Modélisation prédictive des peuplements de poissons

Philippe Boët (*Cemagref* Antony, UR qualité et fonctionnement hydrologique des systèmes aquatiques)

Thierry Fuhs (*Cemagref* Antony, Laboratoire d'ingénierie pour les systèmes complexes)

Pour approfondir notre connaissance du peuplement piscicole du bassin de la Seine et préciser davantage l'importance relative des variables de milieu dans les mécanismes de structuration des communautés, la modélisation est une étape indispensable. Une telle modélisation permettrait notamment de simuler l'impact de différents aménagements. Celle-ci se heurte toutefois à la complexité des systèmes étudiés. Ces derniers sont en effet constitués de nombreuses composantes dont l'ensemble des interactions est encore mal connu. Les relations entre les poissons et les descripteurs physiques de l'habitat ne sont, par exemple, pas *a priori* linéaires. Parmi les différentes possibilités qui s'offrent, nous avons donc recherché des modèles non-linéaires et robustes, c'est-à-dire peu sensibles au bruit des données, mieux adaptés à nos données.

Les réseaux de neurones fournissent un exemple de tels modèles. Nous présentons ici les résultats d'une première utilisation de cet outil séduisant par sa simplicité d'application. Nous détaillons certaines difficultés pour sa mise en œuvre effective. Nos premiers résultats sont satisfaisants. Cette approche s'avère pertinente pour la prédiction des poissons à l'échelle d'un bassin comme la Seine et des possibilités de couplage avec le modèle Riverstrahler notamment s'offrent déjà. D'intéressantes perspectives se dessinent également qui méritent également d'être développées.

1. Présentation des réseaux de neurones artificiels

Comme leur nom l'indique, les réseaux de neurones ont une origine marquée par la biologie du cerveau. L'objectif de McCulloch et Pitts (1943) était de simuler le fonctionnement du cerveau humain à l'aide de composants simples, les neurones formels, interconnectés en grand nombre (figure). L'idée sous-jacente était que le tout est plus puissant que la somme des parties.

McCulloch et Pitts ont en quelque sorte réussi puisqu'ils ont montré que ces réseaux, à l'instar de la machine de Turing, permettaient de représenter toute fonction calculable.

Cette caractéristique est toujours d'actualité, mais ce qui a fait le succès de ces modèles réside dans leur capacité à modéliser des phénomènes non-linéaires. Or, les régressions statistiques habituellement utilisées sont linéaires et ne peuvent donc qu'approcher d'assez loin des phénomènes fortement non-linéaires. Bien entendu, les statisticiens classiques ne sont pas complètement dépourvus devant ces non-linéarités (modèles linéaires généralisés, estimations non-paramétriques, etc.) mais, par leur simplicité d'utilisation, les réseaux de neurones ont constitué depuis leur avènement des compétiteurs crédibles. D'ailleurs, les ponts entre les deux communautés sont de plus en plus actifs.

Voyons d'où vient ce comportement non linéaire.

La fonction de transfert d'un neurone isolé est déjà non linéaire. Elle s'écrit en effet

$$y_i = f\left(\sum a_i x_i\right)$$

où f est une fonction sigmoïde. On peut néanmoins montrer que ce type de fonction est équivalent à un séparateur linéaire dans le cas d'un problème de discrimination, comme en relève la détection de la présence du poisson.

Par contre, lorsque nous combinons de tels neurones formels en plusieurs couches, la fonction de transfert devient beaucoup plus puissante. Les x_i représentent les données du problème, dans notre cas

les caractéristiques de l'habitat. $\hat{y}_i$ représente la prédiction du système, ici une valeur binaire de présence ou d'absence de poissons. En supposant une seule couche cochée, la valeur prédite de chaque $\hat{y}_i$ s'écrit :

$$\hat{y}_k = f\left(\sum_j w^s_{jk} t_j\right) = f\left(\sum_j w^s_{jk}\left(\sum_i w^e_{ij} x_i\right)\right)$$

où t_j sont des valeurs d'activations des neurones de la couche cachée, w^e_{ij} le poids de la connexion entre le neurone d'entrée i et le neurone caché j et w^s_{jk} le poids de la connexion entre le neurone caché j et le neurone de sortie k.

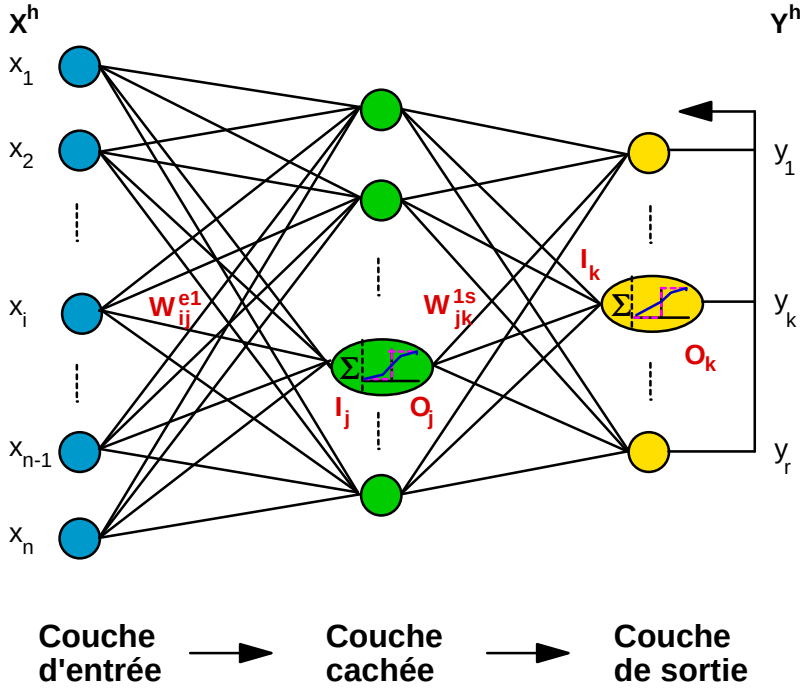

Figure 1. *Principe de fonctionnement d'un réseau de neurones. $X^h = (x_1,..., x_n)$ est le vecteur des entrées, $Y^h = (y_1,..., y_r)$ le vecteur des sorties désirées. W^{e1} et W^{1s} sont les matrices des poids. I_j : entrée du neurone j de la couche cachée, I_k : entrée du neurone k de la couche de sortie, O_j : sortie du neurone j de la couche cachée, O_k : sortie du neurone k de la couche de sortie.*

Des résultats mathématiques ont montré que si l'on considère n'importe quelle fonction continue des entrées vers la sortie, et une précision ε donnée, si petite soit elle, on peut trouver, parmi tous les réseaux à une couche cachée ayant le même nombre d'entrées et de sorties, un réseau situé à une distance inférieure à epsilon de cette fonction. Décrites ainsi, les choses semblent idylliques. Malheureusement, ce résultat n'est qu'un résultat d'existence : il ne donne d'aucune manière le nombre de neurones nécessaires sur la couche cachée pour approcher à moins de epsilon la fonction recherchée ! Et bien entendu, plus la précision demandée est forte, plus grand est le nombre de neurones cachés. Par contre, à architecture fixée, plusieurs algorithmes permettent de calculer les poids des connexions à partir de l'échantillon des (x, y) considéré. Le plus célèbre est la rétropropagation du gradient qui converge effectivement vers un minimum local de l'erreur. Ces algorithmes minimisent l'erreur entre la valeur prédite $\hat{y}_i$ et la valeur observée y. L'erreur quadratique est la plus utilisée :

$$\left(\sum_k y_k - \hat{y}_k\right)^2$$

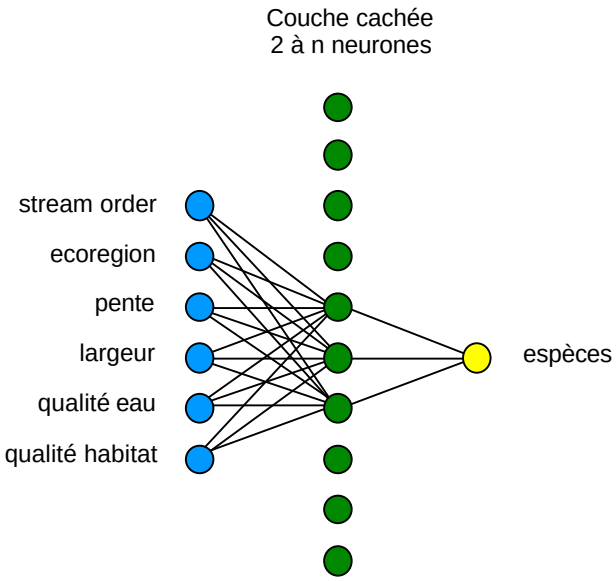

Figure 2. *Schéma des réseaux mis en œuvre.*

2. Matériel et méthode

2.1. Données utilisées et choix des variables d'entrée

Nos travaux s'appuient sur une importante base de données qui couvre l'ensemble du bassin de la Seine. Ces données sont le résultat d'échantillonnages réalisés au moyen de pêches électriques (plus de 700 pêches, réalisées dans 583 stations). Il s'agit d'abondances relatives d'espèces représentant plus de 200 000 poissons, appartenant à 39 espèces.

Ces données sont hétérogènes et bruitées, d'une part parce qu'elles résultent d'échantillonnages répondant à des objectifs différents et, d'autre part en raison de biais liés à la pêche électrique, dont l'efficacité est limitée dans les grands cours d'eau. Elles manquent de précision mais ont l'avantage d'être comparables et de couvrir un vaste espace.

Pour chacune des stations de pêche, une quinzaine de variables décrivent les caractéristiques du cours d'eau et de son environnement immédiat. Parmi celles-ci, nous avons sélectionné les plus pertinentes sur la base de nos travaux antérieurs (Belliard 1994, Belliard *et al*. 1997). Elles sont au nombre de 6 :

- Le *rang fluvial* (Strahler 1957) est un paramètre de position de la station de pêche au sein du gradient amont-aval dans le réseau hydrographique. Par son caractère très synthétique, il rend compte de nombreuses variables physiques et fonctionnelles du cours d'eau.

- L'*écorégion* est la région naturelle, homogène au plan des grands facteurs écologiques (climat, géologie…), où se situe la station de pêche. Nous avons utilisé le deuxième niveau hiérarchique de la carte des régions phyto-écologiques de France au 1/100000, proposée par Dupias et Rey (1985), qui découpe le bassin de la Seine en 7 régions homogènes.

- La *pente* et la *largeur* sont des descripteurs morphologiques classiques en hydrobiologie (p. ex. Huet 1949, Illies & Botosaneanu 1963).

- La *qualité de l'eau* est la note de l'Agence de l'Eau Seine-Normandie. Celle-ci varie de 1 à 5 selon la dégradation de la qualité physico-chimique du cours d'eau.

- La *qualité de l'habitat* est un indice synthétique qui rend compte du degré d'altération de l'habitat physique : lit majeur et lit mineur, nature des berges et du substrat, degré

d'artificialisation de l'écoulement. Cet indice est issu de la synthèse de l'ensemble des schémas de vocation piscicole du bassin de la Seine, réalisé par AREA (1992).

Les descripteurs des stations n'ayant pas toujours été renseignés au moment même de la pêche, seules les données les plus récentes sont utilisées. Ceci conduit à réduire la base de données aux 507 pêches postérieures à 1980.

Enfin, parmi les 39 espèces recensées dans la base de données, seules les 26 plus fréquentes (ie occurrence 9%) sont considérées.

2.2. Mise en œuvre logicielle et architecture

L'objectif de notre modèle est donc de prédire la présence ou l'absence de poissons à partir des caractéristiques du milieu. Le problème posé est un problème de discrimination, pour lequel l'utilisation de réseaux connexionnistes multicouches entraînés par l'algorithme de rétropropagation du gradient a montré son intérêt. Cette démarche a donc été privilégiée au cours de cette approche.

Nous avons choisi d'utiliser l'implémentation MASS décrite par Ripley (1996) et Venables & Ripley (1997), dont les bibliothèques pour les logiciels S-PLUS ou R sont librement distribuées sur l'Internet.

La mise en œuvre effective de cette technique apparemment facile s'avère toutefois délicate. En effet, ces modèles d'apprentissage offrent une très grande richesse de structure mais n'apportent à l'utilisateur aucune aide méthodologique, même empirique, pour dimensionner correctement un réseau en fonction du problème à résoudre.

Le nombre d'unités de la couche d'entrée du réseau est lié aux variables prédictives choisies. Une seule est purement qualitative, l'écorégion, qu'il faut découper en autant d'unités d'entrée que de modalités moins un (ie 6 modalités - 1). Les variables quantitatives étant au nombre de 5, la couche d'entrée du réseau comprend donc 10 unités.

Initialement, notre ambition était de construire un modèle capable de prédire en une seule fois les 26 espèces sélectionnées. Nous aurions ainsi pris en compte les corrélations éventuelles entre espèces. Avec une unique couche cachée de N neurones, ceci représente 37Nx26 paramètres à calculer (les poids des connexions). Au regard des 507 pêches d'exemples disponibles, d'éventuels problèmes de pertinence statistique se posent alors dès que N devient modérément grand (6 à 8).

C'est pourquoi, en sortie, l'étude est effectuée sur une seule espèce à la fois ce qui permet de n'avoir qu'une unité. Ceci, bien entendu, au détriment du nombre de réseaux entraînés, puisqu'il faut autant de réseaux que d'espèces, au lieu d'un seul pour toutes les espèces.

En outre, les interactions entre espèces sont négligées malgré leur éventuelle importance.

Pour la couche cachée, la meilleure valeur du nombre de cellules est recherchée par essais successifs.

2.3. Méthodologie expérimentale

Comme toute inférence statistique, les réseaux de neurones artificiels sont soumis au dilemme biais-variance ou dilemme de l'apprentissage (Geman et al., 1992).

On pourrait en effet penser que l'idéal est de considérer l'espace de recherche le plus grand possible, c'est-à-dire les réseaux ayant le plus grand nombre possible de neurones cachés. Malheureusement, cela n'est pas une stratégie gagnante.

En effet, si nous recherchons un modèle complexe, donc un réseau ayant beaucoup de paramètres, nous aurons plus de chance de trouver un réseau proche de la fonction inconnue et ainsi le biais pourra être faible. Par contre, si cet espace de recherche est très grand, on pourra trouver de nombreux réseaux

qui collent à l'échantillon d'apprentissage, réseaux pouvant être contradictoires sur de nouveaux échantillons : la variance sera alors forte.

À l'inverse, si le modèle recherché est simple, le biais obtenu sera fort (par exemple, un séparateur linéaire pour un problème fortement non-linéaire) mais la variance restera bonne les résultats n'étant que peu modifiés sur un nouvel échantillon.

La capacité de généralisation d'un réseau, c'est-à-dire son pouvoir prédictif sur de nouveaux exemples (de nouvelles pêches) est alors le résultat du compromis obtenu sur la richesse de l'espace de recherches. Un espace trop riche introduit une variance forte entre les échantillons alors qu'un espace trop pauvre nous laisse nous éloigner d'une solution satisfaisante.

Plusieurs techniques sont proposées pour sélectionner le “ meilleur ” modèle, il s’agit en particulier des méthodes de ré-échantillonnage comme le bootstrap (Efron & Tibshirani 1993) ou la validation croisée.

Pour estimer une erreur non biaisée du taux d’erreur en généralisation nous avons donc effectué une validation croisée par 1/5. La variabilité étant relativement importante, nous avons moyenné les résultats d’au moins 10 validations-croisées pour chaque combinaison espèce-nombre d’unités dans la couche cachée.

Comme beaucoup de méthodes de discrimination, les réseaux de neurone tendent à donner un poids important à la classe majoritaire. Les données étant, en général, inégalement réparties entre présence et absence, la classe la plus nombreuse peut alors avoir une influence excessive dans le calage des poids du réseau. Dans le cas extrême de la carpe commune, par exemple, présente dans moins de 10 % des pêches, le réseau prédisait systématiquement son absence... et, bien sûr, se trompait rarement !

Pour éviter ce biais, nous avons choisi de pondérer les exemples de la classe la moins nombreuse afin d'obtenir, pour chaque espèce, une répartition des exemples équilibrée entre présence et absence. Nous avons ainsi affecté aux observations de la classe minoritaire le coefficient suivant :

$$\left\lfloor 2 \times \frac{\#maj}{\#\min} \right\rfloor \Big/ 2$$

Pour chaque espèce, le protocole mis en œuvre est donc le suivant :

1. Pondération différenciée des exemples afin d’équilibrer présence et absence
2. Répartition des exemples en 4/5 et 1/5 tirés au hasard
3. Apprentissage du réseau pour chaque 4/5 de la base
4. Test du modèle sur chaque 1/5
5. Matrice de confusion par ajout des erreurs sur chaque 1/5
6. Classification des erreurs

Les étapes 2 et 3 sont répétées au moins 10 fois pour calculer la moyenne et l’écart-type de l’erreur de classification. Le protocole est mis en œuvre pour chaque taille de couche cachée.

3. Résultats

Le tableau I résume les résultats obtenus avec un réseau comprenant 5 unités en couche cachée.

Les taux d'erreur de prédiction varie de 13% pour le chabot *Cottus gobio* et la truite *Salmo trutta fario*, à 32,4% pour le goujon *Gobio gobio.*

Nous cherché à choisir pour chacune des espèces le meilleur nombre d'unités en couche cachée et ceci s'est avéré plus délicat que prévu. La figure 3 montre le taux d'erreur de prédiction en fonction du nombre d'unités dans la couche cachée pour 4 espèces. Pour chacun des graphes, figurent 3 courbes d'erreur (v. également Tableau 1) : (i) le taux d'erreurs de prédiction pour l'absence de l'espèce, (ii) le taux d'erreurs de prédiction pour la présence de l'espèce, et le taux d'erreur totale des prédictions de l'espèce.

Dans la plupart des cas, l'erreur décroît assez rapidement avec l'augmentation du nombre de neurones, mais l'on observe ensuite un plateau qui ne permet pas de sélectionner de façon rigoureuse la meilleure architecture possible.

À l'avenir, nous pensons résoudre ce problème en pénalisant l'erreur de prédiction commise par le critère d'Akaiken et/ou un critère Bayesien (Schwarz 1978, Sakamoto *et al*. 1986). Ces critères permettent de juger le meilleur compromis entre nombre de paramètres du modèle (ie nombre de poids dans le réseau) et la précision obtenue en prédiction.

4. Discussion

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

Vingt six espèces sont testées, choisies parmi les plus représentatives présentes dans le bassin.

Alors que les données d'entrée sont assez fortement bruitées, les taux de réussite en généralisation varient de 67,6 à plus de 87 % selon les espèces. Ceci représente des performances très appréciables car une erreur de mesure de l'ordre de 10 à 20 % sur ce type de données est très probable.

Les meilleurs résultats sont obtenus pour quatre espèces la truite *Salmo trutta fario*, le chabot *Cottus gobio*, la loche franche *Nemacheilus barbatulus*, et le vairon *Phoxinus phoxinus*. Ces espèces sont aussi parmi celles dont les profils écologiques sont les plus nets sur le bassin de la Seine (Figure 4). Pour une classe i, la valeur du profil est égale à Vi = Fi - Ftot, où Fi est la fréquence relative de l'espèce dans les relevés de la classe i et Ftot la fréquence relative de l'espèce sur l'ensemble des relevés. Tous les profils sont significatifs (χ_2 ; $p<0,001$).

Ces espèces font partie du cortège faunistique classique rencontré dans les zones amont où les habitats restent parmi les moins perturbés.

Tableau 1. *Taux moyen et écart type d'erreur avec 5 neurones en couche cachée. err(0) : erreur de prédiction sur l'absence, err(1) : erreur de prédiction sur la présence, err(tot) : erreur de prédiction totale ; sd : écart type.*

Epèces	err(0)	s.d.(0)	err(1)	s.d.(1)	err(tot)	s.d.(tot)
Abramis brama	0,324	0,028	0,107	0,015	0,267	0,012
Alburnus alburnus	0,247	0,016	0,097	0,014	0,190	0,008
Anguilla anguilla	0,297	0,036	0,256	0,027	0,279	0,016
Barbus barbus	0,189	0,027	0,147	0,019	0,181	0,012
Blicca bjoerkna	0,227	0,040	0,165	0,030	0,214	0,017
Chondrostoma nasus	0,229	0,030	0,162	0,024	0,217	0,018
Cottus gobio	0,132	0,013	0,123	0,009	0,128	0,006
Cyprinus carpio	0,173	0,030	0,182	0,016	0,174	0,013
Esox lucius	0,126	0,046	0,364	0,019	0,224	0,011
Gasterosteus aculeatus	0,227	0,050	0,247	0,027	0,230	0,018
Gobio gobio	0,448	0,047	0,214	0,056	0,324	0,016
Gymnocephalus cernua	0,226	0,058	0,266	0,025	0,230	0,020
Lampetra planeri	0,221	0,048	0,254	0,027	0,225	0,022
Leuciscus cephalus	0,231	0,021	0,195	0,038	0,207	0,016
Leuciscus leuciscus	0,194	0,031	0,206	0,017	0,198	0,014
Lota lota	0,167	0,027	0,115	0,017	0,157	0,013

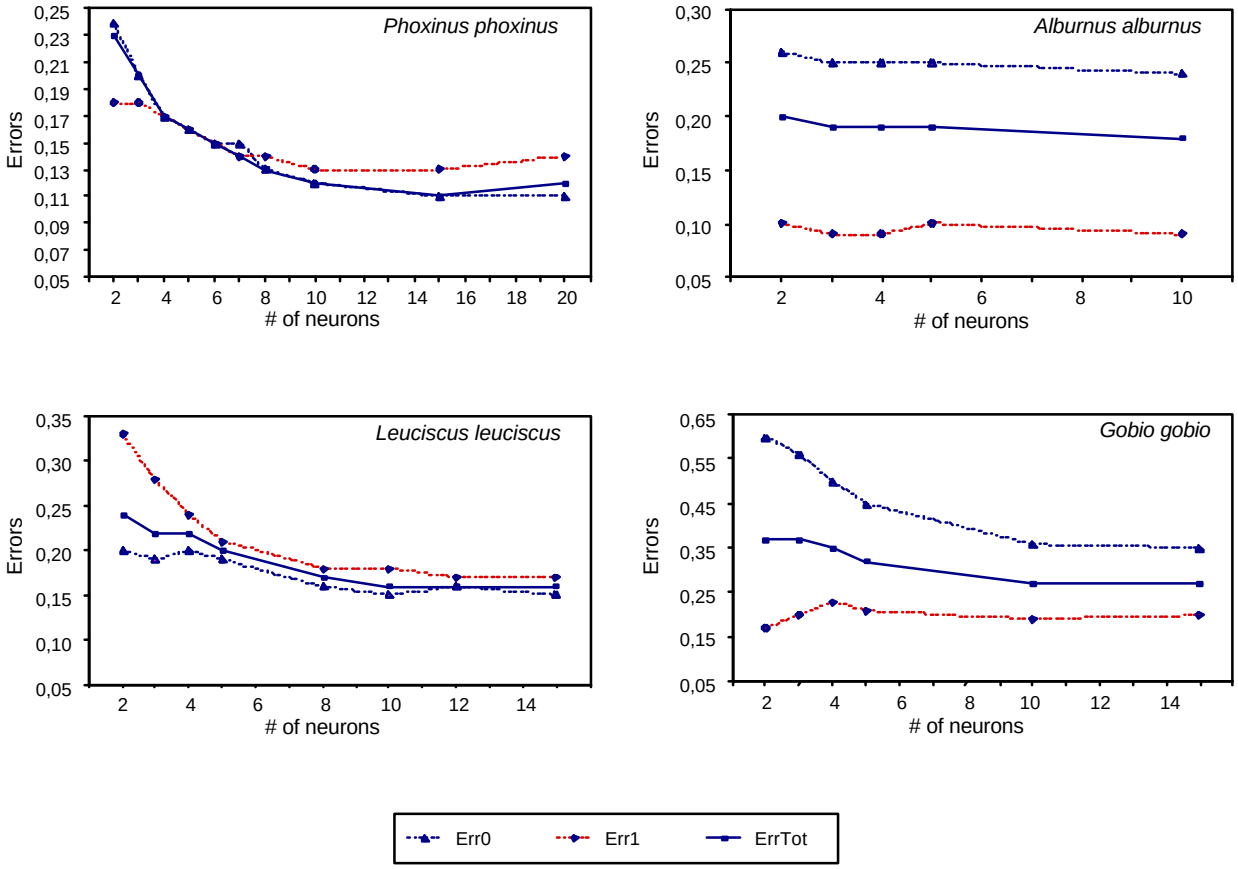

Figure 3. *Comportement de l'erreur de prédiction en fonction du nombre d'unités dans la couche cachée des réseaux. Err0 : erreur de prédiction sur l'absence, Err1 : erreur de prédiction sur la présence, ErrTot : erreur de prédiction totale.*

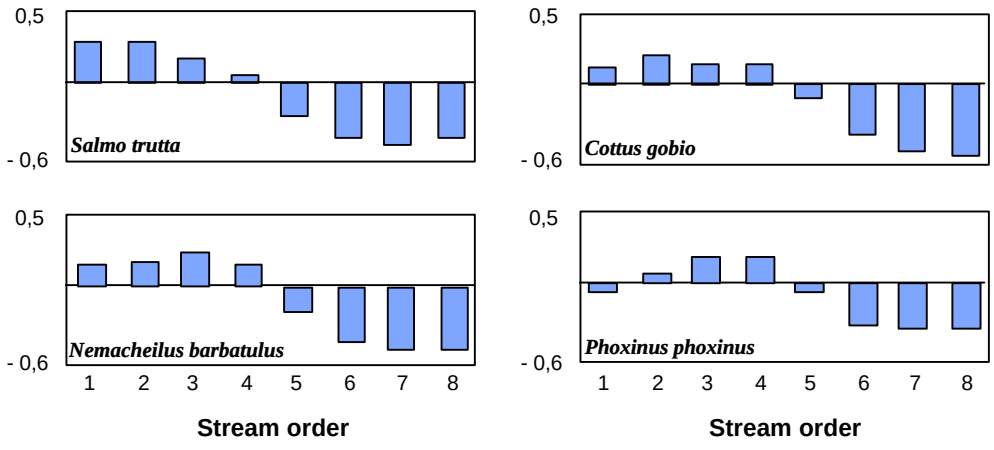

Figure 4. *Profil écologique de la truite Salmo trutta fario, du chabot Cottus gobio, de la loche franche Nemacheilus barbatulus, et du vairon Phoxinus phoxinus en fonction du stream order (d'après Belliard 1994).*

En revanche, le moins bon résultat concerne le goujon *Gobio gobio* (32,4 % d'erreur totale). Dans ce cas, un phénomène subtil peut expliquer les mauvaises performances du réseau. Dans le bassin

de la Seine, cette espèce se distribue en effet en deux groupes distincts (Figure 5). Dans les parties amont, des populations composées de petits individus sont classiquement inféodées à des eaux vives de bonne qualité s'écoulant sur des substrats de sable ou de gravier. À l'aval, dans des zones plus riches en matière organique, se trouvent au contraire de gros goujons pouvant parfois mesurer plus de 25 cm de longueur. Il conviendrait certainement de bien séparer ces deux sous-ensembles pour mieux entraîner les réseaux et améliorer leur qualité de prédiction.

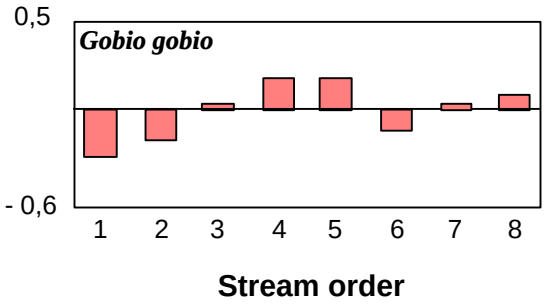

Figure 5. *Profil écologique du goujon Gobio gobio en fonction du stream order (d'après Belliard 1994). Pour une classe i, la valeur du profil est égale à Vi = Fi - Ftot, où Fi est la fréquence relative de l'espèce dans les relevés de la classe i et Ftot la fréquence relative de l'espèce sur l'ensemble des relevés. Tous les profils sont significatifs (c_2 ; p<0,001).*

De même, l'écologie de l'anguille *Anguilla anguilla* peut expliquer le faible résultat observé. En raison de la pollution de la Seine au cours des années 60, cette espèce avait fortement régressé sur le bassin. Actuellement, grâce à l'amélioration générale de la qualité des eaux suite aux efforts de traitement des rejets urbains, l'anguille est à nouveau présente en abondance, mais la répartition de cette espèce migratrice dans les parties amont du bassin notamment est très dépendante des obstacles liés aux barrages qui limitent ses déplacements.

Un examen approfondi des résultats obtenus est également intéressant. Dans le cas du brochet par exemple *Esox lucius* (Tableau 1) l'erreur totale peut paraître relativement faible, mais en fait la prédiction de la présence de ce poisson est mauvaise. Dans la plupart des cas en effet le réseau prédit l'absence de l'espèce alors que celle-ci est en réalité présente dans les échantillons capturés (Figure 6). Cette espèce faisant l'objet de nombreuses opérations de réempoissonnements par les associations de pêcheurs, on peut penser que sa présence n'est que partiellement dépendante des caractéristiques de l'environnement aquatique.

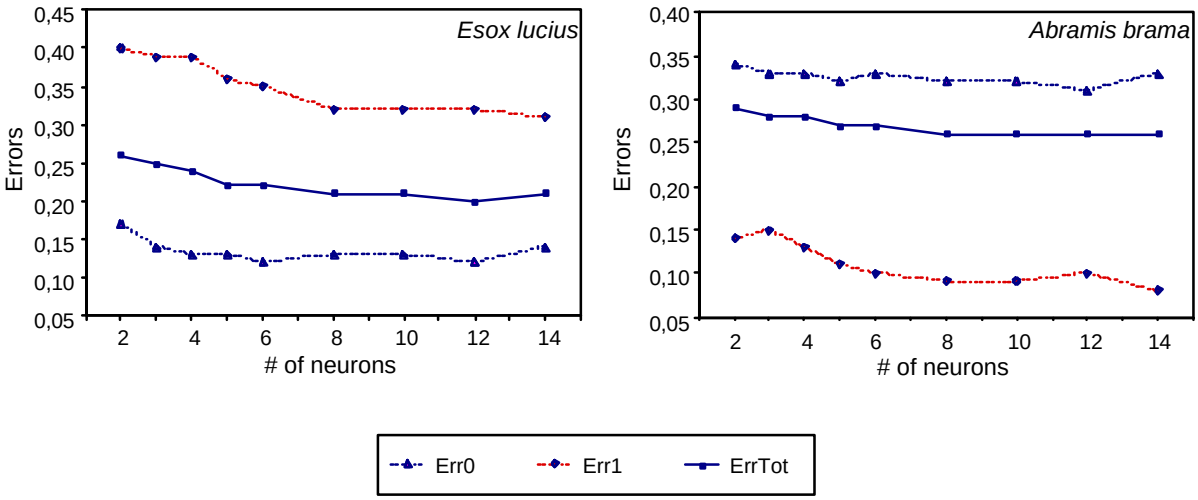

Figure 6. *Comportement de l'erreur de prédiction en fonction du nombre d'unités dans la couche cachée des réseaux. Err0 : erreur de prédiction sur l'absence, Err1 : erreur de prédiction sur la présence, ErrTot : erreur de prédiction totale.*

À l'inverse, dans le cas de la brème commune *Abramis brama*, l'erreur totale apparaît forte, mais ceci est essentiellement dû à la prédiction de la présence de l'espèce par le réseau alors que l'espèce ne figure en fait pas dans les captures (Tableau 1 et Figure 6). En raison de sa morphologie - corps aplati latéralement - et de son comportement benthique, ce poisson vivant généralement au fond des grands cours d'eau est difficile à capturer efficacement à la pêche électrique, même quand il est présent dans le milieu. Une situation similaire se rencontre également dans le cas de la tanche *Tinca tinca*, capable de s'enfouir dans les fonds vaseux. Ces deux exemples illustre l'importance de la représentativité de l'échantillonnage qui pose le problème, déjà largement débattu, de la signification écologique à donner à l'absence d'une espèce dans des relevés faunistiques.

Néanmoins, il est bien sûr possible également que les variables prédictives d'entrée du modèle ne soient pas suffisantes pour prédire efficacement ces différentes espèces et que d'autres caractéristiques soient nécessaire pour affiner leur prédiction.

5. Perspectives : de la boîte noire à la boîte de verre…

La prédiction obtenue à l'aide des réseaux de neurones est de bonne qualité. Les taux d'erreur s'échelonnent de 13% à 32%, ce qui est tout à fait intéressant lorsqu'on les compare aux erreurs de mesure qui peuvent fréquemment atteindre 20%. En ce sens, les réseaux de neurones se révèlent une technique non-linéaire tout à fait pertinente dans la prédiction de la faune ichtyologique à l'échelle du bassin versant.

Néanmoins, leur mise en œuvre n'est pas toujours aisée. Tout d'abord techniquement, car comme il apparaît sur les figures 3 et 6, les données à notre disposition ne permettent pas de trancher facilement quant à la taille optimale des réseaux construits. Ensuite, parce qu'ils apparaissent comme une boîte noire au biologiste : les coefficients du réseau ne peuvent en effet avoir une interprétation biologique.

Ce dernier point nous semble le plus préjudiciable dans la perspective d'une utilisation de ces modèles prédictifs pour l'aide à la gestion au niveau du bassin versant. Il est par exemple délicat de prédire l'impact d'un aménagement sur la présence d'une espèce particulière sans proposer " d'explication " à cet impact, et ce qu'il soit positif ou négatif.

C'est pourquoi, nous avons aussi testé une autre méthode de discrimination appelée " arbres de décision ". L'objectif de cette méthode est le partitionnement récursif de l'espace des observations en sous-domaines les plus homogènes possibles quant à la classe de leurs éléments. La construction se fait en partant de la partition triviale, contenant toutes les observations, à laquelle on attribue la classe majoritaire. On tente alors de séparer cet ensemble suivant une des variables de l'échantillon. Cette variable et la valeur de séparation sont déterminées de manière à engendrer des sous-ensembles plus homogènes quant à la classe de leurs éléments. Un nœud est alors construit qui contient le test sur la variable. Conventionnellement, si le test réussit, l'observation est affectée à la feuille gauche de l'arbre. Ce découpage est poursuivi récursivement sur chaque branche de l'arbre jusqu'à obtention de nœuds totalement homogènes ou d'effectifs trop faibles pour rester représentatifs. Un exemple d'un tel arbre est montré par la figure 7. Celui-ci modélise la présence-absence du brochet.

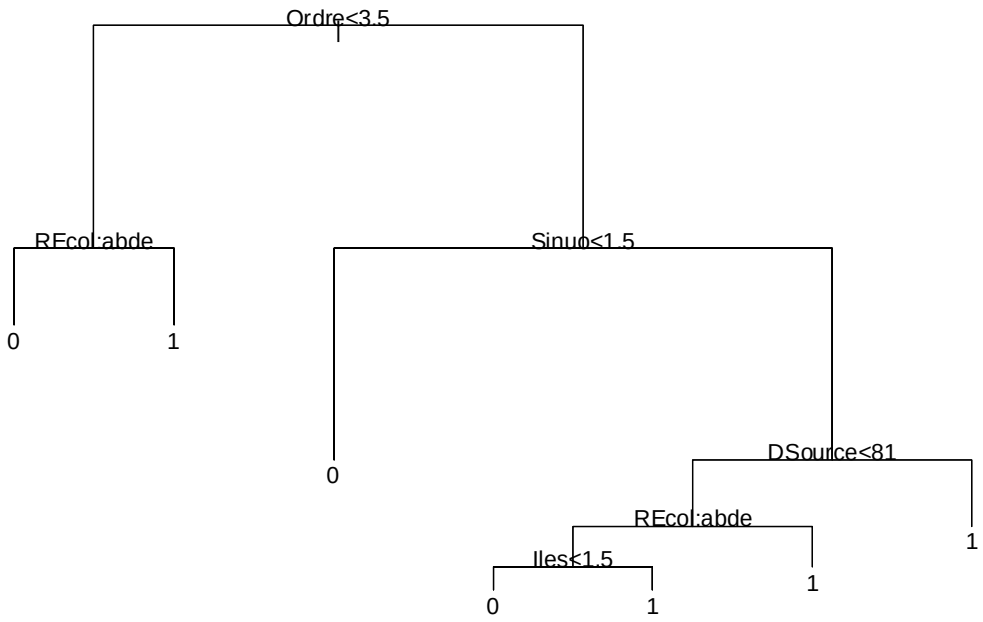

Figure 7. *Modélisation de la présence-absence du brochet au moyen d'arbres de décision. Dans cet exemple seules 5 variables sont retenues par le modèle parmi 19 variables possibles.*

Les avantages de la technique des arbres de décision sont multiples. En premier lieu, les paramètres sont tous explicites : les tests aux nœuds de l'arbre portent chacun sur une seule variable. Si celle-ci est quantitative, le test indique un seuil ; si elle est qualitative, le test indique l'appartenance à un sous-ensemble des modalités possibles de la variable. Dans les deux cas, ceci est immédiatement interprétable par le biologiste. Ainsi, le modèle s'apparente beaucoup plus à une boîte de verre ! En deuxième lieu, les arbres de décision font la sélection des variables les plus pertinentes. Si une variable n'apparaît en aucun nœud de l'arbre, c'est qu'elle est peu pertinente pour la discrimination recherchée. Cette sélection se fait automatiquement lors de la construction de l'arbre, puisque seules les variables discriminantes y sont retenues. Enfin, la qualité de la prédiction est aussi bonne voire meilleure que celle des réseaux de neurones. Ceci apparaît sur le tableau 2, lequel recense les résultats obtenus pour la présence du brochet. L'erreur de prédiction y est très comparable à celles des réseaux de neurones, mais l'arbre correspondant est plus compact que le réseau correspondant qui contient 8 unités sur la couche cachée.

Tableau 2. *Prédiction de la présence-absence du brochet : comparaison des méthodes réseaux de neurones et arbres de décision.*

Méthode	Variables	Err. moy	Err. min	Err. max	Taille
R.N.	6	20,8	13,1	32	8 cachés
			(0)	(1)	
Arbres	6 sur 6	20,3	13,4	40	15 feuilles
Arbres	5 sur 19	20,1	10	28,5	7 feuilles

6. Conclusions

Compte tenu de la nature des données traitées et du caractère très synthétique des variables d'entrée, ces premiers modèles de prédiction au moyen des réseaux de neurones s'avèrent très satisfaisants. Ils sont déjà très proches de ceux obtenus à l'aide de méthodes classiques, comme par exemple les analyses discriminantes et les régressions multiples utilisées par Pouilly (1994) ou Capra (1995). Mais ces derniers, travaillant à l'échelle du micro-habitat, disposent de données très fiables de description de l'habitat et d'échantillonnage de la faune en place.

Ces essais sont donc très encourageants si l'on considère qu'à l'heure actuelle il n'existe guère de modèles prédictifs de poissons à l'échelle d'un bassin fluvial. La zonation piscicole de Thienneman (1925), reformulée par Huet (1949, 1959) et connue comme la " règle des pentes ", permet seulement de distinguer quatre zones (à truite, ombre, barbeau, ou brème) le long du gradient amont-aval d'un cours d'eau. Plus récemment, Verneaux (1977, 1981) a proposé un calcul permettant d'approcher le groupement d'espèces caractéristiques, ou biocœnotype théorique, d'un secteur de cours d'eau, en fonction de différents paramètres, malheureusement pas toujours faciles à renseigner, comme par exemple " la température maximale moyenne du mois le plus chaud ".Néanmoins, ces résultats semblent encore susceptibles d'améliorations.

À notre connaissance, Oberdorff *et al*. (1998) viennent d'appliquer avec succès des procédures de régressions logistiques pour élaborer des modèles qui décrivent la présence des espèces piscicoles et la richesse des peuplements dans les grands cours d'eau français. La variabilité expliquée est dans ce cas aux alentours de 73%. Seuls, Mastrorillo *et al*. (1997, 1998) ou Guégan *et al*. (1998) ont déjà démontré les capacités des réseaux de neurones artificiels pour prédire la richesse spécifique des communautés de poissons à large échelle en fonction de quelques paramètres synthétiques du milieu.

Actuellement, nos résultats se concentrent sur la prédiction de présence ou d'absence d'une espèce donnée, alors que l'ambition initiale est de considérer directement tout le peuplement. Nous pensons pouvoir dépasser cette étape prochainement avec de nouvelles données complémentaires et pouvoir envisager des prédictions robustes à cette échelle du peuplement (ie richesse spécifique, abondances relatives des espèces).

L'intégration des données à un système d'information géographique est démarrée. Il s'agit à terme de spatialiser les modèles en tenant compte non seulement des coordonnées géographiques des tronçons mais aussi de la topologie arborescente du chevelu hydrographique. Cette intégration permettra ensuite de visualiser les peuplements à l'échelle du bassin puis de simuler l'impact des aménagements et de la gestion d'ouvrages existants sur ces peuplements.

Néanmoins il semble d'ores et déjà possible d'étudier, avec nos premiers modèles, les conséquences des 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, peuvent traduire une perturbation (physique ou chimique) et sont susceptibles de constituer un premier élément du diagnostic d'un éventuel facteur de déséquilibre du peuplement piscicole en place ; encore très synthétiques actuellement, comme par exemple la note de qualité de l'eau, ces variables pourraient être décomposées afin d'affiner un tel diagnostic.

Nous envisageons également de prendre comme entrées de ces modèles les données fournies par le modèle Riverstrahler. Il s'agirait là d'un tout premier couplage du compartiment poissons avec les modèles de processus établit à l'échelle du bassin de la Seine.

7. Références

AREA, 1992. Cartographie de synthèse des schémas départementaux de vocation piscicole et des ZNIEFF humides du bassin Seine -Normandie. Rapport DIREN Ile-de-France Délégation de bassin Seine-Normandie, Agence de l'Eau Seine-Normandie, décembre 1992, 64 p.

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.

BELLIARD J., BOËT P. & TALES E., 1997. Regional and longitudinal patterns of fish community structure in the Seine River basin, France. *Environ. Biol. Fish.*, 50, 133-147.

CAPRA H., 1995. Amélioration des modèles prédictifs d'habitat de la truite fario : échelles d'échantillonnage ; intégration des chroniques hydrologiques. *Thèse Doc. Univ. Claude Bernard - Lyon I*, 281 p.

DUPIAS G. & REY P., 1985. *Document pour un zonage des régions phyto-écologiques.* CNRS, Février, 39 p. + carte.

EFRON B. & TIBSHIRANI R.J., 1993. *An introduction to the bootstrap.* Chapman & Hall.

FAUSCH K.D., LYONS J., KARR J.R. & ANGERMEIER P.L., 1990. *Fish communities as indicators of environmental degradation. in* : S.M. Adams (Ed.), Biological indicators of stress in fish, American Fishery Society Symposium, 8, p. 123-144.

GEMAN S., BIENENSTOCK E. & DOURSAT R., 1992. Neural networks and the bias/variance dilemma. *Neural Computation*, 163-174.

GUEGAN J.-F., LEK S. & OBERDORFF T., 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. *Nature*, 391 (22), 382-391.

HUET M., 1959. Profiles and biology of western european streams as related to fish management. *Trans. Am. Fish. Soc.*, 88 (3), 155-163.

ILLIES J. & BOTOSANEANU L., 1963. Problèmes et méthodes de la classification et de la zonation écologique des eaux courantes, considérées surtout du point de vue faunistique. *Mitt. Intern. Ver. Limnol.*, 12, 1-57.

MASTRORILLO S., DAUBA F., OBERDORFF T., GUEGAN J.-F. & LEK S., 1998. Predicting local fish species richness in the Garonne River basin. *C.R. Acad. Sci. Paris, Sciences de la vie / Life Sciences*, 321, 423-428.

OBERDORFF T., PONT D., HUGUENY B., BOËT P., PORCHER J.-P. & CHESSEL D., 1998. *A probabilistic model characterizing riverine fish communities of French rivers : a framework for the adaptation of a fish based index. in* : M. Jungwirth, S. Schmutz & M. Kaufmann (Eds), Assessing the ecological integrity of running waters, Intern. Symposium, Vienna, Austria, 9-11 november 1998.

POUILLY M., 1994. Relations entre l'habitat physique et les poissons des zones à cyprinidés rhéophiles dans trois cours d'eau du bassin rhodanien : vers une simulation de la capacité d'accueil pour les peuplements. *Thèse Doc. Univ. Claude Bernard - Lyon I*, 256 p.

RIPLEY B.D., 1996. *Pattern recognition and neural networks.* Cambridge University Press, Cambridge, 403 p.

SAKAMOTO Y., ISHIGURO M. & KITAGAWA G., 1986. *Akaike Information Criterion Statistics.* Reidel Publishing Company.

SCHWARZ G., 1978. Estimating the dimension of a model. *Annal. Stat.*, 6, 461-464.

SOUCHON Y. & TROCHERIE F., 1990. *Technical aspects of French legislation dealing with freshwater fisheries (June 1984) "Fisheries orientation schemes" and "fishery resources management plans". in* : W.L.T. Van Densen, B. Steinmetz & R.H. Hughes (Eds), Management of freshwater fisheries, Proceeding of EIFAC Symposium, Göteborg, Sweden, 31 May - 3 June 1988, Pudoc Wageningen, p. 190-214.

VENABLES W. & RIPLEY B.D., 1997. *Modern Applied statistics with S-Plus.* (2nd ed.) Springer-Verlag, New-York, 548 p.

VERNEAUX J., 1977. Biotypologie de l'écosystème "eau courante". Déterminisme approché de la structure biotypologique. *C.R. Acad. Sc. Paris*, t. 284 (sér. D), 77-79.

VERNEAUX J., 1981. LES POISSONS ET LA QUALITE DES COURS D'EAU. ANN. SCI. UNIV. FRANCHE-COMTE, BESANÇON, BIOL. ANIM., 4EME SER. (FASC. 2), 33-41.

Sommaire général

Introduction du thème : Biodiversité

Inventaire des données disponibles sur le développement des macrophytes dans les cours d'eau d'amont du réseau hydrographique de la Seine.

Modélisation prédictive des peuplements de poissons

Caractéristiques des milieux stagnants et rôle sur le fonctionnement de la Seine.