

MATHEMATIQUES ET SCIENCES DE LA VIE

par

R. Tomassone

I. QUELLES MATHÉMATIQUES POUR LES BIOLOGISTES ? COMMENT ET QUAND LES ENSEIGNER ?

par R. Tomassone, Institut National Agronomique

L'utilisation des mathématiques en biologie soulève au moins deux problèmes spécifiques : la formation des biologistes au bon emploi des mathématiques et la nature de l'intervention des mathématiciens dans un travail de biologie. Elle soulève la question de l'équilibre entre l'utilité du résultat biologique et la qualité du travail mathématique.

1. La situation

Comme les physiciens au début du XX^{ème} siècle, les biologistes ont de plus en plus besoin de mathématiques, que ce soit pour un travail de recherche ou un travail d'ingénieur.

En France, les biologistes sont généralement des universitaires, des médecins ou des agronomes ; leurs domaines de travail s'étendent des disciplines fondamentales aux industries pharmaceutiques ou agro-alimentaires, en passant par la gestion de l'environnement. Leur formation initiale, à de très rares exceptions, est axée sur la biologie : s'ils ont choisi d'étudier la biologie c'est bien sûr par goût, mais aussi, et souvent simultanément, pour ne pas faire de mathématiques. Cette situation générale n'exclut naturellement pas les cas particuliers : de même qu'il est possible de citer l'existence de vieux ivrognes nonagénaires en bonne santé, il est toujours possible de citer le nom

d'un biologiste remarquable particulièrement compétent dans les domaines mathématiques les plus variés. Mais l'existence d'un vieil orme vigoureux n'empêche pas le dépérissement général de tous les ormes ! La situation mondiale, avec vraisemblablement des variations locales, n'est pas à notre connaissance différente de celle que nous connaissons en France.

Nous arrivons donc à la constatation assez générale, que si bon nombre de biologistes, au cours de leur carrière professionnelle, reconnaissent l'impérieux besoin de maîtriser certains outils mathématiques, ils n'ont pas reçu dans leur formation initiale les moyens matériels et intellectuels pour y parvenir.

2. Quelles mathématiques ?

A cette question, nous ne savons pas donner de réponse satisfaisante, si nous ne précisons pas davantage non seulement le domaine de la biologie, mais aussi l'application dans ce domaine : aucune méthode mathématique n'est a *priori* exclue, même si nous savons par expérience que certaines ont déjà fait leurs preuves. Ainsi, par nature et par tradition, les méthodes statistiques ont un rôle central : la nécessité de prendre en compte, et si possible de maîtriser l'aléatoire justifie cette réalité incontournable.

Mais la connaissance de la géométrie s'avère de plus en plus indispensable ; le développement de techniques graphiques sur ordinateur est une raison simple. Mais ce n'est que récemment que l'on s'est rendu compte que le fait de travailler, en statistique, sur un espace de paramètres ayant une structure géométrique propre conduisait à un rapprochement de deux domaines mathématiques jusqu'alors distincts. Cette constatation n'est pas sans conséquence sur les utilisateurs des statistiques habitués à une toute autre approche.

De même, la connaissance théorique et la maîtrise technique des systèmes différentiels sont nécessaires pour toute étude de systèmes dynamiques, qu'elle soit faite au niveau cellulaire pour mieux comprendre les paramètres physiologiques agissant sur la production laitière d'une vache, ou au niveau d'un écosystème prairial pour mieux essayer d'en contrôler l'évolution. Mais la connaissance de ce domaine fait appel à des champs mathématiques si variés qu'il est rare qu'un mathématicien lui-même puisse bien les maîtriser tous : analyse numérique, planification expérimentale, simulation, modélisation stochastique. Il serait donc malvenu, ici, de faire un quelconque

reproche à des biologistes. La forte non linéarité des relations crée des problèmes complexes au niveau mathématique lui-même.

Enfin des champs nouveaux s'ouvrent régulièrement aux applications ; il y a quelques années la théorie des catastrophes a montré la voie ; celle des fractals constitue un exemple plus récent. Même si ces champs n'offrent quelquefois qu'une approche qualitative, ou que très partiellement quantitative, de certains phénomènes vivants, ils n'en constituent pas moins les uniques moyens de formalisation actuellement disponibles pour décrire et tenter de comprendre ces phénomènes.

Tout le monde sera d'accord pour accepter l'idée que les domaines que nous venons de citer sont potentiellement importants pour les biologistes. Ceci n'exclut pas que d'autres domaines puissent l'être ou le devenir, mais sans doute actuellement avec une moindre importance.

3. Un travail à deux !

Après cette évocation des mathématiques à connaître et à maîtriser, il est compréhensible qu'un biologiste soit rarement capable d'être simultanément excellent dans son domaine et dans plusieurs domaines mathématiques. Certes, il peut très souvent - et grâce à de timides évolutions dans l'enseignement, les progrès sont réels depuis quelques années - bien utiliser un outil particulier comme les statistiques. Mais, ce qui est humainement impossible réside dans la difficulté à faire un diagnostic raisonnable pour *délimiter un autre domaine mathématique* qui puisse lui être utile.

Parallèlement, il est difficile au mathématicien de s'investir totalement dans plusieurs domaines biologiques. Pour une application particulière, il lui faut posséder un minimum de connaissances de base et connaître les difficultés expérimentales qui lui sont inhérentes. S'il ne fait pas un effort minimal de compréhension, il restera "le mathématicien", dans le pire des cas "*le calculateur*". Ceci peut suffire pour une application, mais ce ne sera que rarement une incitation suffisante pour poursuivre une collaboration à long terme. Il faut, en effet, que le mathématicien puisse compenser l'éventuelle "médiocrité" de son apport théorique propre par "l'utilité" pratique du résultat auquel celui-ci peut permettre d'aboutir.

Il s'ensuit donc que tout travail non routinier implique une collaboration régulière du biologiste et du mathématicien ; mais il faut une plage commune de compétences, et chacun doit aller vers l'autre.

a - Le biologiste vers le mathématicien

Pour évaluer l'intérêt d'un nouvel outil, le biologiste devra pouvoir lire, pour les comprendre, des mathématiques éventuellement récentes. Comment le préparer à cette lecture ? Raisonnons sur un schéma particulièrement simple ; toute étude biologique peut être décrite par un processus en quatre étapes, où l'importance des connaissances mathématiques est variable :

1/ A partir de connaissances a priori, définition d'un modèle du sujet étudié. Ce modèle implique une parfaite définition de ses conditions d'application ; en particulier on écrira souvent :

OBSERVATION = PARTIE CONTROLÉE + PARTIE ALÉATOIRE

ou plus formellement : $y_i = f[a; x_i] + e_i$

où y_i représente un vecteur d'observations,
 $f[;,]$ une fonction analytique connue,
 a un vecteur de paramètres,
 x_i un vecteur de variables contrôlées,
 e_i un vecteur aléatoire.

L'indice i indique que ce modèle devra être vrai pour tout ensemble d'observations $\{y_i, x_i\}$. Il est alors important de préciser les suppositions portant sur ces différents termes : définition de la forme analytique particulière de $f[;,]$, définition de la nature de la distribution du terme aléatoire. Ainsi, un modèle aussi simple que celui de la régression linéaire simple s'écrit :

$$y_i = a_1 + a_2 x_i + e_i$$

La forme analytique est linéaire en fonction des paramètres $a = [a_1, a_2]^t$; la distribution du terme aléatoire est souvent Normale.

2/ Une expérience représente une réalisation du modèle.

3/ Une méthode mathématique permet de calculer, on dit estimer, les valeurs optimales \hat{a}_1 et \hat{a}_2 des paramètres et les caractéristiques de la distribution des termes aléatoires.

4/ L'étape ultime, dite de validation, consiste à s'assurer que les suppositions sur lesquelles est fondé le modèle sont acceptables au vu des données de l'expérience. S'il n'en est pas ainsi, il est indispensable de recommencer ce processus.

Généralement, de la part du biologiste, les quatre étapes du processus précédent ne demandent *pas le même type de réflexion* :

- la première est une phase de formalisation indispensable qui l'oblige à définir deux champs différents de connaissances, celui qu'il connaît et celui qu'il veut découvrir.

- la seconde est essentiellement expérimentale, elle lui demande de mettre en oeuvre des techniques propres à sa discipline.

- la troisième a souvent caché les problèmes les plus importants auxquels il doit réfléchir ; essentiellement fondée sur le calcul numérique, dans son mode habituel de pensée elle constitue un *masque* pour les autres. L'informatique y joue souvent un rôle si essentiel qu'elle conduit généralement, dans son esprit, à une *confusion* entre la réflexion (= le modèle) et la réalisation concrète (= le traitement des données).

- la quatrième revêt deux formes l'une expérimentale (il refait une seconde expérience indépendante pour confirmer les premiers résultats), l'autre mathématique grâce à des techniques récentes d'auto-validation de résultats expérimentaux.

Autant il est important que le biologiste soit maître des deux premières phases et de l'éventuelle partie expérimentale de la dernière, autant ses connaissances sur la troisième peuvent être succinctes, et sûrement peu techniques. Par exemple, il doit savoir exploiter les résultats d'un calcul, mais il est souvent inutile qu'il sache le réaliser matériellement.

b - Le mathématicien vers le **biologiste**

Il est essentiel que le mathématicien *n'adopte pas une attitude dominante* ; s'il veut jouer un rôle, ce qui nous semble essentiel, il doit d'abord accepter de ne pas parler de ses propres outils, et s'imprégner de la problématique du biologiste et de ses objectifs. Ceci ne signifie pas qu'il ne doive pas les critiquer, ce peut même être une partie importante de son apport. Mais sa critique doit, en premier lieu, porter sur la cohérence interne du modèle par rapport aux objectifs. C'est là qu'il peut suggérer l'emploi d'autres approches, ce qu'il est seul à pouvoir proposer car sa connaissance des mathématiques lui permet de le faire. Une fois l'expérience faite, il doit quelquefois s'assurer que le biologiste ne tire pas davantage d'information des résultats mathématiques qu'ils n'en recèlent en réalité. Ce supplément

d'informations peut, bien sûr, se révéler utile pour reconstituer un nouveau stock de connaissances, nécessaire pour bâtir une nouvelle recherche.

Mais tout au long de son dialogue, il doit être en éveil permanent ; c'est au cours de ce dialogue que sa capacité d'imagination peut lui suggérer de nouveaux axes de recherche mathématique. Ces axes pourront alors être développés avec toute la cohérence interne propre aux mathématiques ; mais, au courant de la problématique du biologiste et de ses contraintes expérimentales, il pourra construire des *schémas réalistes et applicables*.

Naturellement, le succès n'est pas garanti. Ceci peut même être préjudiciable à sa propre carrière au sein de la communauté des mathématiciens : ceux qui pourront le juger ultérieurement, les mathématiciens, n'auront pas fait le même parcours "intellectuel" que lui. Nous avons donc décelé un problème crucial qui conduit actuellement à une impasse : trop souvent les bons mathématiciens ne travaillent pas avec des biologistes, car *le travail de ces derniers ne peut pas être valorisé dans la communauté des mathématiciens*. Seuls s'orientent vers la biologie, nous l'avons dit, ceux qui n'ont pas les capacités de faire de "bonnes mathématiques". A cette constatation s'ajoute le fait qu'une bonne mathématique appliquée n'est pas le fait d'une personne seule, mais qu'elle est le résultat d'un travail d'équipe. Le critère de jugement constitué par une thèse, travail individualiste par excellence, est évidemment soumis à cet aléa. En outre, une bonne thèse en mathématiques appliquées est aussi bien souvent le résultat d'un assemblage judicieux de différentes techniques mathématiques, et non pas la découverte d'une nouvelle méthode entièrement originale. Les mathématiciens sont-ils, dans leur ensemble, décidés à l'admettre ?

4. Bilan provisoire

Pour le biologiste une amélioration importante passe par la formation, qui doit pouvoir se faire tout au long de sa vie professionnelle. Doit-elle être faite par des biologistes ou par des mathématiciens ? Nous n'avons pas de réponse absolue ; les deux cas de figure peuvent très bien coexister. Il serait bon toutefois que les *manuels de référence* indispensables pour l'enseignement soient réalisés par de très bons mathématiciens, capables de se faire comprendre par des biologistes, et peut-être en collaboration avec eux. Leur existence dans la *langue maternelle* du biologiste, française pour ce qui nous concerne, est capitale ; un très grand effort doit être fait

dans ce domaine. Mais il n'est pas certain que la logique valable pour l'apprentissage d'un mathématicien le soit pour un biologiste. Ainsi, s'il est naturel que l'apprentissage de la statistique commence, pour un statisticien, par la maîtrise de la théorie des probabilités, il n'en est pas toujours de même pour un biologiste. Grâce à l'informatique, il est possible de "simuler" de nombreuses réalisations expérimentales, à partir desquelles la pratique statistique peut s'acquérir. Ceci ne supprime pas un retour ultérieur vers un enseignement plus "logique", mais ce retour correspond alors à un besoin du biologiste, et non plus à un dogme établi indépendamment de lui.

Pour le mathématicien, il est capital que la mathématique appliquée soit reconnue non comme un pis aller destiné à ceux qui ne peuvent pas suivre la voie royale des "vraies mathématiques", mais comme une branche qui doit attirer les meilleurs. Il en va de l'intérêt des biologistes et de la survie des mathématiciens qui ne doivent pas rester solitaires ; leur isolement n'est profitable à aucune des deux communautés scientifiques.

En conclusion nous pensons que, dans les relations entre mathématiques et sciences de la vie, nous devons faire cohabiter deux types de personnes ayant un objectif commun ; cette indispensable cohabitation implique deux constatations importantes :

- la formation des biologistes aux mathématiques ne suit pas obligatoirement des schémas classiques, en particulier ceux de la formation des mathématiciens.

- le jugement de la qualité du travail des mathématiciens dans ce cadre peut difficilement être fait selon les mêmes critères que celui des mathématiciens entre eux.

II. A PROPOS DU CONTINUUM STATISTIQUE-MODELISATION EN ECOLOGIE

par Jean-Dominique Lebreton, C.E.F.E./C.N.R.S

1. Introduction

Le caractère de plus en plus quantitatif des sciences expérimentales se traduit en Biologie par une "mathématisation" croissante, selon deux courants *a priori* distincts.

Le premier, déjà ancien, consiste en l'emploi généralisé des méthodes statistiques pour traiter des données expérimentales, quasi-expérimentales, ou d'observation. Le second - dont la généralisation au moins est récente - consiste en un développement marqué de la modélisation. Le but de cette note est d'illustrer, à partir d'une expérience d'enseignant, de chercheur et de consultant dans des groupes de biomathématique à Lyon puis à Montpellier, comment statistique et modélisation cohabitent pour les biologistes, à l'aide d'exemples pris en écologie (au sens large, voir par exemple CALOW, 1987). Nous discuterons également les limites de cette dichotomie et le rôle que sont amenés à jouer les biomathématiciens.

Du fait même que l'essor progressif des mathématiques en Biologie est nettement dessiné, je n'hésiterai pas à présenter, au risque de paraître négatif, ce qui me semble être les difficultés du moment, et les remèdes que l'on peut envisager d'y apporter, sans revenir sur divers points classiques (rôle des approches inférentielle et descriptive, contraintes de la pluridisciplinarité) ni tenter d'être exhaustif. Ces réflexions qui n'engagent que leurs auteurs, pourraient avoir une portée générale, bien qu'elles concernent une branche particulière de la biologie.

2. Utilisation de la statistique en biologie

La plupart des biologistes reçoivent actuellement au cours de leurs études une formation aux techniques statistiques d'analyse des échantillons. La situation dans notre pays reste cependant très inégale pour ce qui est de la formation reçue en premier et second cycle. En outre, la plupart des biologistes "en poste" ont acquis leur formation statistique sur le tas.

Le "menu" classique porte sur la statistique descriptive, les tests de comparaison de moyennes et les notions de base de corrélation-régression, c'est-à-dire en gros le contenu de l'ouvrage de VESSEREAU (1967) dans la collection "Que sais-je". Il s'y ajoute, selon les cas, des connaissances en analyse multivariée et/ou en analyse de variance. Sans pouvoir étayer cette remarque de données chiffrées, j'aurais tendance à penser que les techniques accessibles aux biologistes sont en général bien utilisées, y compris dans la définition de plans d'expérience ou d'observation, avec bien entendu une concentration sur un "noyau dur" de techniques. C'est dire qu'il y a aussi un sous-emploi de nombreux tests spécialisés : à titre d'exemple, on trouve ainsi dans RAO (1972 pp. 578 sqq) un test pour déterminer si des

individus supplémentaires appartiennent à l'une, l'autre, ou aucune de deux populations d'où sont extraits deux échantillons soumis à une analyse discriminante de référence. Ce test, qui ferait le bonheur de plus d'un paléontologue rencontrant sans cesse de nouveaux taxons dans ses échantillons, est inaccessible en pratique parce que publié dans un ouvrage trop spécialisé.

Des logiciels comme SAS (SAS, 1982) ou BMDP (DIXON et BROWN, 1979) favorisent la diffusion lente de techniques sophistiquées, par l'intermédiaire de biologistes qui ont acquis leur autonomie dans l'utilisation de ces logiciels.

L'acquisition souvent individuelle des connaissances, et la difficulté même des concepts de la statistique - difficulté qu'on a tendance à sous-estimer une fois franchi le pas - font qu'on ne peut attendre des biologistes qu'ils acquièrent une vue unitaire d'un champ donné de la statistique. La structure même de l'enseignement des tests "de base" - et c'est une étape dont on conviendra qu'il est difficile de se passer dans la mesure où elle confère une large autonomie aux biologistes - conduit fréquemment à présenter les tests statistiques comme des recettes, plus que comme la mise à l'épreuve de modèles, c'est-à-dire de relations basées sur des hypothèses, vérifiables ou non, vérifiées ou non. On parlera ainsi d'analyse de variance et de régression plutôt que de modèle linéaire, de test G^2 plutôt que de modèles logistiques-linéaires. On peut noter également que ce mode d'apprentissage de la statistique, et le contenu de bien des ouvrages, induisent des pratiques qui deviennent dominantes sans être soumises à examen critique : de nombreux biologistes sont ainsi littéralement obsédés par l'hypothèse de normalité en analyse de variance, mais ignorent totalement ou presque l'hypothèse d'homoscédasticité, pourtant plus à même le plus souvent de détruire la puissance de l'analyse.

Les deux conséquences les plus marquantes de cet état des relations des biologistes avec la statistique - où nous avons en tant que statisticiens une responsabilité évidente - me semblent en définitive être :

a) Une tendance à l'emploi de trop de tests, trop souvent univariés, sur les mêmes données, ou de collections de tests disjoints sur des sous-ensembles d'un corpus de données. Les corollaires sont une absence de contrôle du risque de première espèce, et une perte souvent considérable de puissance.

b) Une perte de la puissance modélisatrice de l'analyse statistique. L'avènement de logiciels mettant la notion de modèle en

avant (par ex. GLIM ; BAKER et NELDER, 1978) et d'une plus grande flexibilité des méthodes d'analyse multivariées (voir par ex. SABATIER, 1987) devrait permettre de lutter, à travers la consultation statistique, contre ce second point.

3. Utilisation de la modélisation en biologie

A l'opposé, la modélisation est utilisée soit sous l'angle de l'Analyse des Systèmes, avec souvent des systèmes d'équations déterministes, différentielles ou de récurrence, d'un volume important, soit à l'autre extrême sous l'angle des modèles théoriques très compacts construits dans une perspective fortement hypothético-déductive, notamment en biologie de l'évolution. Dans les deux cas, il s'agit d'une mathématisation dont on peut dire un peu abruptement qu'elle tend à reproduire celle des sciences physiques. Entre ces deux extrêmes existent bien entendu de nombreuses situations intermédiaires.

Le modèle est dans le premier cas un outil de simulation visant à représenter l'évolution temporelle d'un système complexe, et l'hydrobiologie par exemple utilisera des modèles de flux spatio-temporels bien proches de ceux de l'hydrodynamique (voir par ex. PARKER, 1968). Un des exemples les plus classiques de gros modèles en Ecologie est certainement le modèle ELM ("Ecosystem Level Model") construit pour représenter le fonctionnement à l'échelle de quelques centaines de jours de steppes arides d'Amérique du Nord (INNIS, 1978). Ce modèle, dont les limitations des performances sont bien comprises (WOODMANSEE, 1978), comporte une quarantaine d'équations aux différences, et n'a évidemment d'existence qu'à travers un programme d'ordinateur. Les analyses de sensibilité renseignent beaucoup plus que les résultats bruts sur la structure du Modèle, et soulignent pour les biologistes les domaines où doivent se porter les efforts. Néanmoins, si le degré de non-linéarité de tels modèles reste probablement limité, on peut craindre des interactions numériques entre des parties très éloignées du modèle (MAGUIRE, 1974).

Dans le second cas, les modèles sont comme nous l'avons souligné étroitement associés à une démarche hypothético-déductive : le modèle découle d'une théorie, et conduit à des prédictions qui permettent par confrontation avec le monde réel de réfuter ou non la théorie. Ces modèles sont le plus souvent construits et traités par les biologistes eux-mêmes, souvent avec l'aide de simulations. La performance de tels modèles par rapport aux questions étudiées est alors importante, et c'est ce qui explique leur important développement : on pourrait dire que, paraphrasant CLEMENCEAU, les biologistes considèrent que la

modélisation est une chose trop sérieuse pour la laisser aux modélisateurs. 31 des 82 notes ou articles parus en 1987 dans la revue *American Naturalist* portent ainsi sur le développement d'un modèle mathématique. Deux autres périodiques s'intitulent *Journal of Theoretical Biology*, et *Theoretical Population Biology*. La fécondité de cette approche est indéniable, ne serait-ce que parce que le débat sur différentes théories est partiellement clarifié par l'écriture sous forme mathématique d'un certain nombre d'hypothèses : il s'agit là d'un avantage classique des modèles mathématiques sur les modèles dialectiques (LEGAY, 1973). La confrontation avec le monde réel reste souvent qualitative, ou fait l'objet d'une analyse statistique classique de données visant à mettre à l'épreuve une des déductions du modèle. Il est vrai que la confrontation directe avec des données est rendue difficile par le caractère strictement déterministe de bon nombre de ces modèles.

L'absence d'étude mathématique, au profit de calculs strictement numériques, est une autre faiblesse fréquente, d'autant que la diversité des comportements de systèmes dynamiques même très simples est un paradoxe difficile à admettre pour le non-mathématicien. Il est en particulier difficile de convaincre les biologistes que de tels calculs n'explorent au mieux qu'une partie des situations, avec des risques d'erreurs inhérents à nos moyens de calcul :

Le calcul de la série $\sum_{i=1}^p \frac{1}{i}$, pour p croissant (sur ordinateur compatible PC en Basic) indique ainsi une stabilisation à 15.40638. D'autres programmes, dans d'autres langages, sur d'autres machines, indiqueraient une stabilisation à une autre valeur alors que cette série est bien connue pour être divergente : en dessous du seuil d'*underflow*, $\frac{1}{i}$ est remplacé par 0 ...

On peut citer également dans ce contexte l'important retentissement des modèles de récurrence non-linéaires, utilisés comme modèles en temps discret de la dynamique des populations (voir un résumé dans LEBRETON et MILLIER, 1982) : les comportements chaotiques revêtent entre autres un intérêt tout particulier dans la mesure où ils ressemblent étrangement aux "gradations" de populations d'insectes, c'est-à-dire à des explosions de population aperiodiques. Il convient tout d'abord de rappeler que le calcul de ces comportements sur ordinateur ne saurait être chaotique puisque nos machines ne travaillent que sur un petit sous-ensemble des rationnels.

En outre, *in natura*, des conditions de milieu exceptionnelles concourent le plus souvent à de telles explosions, et l'on voit donc bien que les modèles les plus pertinents devraient prendre en compte la variabilité de l'environnement.

Un avantage des modèles stochastiques est donc leur plus grande pertinence, mais aussi leur confrontabilité plus aisée aux données. Les difficultés techniques qui ne manquent pas de se faire jour peuvent être résolues de trois façons :

- a/ par la simulation : nous venons d'en souligner les dangers si elle est utilisée en dehors de toute étude mathématique préalable ;
- b/ la statistique ad-hoc ;
- c/ la collaboration pluridisciplinaire.

La pratique de la statistique ad-hoc sur des données qui relèvent en fait de processus stochastiques est une des voies les plus dangereuses, car elle donne fréquemment naissance à des solutions erronées. Comme ces solutions s'adressent à des problèmes biologiquement importants, il en résulte parfois des pratiques erronées qui perdurent malgré des mises en garde répétées. En voici un exemple en dynamique des populations :

A partir du modèle de croissance en temps discret :

$$N_{t+1} = a N_t^b = a N_t^{b-1} \times N_t \quad (1)$$

$$\text{on a} \quad \log N_{t+1} = \log a + b \log N_t \quad (2)$$

(N_t est l'effectif d'une population au temps t).

Si $b=1$, il y a croissance exponentielle. $b \neq 1$ indique au contraire une croissance hypoexponentielle, c'est-à-dire une régulation : le taux de multiplication $a N_t^{b-1}$ devient une fonction décroissante N_t .

A partir de (2), divers auteurs (voir un résumé dans **EBERHARDT**, 1970) ont proposé au début des années 60 de tester l'hypothèse $b=1$ en comparant la pente estimée par régression de $\log N_{t+1}$ sur $\log N_t$ à 1.

C'est oublier que, si $\log N_t$ est soumis à des erreurs additives iid ϵ_t , de variance σ^2 , on a :

$$\log N_{t+1} = \log a + b \log N_t + \epsilon_t - \epsilon_{t-1}$$

$$\log N_t = \log a + b \log N_t + \epsilon_{t-1} - \epsilon_t$$

$$E((\epsilon_t - \epsilon_{t-1})(\epsilon_{t-1} - \epsilon_t)) = -2 \sigma^2,$$

ce qui viole l'hypothèse d'indépendance des individus dans l'échantillon soumis à la régression.

EBERHARDT (1970) démontre que $E(b) < 1$ lorsque le modèle de régression usuel est appliqué ainsi à des effectifs. Des dizaines d'auteurs ont, avant et après 1970, conclu ainsi à l'existence de fortes régulations dans les populations qu'ils étudiaient.

La *Collaboration pluridisciplinaire* présente quant à elle diverses contraintes ; il est bien connu qu'elle exige un état d'esprit particulier des deux parties : le mathématicien devra notamment se soumettre aux objectifs biologiques, le biologiste aux contraintes des mathématiques. Le mathématicien devra admettre des modes de variabilité complexes, bien différents de bruits "blancs", ou même "roses" (cf. CHESSON, 1978). Ce type de collaboration bute fréquemment sur la rareté des biornathématiciens. Il s'agit en fait le plus souvent d'une chaîne pluridisciplinaire plus que de la collaboration de deux personnes seulement.

Ajoutons enfin que l'enseignement de la modélisation en biologie est difficile, car il ne peut éviter de toucher à l'épistémologie (cf. LEGAY, 1973), et repose sur des techniques très polymorphes (on trouvera un aperçu des techniques utilisées en écologie dans JEFFERS, 1977).

4. Discussion

Il me semble donc que les développements de l'utilisation de la statistique et de la modélisation en biologie devraient s'attacher à promouvoir :

- a) la notion de modèle en statistique
- b) les aspects stochastiques en modélisation.

L'ajustement non-linéaire d'une courbe de croissance, ou la construction de modèles permettant d'estimer des taux de survie sont de bons exemples de situations intermédiaires qui peuvent être entièrement présentées sous l'angle statistique ou sous l'angle modélisation (bien des modèles de survie ont d'ailleurs initialement été construits comme modèles déterministes).

Il s'agirait donc de placer statistique et modélisation non pas comme des techniques concurrentes, ni comme des techniques complémentaires, mais comme des constituants d'un *continuum*, malgré la distance qui sépare un test t d'un système d'équations différentielles.

Remerciements

Je remercie R. VARRO qui a attiré mon attention sur l'exemple de série divergente cité dans le texte, et N. BARBICHON.

Bibliographie

- BAKER, R.J. et NELDER, J.A., 1978 - The **GLIM** System, Release **3**, Generalized interactive modelling. Numerical algorithm group, Oxford.
- CALOW, P., 1987 - "Towards a definition of functional ecology". *in Functional Ecology*, 1 : 57-61.
- CHESSON, P., 1978 - "Predator-prey theory and variability". *in Annu. Rev. Ecol. Syst.*, 9 : 323-347.
- DIXON, W.J. et BROWN, M.B. (Eds) 1979 - Biomedical computer **programs P-series**. Univ. of California Press, Berkeley, 88 pp.
- EBERHARDT, L.L., 1970 - "Correlation, regression, and density dependence". *in Ecology*.
- INNIS, G.S., 1978 - **Grassland simulation model**. Ecological studies n°26, Springer-Verlag, New-York.
- JEFFERS, J.N.R., 1977 - **An** introduction to systems analysis : with **ecological** applications. Arnold, Londres.
- LEBRETON, J.D. et MILLIER, C. (Eds.) 1982 - Modèles dynamiques déterministes en biologie. Masson, Paris.
- LEGAY, J.M., 1973 - La méthode des modèles, état actuel de la méthode expérimentale. Informatique et Biosphère, Paris.
- MAGUIRE, B., 1974 - "Mega problems of megamodel builders". *in Simulation.*, 22.
- PARKER, R.A., 1968 - "Simulation of an aquatic ecosystem". *in Biometrics*, 24 : 803-82.
- RAO, C.R., 1972 - Linear **statistical inference** and its applications. Wiley, New York.

SABATIER, R., 1987 - Méthodes factorielles en analyse de données : approximations et prise en compte de variables concomitantes. Thèse Doct. ès-Sciences, Univ. des Sc. et Tech. du Languedoc, Montpellier.

S.A.S. 1982 - **S.A.S** User's Guide, **Statistics**. SAS Institute Inc., Cary, North Carolina.

VESSEREAU, A., 1967 - La Statistique, Que sais-je ?.

WOODMANSEE, R.G., 1978 - Critique and analyses of the Grassland Ecosystem model ELM. pp. 257-281, in INNIS.

III. ELISA ET LE STATISTICIEN par E. Jolivet, INRA, département de Biométrie.

1. Introduction

C'est un lieu commun d'affirmer que les mathématiques se sont nourries des problèmes posés par les sciences expérimentales. La biologie n'a pas, à cet égard, un statut privilégié. Nous pouvons néanmoins souligner deux particularités : la conscience notoire du fait que l'interfécondation des mathématiques et des sciences de la vie n'est encore guère avancée, la complicité toute spéciale existant entre la biologie et la statistique. Cette dernière caractéristique provient essentiellement de la variabilité universellement constatée dans le monde vivant et de notre inaptitude à en rendre compte autrement que par des modèles probabilistes. Notre propos est ici de montrer, sur un exemple simple, comment la prise en compte rigoureuse du caractère imprécis des connaissances relatives à un phénomène biologique peut conduire le statisticien à fournir des méthodes bien adaptées.

2. L'exemple des dosages

C'est jusque dans les méthodes de mesure que le biologiste rencontre des difficultés à expliquer de manière suffisamment précise les phénomènes mis en jeu. La description mathématique, indispensable puisqu'il s'agit de quantifier et de comparer, reste fruste soit par impuissance à rendre compte en détail de ce qui se passe, soit par nécessité, car pour être opératoire, le modèle mathématique ne doit pas être trop complexe. Cette situation se rencontre en particulier dans les méthodes de dosage modernes, comme les dosages radio-immunologiques, couramment appelés RIA, sigle correspondant à

l'expression anglaise *radio immuno assays*, ou les tests immuno-enzymatiques, appelés ELISA, non par romantisme, mais comme abréviation de *enzyme linked immuno sorbent assays*. Dans l'un et l'autre cas, il s'agit de méthodes de dosage indirectes, du fait que ce n'est pas la concentration du produit testé lui-même que l'on mesure, mais celle d'un produit obtenu à la suite de réactions chimiques, dont le produit testé est l'un des précurseurs.

Détection et dosage d'hormones, d'anticorps sont en particulier les champs d'application de la méthode ELISA, à laquelle nous nous restreignons maintenant. Le principe en est le suivant : le milieu contenant le produit à doser est dilué plusieurs fois, et pour chaque dilution, le mélange est mis en contact avec un milieu réactif. Le réactif est contenu dans les puits d'une plaque. Chaque puits correspond donc à une dilution fixée. On mesure ensuite la densité optique du mélange contenu dans chacun des puits après réaction. Cette densité optique est un indicateur de la concentration de produit à doser introduite dans le puits. D'une façon générale, les données obtenues se présentent comme la densité optique mesurée en fonction du logarithme de la dilution. Portées sur un graphique (voir figure 1)⁽¹⁾, elles se répartissent suivant une courbe en S que l'on souhaiterait évidemment caractériser par quelques paramètres interprétables et surtout comparables lorsque l'on passe d'une situation expérimentale à une autre.

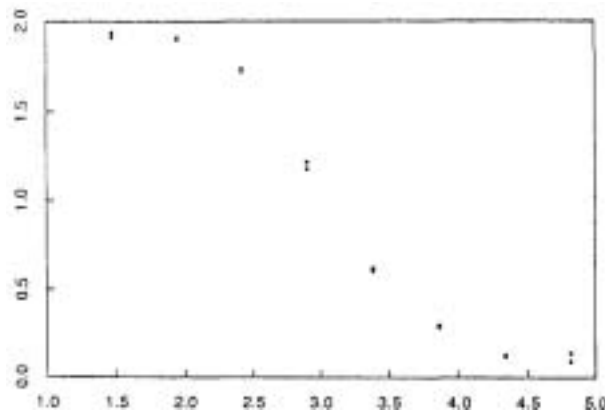


Figure 1. Données issues d'un test ELISA pour la recherche d'anticorps dans un sérum de vache

(1) Ces données sont issues de [1].

3. La régression non-linéaire

Pour représenter et étudier ce type de phénomène, le modèle mathématique suivant offre un cadre correct. Si nous appelons Y la densité optique mesurée, elle peut être reliée à la dilution $z=e^x$, donc finalement à la quantité de produit introduite dans le puits, par une relation

$$Y = f(x, \theta) + \epsilon$$

où f est une fonction déterminée de x , dépendant de façon non-linéaire d'un vecteur de paramètre Q et où ϵ est une variable aléatoire.

f est donc ici la relation fonctionnelle existant entre la quantité initiale du produit introduite dans un puits et la densité optique du milieu après réaction. Il n'est pas possible d'établir correctement sa forme, le phénomène à décrire étant fort complexe. On choisira alors de donner à f une forme classique de courbe en S, par exemple

$$f(x, \theta) = \theta_2 + \frac{\theta_1 - \theta_2}{1 + \exp(\theta_3(x - \theta_4))}$$

ϵ est l'erreur, l'écart aléatoire existant entre la mesure lue sur l'appareil et la vraie valeur pour une dilution x donnée. Cet aléa provient des fluctuations de l'appareil, des erreurs de lecture, mais sans doute surtout de la dilution elle-même. Le choix courant est de supposer qu'il s'agit d'une variable aléatoire gaussienne d'espérance nulle et de variance, inconnue, σ^2 .

Le but sera par exemple de comparer les niveaux d'anticorps chez une même vache à plusieurs moments de l'année [5].

Dans le cadre du modèle que nous venons de décrire, des résultats classiques, essentiellement asymptotiques, sont connus depuis presque vingt ans [7].

L'étude du modèle statistique de régression non-linéaire a évidemment beaucoup progressé depuis. Nous voulons aborder ici quelques uns des points où l'inspiration est clairement issue de situations réelles du type de celles que nous venons d'évoquer.

a - Modèle de l'espérance

Puisque nous savons pertinemment que le modèle f choisi pour l'espérance est faux, pourquoi ne pas tenir compte de cette réalité, la formaliser et en tirer les conséquences ?

Nous supposons donc que le vrai modèle de l'espérance est une fonction g , que nous approchons par la fonction de l'ensemble $\{f(x, \theta)\}_{\theta \in \Theta}$ dont l'écart aux données est le plus faible. Il y a évidemment plusieurs choix d'écart possibles : maximum de vraisemblance, ou minimisation d'un contraste, comme les moindres carrés. Pour ce dernier choix, asymptotiquement, c'est-à-dire pour un effort expérimental suffisant, on montre [2] que la fonction f ainsi sélectionnée est bien la plus proche de g . De plus, les fluctuations du vecteur estimateur de θ autour de la valeur $\theta(g)$, qui rend f la plus proche de g , sont décrites approximativement par une loi de probabilité gaussienne.

Cette approche est en fait une alternative à la régression non-paramétrique, où l'on cherche à estimer la fonction de l'espérance non plus dans une famille du type $\{f(x, \theta)\}_{\theta \in \Theta}$, mais dans un espace beaucoup plus vaste, en précisant par exemple que f appartient à un espace de Sobolev. La méthode du modèle inadéquat est dans un certain sens plus proche de la pratique quotidienne des utilisateurs de la statistique. Cependant, les méthodes non-paramétriques, issues du même souci de se placer dans un cadre d'hypothèses moins contraignant et plus conforme à la réalité, font l'objet depuis quelques années de recherches importantes, dans l'ensemble de la communauté scientifique. Une perspective particulièrement importante et intéressante, du point de vue de l'application à la biologie, est la comparaison non-paramétrique de courbes.

b - Modèle de l'aléa

En ce qui concerne la loi des erreurs, l'hypothèse gaussienne n'est absolument pas cruciale, dans la mesure où bon nombre de résultats sont encore vérifiés si l'on s'en passe. En revanche, l'hypothèse selon laquelle les erreurs sont équidistribuées est souvent grossièrement fautive, la variance des observations évoluant avec la moyenne. Là encore, il est important de prendre cette information en compte. Suivant le plan d'expérience, il est possible d'aborder ce problème de façon paramétrique ou non-paramétrique, de supposer le modèle de la variance inadéquat ([4], [8]), de s'interroger sur

l'influence d'une estimation plus ou moins exacte des paramètres de la variance sur ceux, la plupart du temps d'un plus grand intérêt pratique, de l'espérance [3]. Là encore, le souci essentiel est de ne se fonder que sur des hypothèses raisonnables.

c - Défauts de la technique de mesure

Revenons un instant à la technique ELISA. Afin d'obtenir la meilleure précision relative, le souhait de l'expérimentateur est d'étalonner son appareil de manière à ce que les densités optiques mesurées s'étalent le plus possible sur la plage de sensibilité de l'appareil. Comme le résultat de l'expérience n'est pas connu à l'avance, il arrive assez souvent que, pour les faibles dilutions, qui correspondent aux fortes densités optiques, l'appareil soit saturé. On sait alors seulement que la donnée est supérieure à la graduation maximale, soit ζ . Ignorer ces données, ou les remplacer par ζ n'est pas correct : c'est ce que suggère l'intuition et prouve la théorie. Si l'on suppose connues et la fonction f , et la loi de ϵ_n , la méthode du maximum de vraisemblance s'impose et s'applique sans guère de difficultés. En revanche, si l'on souhaite se placer dans un cadre de suppositions plus réaliste, il est souhaitable de lever au moins les contraintes sur ϵ_n , en supposant par exemple qu'il s'agit de variables centrées équidistribuées, sans faire plus d'hypothèses sur leur loi. Les résultats obtenus dans le cas de la régression linéaire [6] doivent encore être étendus au cas de la régression non-linéaire.

4. Conclusion

Le caractère anecdotique et illustratif de ce qui précède n'échappera à personne. De plus, même sur cet exemple, c'est un aspect bien partiel des choses qui a été présenté. Surtout, que l'on ne me prête pas l'intention d'affirmer que l'observation de la nature est l'essentielle source de progrès de notre discipline. Quiconque s'est peu ou prou affronté à une activité de recherche en mathématique sait bien que cette science puise surtout en elle les éléments de son développement, même lorsqu'elle s'intéresse aux applications. Quoi qu'il en soit, j'aimerais terminer par deux propositions.

La première n'est guère originale, qui affirme que le mathématicien doit bien s'imprégner de la réalité des applications qu'il traite, afin d'apporter des éléments de réponse convenables aux questions concrètes qui lui sont posées. Ce but est d'autant plus

difficile à atteindre en ce qui concerne les sciences de la vie que, en France, les parcours éducatifs des biologistes et des mathématiciens sont très tôt disjoints.

La seconde est peut-être un peu plus nouvelle. Au lieu de nous effaroucher, comme parfois les élèves de l'École qui nous accueille aujourd'hui, ou même comme des mathématiciens plus chevronnés, du caractère souvent imprécis des connaissances actuelles de la biologie, ou d'autres sciences expérimentales, considérons cet état de fait comme une donnée du problème, et traitons-le de la manière la plus exacte possible. J'espère avoir montré à travers l'histoire d'ELISA, peut-être un peu romancée pour qu'elle soit plus convaincante, que des statisticiens s'étaient lancés avec succès dans cette voie. Est-ce un exemple à suivre dans d'autres domaines des mathématiques ?

Remarque : Les travaux de mes collègues Olaf Bunke (Humboldt Universität, Berlin), Sylvie Huet et Antoine Messéan (Biométrie, INRA) m'ont fourni l'essentiel du matériel à partir duquel j'ai préparé cet exposé.

Références

- [1] F. BERTETTO. *Mise au point d'une méthode ELISA destinée au titrage des anticorps circulant contre le coronavirus bovin, avec modélisation mathématique des résultats*. Thèse, Ecole Nationale Vétérinaire, Maison-Alfort, 1986.
- [2] O. BUNKE. *Assessing the performance of regression estimators and models under nonstandard conditions*. In *Seminarbericht Nr 89*, 1987.
- [3] R.J. CARROLL and D. RUPPERT. *Robust estimation in heteroscedastic linear models*. *The Annals of Statistics*, 10:429-441, 1982.
- [4] S. HUET. *Maximum likelihood and least squares estimators for a nonlinear model with heterogeneous variances*. *Statistics*, 17:517-526, 1986.
- [5] S. HUET and J. LAPORTE. *Statistical methods for the comparison of antibody levels in serums assayed by enzyme linked immuno sorbent*. Rapport technique, INRA, Département de Biométrie, 1987.

- [6] I.R. JAMES and P.J. SMITH. Consistency results for linear regression with censored data. *The Annals of Statistics*, 12:590-600, 1984.
- [7] R.I. JENNRICH. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40:633-643, 1969.
- [8] A. MESSEAN. *Application de la géométrie différentielle à la statistique du modèle non-linéaire*. Thèse, Université Paris-Sud, Orsay, 1984.

IV. QUELS OUTILS MATHÉMATIQUES POUR LE GENIE DES PROCÉDES BIOTECHNOLOGIQUES ? par A. Cheruy, Laboratoire d'Automatique de Grenoble, CNRS.

Avec l'essor des biotechnologies, de nouveaux procédés mettant en oeuvre des microorganismes se développent au niveau industriel. Le but ultime des efforts actuels de recherche et développement est d'aboutir à des procédés performants, c'est-à-dire assurant d'une manière reproductible et stable une productivité maximale et au moindre coût. Pour ce faire, il faut conjuguer trois approches :

- l'approche "biologique" où l'on cherche à améliorer les performances par l'utilisation de souches et milieux de culture appropriés ;

- l'approche "technologique" où l'on recherche les modes de fonctionnement et les technologies les plus efficaces, et les plus rentables (procédé *batch* ou continu, bioréacteur infiniment mélangé ou à cellules fixées ...);

- l'approche "mathématique" où l'on cherche à maximiser la productivité par une conduite optimale (et souvent automatisée) du bioréacteur déterminée à l'aide d'un modèle mathématique de la dynamique du procédé.

En nous appuyant sur notre expérience personnelle, nous allons essayer de montrer en quoi consiste cette dernière approche qui relève essentiellement de l'Automatique, et préciser ses besoins en outils mathématiques.

1. L'approche mathématique **des** bioprocédés

Comme pour n'importe quel type de procédé, cette approche comprend 2 étapes :

- 1) une étape de modélisation où l'on cherche à caractériser mathématiquement le comportement dynamique du procédé.
- 2) une étape de commande où l'on utilise le modèle mathématique précédemment défini pour déterminer une commande à des fins d'optimisation ou de régulation...

Dans la première étape, on fait appel aux outils mathématiques de l'analyse des systèmes et dans la seconde, à ceux de la théorie du contrôle. Mais leur application aux bioprocédés pose des problèmes spécifiques que nous allons analyser au niveau de chaque étape.

Auparavant, il convient de souligner que cette approche mathématique des bioprocédés est rarement menée de pair avec les approches biologique et technologique ; en pratique elle intervient après les 2 précédentes et ses attendus sont souvent sous-estimés. En effet, le biologiste et le technologue ont déjà, chacun à leur niveau, recherché de manière expérimentale des conditions optimales de fonctionnement qui paraissent, **a priori**, difficiles à améliorer. Or les bioprocédés comportent tellement de paramètres (biologiques et technologiques) qu'il est difficile de tous les prendre en compte expérimentalement, d'autant que le comportement dynamique de ces procédés est généralement non stationnaire. Enfin, indépendamment de la définition des conditions optimales de fonctionnement d'un bioprocédé, son contrôle automatique est quasiment indispensable pour le maintenir et le stabiliser dans ces conditions. Différents exemples rapportés dans la littérature illustrent l'apport de l'approche mathématique : citons simplement un exemple que nous avons étudié personnellement concernant un procédé industriel de biosynthèse d'antibiotique où la productivité a été améliorée de 30% par une telle approche (1).

2. La modélisation des bioprocédés

Le modèle mathématique recherché ici est simplement un outil permettant de calculer une commande et, de ce fait, il doit essentiellement rendre compte du comportement dynamique du procédé en réponse à l'excitation de ses variables d'action.

Classiquement les modèles utilisés sont des systèmes d'équations différentielles (la plupart du temps une équation par variable d'état du procédé). Dans le cas des bioprocédés, ces équations sont non linéaires, à paramètres constants et variant dans le temps, et comportent des termes stochastiques pour rendre compte de perturbations, de bruits de mesure, d'erreur de modélisation par exemple.

En pratique, l'obtention d'un tel modèle est une tâche délicate, laborieuse et aux résultats souvent décevants. Les raisons en sont multiples :

a/ Les phénomènes mis en jeu dans un bioprocédé sont complexes, mal cernés et se traduisent par un comportement dynamique non linéaire et non stationnaire. En plus, il est très difficile de définir les variables caractéristiques de l'état de fonctionnement (variables d'état).

b/ Au niveau expérimental, de nombreuses possibilités d'analyse biologique, physique ou chimique existent ; mais il y a peu de capteurs fournissant des mesures en ligne significatives indispensables pour une commande automatique. D'autre part les mesures ne sont pas toujours reproductibles ni d'excellente qualité et il est souvent difficile d'exploiter leur variabilité. Par ailleurs, l'expérimentation est limitée par des contraintes inhérentes à la nature du procédé.

c/ Mais le principal écueil en modélisation est le manque de méthodologie. Cette méthodologie devrait indiquer la démarche à suivre pour élaborer un modèle en définissant des étapes et des objectifs intermédiaires ; elle permettrait de coordonner efficacement les compétences variées auxquelles il faut faire appel (biologie, mathématique, automatique, génie des procédés, informatique...) et d'orienter les recherches dans chaque discipline. Actuellement on travaille au "coup par coup" et les modèles prennent des formes variées essentiellement en fonction des compétences que l'on a pu réunir. Cependant, les études de cas sont maintenant suffisamment nombreuses et démonstratives pour qu'une certaine synthèse méthodologique puisse être effectuée (2). Cette dernière a été amorcée récemment avec le développement de systèmes d'aide à la modélisation (3,4). Dans ces logiciels, des connaissances et une certaine expertise sont introduites et l'on est amené à préciser leur niveau d'intervention :

- aide à l'analyse des modèles, en particulier de leurs propriétés mathématiques (sensibilité, stabilité, identifiabilité, observabilité, commandabilité ...). Signalons à ce niveau que les outils mathématiques

d'analyse font cruellement défaut car l'on a généralement affaire à des systèmes non linéaires non stationnaires et par exemple, on sait très mal étudier leurs propriétés de commandabilité et d'observabilité, ce qui est un lourd handicap lorsque le modèle est élaboré à des fins de commande.

- aide au traitement des modèles : simulation, identification... avec choix des méthodes et techniques les mieux adaptées.

- aide à l'analyse de données expérimentales : filtrage, analyse statistique ... afin d'extraire au mieux l'information significative.

Pratiquement, à chacun de ces niveaux, on peut faire apparaître un manque d'outils mathématiques appropriés à cause essentiellement des propriétés de non linéarité et de non stationnarité dans le comportement dynamique des bioprocédés, et aussi à cause de la qualité de l'information accessible et disponible par l'expérimentation.

3. La commande de bioprocédés

Il convient tout d'abord de souligner que les performances d'une commande de procédés dépendent à la fois de la méthode de commande utilisée et de la qualité du modèle qui a servi à l'établir. Et de nombreux exemples montrent que, souvent, une commande ne donne pas en pratique les résultats escomptés d'après les simulations parce que le modèle utilisé n'est pas "pleinement" valable.

La détermination d'une commande relève de la théorie du contrôle, mais cette dernière n'est bien développée que pour les systèmes linéaires. Or les bioprocédés sont généralement modélisés par des systèmes non linéaires pour lesquels les résultats théoriques sont peu nombreux (5), aussi fait-on appel à des techniques de linéarisation qui généralement limitent la portée des résultats. Cependant, toutes ne correspondent pas à une approximation : c'est le cas par exemple de la commande "non linéaire linéarisante" récemment proposée, et qui consiste à faire une commande non linéaire d'un procédé, telle que l'ensemble (procédé plus commande) soit linéaire. Des tentatives d'application de cette approche sur les bioprocédés (6) ont montré que la mise en oeuvre pratique est compliquée par la prise en compte de contraintes physiques (positivité de variables représentant des débits ou des concentrations) et les résultats ne sont pas aussi satisfaisants qu'on pouvait l'espérer.

De même pour prendre en compte la non stationnarité des bioprocédés, on a cherché à faire appel à la théorie de la commande adaptative mais cette dernière s'applique essentiellement aux systèmes linéaires monovariables. Aussi, conviendrait-il de développer des outils pour la commande des systèmes non linéaires.

Par ailleurs, pour mettre en oeuvre une commande, il faut disposer de mesures en ligne ; or, nous avons déjà signalé le manque de capteurs en particulier pour les variables d'états biologiques. C'est pourquoi, on cherche généralement à les reconstruire en temps réel à partir des mesures existantes. Cette conception d'observateur est souvent effectuée en s'appuyant sur l'extension de techniques linéaires (ex : filtre de Kalman étendu) et sans que l'on soit à même de garantir que le système en question est observable.

4. Conclusion

Une approche mathématique efficace des procédés biotechnologiques passe par le développement d'outils mathématiques pour l'analyse et le contrôle de systèmes non linéaires et non stationnaires. En effet, la non linéarité et la non stationnarité sont 2 caractéristiques essentielles du comportement dynamique des bioprocédés, et comme les outils disponibles concernent surtout les systèmes linéaires, les applications actuelles font appel à des approximations et linéarisations, ce qui limite beaucoup la portée des résultats. Ce besoin d'investigation en Automatique non linéaire n'est pas une demande spécifique du génie des bioprocédés ; mais en biotechnologie, les problèmes se posent de manière plus ardue que pour les autres types de procédés (physico-chimiques par exemple) car les non linéarités sont très marquées et inhérentes à la dynamique. Aussi la démarche généralement pratiquée par les automaticiens, et qui consiste à décomposer un système dynamique non linéaire en un système statique non linéaire auquel on superpose une dynamique linéaire, n'est pas valable pour les bioprocédés et tout particulièrement pour ceux de type *batch* pour lesquels on ne peut pas définir d'état statique : l'état évolue dans le temps et il convient de chercher à contrôler son évolution.

Enfin, dans l'approche mathématique des bioprocédés, soulignons le rôle charnière joué par l'ingénieur (automaticien) entre le biologiste et le mathématicien. En effet, de par sa formation à la fois expérimentale et théorique, il est en général parfaitement apte

à comprendre les 2 types de problématiques et donc souvent le mieux placé pour les coordonner.

Références

- (1) A. CHERUY, A. DURAND, "Optimisation of Erythromycin biosynthesis by controlling pH and temperature : theoretical aspects and practical application", *in* Biotech. **Bioeng.**, 9:303-320, 1979.
- (2) A. CHERUY, "Méthodologie de la modélisation", in **Ecole** d'été sur "Modélisation et génie des systèmes biologiques", CNRS-INRIA, Sophia-Antipolis, 13-15 Sept. 1985.
- (3) A. PAVE, F. RECHENMAN, "Computer aided modelling in biology, an artificial intelligence approach", in **A.I. Applied to Simulation**, Ed. KERCKHOFFS, VANSTEENKISTE, ZEIGLER SCS Simul Serie, 18, 52-66, 1986.
- (4) R. MONTELLANO, A. CHERUY, "Computer aided design in modelling of bioprocesses", **4th Eur. Congress on BIOTECHNOLOGY** 1, 289-293, 1987.
- (5) J.P. GAUTHIER, **Structure des systèmes non linéaires**, Editions du CNRS 1984.
- (6) D. DOCHAIN, "On line parameter estimation, adaptative state estimation and adaptative control of fermentation processes", Thèse, Université Catholique de LOUVAIN, 1986.

V. MODELISATION EN BIOLOGIE : PROBLEMES LIES A L'INTERPRETATION DES OBJETS MATHEMATIQUES par Alain Pavé, Laboratoire de Biométrie, Université Claude Bernard.

Feller en 1940, remarquait qu'un modèle simple, le modèle logistique, était fréquemment choisi pour représenter des situations biologiques diverses, sur la base de "bons ajustements" aux données expérimentales. Il montrait notamment que d'autres modèles, avec les techniques de l'époque, s'ajustaient mieux à ces données, et partant de là il s'interrogeait sur le choix systématique du modèle logistique par les expérimentateurs.

En fait, on peut prendre un point de vue "orthogonal" à la position de Feller, et se demander pourquoi un même objet mathématique est susceptible de représenter des situations biologiques

diverses. Par exemple, toujours en prenant le modèle logistique, on remarque qu'il est utilisé comme modèle de divers phénomènes de croissance (aussi bien la croissance de populations microbiennes ou de populations humaines, que la croissance individuelle de vertébrés...), ou même de décroissance. Le modèle peut être vu uniquement sur le plan descriptif, alors la discussion s'arrête et renvoie aux remarques de Feller. On peut aussi imaginer qu'il représente, au niveau phénoménologique, des mécanismes plus profonds, comme le disait fort justement J. Monod : "le contenu d'une expression mathématique est toujours beaucoup plus riche que ne le croit en général son auteur". Notre propos est d'étudier ce deuxième point de vue et notamment de proposer des outils d'analyse des modèles conduisant à une interprétation, et de là à une justification au niveau de la cohérence du choix d'un modèle pour représenter un phénomène biologique donné. On apporte ainsi une dimension sémantique qui disparaît classiquement dans l'objet mathématique, ce dernier pouvant "vivre sa vie" indépendamment de toute interprétation hors du champ mathématique. Enfin on peut remarquer que ces préoccupations rejoignent par ailleurs des questions posées en I.A. sur les notions de connaissances superficielles et de connaissances profondes.

1. Différents niveaux de signification

Nous discuterons essentiellement de l'aspect formel, il faut cependant remarquer qu'une formule, outre son intérêt propre, n'est intéressante que dans la mesure où elle se situe dans un cadre permettant sa manipulation, sa transformation, mais aussi de la mettre en relation avec d'autres objets comme des objets géométriques, par exemple graphiques.

A un moment donné, si on se place dans une optique de modélisation, c'est-à-dire de représentation formelle d'un objet ou d'un phénomène, on choisira ou on construira une formule, en particulier une formule mathématique, sensée représenter cet objet ou ce phénomène dans un système formel. Il y a lieu d'examiner ce que ce symbolisme peut nous apporter, quel degré de signification on peut lui associer et quelle interprétation on peut en donner dans le champ d'application.

Pour une formule on peut distinguer grossièrement trois niveaux à contenu sémantique croissant :

a/ La formule vue comme un objet formel : à ce niveau les symboles sont équivalents, ils n'ont pas d'interprétation particulière. C'est-à-dire qu'on peut appliquer à cette formule toutes les opérations permises, en particulier les opérations algébriques et arithmétiques. Les symboles littéraux jouent le même rôle. Par exemple, la formule

$$r \ x \ (1 - \frac{x}{K}) \quad (1)$$

peut tout simplement être évaluée numériquement en associant des valeurs numériques aux symboles. On peut aussi lui faire subir tout un ensemble de transformations algébriques donnant des expressions équivalentes (i.e. qui donneront la même évaluation), ou encore lui appliquer l'opération de dérivation relativement à un, ou plusieurs, symboles littéraux.

Enfin, et à la limite, une telle expression peut être vue, au niveau le plus bas, comme une simple chaîne de caractères, surtout si on en prend la représentation informatique linéaire :

$$r \ * \ x \ * \ (1 - x/K)$$

Dans cette chaîne tous les symboles sont équivalents, on ne la considèrera comme expression arithmétique que si on interprète les symboles, en particulier si on distingue les opérandes, opérateurs et parenthèses. On leur attribue alors un statut particulier, c'est un premier niveau d'interprétation.

b/ La formule vue comme un objet mathématique, écrivons maintenant

$$x' = r \ x \ (1 - \frac{x}{K}) \quad (2)$$

alors toute personne ayant un peu de culture mathématique comprendra qu'il s'agit d'une équation différentielle. En supposant en outre $x' = \frac{dx}{dt}$, on saisit que cette équation définit implicitement un ensemble de relations entre une variable "dépendante" x et une variable "indépendante" t , les autres symboles littéraux représentant des paramètres. Une relation, ou fonction, particulière pourra être déterminée si on se donne une condition initiale, par exemple à $t=0$ $x=x_0$. Outre les transformations autorisées par le niveau formel, on peut essayer d'explicitier la fonction $x=f(t)$ en cherchant la solution formelle de cette équation, on trouve :

$$x = \frac{K}{1 + \frac{K - x_0}{x_0} e^{-rt}} \quad (3)$$

On sait également tracer le graphe de cette fonction. Tout ceci concerne l'étude qualitative, ou mathématique, de l'objet concerné.

e/ L'objet mathématique vu comme modèle.

L'objet mathématique devient un modèle s'il représente une situation du monde physique ou biologique. Par exemple les formules (2) et (3) sont liées au modèle logistique très utilisé pour représenter des phénomènes biologiques, notamment la croissance de populations, ou d'organismes. Alors

* les variables prennent une signification physique, chimique ou biologique, ayant une unité. Par exemple, dans les formules ci-dessus x peut représenter la taille d'une population, ou une mesure équivalente, t est le temps. Par ailleurs, en écologie il est habituel d'interpréter r comme un taux de croissance et K , taille maximale de la population, comme la capacité limite d'exploitation du milieu par cette population.

* en fonction de l'interprétation des variables et paramètres, l'intervalle de signification de la fonction est précisé. Il est inclus évidemment dans l'intervalle de définition. Par exemple, on ne considère que les temps positifs, que les solutions positives...

Par ailleurs, les relations (multiplicatives, additives...) entre les variables dans les termes constituant la formule peuvent représenter certains mécanismes plus élémentaires que le phénomène observé et modélisé. Nous conviendrons d'appeler ces mécanismes "processus" pour les systèmes dynamiques. Un phénomène donné peut être la conséquence d'un processus ou de la combinaison de plusieurs d'entre eux.

Ainsi, il existe plusieurs façons de voir une formule, de plus en plus précise depuis l'objet formel, jusqu'au modèle, et même dans ce dernier cas l'interprétation peut être plus ou moins précise. En fait, une relation hiérarchique peut être définie entre ces points de vue, la formule prenant de plus en plus de signification et perdant par là-même de sa généralité (restriction des domaines de définition, signification des variables, interprétation du contenu même de la formule). Il s'agit d'une relation hiérarchique de spécialisation.

2. Interprétation mécaniste

Reprenons l'interprétation mécaniste des termes d'une formule. On peut imaginer deux niveaux : le niveau phénoménologique, celui de l'observation et de la réponse globale du modèle, et le niveau explicatif, celui des processus. Sachant que ce qui est phénoménologique à un niveau peut être explicatif à un niveau plus élevé, et devenir ainsi un processus d'une situation plus globale. Inversement, un processus peut être un phénomène à un niveau plus fin, lui-même interprétable en processus de ce niveau inférieur. En intelligence artificielle on distingue, sur une base voisine, les connaissances dites superficielles (niveau phénoménologique) des connaissances profondes (niveau explicatif).

En fait, la profondeur de l'interprétation dépend avant tout de la façon dont on souhaite utiliser le modèle, c'est-à-dire l'objectif de la modélisation. Ainsi, on peut distinguer les modèles suivant l'objectif, on parle par exemple de modèles descriptifs, de modèles de mécanismes, de modèles théoriques... L'intersection n'étant pas vide entre ces ensembles : un même modèle pouvant répondre à plusieurs de ces objectifs. Cependant, un modèle de mécanisme devra avoir de bonnes qualités descriptives, au moins qualitativement, un modèle descriptif devra avoir de bonnes propriétés quantitatives. Bien que ce dernier type de modèle n'ait d'ambition qu'au niveau phénoménologique, on peut s'interroger pour certains d'entre eux, sur le point de savoir si la bonne qualité d'une description n'est pas liée uniquement à une bonne souplesse de l'objet mathématique (comme les fonctions polynomiales), mais plus profondément à quelques types de processus qu'ils peuvent représenter. On parlera alors d'interprétation mécaniste. Cette façon d'aborder le problème de modélisation est assez fréquente en biologie, souvent on choisit d'abord un modèle qui décrit bien les données avant de s'interroger sur sa signification... Cette démarche est certes critiquable, il est donc bon d'en cerner les limites.

L'objet essentiel de cette contribution est de tenter de faire le lien entre ces deux niveaux. Ainsi, nous avons examiné des objets mathématiques (équations différentielles) utilisés en dynamique des populations et dans l'étude de la croissance d'organismes. Nous avons tenté de préciser les types d'interprétation qu'on en peut proposer. A ce propos, nous avons montré qu'une formulation en termes de schémas fonctionnels est très utile aussi bien pour l'interprétation que la construction. Ces problèmes seront abordés en prenant toujours comme exemple le modèle logistique.

En fait, cette réflexion a été menée dans le cadre du projet EDORA d'élaboration d'un système informatique d'aide à la modélisation en biologie. Ce système devrait être intégrer des connaissances diverses sur des modèles utilisés en biologie, en particulier des connaissances relatives à leur interprétation en termes de phénomènes et de processus (Pavé et Rechenmann, 1986, Pavé, 1986, Houiller, 1987).

Schémas fonctionnels

Depuis longtemps, la pratique scientifique utilise des schématisations intermédiaires entre l'énoncé discursif de connaissances *a priori*, ou d'hypothèses, sur la structure ou le fonctionnement d'un système. Pour certaines de ces schématisations des liaisons très fortes existent avec des formulations mathématiques, tout particulièrement avec certaines classes d'équations différentielles, aussi bien pour la génération d'expressions *a priori* que pour leur interprétation *a posteriori*. Citons les diagrammes en boîtes et flèches pour les systèmes à compartiments, les *bond graphs*, les diagrammes de Forrester en systémique, la représentation des réactions chimiques... En fait, cette dernière représentation peut être adaptée à des cas plus généraux, en particulier à la dynamique des populations, dans la mesure où les quantités étudiées sont le bilan de processus élémentaires connus (Garfinkel, 1962, 1968). Nous avons donc utilisé cette représentation bien connue par ailleurs.

Le modèle logistique et ses interprétations

Ce modèle, comme nous l'avons déjà signalé, est certainement le modèle le plus connu en biologie. Il fut proposé au milieu du XIXe siècle par Verhulst (Verhulst, 1938) pour décrire la croissance de populations humaines (en l'occurrence la population de la Belgique). Verhulst discutait du modèle exponentiel de Malthus ($x' = ax$), il supposait que la régulation de la taille d'une population puisse venir de contraintes (biologiques ?) internes à cette population. Il proposait de tenir compte de ces contraintes par l'introduction d'un terme de freinage linéaire $a = 1 - \frac{x}{K}$). Le succès de ce modèle est certainement dû à la simplicité de sa formulation, à l'interprétation des paramètres en termes biologiques, et à la grande diversité des situations que ce modèle peut décrire. Comme l'écrivait Lotka (1925) : "*it has been found to fit very acceptably a number of observed examples of population growth*", observation que reprenait et critiquait Feller en 1940. Comme nous allons le voir, cette diversité est sans doute en partie explicable par les différents schémas

fonctionnels qui peuvent générer ce modèle, et donc la variété de processus et de combinaisons de processus qu'il peut représenter.

Interprétation de la formule en écologie

Dans un premier temps considérons la formulation différentielle [2] :

$$x' = rx(1 - \frac{x}{K})$$

Comme il a déjà été signalé x représente la taille, ou la densité d'une population, ou toute mesure équivalente. Les paramètres K et r sont interprétés respectivement comme la capacité de la population d'exploiter le milieu et comme un taux de croissance lié à la fertilité (proportionnel à la durée de génération). Ce modèle permet ainsi de discuter, et de comparer des populations. Classiquement on parle de populations à stratégies r pour celles dont le taux de croissance est élevé, et de populations à stratégies K pour celles qui exploitent au mieux le milieu. De nombreuses discussions ont été conduites sur ce concept de stratégie r et K , tant sur le plan théorique que pratique. Des interprétations sensiblement différentes ont aussi été proposées, notamment pour les populations humaines.

b - Interprétation à l'aide de schémas fonctionnels

Par ailleurs, on peut tenter d'analyser plus finement ce modèle en essayant de le relier à des schémas fonctionnels interprétables, évidemment, en termes biologiques. Pour ceci on a cherché les écritures équivalentes à (2) par transformations algébriques telles que les équations ainsi obtenues pouvaient être générées à partir de schémas fonctionnels type chimiques. C'est ainsi qu'on a pu constituer la figure 1.

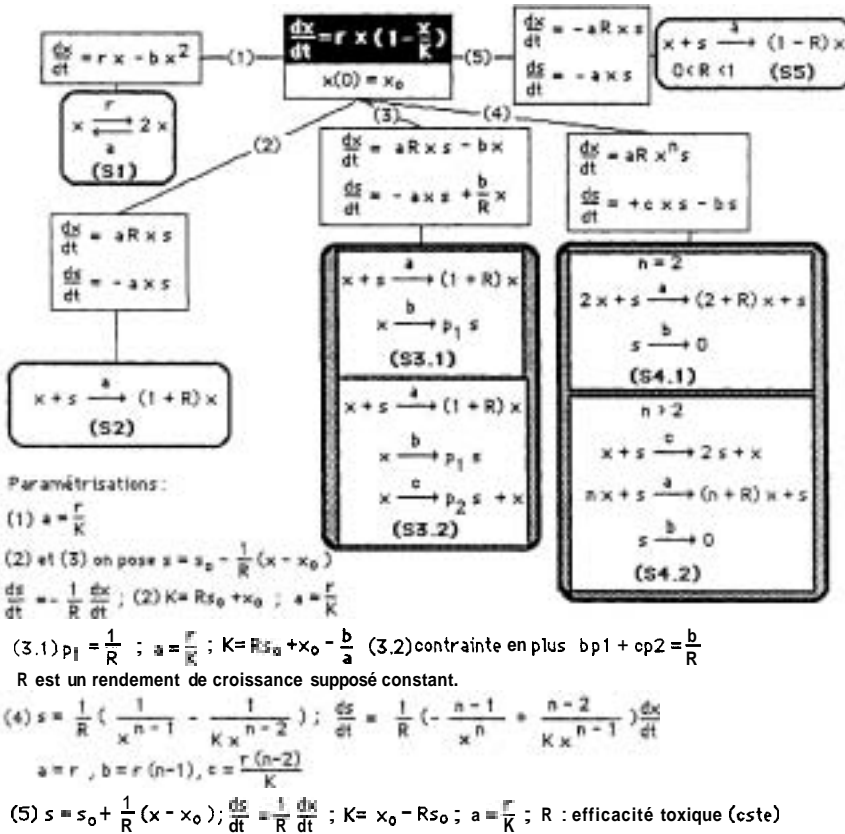


Figure 1. Interprétation du modèle logistique :

différentes expressions équivalentes et schémas fonctionnels associés, chaque situation est commentée dans le texte. Les paramètres et variables ne prennent que des valeurs positives.

Les différents schémas obtenus peuvent être interprétés de la façon suivante :

(S1) représente une croissance limitée par un processus de compétition intraspécifique (voire prédation intraspécifique)

(S2) peut s'interpréter comme la croissance d'une biomasse x dans un milieu limité en substrat (ou ressources) s .

(S3.1) et (S3.2) représentent la croissance d'une population sur un milieu limité en substrat, la biomasse est soumise à une dégradation (ou processus de mortalité) qui régénère une quantité équivalente de substrat à celle consommée pour produire la biomasse pour (S3.1). Cette hypothèse est peut être trop forte dans la mesure où l'on sait que les produits de dégradation ne sont, en général, pas réutilisables en totalité comme substrat (du moins pas directement), c'est-à-dire que p_1 est plus petit que $1/R$. Une façon d'améliorer cette représentation est de supposer que la biomasse (des individus d'une population) est capable "d'exploiter" le milieu pour produire du substrat nécessaire à sa croissance et à son maintien. Le schéma fonctionnel (S3.2) tient compte de cette situation.

(S4.1) et (S4.2), ces schémas décrivent la croissance d'une biomasse x en présence d'un facteur de croissance de type catalyseur s qui se dégrade spontanément (décroissance exponentielle de s décrite par la dernière réaction), si $n > 2$ il y a production de facteur de croissance par la biomasse x .

(S5) s'interprète comme l'action d'une substance toxique sur la biomasse, cette substance étant elle-même dégradée par cette biomasse (par exemple l'action d'un antibiotique sur une population bactérienne, antibiotique simultanément métabolisé par cette même population).

c - Commentaires

Les notions de stratégies r et K peuvent se rediscuter, dans le cadre restreint que nous proposons, en termes de rendement de croissance (R), de vitesse de croissance (caractérisée par la constante $a = \frac{r}{K}$), de mortalité (caractérisée par la constante b), pour le schéma (S3), et de s_0 (quantité totale de substrat, ou plus généralement de ressources disponibles pour une population donnée).

L'interprétation (S4) faisant intervenir un facteur de croissance, nous semble plus satisfaisante pour représenter des courbes de croissance d'organismes (notamment d'organismes supérieurs comme les

vertébrés), dans ce domaine c'est un modèle concurrent du modèle de Gompertz (Pavé et al., 1986).

Pour le schéma **(S5)** on peut penser, par exemple, à l'action d'un antibiotique sur une population bactérienne si celui-ci est simultanément dégradé par cette population.

Enfin, il faut retenir que K (position de l'asymptote horizontale) dépend dans les interprétations **(S2)**, **(S3)** et **(S5)** de x_0 et de s , les conditions initiales, ce qui limite le champ des possibilités. Notamment pour **(S2)** on ne peut observer qu'une solution croissante et pour **(S5)** qu'une solution décroissante.

3. Conclusion

Il ne faut voir une tentative d'interprétation que comme indicative. Rien ne prouve, d'une part qu'un phénomène qui semble "logistique" puisse être relié à l'un des mécanismes décrits, d'autre part que la liste même des interprétations proposées soit exhaustive. Cependant, il est clair qu'une lecture biologique d'une formule est d'autant plus satisfaisante qu'elle est suggestive et explicative, et qu'elle peut conduire ainsi à des améliorations ultérieures du modèle. Aussi bien pour l'interprétation que pour la modification, ou même la construction de novo, les schémas fonctionnels semblent efficaces dans le champ de la dynamique des populations. On notera que cette approche a été testée pour d'autres modèles, et que la modélisation a été, dans certains cas, un complément efficace à un dispositif expérimental (cf. par exemple, Steinberg et al., 1987).

Cet exemple est lui-même une bonne illustration de concepts introduits en intelligence artificielle à propos de connaissances profondes et de connaissances superficielles. Inversement la nécessité de formaliser la connaissance est un excellent moyen de préciser les concepts dans un domaine d'application, et même de détecter les trous de connaissances. Nous en faisons l'expérience dans le projet EDORA, en particulier pour préciser la notion de modèle, mais aussi pour un ensemble de modèles en les analysant le plus finement possible.

Enfin, nous apportons un élément supplémentaire à la discussion lancée par Feller, après tout le modèle logistique est sans doute acceptable dans de nombreuses conditions, cependant il n'est peut être pas inutile de vérifier que son emploi est compatible avec le phénomène biologique étudié.

Références

- Feller W. - On the logistic Law of Growth and its empirical verification in Biology, 1940, repris dans : Oliveira-Pinto F., Conolly B.W. - *Applicable Mathematics of Non-Physical Phenomena*, 1982, Ed. Ellis Horwood & J. Wiley, Chistester.
- Garfinkel D. - Digital computer Simulation of an Ecological System based on a modified Mass Action Law. *Ecology*, 45, 502-507, 1962.
- Houllier F. - Construction et interprétation de modèles dynamiques : exemples forestiers. *Cahiers d'Edora*, 1, 1988.
- Lotka A.J. - *Elements of Physical Biology*. Williams & Wilkins, 1925, Baltimore.
- Monod J. - *Recherches sur la croissance de cultures bactériennes*. Thèse Doct. ès Sciences, 1942, Herman, Paris.
- Oliveira-Pinto F., Conolly B.W. - *Applicable Mathematics of Non-Physical Phenomena*. 1982, Ed. Ellis Horwood & J. Wiley, Chistester.
- Pavé A., Rechenmann F. - "Computer Aided Modelling in Biology : an Artificial Intelligence Approach". In *Artificial Intelligence Applied to Simulation*, Eds. Kerckhoffs, Vansteenkiste G.C., Zeigler B.P., Soc. for Comput. Simul., 18, 1986, 53-66.
- Pavé A. - "Schémas fonctionnels et modélisation. Etude de modèles de la dynamique des populations". Actes du Colloque *Biométrie-Econométrie Sophia* Antipolis, 1985, Eds Demongeot J. et Malgrange P., Presses de l'université de Dijon.
- Pavé A. - "Utilisation et interprétation du modèle de Gompertz", application à l'étude de la croissance de jeunes rats musqués (*Ondatra zibethica* L.), *Biom. Praxim.*, 1986, 26, 123-140.
- Steinberg C., Faurie G., Zegerman M., Pavé A. - "Régulation par les Protozoaires d'une population bactérienne introduite dans le sol. Modélisation mathématique de la relation prédateur-proie". In *Rev. Ecol. Biol. Sol*, 1987, 24, 1, 49-62.
- Verhulst P.F. - "Notice sur la loi que la population suit dans son accroissement". *Corr. Math. Phys.*, 10, 1838. Trad. anglaise : A Note on the Law of population Growth. In Smith D. & Keyfitz N. : *Mathematical Demography*, Biomath., Vol. 6, Springer-Verlag, Berlin, 1977.