

Modélisation prédictive des peuplements piscicoles, à l'échelle du bassin de la Seine

Philippe Boët, Thierry Fuhs, Guillaume Gorges & Laurent Toupotte (*Cemagref Antony*)

Il s'agit de prédire le peuplement piscicole en fonction des caractéristiques physiques, chimiques et biologiques du milieu fluvial et ses annexes en faisant appel à des techniques essentiellement statistiques. Les modèles prédictifs sont intégrés à un système d'information géographique permettant de visualiser les peuplements à l'échelle du bassin. À terme l'ambition est d'élaborer un outil capable de simuler l'impact d'aménagements ou de modifications des modalités de gestion d'ouvrages existants sur les peuplements, par exemple, la chenalisation au gabarit européen, l'installation d'un nouveau barrage-réservoir, voire l'effet de rejets chroniques.

Au cours de l'exercice 1999, le développement de modèles d'arbres de décision, couplés à une base de données spatialisées au moyen d'un SIG, a débouché sur un outil opérationnel. Mais sa mise en œuvre se heurtait cependant, d'une part, à l'insuffisance de données disponibles à l'échelle de l'ensemble du bassin de la Seine, et d'autre part, à la disparité et à l'hétérogénéité des données utilisées.

Cette année a donc essentiellement été consacrée à couvrir l'ensemble du bassin hydrographique, ainsi qu'à valider, homogénéiser et sécuriser les données disponibles. À cette fin, ont été menées parallèlement, l'importation et l'enrichissement du référentiel géographique commun au PIREN-Seine, ainsi que l'intégration des données descriptives des tronçons de cours d'eau et des données de pêches dans une base Oracle™.

Sur cette base, de nouveaux modèles ont été construits prenant en compte tout le réseau hydrographique du bassin de la Seine. Un essai de simulation a également été tenté.

1. Structuration spatiale des données

1.1. Données géoréférencées

La base géographique sur laquelle reposent les modèles prédictifs est actuellement constituée de 6 couches d'informations :

- le réseau hydrographique (échelle 1 : 50 000), fourni par l'Agence de l'Eau Seine-Normandie et commun à la base PIREN ;
- les stations de pêches, localisées et spatialisées ;
- les bassins versants, définis et numérisés sur la base de la couverture IGN du bassin (1 : 100 000) ;
- l'occupation du sol, définie par la base de données Corine Landcover (1 : 100 000) ;
- un modèle numérique de terrain (Gtopo 30) de résolution 650 m ;
- les écorégions de Dupias et Rey (1985) à l'échelle 1 : 100 000.

Ces différentes couches permettent de calculer diverses variables dérivées, en particulier, le rang fluvial (Strahler 1957) et la pente des tronçons de cours d'eau, la surface des bassins versants, la distance entre les différentes stations de pêche, ainsi que la distance de chaque station de pêche à l'exutoire final.

1.2. Constitution d'une base de données Oracle™

L'ampleur des données ainsi rassemblées a conduit à la nécessité impérieuse d'optimiser leur gestion au moyen d'une base de données relationnelle. Toutes sont maintenant rassemblées dans une base Oracle™ : variables dérivées du système d'information géographique, données physiques décrivant localement les caractéristiques des stations de pêche et données biologiques afférentes. L'intérêt de la constitution d'un tel outil est de disposer maintenant d'un référentiel validé et sécurisé ; celui-ci est en outre conçu pour être interrogeable à distance via un navigateur Internet.

Une centaine de variables, réparties en 7 tables, sont ainsi renseignées. Nous présentons ici les principales, utilisées ci-après pour la construction des arbres de décision.

La longueur du tronçon de cours d'eau, calculée par le SIG, est la longueur entre le nœud amont et le nœud aval de la polyligne représentant le tronçon de cours d'eau considéré.

La sinuosité du tronçon intègre la complexité du chevelu hydrographique dans le secteur de cours d'eau considéré, son degré d'artificialisation éventuelle (rectification du cours d'eau) ou l'augmentation potentielle du linéaire d'habitat pour la faune piscicole. La sinuosité est obtenue par le rapport de la longueur du tronçon sur la distance euclidienne du nœud amont au nœud aval du tronçon.

La pente du tronçon est également un descripteur morphologique du cours d'eau. Elle conditionne la vitesse d'écoulement des eaux et donc la répartition des espèces piscicoles (Huet 1949). Elle est calculée à partir des altitudes des nœuds amont et aval, puis lissée sur la longueur du tronçon.

Le rang fluvial de Strahler (1957) permet de hiérarchiser les tronçons de cours d'eau ; il rend compte de la taille de la rivière et de la densité du réseau hydrographique. Son calcul est obtenu par un parcours de graphe, à partir de la couverture du réseau hydrographique.

La surface de bassin versant décrit aussi la taille des cours d'eau, laquelle intervient de façon prépondérante dans la succession amont aval des peuplements piscicoles. À partir de la couverture de tous les bassins versants d'ordre 1, numérisée par le LGA, les surfaces élémentaires sont ensuite utilisées pour calculer les tailles des bassins versants de chacun des tronçons de cours d'eau du réseau hydrographique.

L'occupation du sol est un indicateur global du degré d'artificialisation du bassin versant et de l'altération éventuelle de l'habitat. À partir de la base Corine Landcover, cinq types d'occupation du sol sont retenus : les zones urbanisées et industrielles ; les surfaces agricoles ; la surface forestière ; la surface en eau et les zones humides. Le pourcentage de chaque type d'occupation du sol est calculé pour tous les tronçons de cours d'eau.

Les écorégions sont les régions naturelles drainées par le réseau hydrographique. Elles sont homogènes du point de vue des conditions physiques et chimiques et sont donc susceptibles d'influencer la composition de la faune piscicole. Les cartes utilisées sont les régions phyto-écologiques définies par Dupias et Rey (1985).

La qualité d'eau est issue d'une simulation par le modèle SENEQUE, considérant l'année hydrologique 1991 comme référence. Elle est intégrée dans la base de donnée sous la forme de la note obtenue selon les critères du SEQ-Eau de l'Agence de l'Eau et varie de 1 à 5 avec la dégradation de la qualité physico-chimique du cours d'eau. Cette note étant rapidement pénalisée dès qu'un paramètre est affecté, nous avons également utilisé la note obtenue pour le seul paramètre oxygène dissous, laquelle paraissant mieux rendre compte de la variation amont-aval de la qualité d'eau.

Les données biologiques sont enfin constituées des résultats de plus de 1100 pêches électriques, réalisées sur 578 stations. Sont renseignés, les effectifs de chaque espèce de poisson.

2. Méthode des arbres de décision

Les arbres de décision constituent une méthode statistique applicable aux domaines de la discrimination et de la régression. Le problème de la prédiction de la présence d'une espèce en fonction des caractéristiques de la portion de rivière considérée étant un problème de discrimination, nous nous limiterons à ce cas dans ce qui suit.

La discrimination statistique consiste à partir d'un échantillon de données représentatives du phénomène considéré, à construire un modèle permettant de prédire la classe d'une nouvelle observation. Les pêches électriques qui décrivent les caractéristiques physico-chimiques et morphologiques de la station ainsi que les espèces détectées sont un tel échantillon. La classe d'une observation (une variable qualitative) est ainsi la variable binaire indiquant la présence ou l'absence d'une espèce donnée.

2.1. Construction des arbres de décision

L'algorithme de construction des arbres de décision se déroule en deux phases successives.

- La première phase est un algorithme glouton de partitionnement récursif binaire. À chaque pas, est déterminée la meilleure séparation linéaire des observations de l'échantillon d'apprentissage puis récursivement sur chaque sous-ensemble de l'échantillon ainsi formés. La récursion stoppe dès que le sous-ensemble est homogène (toutes les observations sont de même classe), ou si son effectif est trop faible, ce seuil étant décidé par l'utilisateur.
- La seconde phase est « l'élagage » de l'arbre. Il s'agit de nettoyer l'arbre de ses branches les moins significatives statistiquement. En effet, tout modèle statistique construit à partir d'un échantillon est sujet à ce qu'on nomme « le dilemme de l'apprentissage ». Cela signifie que si le modèle est très complexe (dans notre cas, un arbre avec de nombreuses feuilles et branches), il sera trop « proche » de l'échantillon ayant servi à le construire et généralisera mal à des observations étrangères. À l'inverse, un modèle trop fruste (peu de branches et de feuilles) ne distinguera pas suffisamment les contours de la surface de décision (frontière entre les observations des différentes classes).

Breiman *et al.* (1984) ont montré qu'il existait un élagage optimal de tout arbre de décision. Cette propriété vient du fait qu'il est possible de classer les nœuds de tout arbre de décision suivant sa résistance à la pénalisation du critère d'erreur. Notons $R(T)$ ce critère non pénalisé, et $R_\alpha(T)$ le critère pénalisé obtenu par $R_\alpha(T) = R(T) + \alpha \text{NombreNœuds}(T)$. Il est immédiat que si nous faisons augmenter α , la taille de l'arbre le pénalisera de plus en plus. Breiman *et al.* ont alors montré qu'il existait une suite optimale d'arbres emboîtés issus de l'arbre construit dans la première phase et dont les éléments (des sous-arbres de cet arbre initial) correspondent à un α donné.

Muni de ce résultat théorique essentiel, il nous faut maintenant choisir le meilleur arbre de cette suite qui maximise la capacité de généralisation, c'est-à-dire la qualité du modèle sur de nouvelles données. Bien entendu, si nous estimons l'erreur à l'aide de l'échantillon initial, il est évident que c'est l'arbre initial qui est optimal par sa construction même. Mais l'erreur ainsi calculée appelée erreur empirique n'est qu'une approximation qui sous-évalue la véritable erreur de discrimination donnée par l'arbre. Une meilleure estimation de cette véritable erreur consiste à utiliser la « validation croisée ». Cette méthode consiste à découper l'échantillon initial en N parties égales ($N=2, 3, 5, 10$ sont des valeurs fréquemment choisies), à construire un arbre et la suite de ces arbres élagués sur chacun des N échantillons constitués de $N-1$ des parties ci-dessus. Alors, au lieu de calculer l'erreur empirique sur ces échantillons, nous la déterminons sur la N ème partie non utilisée dans le développement de l'arbre. En moyennant ensuite sur les N suites d'arbres, nous obtenons une estimation meilleure du critère pénalisé $R_\alpha(T)$ pour tout α . Breiman *et al.* ont montré que celui-ci est une fonction linéaire par morceaux de α , qui hormis les cas pathologiques, est convexe : elle diminue tout d'abord lorsque α augmente pour atteindre

un minimum, puis augmente ensuite. Le paramètre α correspondant au minimum est le paramètre déterminant finalement le meilleur arbre élagué.

2.2. Mise en œuvre

Les modèles de prédiction de la présence-absence des espèces piscicoles sont construits à partir des résultats de 1068 pêches électriques, réalisées dans 578 stations (**Figure 1**).

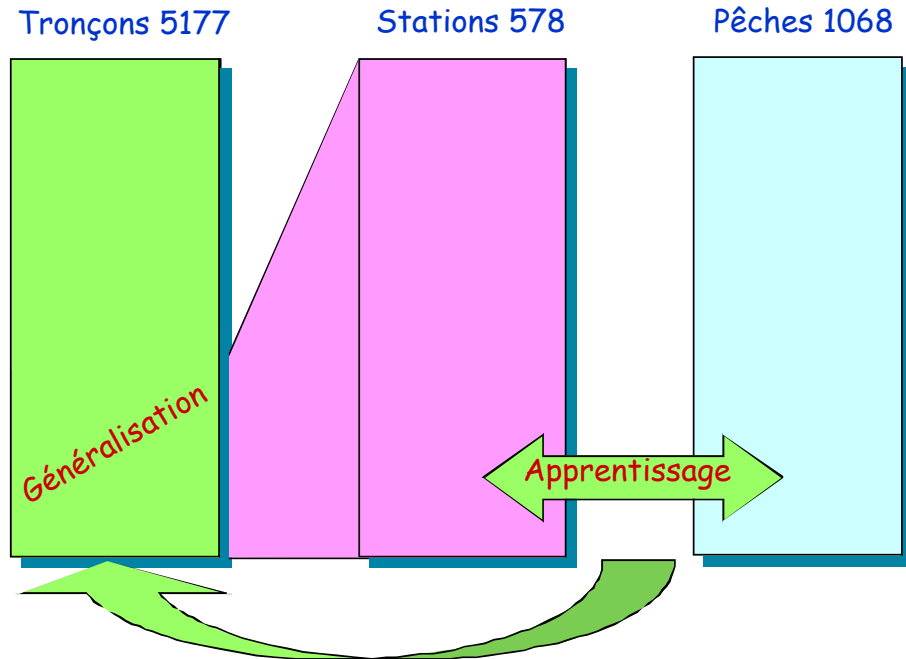


Figure 1. Méthodologie de construction des modèles de prédiction de présence-absence des espèces piscicoles sur le bassin.

Les routines mises en œuvre utilisent la bibliothèque RPART de S-Plus (Therneau & Atkinson 1997). Le meilleur arbre élagué est choisi par validation croisée en découpant le tableau initial des observations en 20 parties égales.

16 arbres correspondant aux principales espèces présentes sur le bassin ont ainsi été construits ; les erreurs totales en généralisation sont globalement modestes (**Tableau 1**).

Les performances des arbres de décision doivent toutefois être appréciées dans leur capacité à prédire tant la présence que l'absence des espèces. Par exemple, les très faibles valeurs concernant le poisson-chat ou le barbeau apparaissent ainsi peu significatives. En effet, comme ces deux espèces sont en fait peu fréquentes dans les observations considérées (10,8 % et 15 % respectivement), les erreurs commises relativement à leur présence sont en réalité très élevées (51,7 % et 46,3 % respectivement).

La modélisation du brochet est également meilleure pour l'absence (erreur de 12,9 %) que pour la présence (erreur de 25,1 %) qui tend à être sous-estimée au regard du contenu réel des pêches, mais ceci s'explique vraisemblablement par les nombreux empoisonnements dont l'espèce fait l'objet.

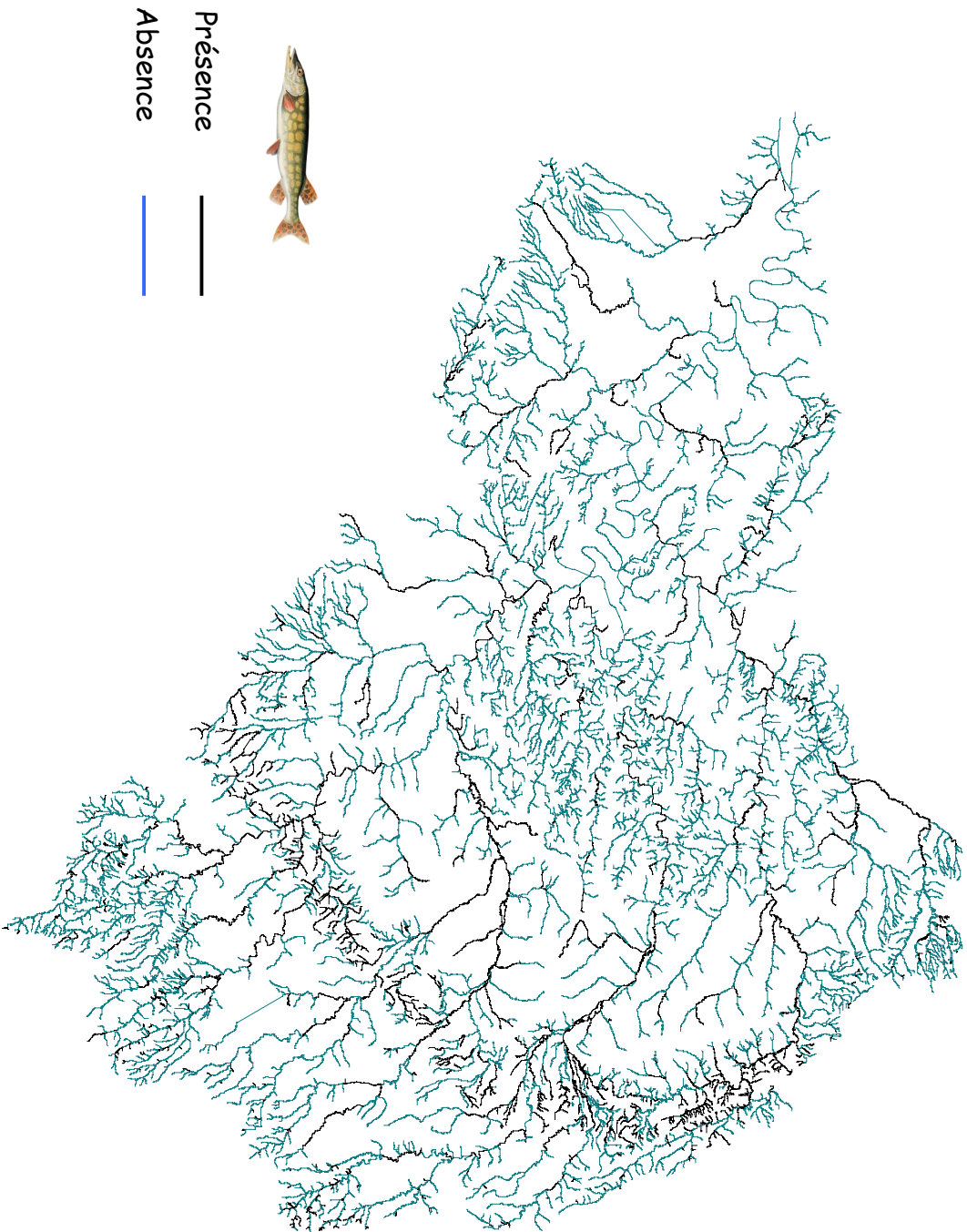


Figure 2. Exemple de généralisation de la prédiction du brochet à l'ensemble du réseau hydrographique de la Seine.

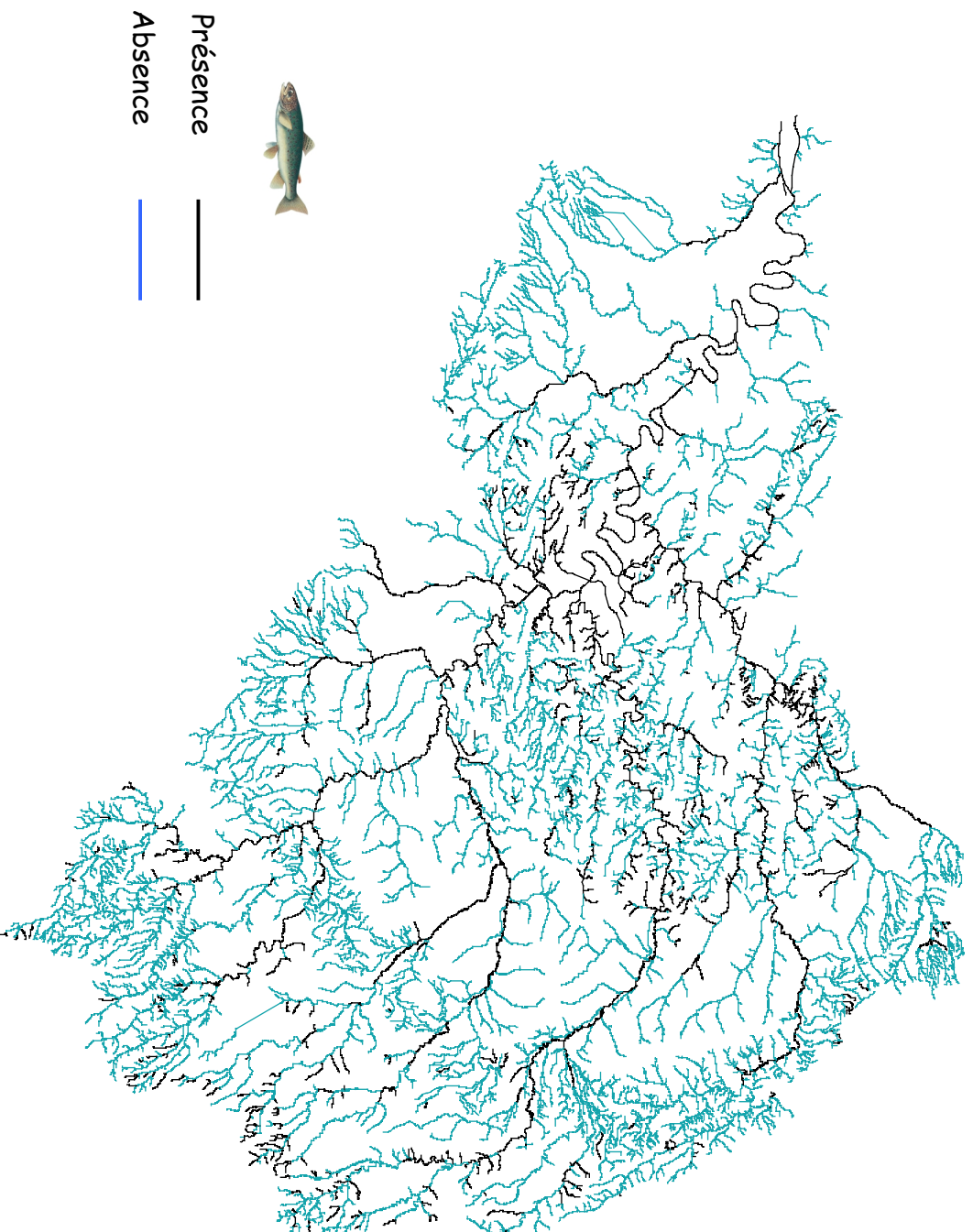


Figure 3. Exemple de généralisation de la prédiction de la truite à l'ensemble du réseau hydrographique de la Seine.

Tableau 1. Erreur totale des modèles des arbres de décision appliqués à la présence absence de 16 espèces de poisson de la Seine.

Espèce	Code	Erreur total
Poisson-Chat	PCH	0,067
Barbeau	BAF	0,095
Truite	TRF	0,114
Gardon	GAR	0,137
Chabot	CHA	0,146
Vairon	VAI	0,155
Hotu	HOT	0,168
Vandoise	VAN	0,171
Brochet	BRO	0,176
Loche franche	LOF	0,180
Brème bordelière	BRB	0,186
Perche	PER	0,187
Ablette	ABL	0,196
Chevesne	CHE	0,230
Anguille	ANG	0,233
Goujon	GOU	0,342

En revanche les prédictions pour la truite sont bonnes dans les deux cas, avec des erreurs de seulement 10,9 % et 12,5 % pour l'absence et la présence respectivement.

Tous ces résultats sont encore préliminaires car ils doivent être analysés plus en détail. En particulier le dépouillement et l'analyse précise des différentes variables prises en compte par les arbres de décision pour chaque espèce reste à faire.

Les illustrations de la généralisation des prédictions de la truite et du brochet à l'échelle du bassin de la Seine ne sont donc présentées qu'à titre d'exemple (**Figure 2** et **Figure 3**).

Dans un même esprit, un premier essai de simulation de scénario est également montré : il s'agit de la prise en compte de données de qualité de l'eau fournies par le modèle SENEQUE considérant des conditions optimales de traitement sur le bassin (abattement des phosphates et des nitrates). Une telle amélioration des conditions de milieu se traduit par une modification sensible prédite par le modèle, de la répartition du chevesne à l'aval de l'agglomération parisienne, comparé à la situation actuelle.

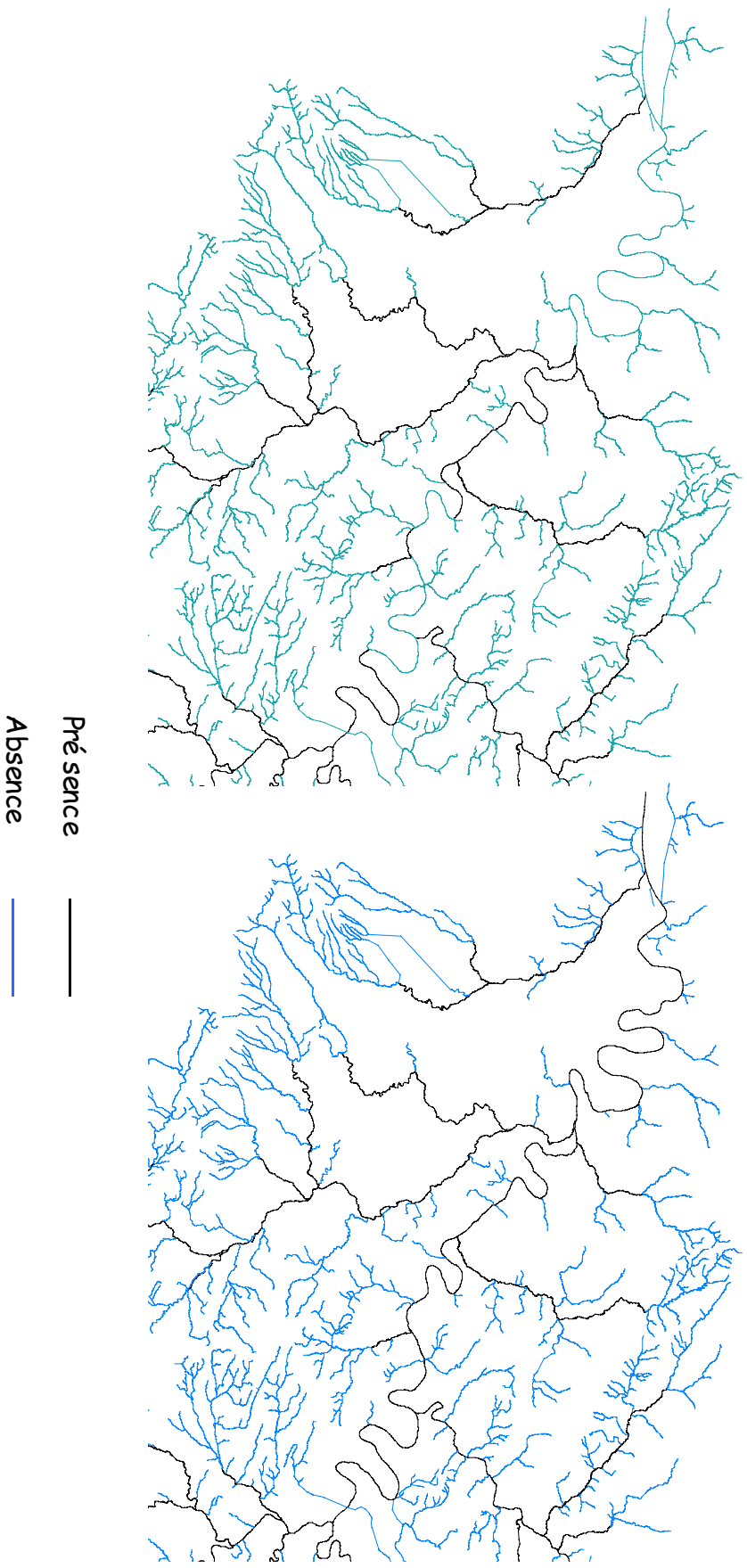


Figure 4. Exemple de prédictions du chevesne : dans les conditions de milieu actuelles (à gauche) ; en simulant des conditions idéalisées d'un traitement optimal des phosphates et des nitrates (à droite). L'espèce est alors prédite sur tout le cours aval de l'agglomération parisienne.

3. Conclusion et perspectives 2001

À cette étape, nous disposons donc d'une « l'architecture » générale qu'il convient en partie de consolider mais dont il nous faut surtout aménager maintenant « l'intérieur », pour finaliser nos modèles de prédiction des peuplements de poissons à l'échelle du bassin de la Seine.

Quelques variables explicatives doivent encore compléter cet ensemble, comme la température et l'hydrologie, paramètres majeurs de la distribution des espèces, voire la base INRA des pratiques agricoles, laquelle permettrait d'intégrer davantage l'anthropisation du bassin versant.

Il s'agit également de figurer l'incertitude des prédictions dans la représentation spatialisée des résultats.

Mais l'essentiel du travail doit maintenant porter sur l'analyse systématique et détaillée des réponses de chacune des espèces piscicoles aux variables des modèles. Doivent également être prises en compte, les abondances relatives des espèces afin de prédire les peuplements piscicoles de manière plus quantitatives. Les données récentes issues du suivi du Réseau Hydrobiologique et Piscicole du Conseil supérieur de la pêche sont enfin à intégrer et à utiliser pour valider les modèles.

Par ailleurs, une application détaillée des modèles sera également mise en œuvre sur le bassin supérieur de la Seine (Seine amont, Yonne, Marne), dans le cadre de l'évaluation fonctionnelle de la typologie des zones humides (rôle habitat).

L'exploration de scénarios d'aménagement, tant rétrospectifs que prospectifs, sera finalement menée en relation avec les actions transversales envisagées au sein du programme.

L'objectif est bien de rendre opérationnel un outil susceptible d'apporter des éléments d'aide à la gestion globale du bassin de la Seine, pouvant à terme s'appliquer également à d'autres contextes.

4. Références

- Breiman L., Friedman J.H., Olshen R.A. & Stone C.J., 1984. *Classification and regression trees*. Chapman & Hall, New York, 358 p.
- Dupias G. & Rey P., 1985. *Document pour un zonage des régions phyto-écologiques.*, CNRS, Février, 39 p. + cartes.
- Efron B. & Tibshirani R.J., 1993. *An introduction to the bootstrap*. Chapman & Hall.
- Fuhs Th., 1998. Les membranes linéaires par morceaux : une approche géométrique de la boucle abduction-induction dans les arbres et listes de décision. *Thèse Université Caen*.
- Huet M., 1949. Aperçu des relations entre la pente et les populations piscicoles des eaux courantes. *Scheiw. Z. Hydrol.*, 11 (3-4), 332-351.
- Quinlan J.R., 1987. C4.5 : programs for machine learning. Morgan Kaufmann.
- Strahler A.N., 1957. Quantitative analysis of watershed geomorphology. *Trans. Amer. Geophys. Union*, 38, 913-920.
- Therneau T.M. & Atkinson E.J., 1997. An introduction to recursive partitioning using the RPART routines. Technical report Mayo Foundation, September 3, 1997, [Distributed in Poscript with the rpart package], 52 p.