

**PREDICTION DE LA COMPOSITION SPECIFIQUE DU
PEUPEMENT PISCICOLE EN MILIEU FLUVIAL PAR DES
METHODES CONNEXIONNISTES
APPLICATION AU BASSIN DE LA SEINE**

Philippe Boët^a et Thierry Fuhs^b

I- INTRODUCTION

Les communautés ichtyologiques sont l'expression de processus biologiques fondamentaux (reproduction, nutrition, abri...) s'exprimant à différentes échelles de temps et d'espace. Elles peuvent être considérées à juste titre comme de bons indicateurs de l'état de santé des écosystèmes aquatiques (Fausch, 1990 et al.).

L'identification, la hiérarchisation et l'évaluation des différents facteurs-clés responsables de l'état actuel de ces communautés sont des outils indispensables à la conservation ou à la restauration des peuplements et des milieux.

Une première synthèse concernant le bassin de la Seine a permis de dégager les facteurs majeurs contribuant à l'organisation actuelle du peuplement piscicole à l'échelle de l'ensemble du réseau hydrographique (Belliard, 1994), les caractéristiques de milieu liées à l'organisation longitudinale et régionale du bassin s'avérant déterminantes.

Ces travaux s'appuient essentiellement sur l'exploitation d'une importante base de données déjà acquises, regroupant des résultats d'inventaires *in situ* réalisés par pêche à l'électricité (plus de 700 pêches réparties sur 583 stations, représentant plus de 200 000 poissons appartenant à 39 espèces). La nature de ces données est très hétérogène. D'une part, parce qu'elles résultent d'échantillonnages mis en oeuvre pour répondre à des objectifs variés, et d'autre part, à cause des biais inévitables inhérents à la technique de capture utilisée, notamment dans le cas de grands cours d'eau.

Si elle s'avérait efficace, la mise en relation des descripteurs du milieu et des espèces piscicoles présentes, permettrait de préciser davantage l'importance relative de ces variables de milieu dans les mécanismes de structuration des

a - Cemagref Antony, division qualité des eaux, 14 avenue de Saint-Mandé, 75012 Paris

b - Cemagref Antony, division électronique et intelligence artificielle, Parc de Touvoie, BP 121, 95185 Antony Cédex

communautés et de tester l'impact de différentes perturbations. Pour y parvenir, la modélisation serait d'un grand recours mais celle-ci se heurte à l'extrême complexité des systèmes étudiés.

C'est pourquoi nous avons tenté d'utiliser des réseaux connexionnistes, capables de résoudre des problèmes non-linéaires et robustes vis-à-vis de jeux de données bruitées ou incomplètes, afin de prédire la composition d'un peuplement piscicole en fonction des caractéristiques du milieu étudié.

II- METHODOLOGIE

L'objectif de notre modèle prédictif est donc de relier caractéristiques du milieu et présence-absence de poissons. Il s'agit d'un problème de reconnaissance des formes, pour lequel l'utilisation de réseaux connexionnistes multi-couches associés à l'algorithme de rétropropagation du gradient ont montré leur intérêt (Rumelhart & McClelland, 1986). Nous avons donc privilégié cette démarche dans notre approche initiale.

Néanmoins, la mise en oeuvre effective d'une telle méthodologie présuppose un examen plus attentif des données disponibles. Idéalement, en effet, un réseau multi-couches devrait prendre en entrée les paramètres de milieu (une quinzaine de variables) et en sortie la présence-absence de toutes les espèces (39 au total). Ceci représente $15 \times N \times 39$ paramètres à calculer (les poids des connexions), où N est le nombre d'unités de la couche cachée, ce qui est illusoire avec seulement 700 pêches disponibles.

C'est pourquoi, nous avons diminué la taille des réseaux de deux manières :

- en entrée, par l'analyse préalable des corrélations entre variables qui nous a conduit à n'en retenir que 8 ;
- en sortie, en effectuant l'étude espèce par espèce, ce qui a permis de n'avoir qu'une unité en sortie, au détriment bien entendu du nombre de réseaux entraînés.

Pour la couche cachée, nous avons recherché par tâtonnement la meilleure valeur du nombre d'unités.

Il a en outre été nécessaire de se prémunir des biais d'échantillonnage. Ceux-ci peuvent être de deux types :

- d'une part, les données sont en général mal réparties entre présence et absence. La classe la plus nombreuse a alors une influence artificiellement plus grande dans le calage des poids du réseau.

- d'autre part, le découpage de la base des pêches en base de test et base d'apprentissage ne peut être *a priori* homogène du fait du nombre limité de pêches. Ceci peut entraîner des taux de généralisation inutilisables car sans rapport avec la distribution sous-jacente.

Pour le premier biais, nous avons choisi de multiplier aléatoirement les exemples de la classe la moins nombreuse de manière à obtenir pour chaque espèce, une répartition équitable entre présence et absence (en fait un rapport entre 0,4 et 0,6 a été considéré comme satisfaisant).

Pour le second, nous avons effectué une "validation croisée par blocs". Cela signifie que nous avons d'abord découpé l'ensemble de la base en trois parties de cardinal égal. Puis, nous avons entraîné chaque réseau sur les deux tiers de la base et l'avons ensuite testé sur le troisième tiers pour estimer sa performance en généralisation. Ceci était répété trois fois, en permutant les trois parties. Nous avons enfin calculé le taux de généralisation moyen et l'écart-type associé.

III- RESULTATS

Eprouvée à l'échelle du bassin de la Seine et en fonction de descripteurs très globaux de la qualité du milieu aquatique (8 variables synthétiques d'entrée), la prédiction de la présence d'espèces par des réseaux connexionnistes multicouches s'avère pertinente.

Bien qu'il soit en théorie possible de traiter toutes les espèces ensemble, le calage des paramètres, qui, outre des temps de calcul considérables, aurait exigé un plus grand nombre d'exemples, a entraîné la nécessité de construire un réseau par espèce.

18 ont été testées, choisies parmi les plus représentatives sur le bassin.

Le nombre de neurones requis dans la couche cachée varie selon les espèces (de 3 à 8), révélant ainsi différents paysages d'erreur, constitués de régions assez stationnaires pour de nombreuses espèces ou plus tourmentées pour certaines, paraissant pouvoir se rapporter à leur écologie.

Alors que les données d'entrée sont assez fortement bruitées, les taux de réussite en généralisation varient de 70 à plus de 85 % selon les espèces, ce qui représente des performances très appréciables, excepté pour le goujon et la truite arc-en-ciel, à cause de leurs particularités écologiques.

IV- CONCLUSIONS ET PERSPECTIVES

Ces résultats sont très encourageants car à l'heure actuelle, il n'existe pas encore de modèles prédictifs "poissons" à l'échelle du bassin fluvial.

Ces résultats se concentrent sur la prédiction d'une espèce donnée, alors que nous aspirions à travailler directement sur le peuplement. Ceci ne pouvant être fait avec toutes les espèces en même temps, il faudrait retraiter les données des pêches afin d'identifier les différents types de peuplement en place et entraîner ensuite des réseaux multicouches où la sortie ne serait plus une espèce particulière mais un peuplement donné. Les problèmes sont donc avant tout de définir écologiquement ces peuplements.

Déjà très satisfaisants compte tenu de la nature des données traitées, ces résultats devraient pouvoir être encore améliorés. En particulier, malgré l'influence connue des facteurs régionaux sur la composition du peuplement, le paramètre "région écologique" n'a pas été pris en compte à cause des problèmes de codage liés à son caractère purement qualitatif. Mais il conviendrait cependant de vérifier la qualité des résultats en ajoutant ce paramètre.

En outre, l'importance relative de certaines variables d'entrée vis-à-vis des différentes espèces mérite encore d'être précisée. La considération des classes d'abondance relative des différentes espèces devrait également conduire à de meilleures prédictions.

Néanmoins il serait d'ores et déjà intéressant d'étudier avec un tel réseau connexionniste, les conséquences de changements de milieu d'origine naturelle ou anthropique sur la composition des peuplements de poissons à l'échelle du bassin hydrographique. Parmi les variables d'entrée certaines décrivent en effet la morphologie du milieu ou sa position dans le gradient amont-aval et ont un caractère figé. D'autres au contraire traduisent une perturbation (physique ou chimique) et sont susceptibles de constituer un premier élément du diagnostic d'un éventuel déséquilibre du peuplement piscicole en place ; encore très synthétiques actuellement, comme par exemple la note de qualité de l'eau, elles pourraient être décomposées afin d'affiner un tel diagnostic.

A terme, des applications concrètes pourraient être envisagées comme par exemple établir le peuplement théorique de référence en un lieu donné, permettant de mesurer ensuite d'éventuels décalages, ce qui constituerait alors un précieux outil de gestion.

BIBLIOGRAPHIE

- BELLIARD (J), 1994 - Le peuplement ichtyologique du bassin de la Seine : rôle et signification des échelles temporelles et spatiales. Thèse Doct. Paris VI, : 197 p.
- FAUSCH (K.D.), LYONS (J.), KARR (J.R.) & ANGERMEIER (P.L.), 1990 - Fish communities as indicators of environmental degradation. p. 123-144, In : S. M. Adams (Ed.), *Biological indicators of stress in fish*, American Fishery Society Symposium 8.
- RUMELHART (D.E.) & MCCLELLAND (J.L.), 1986 - *Parallel distributed Processing*, MIT Press, Cambridge, MA.