Modélisation prédictive des peuplements à l'échelle du bassin

Philippe Boët, Thierry Fuhs et Guillaume Gorges (*Cemagref*)

Enjeux et objectifs

Le poisson présente un caractère écologique intégrateur du fonctionnement du système fluvial. Il revêt en outre une importance sociale forte. La prédiction de la faune ichtyologique au niveau d'un bassin est donc une question importante pour les utilisateurs et les gestionnaires de la ressource en eau.

À partir de modèles de prédiction des communautés piscicoles nous proposons d'élaborer un outil de simulation capable de prédire les peuplements en fonctions de scénarios d'aménagements, de gestion des ouvrages, ainsi que de pollutions.

Il s'agit de prédire le peuplement piscicole en fonction des caractéristiques du milieu fluvial et ses annexes en faisant appel à des techniques essentiellement statistiques. Une intégration à un système d'information géographique permet ensuite de visualiser les peuplements à l'échelle du bassin.

Les données disponibles permettent d'envisager des modèles de prédiction de la richesse spécifique, de l'occurrence des espèces, ainsi que de la composition qualitative et quantitative des communautés. Ces différents modèles sont intégrés à un système d'information géographique capable de fournir une simulation cartographique dynamique.

À terme, il sera aussi possible de simuler l'impact des aménagements et de la gestion d'ouvrages existants sur ces peuplements, par exemple la chenalisation au gabarit européen, l'installation d'un nouveau barrage-réservoir ou l'effet de rejets chroniques.

Cette approche s'appuie sur des travaux préparatoires (Boët & Fuhs 1995), approfondis plus récemment (Boët & Fuhs, *in press*) sur la prédiction de la présence-absence d'espèces. Nous présentons ci-après son état d'avancement en prenant la truite et le gardon comme exemple. Des développements sont en effet encore nécessaires pour finaliser ce travail.

Spatialisation des données

Nous utilisons des données nombreuses et variées déjà rassemblées à l'occasion de travaux antérieurs (Figure 1).

Ces données couvrent une grande partie du bassin de la Seine. Elles résultent d'échantillonnages réalisés au moyen de pêches électriques : plus de 1000 pêches dans 570 stations.

Elles 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 (Figure 1).

Leur intégration à un système d'information géographique est aujourd'hui réalisée (Figure 2).

À cette étape, déjà, la simple projection des espèces présentes dans le contenu des pêches permet de visualiser leur répartition sur le bassin et présente un intérêt, comme l'illustrent les figures 3 et 4.

Arbres de décision

Notre objectif est de modéliser les relations entre les poissons et les descripteurs physiques de l'habitat afin de préciser l'importance relative des variables de milieu dans les mécanismes de structuration des communautés. Ces relations sont complexes et *a priori* non-linéaires, il convient donc d'utiliser des méthodes adaptées : non-linéaires et robustes, c'est-à-dire peu sensibles au bruit des données, convenant mieux à nos données.

C'est pourquoi, nous avons choisi une méthode de discrimination appelée « arbres de décision ».

Principe

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 5. Celui-ci modélise la présence-absence de la truite.

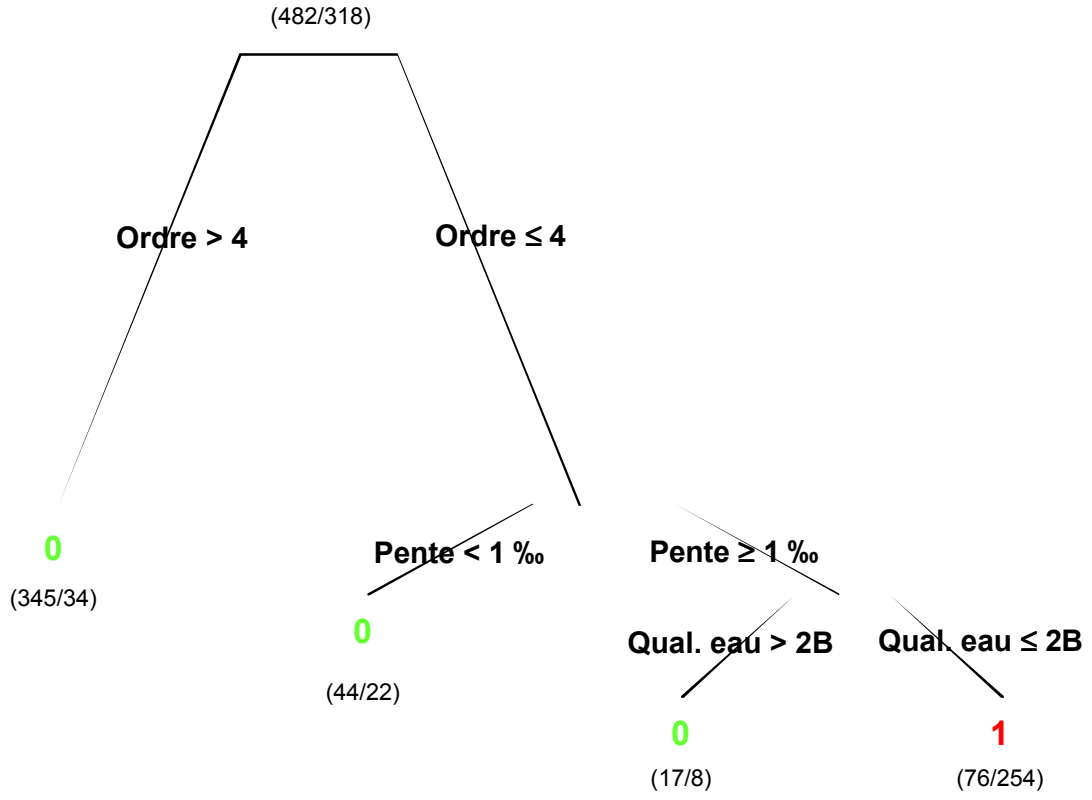

Figure 5. Exemple d'arbre de décision appliqué au cas de la présence-absence de la truite.

Les avantages de la technique des arbres de décision sont multiples.

1. 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.
2. 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.
3. Enfin, la qualité de la prédiction est généralement bonne.

Validation

Comme toute méthode statistique inférentielle basée sur un échantillon de taille finie, les arbres de décision présentent une erreur sous-évaluée. En particulier, un arbre développé au maximum, c'est-à-dire dont toutes les feuilles sont homogènes a une erreur nulle. Mais appliqué à un autre échantillon, il va présenter une erreur certainement plus importante. C'est pourquoi il est nécessaire de choisir un arbre ayant de bonnes capacités de *généralisation,* c'est-à-dire qui reste pertinent vis-à-vis d'un nouvel échantillon.

Pour cela, les arbres de décision utilisent la validation croisée et un critère d'homogénéité pénalisant la taille de l'arbre (c'est-à-dire son nombre de nœuds). L'arbre retenu est alors celui qui réalise le meilleur compromis entre erreur et taille.

Essais préliminaires

Tous les tronçons de cours d'eau n'étant pas renseignés de la même façon, nous avons réduit le nombre des données d'entrée à 6, pour des raisons d'homogénéité :

- 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 à 800 pêches postérieures à 1980, effectuées sur 533 stations.

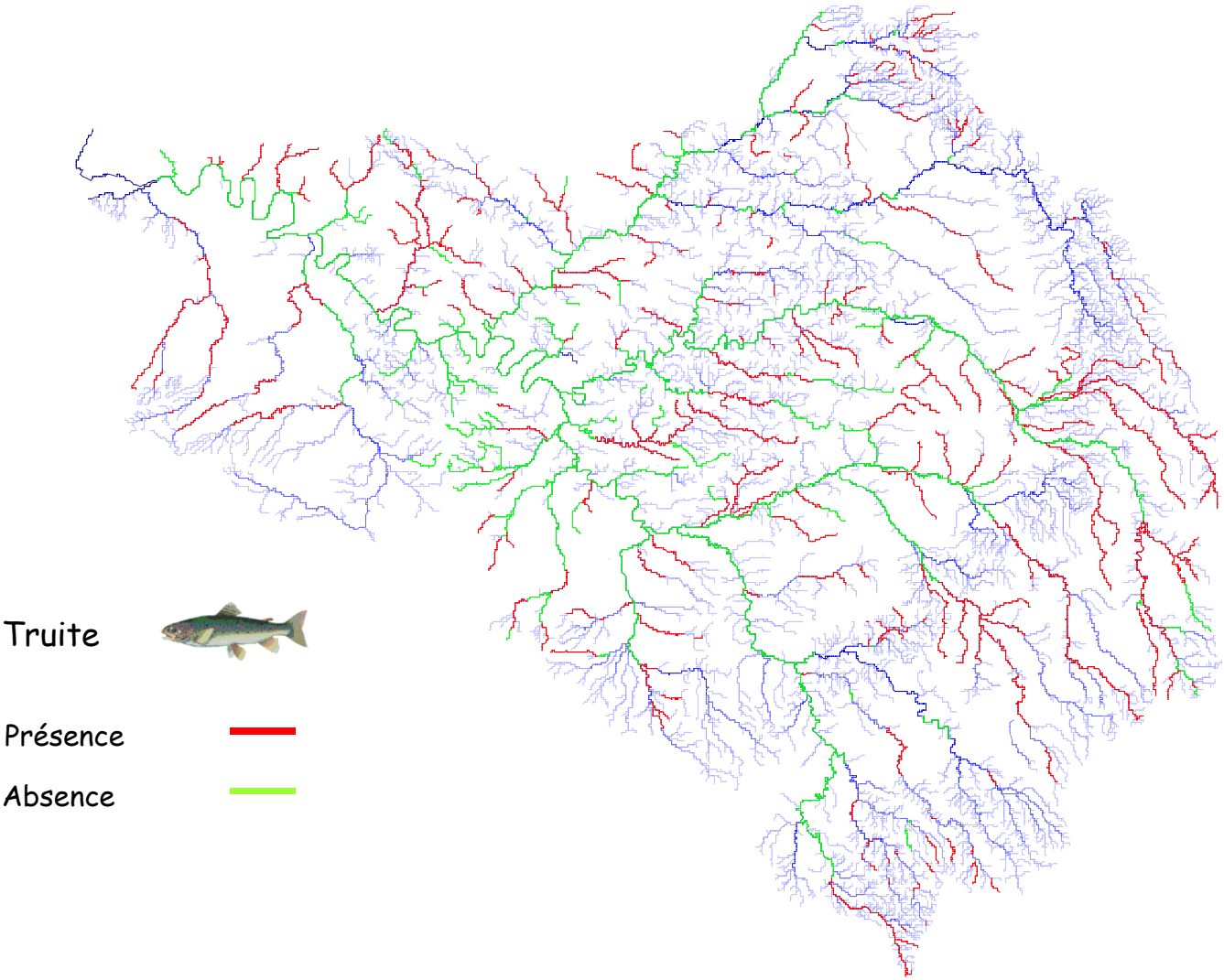

Figure 6. Exemple de généralisation de la prédiction de la truite au linéaire du réseau hydrographique de la Seine.

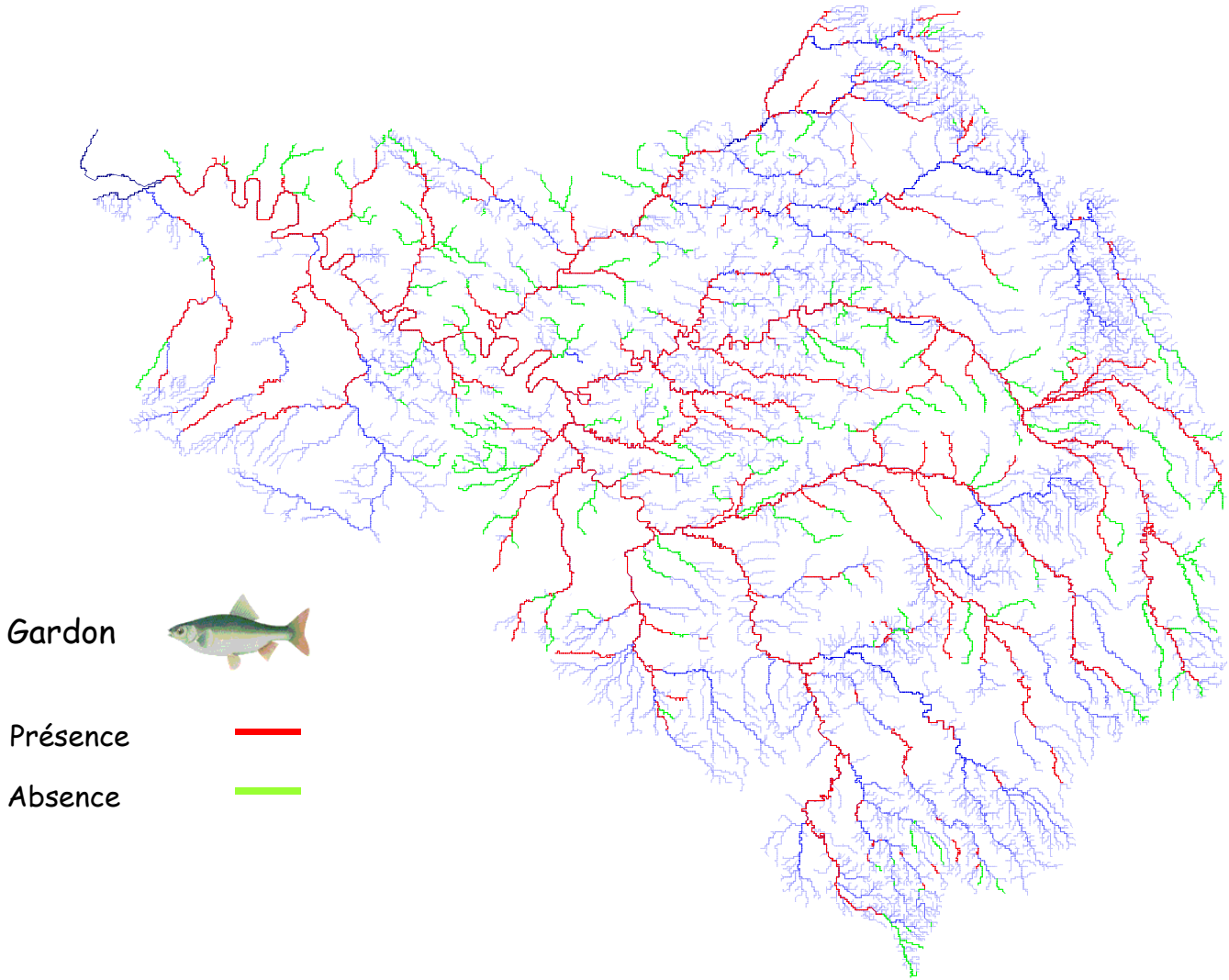

Figure 7. Exemple de généralisation de la prédiction du gardon au linéaire du réseau hydrographique de la Seine.

Enfin, parmi les 39 espèces recensées dans la base de données, seules les 27 plus fréquentes (ie occurrence ≥ 9%) sont considérées.

Les figures 6 et 7 illustrent les résultats obtenus pour la truite et le gardon. Pour ces deux espèces, l'erreur totale en prédiction est respectivement de 0,18 et 0,24, ce qui reste modeste au regard des variables explicatives prises en comptes.

Perspectives

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, aboutit aujourd'hui à un outil opérationnel permettant de prédire les espèces de poissons en fonction des caractéristiques des habitats aquatiques.

En l'état cependant, ce n'est encore qu'une « maquette » logicielle qui est réalisée. Celle-ci offre de nombreuses perspectives mais nécessite toutefois d'être maintenant finalisée et validée.

Pour mener à bien ce travail, plusieurs points restent à traiter :

- d'abord, il s'agit de poursuivre le « nettoyage » des données disponibles et de les intégrer à un SIG commun au PIREN-Seine (ou, au moins, à la base d'un SIG commun, notamment concernant le réseau hydrographique), ainsi qu'à un MNT ;
- puis, il faut généraliser les modèles d'arbre de décision à l'ensemble du bassin. Ceci implique de collecter d'une part, dans des bases existantes (e. g. Corinne LandCover, nouvelles données AREA, base INRA, etc.), et de construire d'autre part (p. ex. surface de BV, pente du cours d'eau, etc.) de nouvelles données explicatives ;
- ensuite, il sera alors possible de procéder au lissage spatial des prédictions par la mise en œuvre de statistiques spatiales gérant véritablement l'arborescence du réseau ; ce dernier point n'est pas traité en standard par les SIG.
- enfin, des simulations tant rétrospectives que prospectives pourront être testées, utilisant par exemple des prédictions de la qualité de l'eau, issues du modèle Sénèque, ou se fondant sur des hypothèses de modifications des habitats (structure physique, régime hydrologique, etc.).

Références

Boët Ph. & Fuhs Th., 1995. Les réseaux de neurones pour prédire la biodiversité des poissons en eau courante. *Ingénieries - EAT*, 4, 3-12.

Boët Ph. & Fuhs Th., *in press*,. Predicting presence of fish species in the Seine River basin using artificial neural networks. *In* : S. Lek & J.-F. Guégan (Eds), *Applications of artificial neural networks to ecological modelling* (chapter 9), Springer Verlag.

Breiman L., Friedman J.H., Olshen R.A. & Stone C.J., 1984. *Classification and regression trees.* Chapman & Hall, New York, 358 p.

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.*

Quinlan J.R., 1987. C4.5 : programs for machine learning. Morgan Kaufmann.

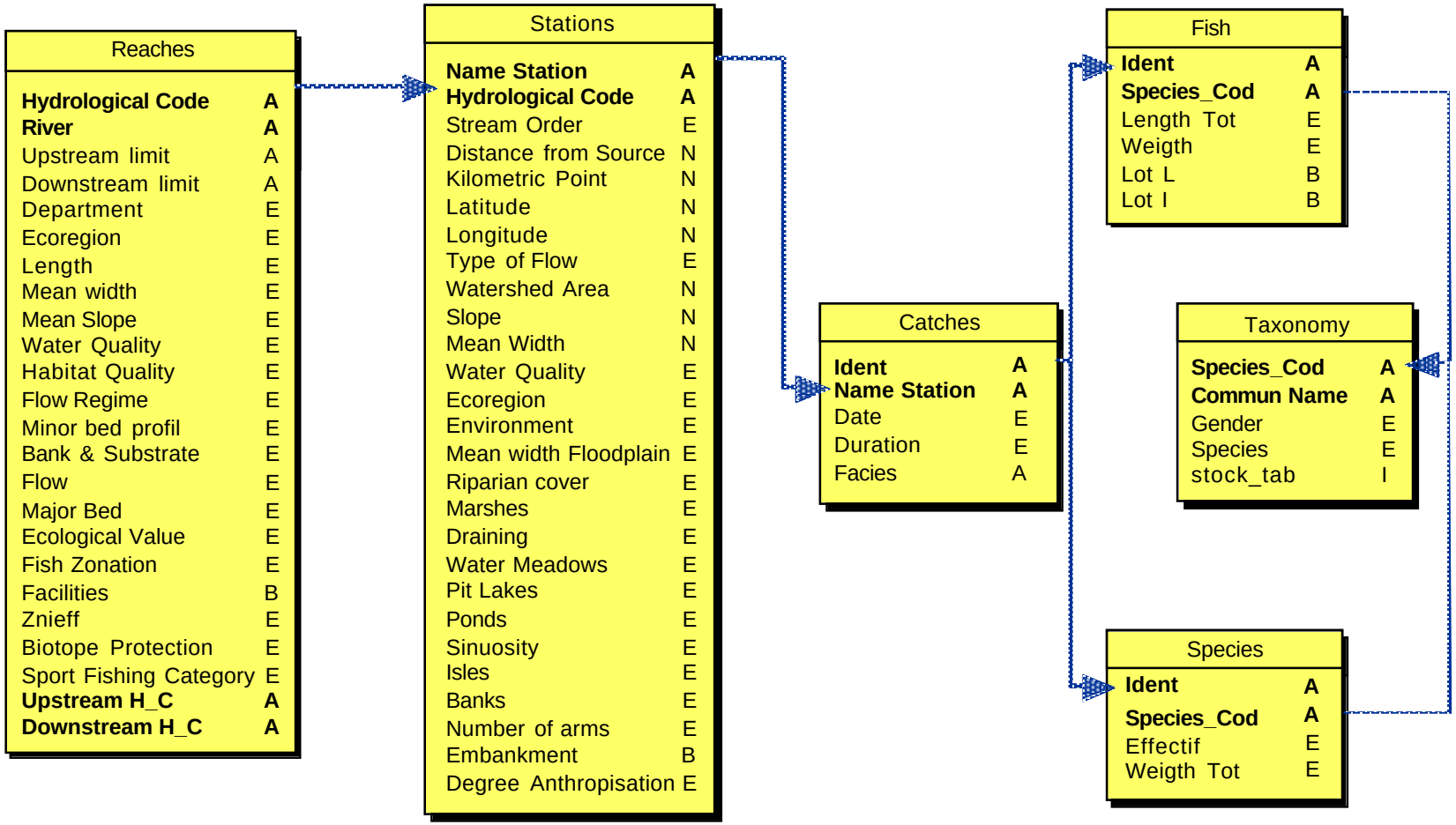

Figure 1. Schéma de la base de données.

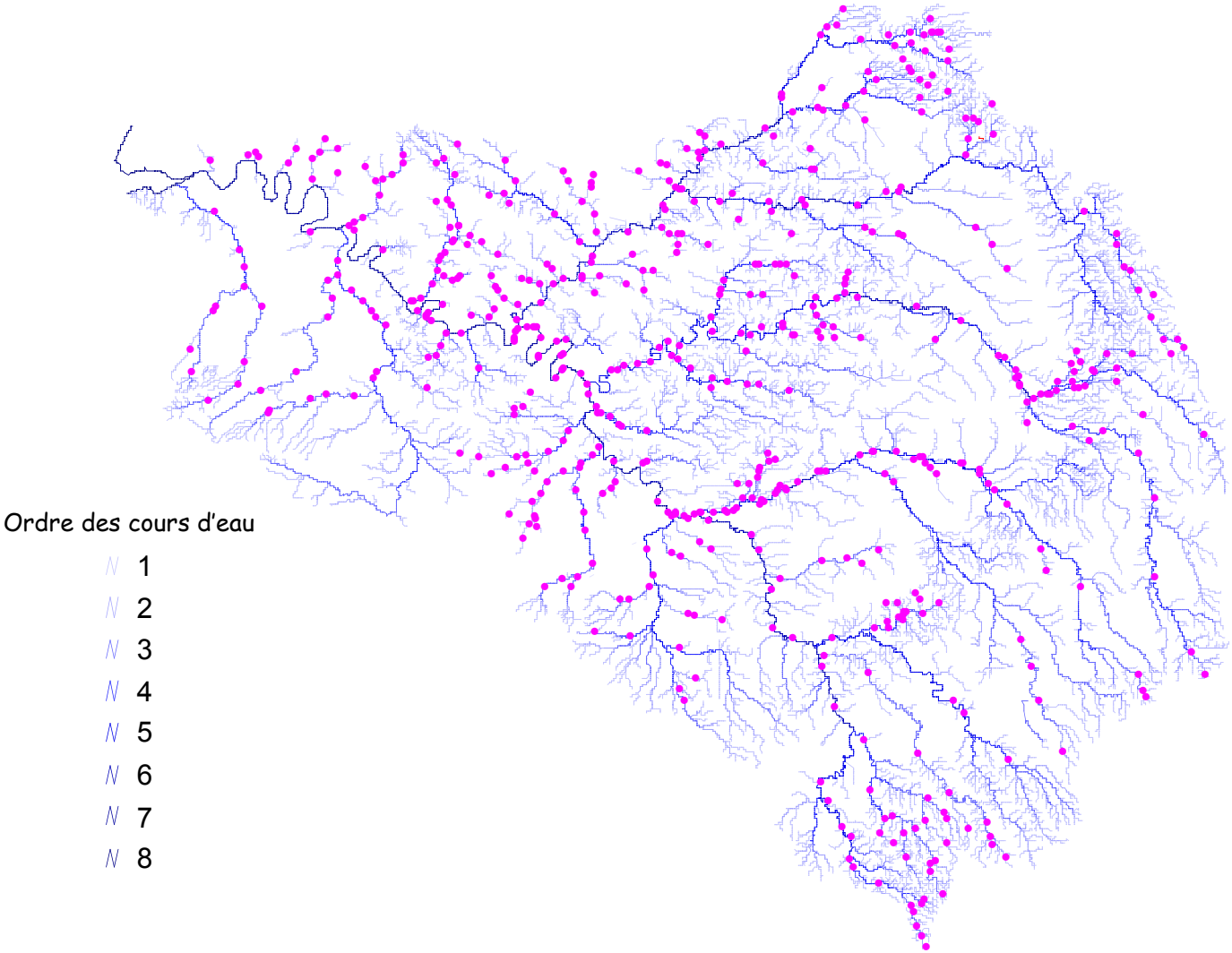

Figure 2. Carte du réseau hydrographique de la Seine et localisation des stations de pêches.

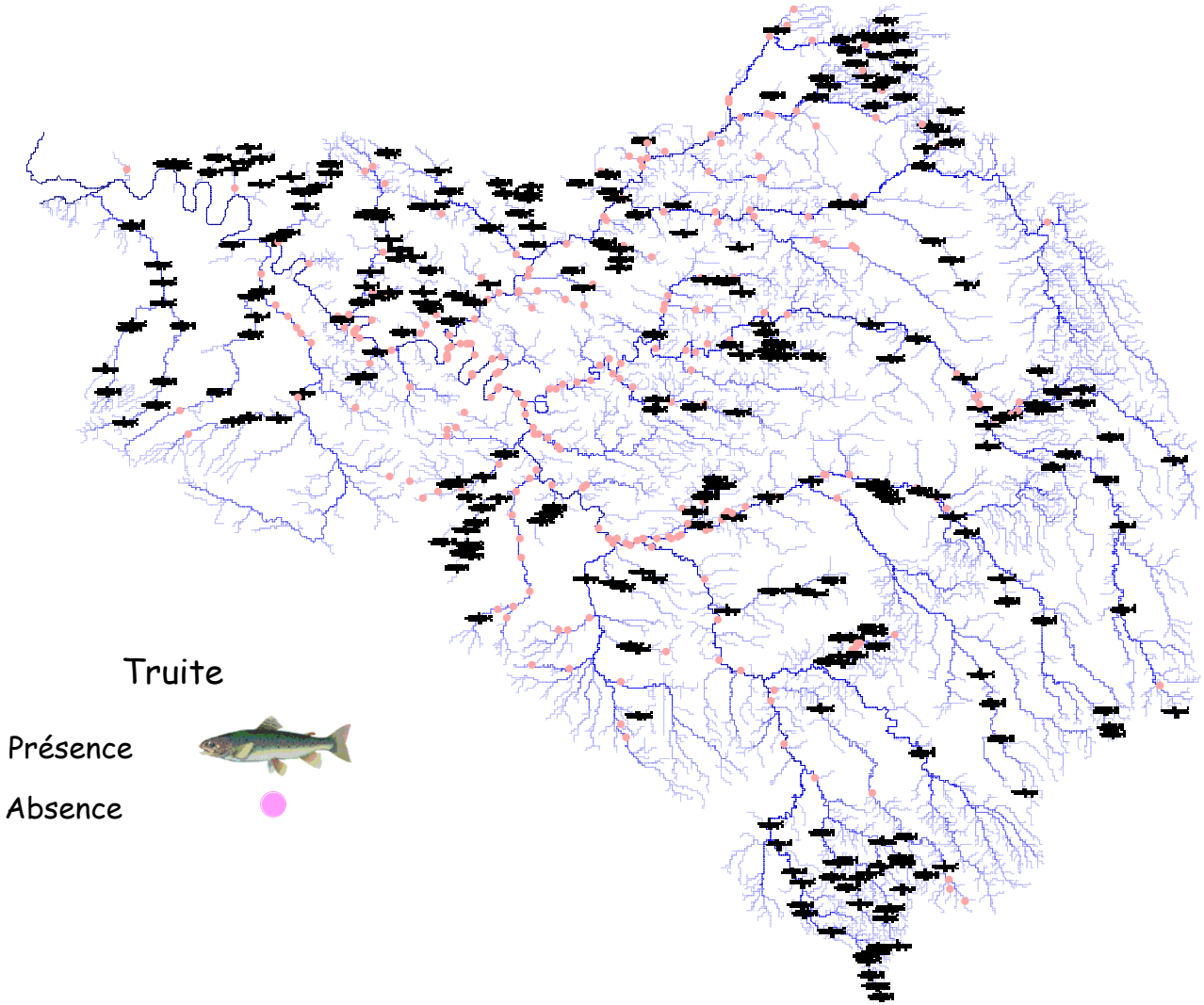

Figure 3. Exemple de la répartition de la truite dans le contenu des pêches.

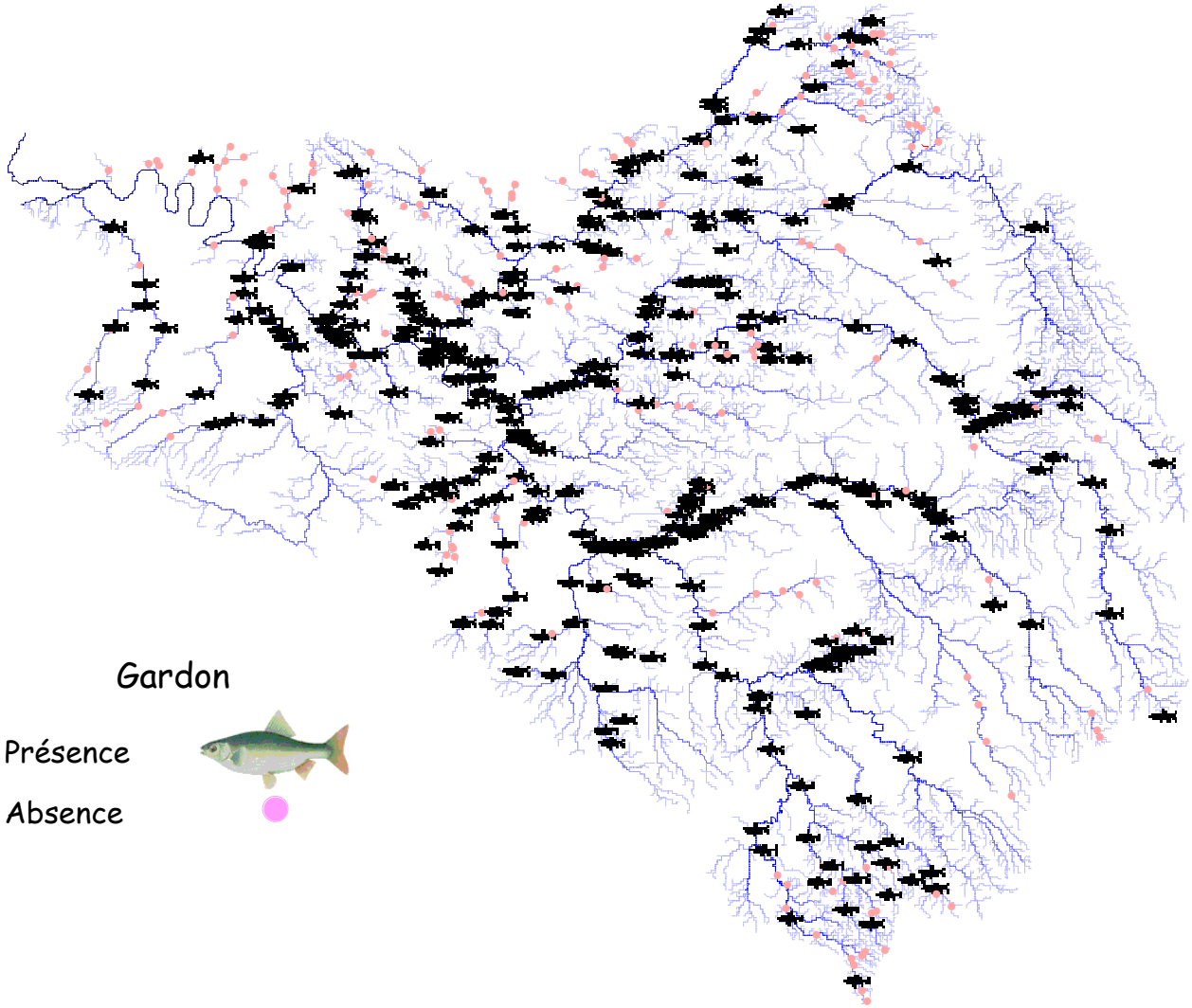

Figure 4. Exemple de la répartition du gardon dans le contenu des pêches.