et à prendre les q premiers.

Contrainte 2 :

La stratégie optimale consiste à prendre tout individu pour lequel $E(u_i/Y) > \lambda$;

λ étant déterminé par la contrainte $E\left(\sum_{i=1}^{N} F_i(Y)\right) = \alpha$

On est donc encore amené à calculer l'espérance conditionnelle ce que l'on ne peut pas faire dans le cadre du modèle (1) si on ignore la valeur de β.

b) *Cas où β est inconnu*

De la même façon que pour le critère quadratique Ω_1, on peut chercher la stratégie de sélection optimale dans la classe des stratégies invariantes par rapport à une translation des paramètres fixes β.

Ce qui veut dire par exemple pour le modèle (2) que le choix des individus u_{ij} sélectionnés est le même quelles que soient les valeurs des effets élevages β_j.

On obtient de la même façon que pour Ω_1 (voir (4)) que le critère optimal de sélection est :

— sous la contrainte 1, classer les individus selon $\widehat{u}_i^B$ et prendre les q premiers;

— sous la contrainte 2, prendre tout individu pour lequel $\widehat{u}_i^B > \lambda$. Si on appelle $F_i'(y)$ l'indicatrice qui prend la valeur 1 lorsque $\widehat{u}_i^B > \lambda$, le calcul de λ se fait avec :

$$\sum_{i=1}^{N} E(F_i'(Y)) = \alpha$$

Mais, la fonction $f^S(Y)$, n'est pas meilleure que $\widehat{u}^B$ pour l'objectif Ω_2, et peut entraîner un gain nettement inférieur pour certaines valeurs de β.

Il serait intéressant, mais l'étude n'a pas été faite ici, de voir ce que peut apporter l'hypothèse supplémentaire de normalité des β_j, sans en préciser les paramètres.

CONCLUSION :

Un certain nombre de sélectionneurs ont pris l'habitude de raisonner en utilisant le risque quadratique Ω_1. Cette attitude est légitime lorsqu'ils désirent connaître l'ensemble des individus étudiés mais devient dangereuse lorsque leur objectif est effectivement de sélectionner une partie de ces individus. En effet, les critères optimaux pour ces deux objectifs ne coïncident pas dans les modèles mixtes.

RÉFÉRENCES BIBLIOGRAPHIQUES

[1] C.R. HENDERSON (1963) : Selection index and expected genetic advance in statistical genetic and plant breeding. NAS. NAC : 141.

[2] B. GOFFINET (1981) : Non admissibilité du B.L.U.P. Document interne.

[3] N.G. COCHRAN (1951) : Improving by means of selection. Proc. Send Berkely Symposium 449-470.

[4] J.M. ELSEN, B. GOFFINET (1981) : Critère optimal de sélection. Quelques résultats généraux. Document interne.

RÉGIONS DE CONFIANCE DANS LE MODÈLE NON-LINÉAIRE

A. MESSEAN

Laboratoire de biométrie, INRA-CNRZ (1)

L'utilisation de plus en plus fréquente de modèles non-linéaires dans la représentation de phénomènes biologiques nous a amené à aborder les problèmes numériques et statistiques de ce type de modèle.

Nous présentons et comparons ici trois méthodes de construction de régions de confiance dans le cadre d'ajustements par moindres carrés non-linéaires : test d'adéquation du modèle, linéarisation du modèle et test du rapport de vraisemblance.

More and more frequently, nonlinear models are used to represent biological systems, and so we are studying the numerical and statistical problems of such models.

Three methods for the construction of confidence regions in nonlinear least squares problems, are presented and compared : lack of fit method, linearization method and likelihood ratio test.

Die immer haüfigere Verwendung von nichtlinearen Modellen für die Darstellung biologischer Phänomene hat uns veranlasst die numerischen und statistischen Probleme dieser Modelle zu untersuchen.

Wir zeigen hier und vergleichen drei Methoden zur Konstruktion von Schätzbereiche, im Rahmen der nichtlineare Kleinste-Quadrat-Anpassungsmethoden : Test der Anpassung des Modells, linearisierung des Modells und Wahrscheinlichkeit-Verhältnis-Test.

De nombreux phénomènes biologiques sont représentés par des modèles non-linéaires (dans lesquels les paramètres inconnus interviennent de façon non-linéaire). Et dans des domaines comme la dynamique de populations ou la cinétique enzymatique, le modèle non-linéaire est devenu un outil courant, depuis que le développement des moyens de calcul a permis de résoudre le problème numérique de l'estimation des paramètres.

Malheureusement il n'existe pas encore de théorie générale contrairement au cas du modèle linéaire. Les développements statistiques actuels produisent essentiellement des résultats asymptotiques qui ne sont pas toujours applicables.

Nous allons aborder ici, une composante importante de l'appréciation de la qualité de l'estimation effectuée : la construction de régions de confiance associées aux estimateurs.

Nous nous limiterons aux méthodes de construction les plus communément utilisées.

Après avoir présenté ces méthodes, nous comparerons leurs comportements sur des aspects numériques et statistiques en nous appuyant sur deux exemples issus du domaine biométrique.

---

(1) 78350 JOUY-EN-JOSAS.

Journal de la Société de statistique de Paris, tome 123, n° 2, 1982.

I — PRINCIPALES MÉTHODES DE CONSTRUCTION DE RÉGIONS DE CONFIANCE.

Nous considérons le modèle de régression non-linéaire suivant :

$$y_i = f(x_i, \theta^*) + \varepsilon_i \qquad i = 1,\dots, n$$

où : — f est une fonction connue

— θ^* est le vecteur des paramètres inconnus de dimension p et appartenant à l'espace paramétrique $\Theta \subset \mathbb{R}^p$,

— $(\varepsilon_i)_{i=1}^n$ est une suite d'erreurs aléatoires indépendantes et normalement distribuées : $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

— $(y_i)_{i=1}^n$ sont les observations effectuées aux points $(x_i)_{i=1}^n$

Nous notons d'autre part :

— $F_\alpha(p_1, p_2)$ la valeur critique au niveau α de la variable de Fisher-Snedecor à p_1 et p_2 degrés de liberté;

— $F(\theta)$ le vecteur $n \times 1$ des valeurs du modèle $\{f(x_1,\theta),\dots, f(x_n,\theta)\}$,

— Y le vecteur $n \times 1$ des observations $\{y_1,\dots, y_n\}$,

— X la matrice $n \times p$ des dérivées premières du modèle évaluées en θ.

$$X_0 = \left\{\frac{\partial F}{\partial \theta}\right\}_0$$

L'estimateur des moindres carrés de θ^* prend la valeur $\widehat{\theta}$ qui minimise

$$S(\theta) = \frac{1}{n}\sum_{i=1}^{n} (y_i - f(x_i, \theta))^2$$

L'estimateur naturel de σ^2 est $\widehat{\sigma}^2 = \frac{1}{n} S(\widehat{\theta})$.

Sous certaines conditions de régularité, ces estimateurs sont consistants

$$(\widehat{\theta} \underset{n\to+\infty}{\longrightarrow} \theta^*, \ \widehat{\sigma}^2 \underset{n\to+\infty}{\longrightarrow} \sigma^2)$$

et sont asymptotiquement normalement distribués (R.I. JENNRICH, 1969; H. BUNKE & W.H. SCHMIDT, 1980).

Remarque : Les résultats sont démontrés sans faire appel à l'hypothèse de normalité des ε_i.

Les trois méthodes que nous allons présenter s'appuient sur un test de l'hypothèse $H_0 : \theta^* = \theta$.

Considérons la famille de tests $\Phi(y, \theta)$: $\mathbb{R}^n x \theta \to \{0,1\}$ définie par

$$\Phi(y, \theta) = \begin{cases} 0 & \text{si l'hypothèse } H_0 \text{ est acceptée} \\ 1 & \text{si l'hypothèse } H_0 \text{ est rejetée} \end{cases}$$

Soit la fonction R qui associe à y l'ensemble R_y des valeurs de θ telles que l'hypothèse H_0 est acceptée :

$$R_y = \{\theta \in \Theta / \Phi(y,\theta) = 0\}$$

$P[\theta \in R_y] = P[\Phi(y,\theta) = 0]$ est la probabilité de recouvrement de θ et sera notée $C(\theta)$. En particulier $C(\theta^*)$ est la probabilité nominale de recouvrement et nous avons la définition suivante :

Définition : Une région de confiance R_y est dite :

- exacte si $C(\theta^*) = 1 - \alpha$
- asymptotique si $\lim_{n\to+\infty} C(\theta^*) = 1 - \alpha$

• approximative si $C(\theta^*) \simeq 1 - \alpha$
(où α est le niveau du test choisi).

La comparaison entre différentes régions, c'est-à-dire entre différentes familles de tests, porte d'abord sur la probabilité nominale de recouvrement. Puis entre deux régions « admissibles » nous choisirons la région de volume minimal V,

où $V = \int_\Theta C(\theta)\, d\theta$.

(1) *Test d'adéquation du modèle.*

Soit l'hypothèse $H'_o : (y_i - f(x_i, \theta))_{i=1}^{n}$ constitue un échantillon gaussien de moyenne 0 et de variance σ^2.

Soit la région de $\mathbb{R}^n : A_\theta = \{y \in \mathbb{R}^n / H'_o$ acceptée$\}$ (région d'acceptation). Un test de l'hypothèse H'_o constitue un test de $H_0 : \theta^* = \theta$ et nous avons la correspondance $Ry = \{\theta \in \Theta \,/\, y \in A_\theta\}$.

Le test de H'_o choisi s'appuie sur le résultat suivant.

Si P_θ est une matrice idempotente $n \times n$ de rang p alors P_θ et $I - P_\theta$ représentent deux projecteurs orthogonaux et la statistique

$$E(\theta) = (n-p/p)\,[Y - F(\theta)]'.\, P_\theta[Y - F(\theta)] \,/\, [Y - F(\theta)]'[I - P_\theta][Y - F(\theta)]$$

suit sous H'_o (donc sous H_0) une loi de Fisher-Snedecor $F(p,\ n-p)$ (théorème de COCHRAN; SCHEFFE, 1959).

La famille de tests est alors définie par

$$\Phi(y, \theta) = \begin{cases} 0 \text{ si } E(\theta) \leqslant F_\alpha(p, n-p) \\ 1 \text{ si } E(\theta) > F_\alpha(p, n-p) \end{cases}$$

Il importe de bien choisir le projecteur P_θ et un choix naturel est $P_\theta = X_\theta (X'_\theta X_\theta)^{-1} X'_\theta$

Nous avons alors les caractéristiques suivantes :

— la région est exacte, $C(\theta^*) = 1 - \alpha$

— tous les extremums locaux (en particulier l'estimation $\widehat{\theta}$ des moindres carrés) sont contenus dans cette région.

Remarque : il existe de nombreux procédés s'appuyant sur le même test, mais avec un choix différent pour P_θ (HARTLEY, 1964).

(2) *Linéarisation du modèle.*

L'estimateur des moindres carrés $\widehat{\theta}$ suit asymptotiquement une loi normale de moyenne θ^* et de matrice de variance-covariance $\sigma^2 (X'_{\theta^*} X_{\theta^*})^{-1}$, cette dernière étant estimée de manière consistante par $(S(\widehat{\theta})/n - p)\,(X'_{\widehat{\theta}} X_{\widehat{\theta}})^{-1}$

Alors la statistique $L(\theta) = [(\widehat{\theta} - \theta)' X'_{\widehat{\theta}} X_{\widehat{\theta}} (\widehat{\theta} - \theta)/S(\widehat{\theta})]\,(n - p/p)$ suit asymptotiquement sous l'hypothèse H_0 une loi de Fisher-Snedecor à p et $n-p$ degrés de liberté. (GALLANT, 1975 a).

La région de confiance ainsi bâtie possède les caractéristiques suivantes :

— c'est un ellipsoïde centré sur $\widehat{\theta}$

— la région est asymptotique $C(\theta^*) \underset{n \to +\infty}{\longrightarrow} 1 - \alpha$

Remarque : La procédure est analogue à celle utilisée dans le cadre du modèle linéaire, en utilisant ici une linéarisation de $F(\theta)$ autour du point-solution $F(\widehat{\theta})$:

$$F(\theta) = F(\widehat{\theta}) + X_{\widehat{\theta}}(\theta - \widehat{\theta})$$
$$Y = F(\theta^*) + \varepsilon = F(\widehat{\theta}) + X_{\widehat{\theta}}\theta^* - X_{\widehat{\theta}}\widehat{\theta} + \varepsilon$$

• En posant $Z = Y - F(\hat{\theta}) + X_{\hat{\theta}}\hat{\theta}$, nous obtenons le modèle linéaire associé $Z = X_0\beta + \varepsilon$.

(3) *Isocontours de vraisemblance.*

La dernière méthode s'appuie sur le test du rapport de vraisemblance, en utilisant la statistique $V(\theta) = S(\theta) / S(\hat{\theta})$.

Sous l'hypothèse H_0, $(n\text{-}p/p)(V(\theta) - 1)$ possède une distribution qui peut être approximée par celle d'une loi de Fisher à p et $n\text{-}p$ degrés de liberté. (GALLANT, 1975 b).

La famille de tests est alors définie par :

$$\Phi(y, \theta) = \begin{cases} 0 \text{ si } V(\theta) \leqslant c \\ 1 \text{ si } V(\theta) > c \end{cases}$$

où $c = 1 + (p/n\text{-}p)\, F_\alpha\,(\mathrm{p}, \mathrm{n\text{-}p})$

Les principales caractéristiques de cette région sont les suivantes :

— elle est asymptotique $C(\theta^*) \underset{n \to +\infty}{\longrightarrow} 1 - \alpha$

— elle coïncide avec un isocontour de vraisemblance dans le cas gaussien (sinon avec un isocontour du critère $S(\theta)$);

— elle contient les extremums locaux satisfaisant $V(\theta) \leq c$ (en particulier l'estimation des moindres carrés).

II — Comparaison des méthodes de construction

Avant d'aborder les problèmes de construction numérique et de qualité statistique, nous allons présenter une interprétation graphique des différents procédés.

(1) *Interprétation graphique :*

Afin d'illustrer graphiquement la construction d'une région de confiance, nous nous plaçons dans l'espace des observations $\mathbb{R}^n$

L'ensemble des points $F(\theta)$, quand θ varie dans Θ, représente la surface de réponse du modèle, variété dans $\mathbb{R}^n$ de dimension p.

Le vecteur Y est matérialisé par un point dans l'espace $\mathbb{R}^n$.

L'estimation des moindres carrés consiste à trouver $\hat{\theta}$ tel que $F(\hat{\theta})$ soit la projection de Y sur la surface de réponse.

Le plan tangent en $F(\hat{\theta})$ est la variété définie par $T(\theta) = F(\hat{\theta}) + X_\theta(\theta - \hat{\theta})$.

Les figures II-1, II-2 et II-3 illustrent les différentes constructions de régions de confiance dans le cas particulier $n = 2$ et $p = 1$.

(2) *Construction numérique.*

Entre deux régions « admissibles » (c'est-à-dire pour lesquelles $C(\theta^*) = 1\text{-}\alpha$) et de volumes comparables, nous choisirons, bien sûr, la plus facile à construire.

Le classement par ordre de difficulté de construction croissante est le suivant :

1. Linéarisation du modèle : il suffit de calculer $X'X$ et nous pouvons construire directement l'ellipsoïde de confiance.

2. Isocontours de vraisemblance : après avoir divisé l'espace paramétrique selon un réseau de points, il faut calculer $S(\theta)$ en chacun de ces points, puis tracer le contour critique.

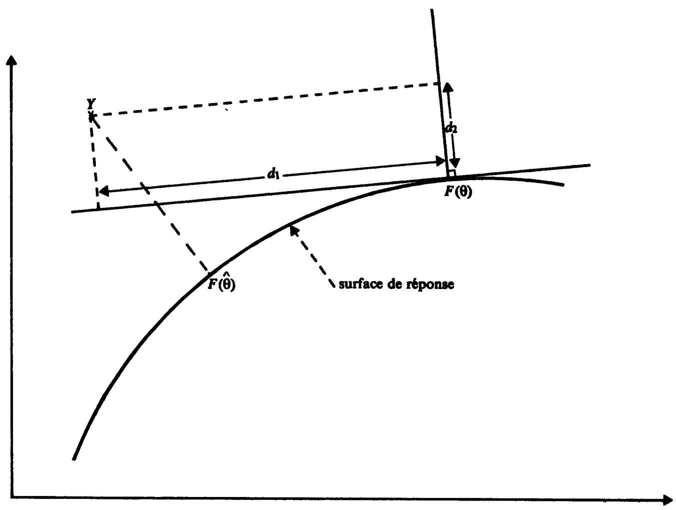


FIGURE II-1

Test d'adéquation du modèle : La région contient l'ensemble des θ tels que

$$d_1^2/d_2^2 \leqslant (p/n-p)\ F_\alpha(p,n-p)$$

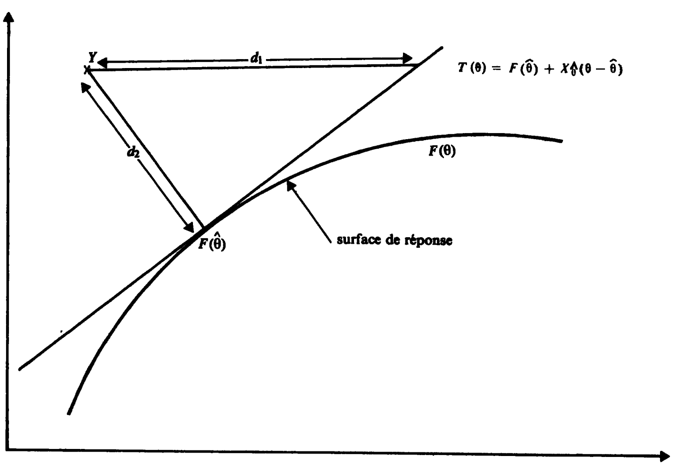


FIGURE II-2.

Linéarisation du modèle : la région regroupe les θ tels que

$$d_1^2/d_2^2 \leqslant (p/n-p)\ F_\alpha(p,n-p)$$

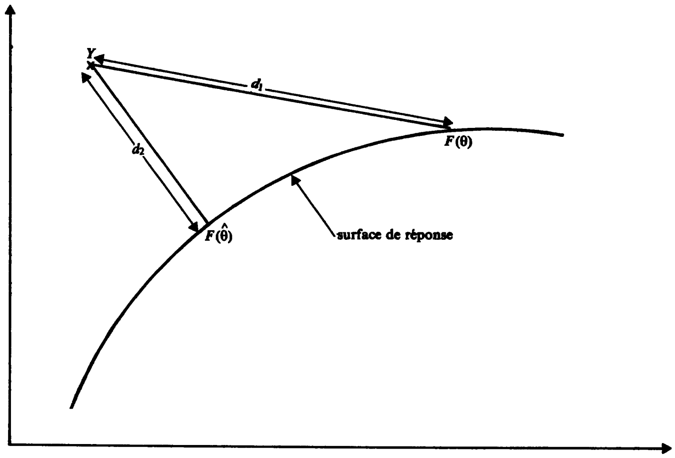


FIGURE II-3.

Isocontours de vraisemblance : la région contient tous les θ tels que

$$d_1^2/d_2^2 \leqslant 1 + (p/n - p)\ F_\alpha(p, n - p)$$

3. Test d'adéquation du modèle : pour chacun des points du réseau précédent, il faut calculer $P_\theta = x_\theta\,(x'_\theta\,x_\theta)^{-1}\,x'_\theta$ puis le numérateur et le dénominateur de E (θ) avant de pouvoir tracer le contour critique.

Cette dernière se révèle spécialement lourde à manipuler en pratique.

(3) *Probabilité nominale de recouvrement.*

Si le modèle $f(x,\ \theta)$ est linéaire, les trois procédés produisent des régions de confiance identiques, les ellipsoïdes classiques du modèle linéaire; en effet les statistiques $E(\theta)$, $L(\theta)$ et $V(\theta)$-1 sont alors égales.

Les méthodes vont donc se différencier en fonction de la non-linéarité du modèle.

Divers auteurs (BEALE, 1960; BATES & WATTS, 1980) se sont attachés à définir des mesures de non-linéarité, estimant l'écart, autour de$F(\hat{\theta})$, entre le modèle $F(\theta)$ et son approximation linéaire $F(\hat{\theta}) + X_0(\theta - \hat{\theta})$. D. BATES et D. WATTS proposent deux mesures de non-linéarité :

— une composante intrinsèque, ne dépendant que du modèle et du plan d'expérience (invariante par transformation des paramètres);

— une composante paramétrique, représentant l'influence du système paramétrique.

Nous avons alors le résultat suivant (MESSEAN, 1982) :

— les isocontours de vraisemblance ont une probabilité nominale de recouvrement (PNR) qui ne dépend que de la composante intrinsèque.

— les ellipsoïdes de confiance ont une PNR qui dépend des deux composantes à la fois.

Quant à la PNR des régions bâties sur le test d'adéquation du modèle, elle ne dépend évidemment pas de la non-linéarité.

Ce résultat est très important puisque l'expérience montre que la composante intrinsèque est le plus souvent négligeable (BATES & WATTS, 1980).

Il apparaît donc que les isocontours de vraisemblance réalisent un compromis intéressant entre :

— la probabilité nominale de recouvrement.
— la facilité de construction numérique.
— la coïncidence avec les contours de vraisemblance.

III — Exemples.

Afin d'illustrer le chapitre précédent, nous avons effectué une comparaison des trois méthodes au niveau de la probabilité nominale de recouvrement. Cette comparaison s'appuie sur deux modèles (logistique et Michaelis-Menten). Le calcul de la probabilité nominale est effectué par simulation : pour chaque modèle nous générons 5 000 échantillons d'observations, ce qui nous donne 5 000 estimations des moindres carrés $\hat{\theta}$. Pour chaque type de région, nous calculons le nombre de fois où θ^* (la vraie valeur) appartient à la région associée à $\hat{\theta}$. Les modèles retenus sont les suivants :

1. *Modèle logistique :*

$$f(x,\theta) = 1 \;/(1 \;+ \exp - \theta_1 (x - \theta_2))$$
$$(x_i)_{i=1}^6 = 0.1,\ 0.3,\ 0.4,\ 0.6,\ 0.7,\ 0.9$$
$$\theta^* = (4.0,\ 0.45)$$
$$\sigma = 0.01.$$

2. *Modèle de Michaelis-Menten.*

$$f(x,\theta) = \theta_1 x/(\theta_2 + x)$$
$$(x_i)_{i=1}^6 = 2.0,\ 0.66,\ 0.4,\ 0.3,\ 0.225,\ 0.2$$
$$\theta^* = (0.1,\ 2.0)$$
$$\sigma = 0.01.$$

Le tableau III.1. nous donne, pour le modèle logistique et le modèle Michaelis-Menten, les probabilités nominales calculées pour un niveau α de 5 % ainsi que les mesures de non-linéarité, Γ^N et Γ^T, proposées par BATES et WATTS (1980).

Tableau III-1.

Probabilités nominales de recouvrement et mesures de non-linéarité.

Modèle	Probabilités nominales de recouvrement			Mesures de non-linéarité	
	Adéquation du modèle	Linéarisation du modèle	Isocontours de vraisemblance	Γ^N	Γ^T
Logistique	0.9512	0.9510	0.9510	0.0152	0.0352
Michaelis-Menten	0.9544	0.8730	0.9534	0.221	2.410

D'autre part les figures III.2. et III.3. représentent pour l'un des échantillons simulés, les régions de confiance effectivement obtenues.

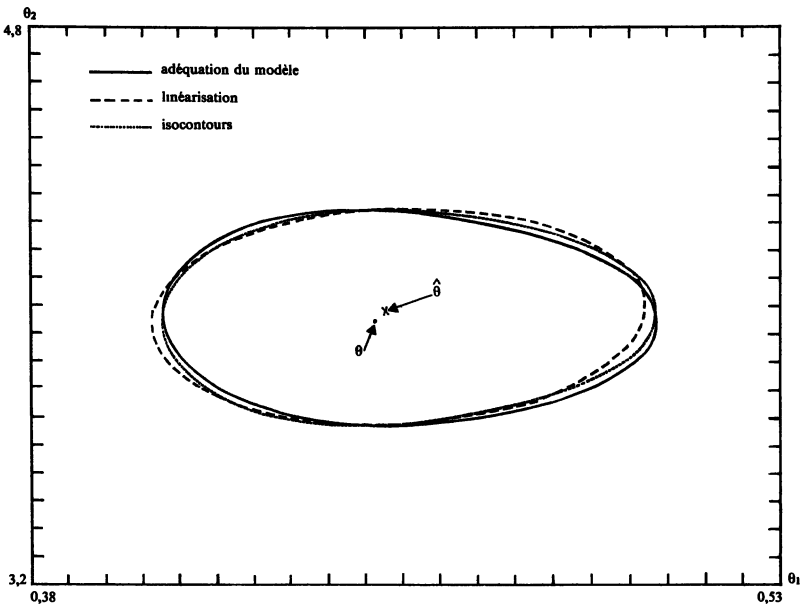


FIGURE III.2.

Régions de confiance à 95 % dans le cadre du modèle logistique pour l'échantillon suivant :
y = *0,171; 0,362; 0,456; 0,649; 0,727; 0,845*

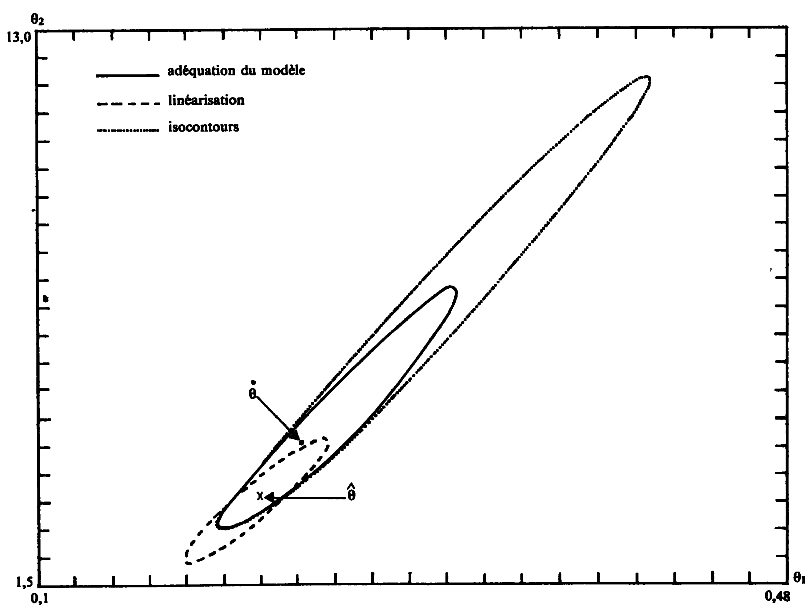


FIGURE III.3.

Régions de confiance à 95 % dans le cadre du modèle de Michaelis-Menten pour l'échantillon suivant : y = *0,0484; 0,0317; 0,0266; 0,0107; 0,0128; 0,0077*

L'examen du tableau III.1 et des figures appelle les commentaires suivants :

— Lorsque la non-linéarité est faible (cas du modèle logistique) les trois méthodes sont équivalentes, conservent la probabilité nominale 1-α (l'écart-type σ_S de la simulation étant égal à 0.003) et produisent des régions pratiquement identiques.

— Lorsque la non-linéarité paramétrique est forte (modèle de Michaelis-Menten) la méthode de linéarisation du modèle est désastreuse tandis que les isocontours conservent la probabilité 1-α.

Remarque : Pour certains échantillons, dans le cas du modèle de Michaelis-Menten, la région produite par les isocontours de vraisemblance, n'est pas fermée; en effet si la probabilité nominale n'est pas affectée par la paramétrisation, sa forme dépend de celle-ci.

CONCLUSION.

Cette étude montre que la méthode des isocontours réalise un excellent compromis entre la conservation de la probabilité nominale de recouvrement et la facilité de construction. De plus la région de confiance produite coïncide avec un isocontour de vraisemblance (dans le cas gaussien) ce qui est statistiquement satisfaisant.

La relative difficulté de construction de ces régions peut être surmontée en exploitant les possibilités de la reparamétrisation; malheureusement les transformations bi-univoques ne s'appliquent pas à toutes les classes de modèles. (BATES & WATTS, 1981).

D'autre part, il faut signaler qu'il existe d'autres méthodes de construction de régions de confiance : JACKKNIFE (DUNCAN, 1980), meilleure approximation de la matrice des variances-covariances des estimateurs (CLARKE, 1980).

RÉFÉRENCES BIBLIOGRAPHIQUES

BATES D.M. & WATTS D.G. (1980) : Relative curvature measures of nonlinearity. *J.R.S.S, B 42,* p. 1-25.

BATES D.M. & WATTS D.G. (1981) : Parameter transformations for improved approximate confidence regions in nonlinear least squares. *The Annals of statistics,* 9, 6, p. 1152-1167.

BEALE E.M.L. (1960) Confidence regions in nonlinear estimation. *J.R.S.S., B 22,* p. 41-88.

BUNKE H. & SCHMIDT W.H. (1980) : Asymptotic results on nonlinear approximation of regression functions and weighted least square. *Mathematische Operationforschung und statistics, Series Statistics,* 11, p. 1-22.

CLARKE G.P.Y. (1980) : Moments of the least squares estimators in a nonlinear regression model. *J.R.S.S., B 42,* p. 227-237.

DUNGAN G. (1978) : An empirical study of jackknife-constructed confidence regions in nonlinear regression. *Technometrics, 20,* p. 123-129.

GALLANT A.R. (1975 a) : *Inference for nonlinear models.* Institute of statistics, Mimeograph series n° 875.

GALLANT A.R. (1975 b) : The power of the likelihood ratio test of location in nonlinear regression models. *J.A.S.A., 70,* p. 199-203.

HARTLEY H.O. (1964) : Exact confidence regions for the parameters in nonlinear regression laws. *Biometrika, 51,* p. 347-353.

JENNRICH R.I. (1969) : Asymptotic properties of non-linear least squares estimators. *Annals of mathematical statistics,* 40, p. 633-643.
MESSEAN A. (1982) : *Contribution à la statistique du modèle non-linéaire.* Thèse de doctorat-ingénieur ORSAY (à paraître).
SCHEFFE H. (1959) : *The analysis of variance.* Wiley.