

UN REGARD SUR L'ANALYSE DE L'INFORMATION EN HALIEUMETRIE

Camille Duby^a

I- INTRODUCTION

Lorsque j'ai accepté de présenter cette session du 2ème Forum Halieumétrique, j'ai été confrontée au problème de comprendre et déterminer l'originalité et la spécificité de l'halieumétrie pour l'analyse de l'information.

Je crois savoir que l'halieumétrie a deux objectifs liés, l'un scientifique, l'autre économique, et de lourdes responsabilités socio-économiques puisque ses résultats servent à éclairer les politiques et donc ont des répercussions importantes sur toutes les populations qui vivent de la pêche.

L'halieumétrie a une très ancienne tradition de modélisation pour l'étude de l'évolution des stocks. Le recueil de l'information pose des problèmes particuliers et difficiles pour essentiellement trois raisons. La première est que l'espace à explorer, même si on se limite à des zones de pêche (concept pas toujours très clair) est gigantesque, hétérogène et de dimension 3, la troisième dimension étant, il me semble, traitée de façon assez discrète. A ces trois dimensions spatiales, il faut ajouter la dimension du temps. La deuxième vient du fait que la ressource est mobile et multispécifique. La troisième est que le milieu est complexe et ses interactions avec la ressource sont peu ou mal connues.

Cette difficulté du recueil de l'information a bien sûr une incidence sur son traitement. Il me semble que, contrairement à d'autres domaines, ce n'est pas le manque de données qui est le problème majeur, mais la qualité des données. J'entends par qualité des données le fait qu'elles soient complètement ciblées sur les questions posées. La quasi impossibilité d'expérimenter (sauf en aquaculture qui a beaucoup de points communs avec l'agriculture) fait que souvent ces données recouvrent une énorme quantité de facteurs plus ou moins parasites car on ne peut pas les contrôler vraiment, ni même parfois les mesurer.

^a - Mathématiques Appliquées à la Biologie - Institut National Agronomique - 16, rue Claude Bernard - 75231 Paris cedex 05

Quand on prend l'ensemble des articles qui vont être présentés au cours de cette session, à première vue, on ne distingue pas de spécificités de l'halieumétrie en comparaison avec la biométrie par exemple, ou même avec la statistique. Les halieumètres sont « à la page » en utilisant des méthodes parmi les plus récentes comme les réseaux de neurones, l'approche symbolique, l'analyse canonique fonctionnelle et l'analyse d'images. Les modèles linéaires généralisés, les modèles additifs interactifs généralisés et des modèles stochastiques sont aussi présentés.

La dimension spatiale apparaît explicitement, en se référant beaucoup aux travaux de géostatistique plus anciens.

Pourtant si on regarde de plus près cet ensemble de travaux, on peut s'apercevoir qu'ils ne vont pas dans tous les sens et on peut dégager des lignes de force qui peuvent être intéressantes pour mieux comprendre les préoccupations des halieumètres.

II- LIGNES DE FORCE DE CES TRAVAUX

La première qui apparaît est le besoin qui semble actuellement fort de se dégager du « linéaire », d'une part pour proposer des modèles qui permettent une meilleure prédiction, d'autre part pour mieux expliquer certains phénomènes. Les méthodes issues de cette volonté recouvrent les modèles non linéaires classiques, les modèles linéaires généralisés, les modèles additifs généralisés, les modèles non linéaires de séries chronologiques et les méthodes de réseaux neuronaux. Les présentations qui en sont faites sont exploratoires et comparatives pour "voir" ce que l'on peut attendre d'elles.

La deuxième ligne de force et qui me paraît très porteuse pour l'avenir de l'halieumétrie, est la volonté d'appropriation des méthodes par certains halieumètres. J'entends par là, une recherche de modélisations de base complètement adaptées au contexte et prenant en compte le plus possible les connaissances sur la ressource, le milieu et les interactions entre la ressource et le milieu, ce qui conduit à une originalité de la démarche d'analyse de l'information. Il est certain que ces recherches sont intimement liées au recueil des données, aidant à vraiment comprendre ce que ces données représentent. Dans ces travaux, l'effort est plus porté sur la prise en compte des spécificités des problèmes posés, l'agencement des méthodes utilisées pour l'analyse découlant assez naturellement de la façon dont les questions sont formalisées.

III- MODELES DE REGRESSION ET RESEAUX NEURONAUX

Nous utilisons ici le terme générique de régression pour parler de tous les modèles statistiques dont le but est de prédire ou d'expliquer une variable, uni ou multidimensionnelle, qualitative ou quantitative, à l'aide d'autres variables.

Si l'on se réfère à la très intéressante revue faite par Bing CHENG et D. M. TITTERINGTON [1994] sur les réseaux de neurones vus dans une perspective statistique et les commentaires sur cette revue, on s'aperçoit que ces deux types de méthodes, régressions et réseaux de neurones sont en fait étroitement liées, les premières étant d'origine statistique et les deuxièmes d'origine informatique. Je ne vais pas ici présenter ces approches. Je vais seulement essayer de faire apparaître leurs proximités.

On peut en effet montrer que si on appelle y la variable de sortie à prédire, x le vecteur des variables d'entrée et w le vecteur des poids associés aux connexions, le traitement avec un réseau neuronal revient à écrire le modèle $y = f(\phi(x, w))$ pour un seul neurone, et pour un réseau à plusieurs couches le modèle $y = f(\phi(u, w))$, en notant u_k la sortie de la k ème des M unités cachées et $u_k = g_k(\psi_k(x, v_k))$ $k = 1, \dots, M$. La plupart du temps ϕ est linéaire (mais parfois il peut être utile de faire un changement de variable), et f est choisie parmi quelques fonctions particulières.

On peut considérer les réseaux de neurones comme un cadre général des modèles de régression (par exemple les modèles de "projection poursuit" introduits par FRIEDMAN [1981]. Dans cette recherche d'analogies, on peut aussi retrouver les problèmes des modèles non linéaires de séries chronologiques.

On peut mieux voir ainsi les possibilités d'exploration du domaine non linéaire par ces méthodes et leurs capacités d'ajustement ; on peut mieux mesurer leur coût en nombre de paramètres en jeu, et par voie de conséquence le nombre de données nécessaires (en général très grand étant donné le nombre de paramètres) pour avoir des estimations convenables ainsi qu'une certaine robustesse des résultats. Et ce point est très important pour comparer les méthodes.

Les réseaux de neurones n'ayant pas au départ de structure aléatoire, le problème de la confiance à attribuer aux résultats, de leur validation est traité avec les méthodes de validation croisée utilisées aussi par les statisticiens. On peut imaginer obtenir aussi des variances ou des régions de confiance pour les prédictions en poussant jusqu'au bout l'idée de validation croisée, en utilisant des méthodes de Jackknife ou de Bootstrap.

Un autre aspect est aussi la recherche de qualités explicatives des modèles qui est souvent souhaitée. Les modèles non linéaires classiques, non pas construits "en aveugle" mais fondés sur des hypothèses de mécanismes plausibles ont souvent cette qualité. Les modèles issus des méthodes neuronales pourront peut-être aussi les avoir à condition qu'on rentre dans le "ventre" de ce qu'ils font, et non pas en les utilisant seulement comme des boîtes noires.

IV- CONCLUSION

L'halieumétrie, à travers le petit échantillon dont j'ai disposé ici, paraît bien vivante. Son avenir me paraît lié à la capacité des halieumètres d'intégrer les méthodes développées ailleurs, dans la problématique de l'halieutique et de ne pas se contenter de les plaquer sans rentrer au cœur de l'information qu'ils possèdent et du sens qu'elle peut avoir.

BIBLIOGRAPHIE

CHENG (B.) and TITTERINGTON (D.M.), 1994 - Neural Networks : A Review from a Statistical Perspective (with comments), *Statistical Science* 9 2-54.

FRIEDMAN (J.H.) and STUETZLE (W.), 1981 - Projection pursuit regression, *J. Amer. Statist. Assoc.* 76 817-823.