Deuxième Forum Halieumétrique, Nantes, 1995
*Session II : **Analyse de l'information***

MODELE NEURONAL VERSUS REGRESSION MULTIPLE DE PREDICTION DES NIDS DE TRUITE

M. Delacoste[a] , S. Lek[b], P. Baran[a], I. Dimopoulos[b] et J.L. Giraudel[b]

I - INTRODUCTION

Les modèles d'habitat (déterministes ou stochastiques) mettant en relation les variables du milieu et les caractéristiques des populations piscicoles (abondance, potentiel de reproduction,...) sont d'excellents outils d'aide à la décision. Faush *et al.*, 1988, recensent 95 modèles de ce type. La plupart sont construits à partir de régressions linéaires. Cette procédure sous entend une linéarité des relations entre variables qui sont plutôt rares en écologie.

Dans ce travail, nous nous proposons de comparer la capacité prédictive de la régression multiple et du réseau neuronal (connu par sa capacité à traiter des relations non-linéaires). Les valeurs prédites par les modèles seront comparées à des valeurs observées des données biologiques : prévision de la densité de frayères (nid) de truites communes *(Salmo trutta* L.*)* à partir de 10 variables d'habitat, dans 6 rivières pyrénéennes (SW France).

II- MATERIEL ET METHODES

A- Données biologiques

29 stations d'études réparties sur 6 rivières des Pyrénées centrales ont été subdivisées en 205 faciès d'écoulement. Les caractéristiques physiques de ces faciès ont été mesurées en janvier, immédiatement après la période de reproduction de la truite commune (*Salmo trutta* L). Ainsi, les caractéristiques mesurées reflètent le plus fidèlement possible les conditions rencontrées par la truite lors de sa reproduction.

a - Laboratoire d'Ingénierie agronomique, équipe Ichtyologie appliquée, ENSAT, 145 avenue de Muret, 31076 Toulouse, FRANCE.
b - Laboratoire de Biologie quantitative, Univ. Paul Sabatier, 118 route de Narbonne, 31062 Toulouse cedex, FRANCE. E-mail : lek@cix.cict.fr

B- Traitement des données

1) Régression linéaire multiple (RM)

La technique de RM progressive a été utilisée. Nous effectuons également la RM avec la totalité des variables (pour une comparaison avec les RN - réseaux des neurones-). Les calculs ont été effectués avec le logiciel Systat.

2) Réseau de neurones (RN)

Nous proposons ici une méthode de modélisation basée sur l'un des principes de réseaux neuronaux, l'algorithme de rétropropagation (Rumelhart *et al.*, 1986). Nous utilisons un réseau de 3 couches : une couche d'entrée de 10 neurones codant 10 variables explicatives, une couche de sortie d'un seul neurone codant la variable dépendante (densité de frayères) et entre les deux une couche intermédiaire dont le nombre de neurones est choisi empiriquement. Tous les neurones d'une couche donnée, sauf ceux de la dernière couche, émettent un axone vers chaque neurone de la couche en aval. A chaque connexion entre les neurones de deux couches successives est associé un poids modifiable au cours de l'apprentissage (itérations successives) en fonction des données en entrée et en sortie.

La technique de rétropropagation s'apparente à un apprentissage supervisé (pour apprendre, le réseau doit connaître la réponse qu'il aurait dû donner). Elle modifie ensuite l'intensité de connexion de manière à minimiser l'erreur de la réponse considérée. Pour éviter les phénomène de surapprentissage (modélisation du bruit), nous avons utilisé un réseau à 8 neurones intermédiaires et nous avons arrêté l'apprentissage à 1000 itérations.

3) Techniques de modélisation

La modélisation se fait selon deux étapes :

- dans le premier temps, nous modélisons avec la totalité des 205 enregistrements disponibles dans le jeu de données.
- dans le deuxième temps, nous avons procédé à des tirages aléatoires pour obtenir chaque fois un ensemble d'apprentissage (¾ des enregistrements, soit 154) et un ensemble de validation (¼ des enregistrements, soit 51) sur la totalité des observations. L'opération a été répétée 5 fois donnant lieu a 5 épreuves que nous étudions en RN et en RM. pour chacun de ces 5 jeux, nous effectuons un calage du modèle avec l'ensemble d'apprentissage et nous testons ensuite ce modèle avec l'ensemble de validation.

III- RESULTATS

A- Calage du modèle

En RM, les valeurs du coefficient de détermination (R^2) indiquent une nette amélioration du modèle après transformation des variables (tableau 1). Cette dernière opération améliore en effet la linéarité entre les différentes données. En incluant toutes les variables explicatives disponibles dans le modèle, nous n'avons qu'une très légère augmentation de R^2.

En RN, les modèles établis avec les mêmes variables, montrent une très nette amélioration des résultats, surtout dans le cas des variables non transformées.

Tableau 1. Coefficients R^2 obtenus par les modèles RM et RN en fonction du nombre de variables explicatives (transformées ou non).

Nb de Variables explicatives	Avec Transformation		Pas de Transformation	
	RM	RN	RM	RN
4	0.643	0.741	0.444	0.928
10	0.650	0.811	0.469	0.958

En RM, on a une sous-estimation des fortes valeurs et surestimation des faibles valeurs (fig. 1). Soulignons également la difficulté pour le modèle RM à prédire des valeurs nulles qui sont traduites sur le graphique par une bande verticale. Il faut noter enfin la prédiction de valeurs négatives, surtout pour les faibles valeurs.

En RN avec 4 variables explicatives, on obtient une sous estimation de nombreuses fortes valeurs. Comme dans le cas de RM, le modèle a des difficultés à prédire des valeurs nulles, mais uniquement lorsqu'on transforme les variables. Dans le cas des variables non transformées, les points sont mieux ajustés. En RN avec la totalité des variables, le problème reste posé avec les variables transformées pour les valeurs nulles, ainsi que certaines faibles valeurs. Par contre, pour le cas des variables non transformées (données brutes), nous obtenons un excellent modèle capable de restituer les valeurs observées sur toute l'étendue de la variable dépendante. Notons enfin que le RN, contrairement à la RM, ne prédit pas de valeurs négatives.

En conclusion, pour obtenir le meilleur modèle, il faut utiliser un réseau de neurones avec la totalité des variables disponibles et sans transformations.

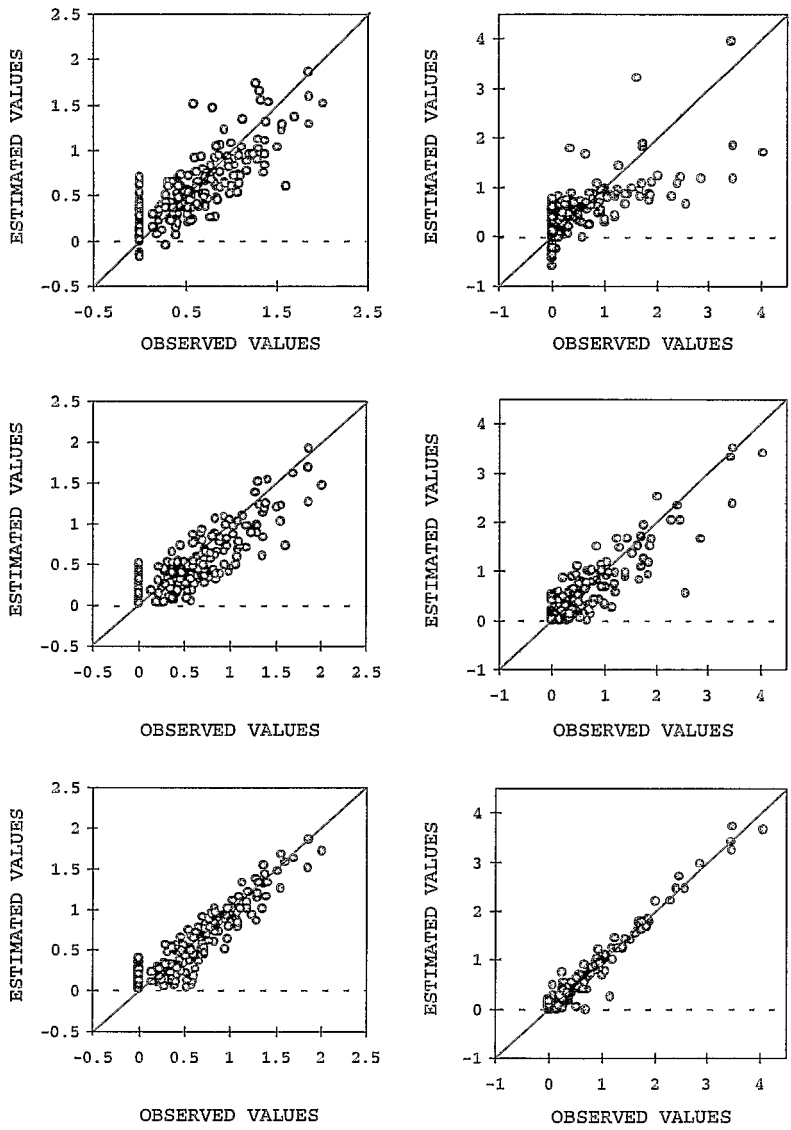

Figure 1. Graphe de corrélation entre les valeurs observées et les valeurs estimées par différents types de modèles. En haut : RM; au milieu : RN avec 4 variables; en bas : RN avec la totalité des variables. A gauche : avec transformation de variables; à droite : sans transformation de variables.

B- Test du modèle

Les résultats obtenus sont présentés dans le tableau 2. La puissance de prédiction des différents modèles déterminés à partir de 5 fractions d'apprentissage a été testée sur 5 fractions test indépendants.

En RM, les R^2 sont faibles dans les 2 jeux. R^2_{learn} est en moyenne de 0.468 pour l'ensemble d'apprentissage et R^2_{test} = 0.371 dans l'ensemble de validation. En RN, le R^2 est élevé aussi bien dans le jeu d'apprentissage que dans le jeu de validation (moyennes de R^2_{learn} = 0.81 et de R^2_{test} = 0.785). Les valeurs de R dans les deux jeux sont élevées et leurs différences sont faibles.

Tableau 2. Coefficient de corrélation du lot d'apprentissage (R_{learn}) et du lot de validation (R_{ltest}).

Numéro Test	RN		RM	
	R_{learn}	R_{ltest}	R_{learn}	R_{ltest}
1	0.892	0.862	0.685	0.487
2	0.914	0.888	0.685	0.628
3	0.904	0.906	0.690	0.626
4	0.883	0.867	0.688	0.566
5	0.905	0.906	0.669	0.740
Moyenne	0.900	0.886	0.684	0.609

IV - CONCLUSION

En écologie, la régression multiple est un des procédés de modélisation prédictive le plus utilisé à l'heure actuelle. Elle est simple à mettre en oeuvre si les relations entre les variables sont linéaires. Si ces relations sont non-linéaires, un travail préliminaire de transformation des variables est nécessaire. On peut également combiner des variables ou en éliminer certaines pour essayer d'aboutir à un modèle capable de donner une meilleure prédiction. Malgré toutes ces transformations, les résultats obtenus sont souvent insatisfaisants (prédiction de valeurs négatives, dépendance des résidus, ...).

Les réseaux neuronaux constituent une nouvelle approche en écologie. Ils sont capables de travailler avec des variables en relations non-linéaires. Ils sont relativement faciles à mettre en oeuvre et ne posent aucune contrainte sur les variables (normalité, linéarité, etc.). Si la transformation des variables permet d'améliorer les résultats en réseaux multiples, les réseaux neuronaux donnent de

meilleurs résultats avec des variables non transformées.

A travers cet exemple emprunté à l'ichtyologie, les réseaux neuronaux apparaissent comme une alternative puissante et performante aux méthodes traditionnelles de régressions multiples.

BIBLIOGRAPHIE

FAUSH (K.D.), HAWKES (C.L.), PARSONS (M.G.), 1988. Models that predict the standing crop of stream fish from habital variables, U.S. Forest Service General Technical Report PNW-GTR - 213p.

RUMELHART (D. E.), HINTON (G. E) and WILLIAMS (R. J.), 1986. Learning representations by back-propagating error. *Nature*, 323, 533-536.