A propos du continuum statistique-modélisation en écologie

Jean-Dominique Lebreton[a]

Introduction

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

Le premier, déjà ancien, consiste en l'emploi généralisé des méthodes statistiques pour traiter des données 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 ex. 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 inférentielles et descriptives, contraintes de la pluridisciplinarité) ni tenter d'être exhaustif. Ces réflexions qui n'engagent que leur auteur, pourraient avoir une portée générale, bien qu'elles concernent une branche particulière de la biologie.

1 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

[a] Centre d'Ecologie Fonctionnelle et Evolutive C.N.R.S., BP 5051 34033 Montpellier CEDEX 1

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. Pour ne prendre qu'un seul exemple, 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 :

1. 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.

2. Une perte de la puissance modélisatrice de l'analyse statistique. l'avènement de logiciels mettant en avant la notion de modèle (par exemple GLIM ; Baker et Nelder 1978) et d'une plus grande flexibilité des méthodes d'analyse multivariées (voir par ex. Sabatier *et al.*, 1989) devrait permettre de lutter, à travers la consultation statistique, contre ce second point.

2 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 évolutive. 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 ; 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, même 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 pertinence de tels modèles par rapport aux questions étudiées est alors importante, et c'est ce qui explique leur important développement : on pourrait dire, en paraphrasant Clémenceau, que 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 certain 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 O...

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 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 apériodiques. 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 :

1. par la *simulation* : nous venons d'en souligner les dangers si elle est utilisée en dehors de toute étude mathématique préalable ;
2. par la *statistique "ad hoc"* ;
3. par 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 engendre fréquemment de graves erreurs. Comme cette approche s'adresse 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 suivant, où N_t est l'effectif d'une population au temps t :

$$N_{t+1} = \mathrm{a}{N_t}^b = \mathrm{a}{N_t}^{b-1} N_t \tag{1}$$

on a :

$$\log N_{t+1} = \log \mathrm{a} + b \log N_t \tag{2}$$

Si $b = 1$, il y a croissance exponentielle. $b < 1$ 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 de 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 $\log N_t$.

C'est oublier que, si $\log N_t$ est soumis à des erreurs additives indépendantes et identiquement distribuées ϵ_t de variance σ^2 on a sous l'hypothèse b=1 :

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

et

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

On a alors :

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

ce qui viole l'hypothèse d'indépendance des unités statistiques de l'échantillon soumis à la régression.

Eberhardt (1970) démontre que $E(\widehat{b}) < 1$ 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 biomathématiciens. 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 (Legay, 1973), et repose sur des techniques très polymorphes (on trouvera un aperçu des techniques utilisées en écologie dans Jeffers, 1977).

3 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 :

1. la notion de modèle en statistique.
2. 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 bon 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'équation différentielles.

Remerciements

Je remercie R. Varro qui a attiré mon attention sur l'exemple de série divergente cité dans le texte.

Références bibliographiques

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. Functional Ecology, 1 : 57-61.

Chesson, P.(1978).- Predator-prey theory and variability. 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. Ecology, 51 : 306-310.

Innis, G.S. (Ed) (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, Londes.

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 memgamodel builders. Simulation, 22.

Parker, R.A. (1968).- Simulation of an aquatic ecosystem. Biometrics, 24 : 803-82.

Rao, C.R. (1972).- Linear statistical inference and its applications. Wiley, New-York.

Sabatier, R., Lebreton, J.D. et Chessel, D. (1989).- Principal Component Analysis with instrumental variables as a tool for modeling composition data. pp 341-352 in COPPI, R. et Bolasco, S. (Eds) Multiway Data Analysis. North Holland, Amsterdam.

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, G.S. (Ed).- Grassland simulation model. Ecological studies n° 26, Springer-Verlag, New-York.