Modèles pressions-impacts. Approche méthodologique, modèles d'extrapolation spatiale et modèles de diagnostic de l'état écologique basés sur les invertébrés en rivière (IBGN)

J.G. Wasson, Bertrand Villeneuve, N. Mengin, H. Pella, A. Chandesris

Cemagref

Appui scientifique à la mise en œuvre de la Directive Cadre Européenne sur l'Eau

Modèles pressions / impacts

Approche méthodologique, modèles d'extrapolation spatiale et modèles de diagnostic de l'état écologique basés sur les invertébrés en rivière (IBGN)

Jean-Gabriel WASSON, Bertrand VILLENEUVE, Nicolas MENGIN, Hervé PELLA, André CHANDESRIS

Département Gestion des Milieux Aquatiques
Unité de Recherche Biologie des Ecosystèmes Aquatiques
Laboratoire d'Hydroécologie Quantitative
Groupement de Lyon
3 bis Quai Chauveau, CP 220
69336 Lyon cedex 09
Tél. 04 72 20 87 87 - Fax 04 78 47 78 75

Décembre 2005

Cemagref

Unité de Recherche Biologie des Ecosystèmes Aquatiques
Département Gestion des Milieux Aquatiques
Laboratoire d'Hydroécologie Quantitative
3 bis quai Chauveau, CP 220
69366 Lyon Cedex 09
Tél. 04 72 20 87 87 - Fax 04 78 47 78 75

Titre : Modèles pressions / impacts. Approche méthodologique, modèles d'extrapolation spatiale et modèles de diagnostic de l'état écologique basés sur les invertébrés en rivière (IBGN)

Auteurs : Jean-Gabriel WASSON, Bertrand VILLENEUVE, Nicolas MENGIN, Hervé PELLA, André CHANDESRIS

Résumé : La mise en œuvre de la Directive Cadre Européenne sur l'Eau, qui fixe comme objectif d'atteindre le « bon état écologique » de tous les cours d'eau en 2015, implique de faire un diagnostic général de l'état des cours d'eau français, puis de mettre en place des actions de restauration. Ces deux phases se traduisent par des besoins précis en termes d'outils opérationnels. L'objectif du présent rapport est de proposer une approche méthodologique pour le développement de modèles pressions / impacts, à la fois prédictifs et explicatifs, qui permettront de cibler les causes d'altérations des cours d'eau, de faire des extrapolations spatiales à l'échelle nationale et régionale, (autorisant la simulation de différentes hypothèses), et de rechercher des facteurs gérables pour la restauration.
Pour développer ces modèles, les données utilisées ont été l'indice IBGN (méthode standardisée basée sur les invertébrés benthiques) comme indicateur de l'état écologique des rivières, et l'occupation du sol, à partir de la base de données CORINE Land Cover, pour l'évaluation des pressions anthropiques impactant les milieux aquatiques. Ceci a nécessité un important travail de développement d'une plate-forme SIG, qui permet de coupler les caractéristiques naturelles et les pressions anthropiques liées à l'occupation générale du bassin versant et celles qui s'exercent au niveau du corridor rivulaire.
Le choix de méthodes d'analyse adaptées aux objectifs a été particulièrement important. Après avoir effectué plusieurs tests, nous avons décidé d'utilisé une combinaison de deux approches : des modèles linéaires performants comme la régression PLS, qui permettent d'extraire les variables explicatives des impacts biologiques observés ; et des modèles non linéaires tels que les arbres de décision, qui permettent de faire une extrapolation spatiale.
Les **modèles d'extrapolation spatiale** ont permis de représenter à l'échelle nationale l'état écologique probable des cours d'eau, à partir de l'IBGN, et avec différentes hypothèses de limites de bon état. Ils ont montré que la probabilité de « bon état » est principalement liée à la proportion de territoires artificialisés (urbanisés) dans le bassin versant. L'agriculture intensive intervient secondairement dans le modèle. Nous avons ensuite testé une deuxième hypothèse en relevant de 1 point d'IBGN la limite de « Bon état » déterminée en première hypothèse, pour tous les types de cours d'eau, ceci afin d'avoir une vison plus précise du risque de non atteinte du « bon état ». Dans ce modèle, un critère agriculture intensive se rajoute au critère artificialisation pour discriminer les stations majoritairement en bon ou mauvais état. On peut en conclure qu'à l'échelle nationale, un grand nombre de sites dont le bassin est occupé par des terres labourées se trouvent en situation limite.
Enfin, les **modèles de diagnostic** ont permis de constater qu'à l'échelle nationale, les impacts liés à l'urbanisation constituent le principal facteur d'altération de l'état écologique, ceux liés à l'agriculture intensive semblent intervenir secondairement. L'agriculture de faible intensité (prairies) joue un rôle qui peut s'avérer négatif ou positif selon le contexte. Les espaces naturels ont, par contre, un effet toujours positif sur les bioindicateurs. Ces modèles sont plus performants et plus explicatifs avec des données détaillées et à l'échelle des hydro-écorégions, ce qui permet de mettre en évidence les pressions au niveau du corridor rivulaire.
En conclusion, le développement de modèles d'extrapolation et de diagnostic, qui permettent une hiérarchisation et une spatialisation des pressions, constitue un apport aux décideurs pour l'analyse des problèmes et la définition de politiques d'action.

Mots-clés : modèle pressions / impacts, IBGN, invertébrés benthiques, extrapolation spatiale, diagnostic, Régression PLS, arbre de décision, occupation du sol, corridor rivulaire, bassin versant, Directive Cadre Européenne sur l'Eau

Keyboard : pressure / impacts relationship, models, IBGN, benthic invertebrates, spatial extrapolation, diagnosis, PLS regression, decision tree, land cover, riparian buffer, watershed, Water Framework Directive

CONVENTION	PROGRAMME DE RECHERCHE	DATE	DIFFUSION
Convention CV03000102 DE/MEDD	HYDRECO (LHQ)	Décembre 2005	tous publics ■ interne ❒ confidentielle ❒

Sommaire

1 Introduction

La mise en œuvre de la DCE (voir encadré) implique clairement deux phases : d'abord un *diagnostic* général de l'état des milieux, qui prépare la mise en place d'un réseau de suivi, puis très rapidement une phase de *restauration* des milieux qui n'atteignent pas le «bon état». Les mesures à mettre en œuvre seront détaillées dans le premier plan de gestion, prévu en 2009, mais les principaux éléments de ce plan devront donc être prêts bien avant. Ces deux phases sont évidemment liées, car *le diagnostic doit orienter les actions de restauration*. Elles se traduisent par des besoins précis en termes d'outils opérationnels : des *méthodes de diagnostic* (typologie, référence, bioindicateurs), et des *modèles pressions / impacts* pour choisir les actions prioritaires de restauration.

Au cours des années 2003 et 2004, le ***Cemagref*** a proposé une approche typologique des cours d'eau, et défini pour les bioindicateurs les plus utilisés des valeurs de référence permettant d'évaluer un état écologique au sens de la DCE.

L'objectif du présent rapport est de proposer une approche méthodologique pour le développement de *modèles pressions / impacts,* appliquée dans un premier temps aux bioindicateurs invertébrés (Indice de Qualité Biologique Normalisé = IBGN).

Après avoir précisé les questions auxquelles ces modèles tentent de répondre, nous détaillerons les aspects méthodologiques concernant la nature des données et les outils mathématiques utilisés. A partir du jeu de données national sur l'IBGN, nous présenterons des exemples de modèles d'extrapolation spatiale du « bon état », et de modèles explicatifs faisant apparaître les causes d'altération des milieux. Les implications pratiques de ces premiers résultats sont discutées en conclusion.

Le contexte de la Directive Cadre

Pour les eaux de surface, la Directive Cadre Européenne sur l'Eau (DCE) fixe pour objectif d'atteindre à l'horizon 2015 le « bon état » pour tous les milieux naturels, de préserver ceux qui sont en « très bon état », et d'atteindre le « bon potentiel » dans les milieux fortement artificialisés. Il s'y ajoute également un objectif « zéro toxique », non daté mais clairement affiché. Mais le point essentiel est ici une obligation de résultat dans le délai imparti, et non plus seulement de moyens, la directive fixant seulement un catalogue des mesures possibles qui restent sous la responsabilité des états membres.

Le bon état est défini d'après la situation la plus déclassante entre un état chimique se rapportant à des normes de concentration de certaines substances particulièrement dangereuses (toxiques), et un état écologique qui repose sur une évaluation des « éléments de qualité » physico-chimiques (paramètres généraux et micro-polluants non inclus dans l'état chimique), et biologiques (peuplements végétaux, invertébrés et poissons). L'objectif de « bon état écologique » est défini comme un écart « léger » à une situation de référence, correspondant à des milieux non ou très faiblement impactés par l'homme.

Selon la définition de la DCE, l'état écologique se réfère « à la structure et au fonctionnement des écosystèmes aquatiques » ; mais son évaluation repose principalement sur la bioindication : les peuplements aquatiques sont les juges de paix. En effet, les normes à appliquer pour les paramètres physico-chimiques généraux doivent être reliées à l'altération des peuplements, et pour les polluants toxiques, elles sont définies sur la base de tests écotoxicologiques. L'évaluation des altérations physiques (ou hydro-morphologiques) n'est explicitement requise que pour identifier les situations de référence et le « très bon état », mais elle est évidemment essentielle en tant qu'élément de diagnostic des causes d'altération.

Ce travail a été financé dans le cadre de la convention entre le **Cemagref** et la Direction de l'Eau du Ministère de l'Ecologie et du Développement Durable (convention CV03000102, DE/MEDD). Il contribue également au programme de la Zone Atelier Bassin du Rhône (ZABR).

1.1 Problématique

1.1.1 Des méthodes de diagnostic ?

Pour établir un diagnostic de l'état des cours d'eau, la première idée qui vient à l'esprit est « d'interroger » directement les bioindicateurs : une analyse pertinente des peuplements permet-elle d'identifier les facteurs qui les perturbent ? La réponse est mitigée, car les relations en cause sont aussi complexes que variées.

Les *structures* de peuplement répondent à trois grandes catégories de *processus* qui déterminent le *fonctionnement* du cours d'eau. Il s'agit de processus *physiques* - qui déterminent la morpho-dynamique fluviale et l'habitat aquatique ; *biogéochimiques*, - qui régissent les flux de matière organique et d'énergie métabolique, ainsi que des variables physico-chimiques essentielles ; et *écologiques*, - qui conditionnent les possibilités de dispersion et de recolonisation des espèces. C'est la réponse biologique à l'altération de ces processus qui détermine l'état écologique.

Cependant, deux facteurs viennent compliquer le problème :

- la variabilité des réponses biologiques en fonction du contexte naturel ;
- les interactions entre différents types d'altérations.

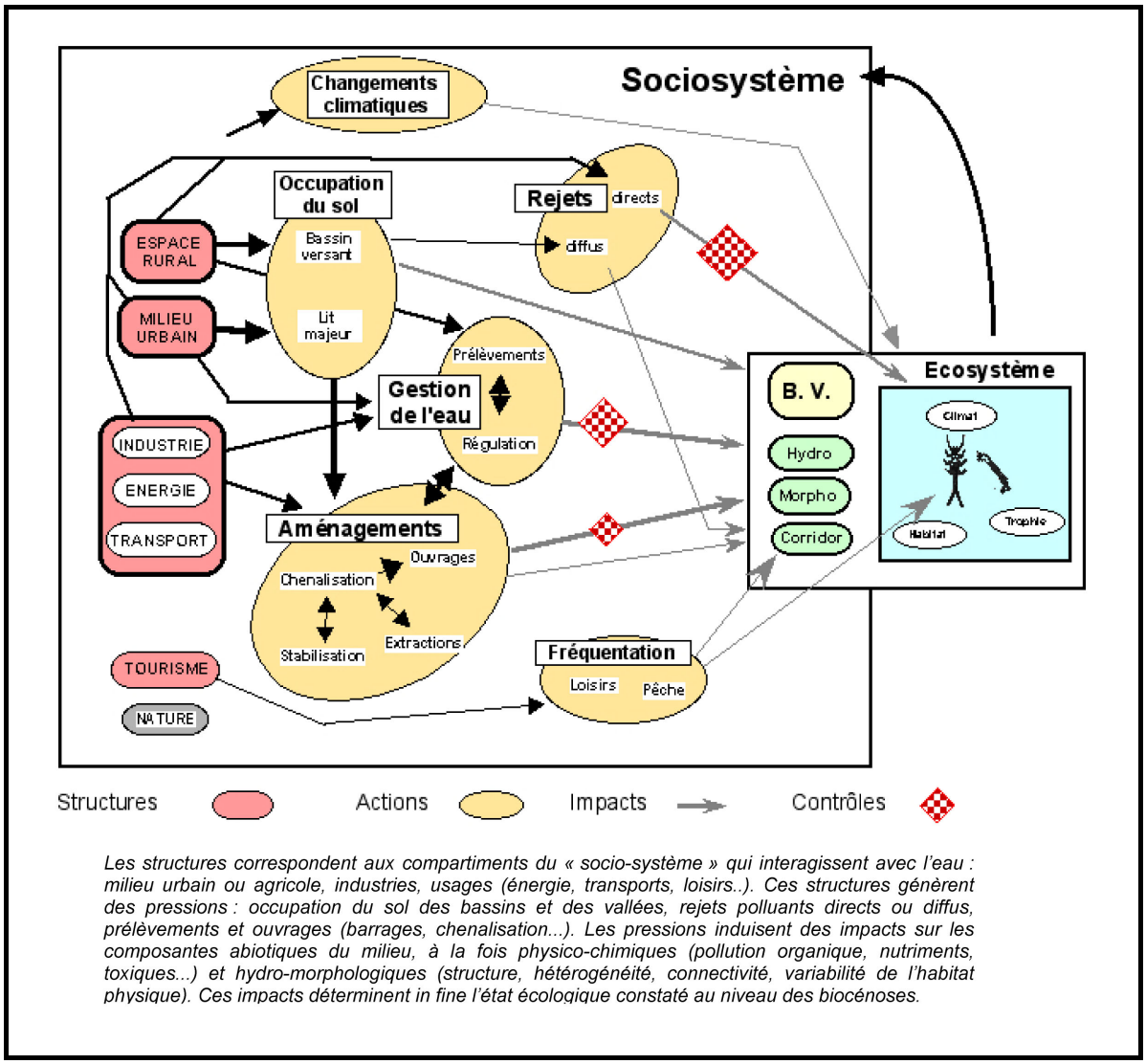
Les bioindicateurs mesurent par définition un état biologique, intégré sur une période qui dépend de la durée de vie des organismes. L'indice IBGN (norme AFNOR NF T 90-350, 1992) basé sur les invertébrés, le plus ancien et le plus utilisé, contient une information relativement pauvre (nombre de taxons, et présence de taxons sensibles à la pollution). Il pourra facilement être complété par des métriques plus fonctionnelles utilisant les traits biologiques des organismes, et des indices de structure des peuplements (**Usseglio-Polatera** et al. 2001). Les méthodes basées sur les diatomées traduisent qualitativement les perturbations des flux de matière organique et de nutriments ; il y manque un aspect quantitatif qui requiert des approches beaucoup plus lourdes pour évaluer la production primaire et le degré d'autotrophie du milieu (**Billen** et al. 1999). En revanche l'indice « poisson » récemment développé (**Oberdorff** et al. 2001) intègre déjà des métriques fonctionnelles.

Ces méthodes, avec des améliorations indispensables pour les invertébrés, permettront de répondre aux exigences de la DCE pour l'évaluation de l'état écologique. Mais elles sont pour le moment insuffisantes pour porter un véritable diagnostic des causes d'altération. Elles donnent au mieux des indications qualitatives sur les facteurs en cause, mais ne permettent pas à elles seules de quantifier les relations de cause à effet.

1.1.2 Modèles pressions / impacts

La question scientifique est ici centrale : il ne s'agit pas simplement de constater les dégâts, il faut identifier et hiérarchiser les causes d'altération pour orienter les investissements de restauration. Des études au cas par cas n'étant pas envisageables, la définition d'une politique d'action passe par une *analyse quantitative des relations entre pressions anthropiques et état écologique, et une extrapolation à des échelles pertinentes pour la gestion*.

Cette approche suppose l'analyse de données tirées des réseaux de suivi, pour développer des modèles à la fois prédictifs et explicatifs reliant les impacts à leurs causes probables. Mais les problèmes à résoudre sont multiples, et passent par l'identification des relations structures / pressions / impacts qui génèrent les états biologiques observés dans les rivières (figure 1). Pour aborder cette complexité, il faut d'abord choisir des objectifs clairs et des échelles de travail. Se poseront ensuite des questions sur les méthodes à mettre en œuvre, en fonction des données disponibles.



Les structures correspondent aux compartiments du « socio-système » qui interagissent avec l'eau : milieu urbain ou agricole, industries, usages (énergie, transports, loisirs..). Ces structures génèrent des pressions : occupation du sol des bassins et des vallées, rejets polluants directs ou diffus, prélèvements et ouvrages (barrages, chenalisation...). Les pressions induisent des impacts sur les composantes abiotiques du milieu, à la fois physico-chimiques (pollution organique, nutriments, toxiques...) et hydro-morphologiques (structure, hétérogénéité, connectivité, variabilité de l'habitat physique). Ces impacts déterminent in fine l'état écologique constaté au niveau des biocénoses.

figure 1. Schéma de fonctionnement du sociosystème

1.2 Objectifs et échelles de travail

Quelle limite de « bon état » ?

L'une des premières questions qui se pose aux services opérationnels est celle de la fixation d'une limite du bon état écologique. La DCE, en définissant le bon état comme un écart « léger » à la référence, laisse une certaine marge d'interprétation ; les définitions « normatives » de l'annexe V ne sont guère précises sur ce point. La question est donc « quel niveau de pression anthropique correspond à différents seuils possibles de bon état » ? *Quel seuil de bon état*, compatible avec les exigences de la DCE, pourrait constituer un *objectif à la fois ambitieux et réaliste* ?

Où agir en priorité ?

Il est évident maintenant que la DCE imposera une politique volontariste, sur des objectifs prioritaires et ciblés. Vient donc immédiatement la question « où agir en priorité » ? Un deuxième niveau de réponse attendu par les opérationnels concerne donc à la fois *la spatialisation et la hiérarchisation des problèmes*.

Cette spatialisation des problèmes concerne deux échelles d'approche :

- Une échelle régionale : quelles sont sur le territoire français les régions les plus impactées ?
- Une échelle linéaire liée à la structure des réseaux hydrographiques : quel est l'impact relatif des pressions s'exerçant sur l'ensemble du *bassin versant* et à proximité immédiate des cours d'eau, au niveau des *corridors rivulaires* ?

Sur quelles causes agir ?

Enfin, il faut préparer une panoplie d'actions qui ne se limitent pas seulement à des mesures techniques. En simplifiant, il ne sera pas toujours suffisant de traiter les conséquences des activités humaines au niveau des écosystèmes aquatiques, il sera souvent nécessaire d'agir sur les causes, en orientant les pratiques des acteurs, voire en modifiant certaines structures. L'identification, à un niveau très général, des grandes *structures socio-économiques responsables des impacts dominants* sur les milieux constitue donc une troisième question scientifique de l'approche pressions / impacts. Ceci permettrait également d'intégrer, au travers de *scénarios d'évolution socio-économique,* un *aspect prévisionnel* de l'évolution probable des problèmes.

Vers des modèles « large échelle » spatialisés…

Ces objectifs nous ont orientés vers le développement de modèles « large échelle » pouvant fournir d'abord des réponses *spatialisées*, c'est-à-dire projetables sur des cartes. Cette visualisation des résultats constitue toujours un atout important pour les décideurs.

… à des échelles nationales et régionales

Mais la géographie de la France est, à l'échelle européenne, très diversifiée. Ceci se traduit par une grande hétérogénéité des milieux, synthétisée par les Hydro-écorégions (HER) (**Wasson** et al. 2002) qui constituent le premier niveau de la typologie.

En effet, les usages de l'eau, l'occupation du sol, et dans une certaine mesure la répartition des populations humaines répondent aux mêmes facteurs géographiques qui ont servi à délimiter les hydro-écorégions. On observe donc une bonne concordance générale entre les HER et la spatialisation des pressions anthropiques. Il en résulte, pour chaque HER, *une spécificité régionale des relations entre la sensibilité des milieux et les pressions qui les affectent.* Il n'est donc pas exagéré de parler de « pathologies régionales » des écosystèmes aquatiques. Cette échelle des HER constitue donc un niveau de résolution pertinent à la fois pour *l'analyse des problèmes et la définition de politiques d'action*.

… intégrant la structure linéaire des cours d'eau.

Enfin, il était important d'intégrer dans les modèles l'analyse des « effets de proximité », en essayant de faire la part des pressions liées à l'occupation générale du bassin versant et de celles qui s'exercent au niveau du corridor rivulaire. Ce point est essentiel pour définir des stratégies de restauration.

Nous avons donc cherché à développer des *modèles pressions / impacts spatialisés et régionalisés*, dans le but de faire apparaître, au moins à l'échelle des principales HER, les *principales causes d'altération des peuplements aquatiques*.

Les objectifs communs à ces modèles pressions / impacts sont donc :

- *l'extrapolation* : spatialisation de l'état écologique probable des masses d'eau, en fonction de différents bioindicateurs et des hypothèses de limite de bon état ; niveau de pressions correspondant au bon état pour les différents peuplements ...
- *le diagnostic* : identifier et hiérarchiser les causes d'altération des milieux, et les « pathologies » dominantes par région.
- *L'aide à la restauration* : identifier les facteurs gérables sur lesquels il faut cibler les actions de restauration, en particulier au niveau des corridors rivulaires.

Notons toutefois que ce dernier objectif suppose, à moyen terme, à la fois une interprétation des résultats obtenus sur les modèles écologiques présentés ici (i.e. explicatifs de la réponse des peuplements aquatiques), et l'intégration d'une réflexion économique et sociale.

1.3 Etat des connaissances d'après la littérature internationale

Avant d'entamer un travail de cette ampleur, il était nécessaire de faire le point des connaissances et des outils existants sur ce sujet.

Une revue bibliographique a été réalisée dans le cadre du projet européen REBECCA[1], pour rechercher les publications internationales traitant des relations à large échelle entre les pressions anthropiques multiples, généralement évaluées par l'occupation du sol, et la réponse des variables biologiques (poissons, invertébrés, végétaux). (**Garcia** et al. in press, **Garcia** et al. 2005). Les principales conclusions de cette recherche sont résumées ci-dessous.

1. Il n'existe pas beaucoup de travaux traitant des relations pressions / impacts à large échelle, et aucun modèle existant ne peut être directement extrapolé aux situations françaises et européennes pour évaluer un état écologique au sens de la DCE.
2. Les travaux existants sont généralement limités à des situations géographiques (bassins, écorégions) relativement homogènes ; la transférabilité des modèles à des contextes socio-écologiques différents est rarement abordée.
3. Pour la raison ci-dessus, il ne ressort pas de conclusion tranchée en ce qui concerne l'impact relatif de l'agriculture et de l'urbanisation, l'une ou l'autre cause apparaissant comme prépondérante selon les cas. De même, il n'est pas possible de tirer des conclusions générales sur l'impact des différentes formes d'agriculture.
4. Il semble également nécessaire de tenir compte du contexte socio-économique et des pratiques environnementales (taux de dépollution, intensification de l'agriculture) pour évaluer les pressions associées à un type d'occupation du sol.
5. L'impact relatif des pressions s'exerçant à l'échelle du bassin versant ou du corridor rivulaire n'est pas clairement établi ; certains travaux mettent en évidence un effet prépondérant de l'occupation du sol des bassins, d'autres aboutissent à la conclusion inverse. Si la restauration des corridors rivulaires est souvent considérée comme une action prioritaire, l'étude de l'impact de cette action à large échelle n'a pas encore été abordée.
6. Enfin, il n'est pas possible d'établir des seuils de pression, par exemple en termes d'occupation urbaine ou agricole du sol, aboutissant à un risque élevé de non atteinte du « bon état écologique » au sens de la DCE.

[1] Site web : http://www.environment.fi/syke/rebecca

En résumé, il n'existait pas au départ de notre travail une méthodologie directement transférable ; mais un nombre suffisant de publications traitant de ces problèmes a permis d'orienter les choix méthodologiques pour apporter des éléments de clarification sur ces différents points.

1.4 Choix méthodologiques

La première question concerne les données disponibles pour développer ces modèles.

Quelles variables biologiques ?

Comme variable de réponse indicatrice de l'état écologique des rivières, nous avons choisi de commencer le travail avec *l'indice IBGN basé sur les invertébrés*, pour lequel on dispose d'un jeu de données national très conséquent (cf. chapitre 2). Malgré ses limitations, cet indice apparaît en termes de fiabilité, de précision, et de réponse aux pressions, très comparable aux différents indices utilisés par les autres pays européens. Il constitue donc un bioindicateur suffisamment pertinent pour développer des modèles pressions / impacts dans le cadre de la DCE. Les travaux antérieurs ont permis de normaliser les variations naturelles de l'indice pour pouvoir évaluer un état écologique à partir des valeurs de référence des différents types de cours d'eau. (ref ?)

Afin de prendre en compte les différents peuplements qui interviennent dans la définition de l'état écologique, des travaux similaires ont commencé avec les diatomées, et une extension aux poissons est en cours, utilisant les données du Réseau Hydrologique Piscicole (RHP) gérée par le Conseil Supérieur de la Pêche (CSP) et en collaboration avec celui-ci.

Quels indicateurs de pressions anthropiques ?

Un premier choix méthodologique, qui découle des objectifs précédents, consiste à privilégier des indicateurs pour lesquels on dispose de données à la fois :

- *spatialement homogènes* au niveau du territoire national, pour permettre l'extrapolation ;
- *représentatives de l'ensemble des pressions* qui s'exercent sur les milieux aquatiques, pour pouvoir hiérarchiser les causes d'altération.

Si l'on se réfère à la figure 2, on peut rechercher ces données

- soit au niveau des paramètres physico-chimiques et hydro-morphologiques de la station, au plus près des variables biologiques, mais l'origine des perturbations est alors inconnue ;
- soit au niveau des diverses pressions qui s'exercent sur les milieux, option qui semblerait préférable, mais nécessite des données homogènes sur les diverses pressions, et une bonne relation topologique avec les stations biologiques ;
- soit au niveau des structures, mais il est alors impossible de faire la part des différentes pressions générées par une structure donnée, telle que l'agriculture ou l'urbanisation par exemple.

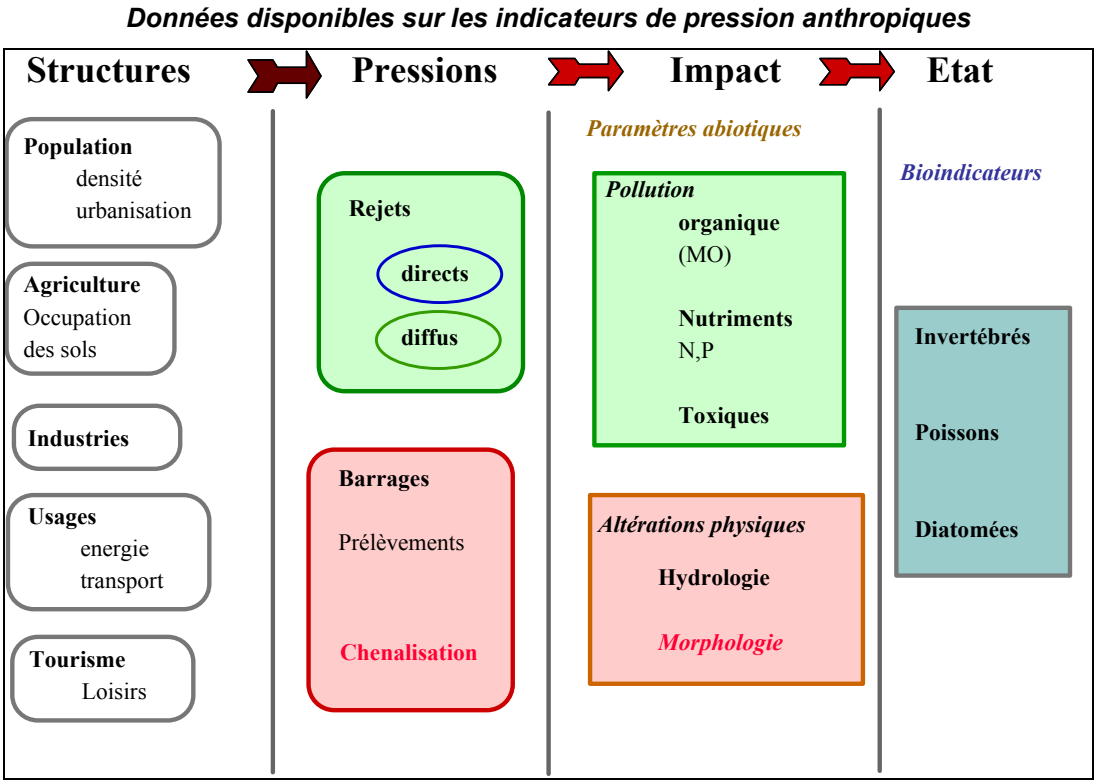
Si l'on considère maintenant les données disponibles, le choix se restreint de lui-même (figure 2).

Au niveau des pressions, seuls les rejets polluants directs sont bien répertoriés par les agences de l'eau ; l'évaluation des pollutions diffuses nécessite une modélisation actuellement impossible à mettre en œuvre à l'échelle nationale. Concernant les

pressions physiques, il n'existe pas de données bancarisées homogènes à l'échelle du territoire métropolitain : on ne trouvait même pas en France, en 2004, un fichier géoréférencé des barrages homogène à l'échelle du territoire national!

Au niveau des paramètres abiotiques, si la physico-chimie de l'eau fait l'objet d'un suivi régulier et assez dense, aucune donnée concernant les impacts hydro-morphologiques ne peut être associée aux stations biologiques.

Il est donc nécessaire de remonter au niveau des structures pour disposer d'une couverture fiable et homogène du territoire à partir des bases de données nationales (IGN, INSEE, RGA) et européennes (CORINE Land Cover). C'est cette option qui a été choisie dans un premier temps (cf. chapitre 2).



D'une manière générale, il n'existe pas de données bancarisées à l'échelle nationale sur les pressions physiques (barrages, endiguements, chenalisation…) ni sur les altérations hydrologiques et morphologiques des stations sur lesquelles on dispose de données biologiques. En revanche, on dispose de couvertures spatiales homogènes au niveau des grandes structures anthropiques permettant d'identifier les principales « forces motrices » qui génèrent des pressions sur les milieux aquatiques.

figure 2. Données disponibles sur les indicateurs de pression anthropique

Quelles échelles de travail ?

Si l'objectif est de fournir une vision d'ensemble des relations pressions / impacts à l'échelle nationale, l'approche par hydro-écorégions permet de stratifier les données en ensembles homogènes du point de vue de la sensibilité des milieux et des pressions qu'ils subissent. En limitant le jeu des facteurs actifs et la variabilité des réponses biologiques, l'approche régionalisée devrait aboutir à des modèles pressions / impacts plus performants, et explicatifs des « pathologies » régionales des écosystèmes.

C'est donc cette voie qui a été explorée, et les modèles développés combinent une *extrapolation spatiale à l'échelle nationale, et la recherche de variables explicatives à l'échelle régionale*, *en intégrant la structure linéaire des cours d'eau*.

Quels outils mathématiques ?

Le choix de méthodes d'analyse adaptées aux objectifs est particulièrement important.

Les premiers tests ont montré que les modèles linéaires classiques étaient peu adaptés à ce type de données ; mais des outils plus performants comme la *régression PLS* permettent d'extraire les variables explicatives des impacts biologiques observés.

En revanche, des modèles non linéaires comme les *arbres de décision* génèrent des classifications moins explicatives, mais qui permettent l'extrapolation spatiale.

C'est donc une combinaison de ces deux approches qui a été utilisée (cf. chapitre 3).

En résumé, fournir un appui scientifique à la mise en œuvre de la Directive Cadre suppose d'aborder la complexité du monde réel à des échelles pertinentes pour le gestionnaire. Nous avons privilégié dans un premier temps une vision à « large échelle » de l'état écologique, basée sur la modélisation des relations entre une réponse biologique révélée par les invertébrés aquatiques, et les structures socio-économiques qui génèrent des impacts. Cette analyse spatiale est couplée à une approche régionale pour simplifier la vision d'un espace fortement hétérogène.

2 Description des données biologiques et des pressions

2.1 Etat écologique des stations

L'état écologique des stations est évalué ici à partir de l'Indice de Qualité Biologique Normalisé (IBGN, norme AFNOR NF T 90-350, 1992) basé sur les peuplements de macro-invertébrés benthiques.

Les données disponibles pour cet indice ont été rassemblées antérieurement par le ***Cemagref*** dans une base de données nommée « GIRAFE ». Le fichier de travail est constitué de 3640 stations, pour lesquelles on dispose de 12318 relevés (ou notes IBGN) couvrant la période 1992-2002.

Les valeurs de référence de l'IBGN et les limites de classes « très bon état » et « bon état » ont été évaluées antérieurement, cette approche est résumée ci-dessous.

*On trouvera dans le rapport **Cemagref**: Détermination des valeurs de référence de l'IBGN et propositions de valeurs limites du "Bon Etat" **Version 2** (2003), http://www.lyon.cemagref.fr/bea/lhq/dossiers_pdf/IBGNFrance.pdf, un descriptif détaillé des données et méthodes utilisées. Toutes les évaluations de l'état écologique utilisées dans le présent rapport sont basées sur la version 2 (2003) des valeurs de référence de l'IBGN.*

L'état écologique de chacune des 3640 stations est déterminé par rapport à son type naturel. La procédure suivante est utilisée :

- Chaque station est affectée à un type en fonction de son hydro-écorégion et du rang du cours d'eau ; à chaque type correspond une note IBGN de référence.
- Les notes IBGN sont alors transformées, selon les préconisations de la DCE, en EQR (pour Ecological Quality Ratio), selon la formule : EQR= (IBGN-1)/(REF-1).
- Cet EQR-IBGN, correspondant à une mesure de « l'écart à la référence », est utilisé dans toutes les analyses suivantes pour déterminer un état écologique indépendant des variations naturelles (typologiques) de l'indice.

Cependant, pour positionner une station donnée dans une classe d'état écologique, il faut tenir compte de l'ensemble des valeurs observées sur cette station. Par analogie avec la méthode utilisée pour définir la limite du très bon état (figure 3), la règle suivante a été retenue :

Pour placer une station dans une classe donnée, il faut que les 3/4 des valeurs observées sur cette station soient au moins égales à la valeur de la limite inférieure de cette classe.

On se basera donc sur le 25ème percentile (ou Q25) de la distribution des valeurs de l'EQR-IBGN observées sur une station pour évaluer sa classe d'état écologique. Cette approche est cohérente avec celle du SEQ Eau, qui utilise le 90ème percentile des paramètres chimiques, correspondant à moins de 10% de dépassement de la norme. Vu le petit nombre de relevés généralement disponibles pour l'IBGN, le Q25 est une valeur plus réaliste.

Comme on ne dispose pas toujours d'un nombre suffisant de relevés pour calculer un Q25, la règle suivante à été appliquée :

- Pour les stations ayant moins de 4 relevés, c'est la valeur la plus faible observée qui est utilisé pour déterminer son état écologique.
- Pour les stations ayant 4 relevés ou plus, c'est le 25ème percentile (ou Q25) de la distribution des IBGN mesurés sur la station qui est utilisé.

Toutefois, la limite la plus importante pour l'application de la DCE est le seuil de « bon état ». Aussi, afin de simplifier la procédure de modélisation, deux classes seulement ont été retenues pour la variable « état écologique » :

- *« **bon état** » (BE), pour les stations classées en bon ou très bon état écologique (couleurs bleue et verte)*
- *« **mauvais état** » (ME) pour les stations des classes inférieures au bon état (couleurs jaune, orange et rouge).*

Détermination de l'état écologique à partir de l'IBGN.

A partir de la base de données GIRAFE, des sites de référence ont été sélectionnés selon des critères de « très faible pression anthropique », en utilisant à la fois l'avis d'experts des DIREN, selon un questionnaire précis, et des variables d'occupation du sol calculées sous SIG. Les valeurs de référence de l'IBGN ont été validées par comparaison avec des données indépendantes du ***Cemagref****.*

Pour chaque type naturel « HER x rang », on calcule la distribution des valeurs IBGN observées sur les sites de référence. La ***médiane*** *de ces observations est la* ***valeur de référence IBGN*** *pour ce type, qui correspond à la valeur la plus probable de l'IBGN en conditions de référence.*

La ***limite du « très bon état »*** *(couleur bleue) est fixée au 25ème percentile de la distribution des valeurs observées sur les sites de référence. On note que par construction, un quart des notes IBGN observées sur les sites de référence sont en-dessous de la limite du « très bon état ».*

Ensuite, en première hypothèse, ***la limite du « bon état »*** *(couleur verte) est fixée en divisant en 4 classes égales l'étendue des valeurs entre la limite du très bon état et la valeur minimum « réaliste » de l'indice, fixée à 1 pour l'IBGN. Enfin chaque limite de classe est exprimée en EQR-IBGN.*

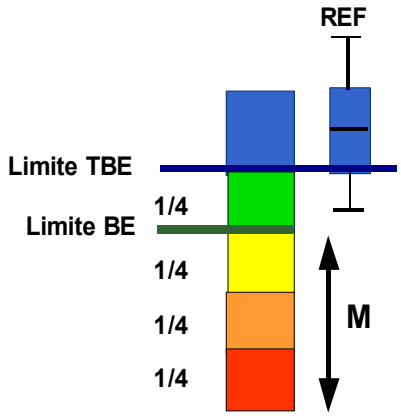


figure 3. Principe de calcul de la limite du bon état pur l'IBGN.

2.2 Calcul des pressions d'occupation du sol

Pour évaluer le degré d'anthropisation des cours d'eau de façon cohérente et homogène à l'échelle du territoire national, la seule possibilité au vu des données existantes est d'analyser l'occupation du sol, représentative des « structures anthropiques » exerçant des pressions sur les milieux aquatiques. Le couplage entre les données biologiques observées sur des stations ponctuelles et les pressions s'exerçant sur les bassins et sur l'environnement immédiat des stations (corridor rivulaire) a nécessité des développements informatiques spécifiques (langage objet) sous SIG.

2.2.1 Données utilisées

La couche d'information géographique utilisée est CORINE Land Cover (CLC) 1990. Elle couvre l'ensemble du territoire et repose sur une nomenclature standard hiérarchisée (voir annexe 1) à 3 niveaux et 44 postes répartis selon 5 grands types d'occupation du territoire (territoires artificialisés, territoires agricoles, forêts et milieux semi-naturels, zones humides et surfaces en eau). La période d'acquisition des images satellitales LANDSAT et SPOT qui ont servi à l'établissement de cette couche va de 1987 à 1994.

Il faut signaler que la couverture CORINE 2000 qui vient d'être diffusée sera utilisée dans les développements ultérieurs des modèles.

2.2.2 Délimitation des bassins versants et des corridors rivulaires

2.2.2.1 Bassin versant de la station

Compte tenu du nombre de stations ponctuelles sur lesquelles nous avons des données biologiques observées, nous avons développé un outil spécifique de calcul de bassin versant. Cet outil est basé sur l'analyse du modèle numérique de terrain dérivé de la BDALTI® de l'IGN au pas de 250 m, sur le découpage en zones hydrographiques et enfin sur le tracé du réseau hydrographique de la BD Carthage® V 3.0.
Une fois la station raccrochée géographiquement sur le tracé du cours d'eau auquel elle appartient, le traitement consiste à découper le modèle numérique de terrain en fonction de la zone amont probable à partir des zones hydrographiques puis à déterminer son bassin versant (figure 4). Etant donné que ce calcul de bassin versant est basé sur le MNT de l'IGN au pas de 250 m, la surface minimale de détection est de quelques kilomètres carrés. En zone de relief peu marqué, l'outil utilisé montre ses limites et le recours à l'expertise manuelle est parfois indispensable.
A partir de l'enveloppe du bassin versant ainsi délimitée, il est possible de calculer le pourcentage de surface des différentes catégories d'occupation du sol CORINE Land Cover.

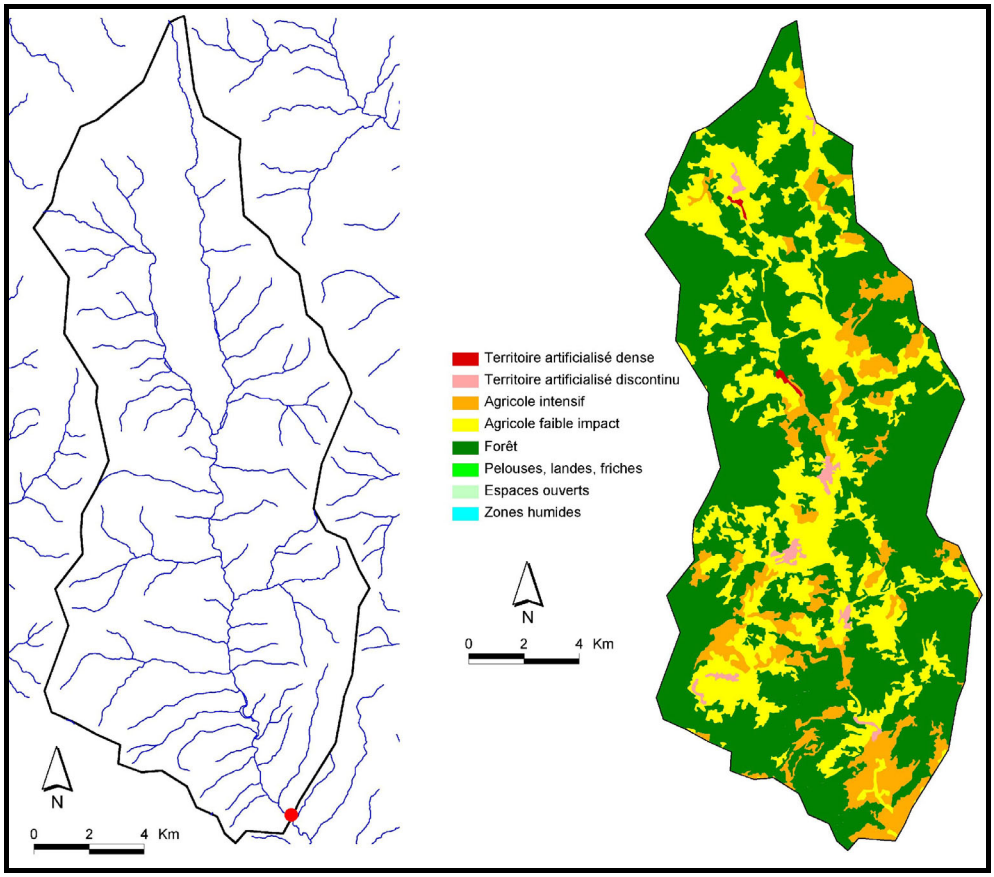


figure 4. Bassin versant calculé à partir de la station (en rouge) et pressions anthropiques dérivées de l'analyse de CORINE Land Cover.

2.2.2.2 Corridor rivulaire de la station

Les hypothèses de travail sont les suivantes :
le tronçon étudié mesure environ 3 kilomètres de long, soit 1,5 km à l'amont et à l'aval de chaque station considérée ; la largeur de la zone à prendre en compte de part et d'autre du cours d'eau dépend tout naturellement de la largeur du lit mineur. Des travaux dans le bassin de la Loire (**Souchon** et al. 2000) ont montré la possibilité d'extrapoler les largeurs moyennes des cours d'eau à partir des rangs de Strahler. La largeur de la zone de corridor rivulaire prise en compte est donc fonction du rang (100 m pour les cours d'eau de rang 1, 2 et 3, 140 m pour ceux de rang 4, 250 m pour ceux de rang 5, 600 m pour ceux de rang 6, 1200 m pour ceux de rang 7 et enfin 2400 m pour ceux de rang 8).

Seuls les arcs de même rang sont retenus pour le calcul de la zone tampon. Il est donc possible, lorsque la station est proche d'une confluence qui change le rang du cours d'eau en aval, de ne considérer que la portion du cours d'eau de même rang que celle sur laquelle est située la station. Le calcul du type d'occupation du sol est exprimé en pourcentage. Ce traitement a été automatisé par l'écriture d'un script en langage Avenue sous ArcView 3.2. Un exemple de résultat est présenté figure 5.
Etant donné que l'échelle de travail de Corine Land Cover est le 1/100 000ème, l'analyse présentée est en limite d'utilisation. Ce traitement permet d'approcher l'occupation dominante du fond de vallée. L'analyse de la ripisylve ou de l'artificialisation des berges nécessiterait d'utiliser une source de donnée plus précise (**Perez-Correa** 2004).

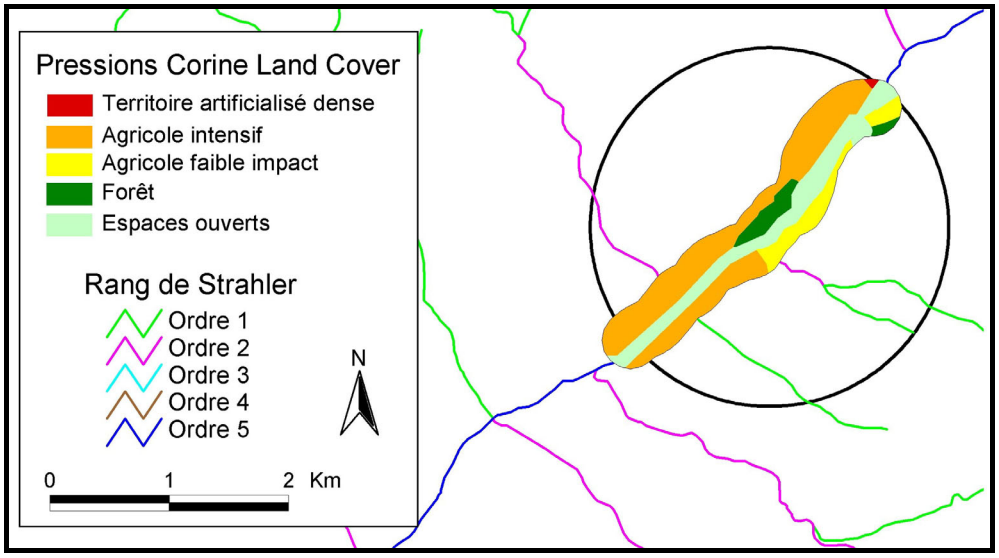


figure 5. Exemple de calcul des pressions anthropiques dérivées de l'analyse de CORINE Land Cover à l'échelle du corridor rivulaire d'une station.

3 Description des méthodes d'analyses de données

Etablir des modèles pressions-impacts nécessite plusieurs étapes techniques comme l'analyse des données de pressions et la sélection des données pertinentes vis à vis de la variable cible, mais il est aussi nécessaire d'utiliser des méthodes prédictives permettant d'extrapoler les résultats des modèles sur une carte.

Dans ce but, nous avons utilisé l'analyse en composante principale, la régression PLS et les arbres de décision.

3.1 Corrélations et analyse en composantes principales (ACP)

L'ACP est une méthode très efficace pour l'analyse de données quantitatives (continues ou discrètes) se présentant sous la forme de tableaux à *M lignes (stations)/ N colonnes (variables de pressions)*.

Elle permet de :

- visualiser et analyser rapidement les corrélations entre les *N variables de pression*,
- visualiser et analyser les *M stations* initialement décrites par *N variables de pression* sur un graphique à deux ou trois dimensions, construit de manière à ce que la dispersion entre les données soit aussi bien préservée que possible,
- construire un ensemble de *P facteurs* non corrélés (*P*<=*N*) qui peuvent ensuite être réutilisés par d'autres méthodes (la régression par exemple).

Les données sont souvent collectées sur des variables qui ne sont pas seulement corrélées, mais sont aussi très nombreuses. Cela rend l'interprétation des données et la détection de sa structure difficile. En transformant les variables originales en un nombre plus petit de facteurs non corrélés, l'Analyse en Composantes Principales (ACP) rend ces deux tâches plus faciles.

3.2 La régression Partial Least Squares (PLS) (Tenenhaus, 1998)

La méthode de modélisation la plus fréquemment utilisée en écologie est sans aucun doute la régression linéaire multiple. C'est une méthode qui permet de mettre à disposition un outil dont le pouvoir de représentation est extrêmement large.

Ce type de modèle est utilisé pour rendre compte des relations linéaires simples existant entre une variable dépendante et des prédicteurs.

Le modèle est le suivant :

$$Y = \alpha + \sum_{j=1}^{p} \beta_j X_J + \varepsilon$$

où Y est la variable indépendante, $(X_1,\ldots,X_P)$ le jeu de prédicteurs et ε le résidu du modèle suivant une loi normale $N(0,\sigma^2)$ et α l'ordonnée à l'origine. Le paramètre estimé β_j est obtenu par la méthode des moindres carrés.

La régression PLS se présente comme une extension de ce modèle linéaire classique. Son objectif est de palier le principal défaut de cette dernière : l'instabilité des coefficients de régression due à la colinéarité des prédicteurs. En effet, lorsque la colinéarité devient forte au sein du jeu de variables prédictives (corrélations fortes, grand nombre de prédicteurs…), les estimations des coefficients de régression fluctuent

énormément d'un échantillon à l'autre. Souvent, l'addition ou le retrait de quelques données dans l'échantillon a des répercussions sur les variables retenues dans le modèle. Aussi, lorsque l'on utilise la régression multiple pour évaluer l'importance relative des prédicteurs, l'interprétation devient aussi délicate que dangereuse. En effet, plus les variables prédictives sont inter-reliées, moins les coefficients de régression seront fiables pour évaluer leur importance relative.

***Algorithme de régression PLS (Tenenhaus 1998)** :*

On cherche à réaliser une régression d'une variable à expliquer y sur des prédicteurs x_1, x_2, .., x_n qui peuvent être fortement corrélées entre elles ou être plus nombreuses que le nombre d'observations. Par ailleurs, les coefficients doivent être interprétables dans le sens où l'on doit pouvoir mesurer la contribution de la variable x_j à la construction de la variable y à l'aide du coefficient de régression.

- *On construit d'abord une composante* $t_1 = w_{11}x_1 + ... + w_{1p}x_p$ *où* $w_{1j} = \frac{\mathrm{cov}(x_j, y)}{\sqrt{\sum_{j=1}^{p} \mathrm{cov}^2(x_j, y)}}$*, ce qui revient à chercher une* ***combinaison linéaire des prédicteurs maximisant la covariance entre les prédicteurs et la variable dépendante****.*

- *On effectue ensuite* ***une régression simple de y sur t_1*** $y = c_1 t_1 + y_1$ *où c1 est le coefficient de régression et y_1 le vecteur des résidus d'où une première équation de régression* $y = c_1 w_{11} x_1 + ... + c_1 w_{1p} x_p + y_1$ *dont les coefficients sont très faciles à interpréter.*

- *On construit ensuite de la même façon une* ***composante t_2****, combinaison linéaire des x_j, non corrélée à t_1 et expliquant bien le vecteur des résidus y_1 :*

  $t_2 = w_{21}x_{11} + ... + w_{2p}x_{1p}$ *où* $w_{2j} = \frac{\mathrm{cov}(x_{1j}, y_1)}{\sqrt{\sum_{j=1}^{p} \mathrm{cov}^2(x_{1j}, y_1)}}$

- *On effectue ensuite* ***une régression de y sur t_1 et t_2*** *:* $y = c_1 t_1 + c_2 t_2 + y_2$ *ce qui permet d'obtenir une équation de régression plus précise que la première.*

- *Cette procédure itérative se poursuit jusqu'à la* ***régression de y sur t_1, t_2, …, t_H*** *où H est égal au nombre de prédicteurs.*

- ***Le nombre de composantes à retenir est ensuite déterminé par validation croisée*** *selon un critère de minimisation de l'erreur, le* ***PRESS*** *(PRédiction Error Sum of Squares) qui correspond à la somme de tous les carrés des erreurs de prévision calculées sur les jeux tests.*

On obtient ainsi une équation de régression comparable à l'équation du modèle de régression linéaire classique. Cependant, le processus de sélection des composantes t implique quelques différences :

- les coefficients des prédicteurs sont interprétables même en cas de corrélation forte entre les prédicteurs,

- Ils sont comparables (signe et amplitude) avec le coefficient de corrélation simple entre un prédicteur et la variable à expliquer,
- Une partie seulement de la variabilité des prédicteurs est utilisée pour construire cette équation de régression.

L'équation finale s'interprète enfin comme une équation de régression linéaire, les coefficients reflètent bien l'effet d'un prédicteur relativement aux autres et le R^2 est un bon estimateur de l'efficacité du modèle.

3.3 Les arbres de décision (Breiman et al. 1984)

3.3.1 Descriptif de la méthode

L'objectif de cette méthode est le partitionnement récursif de l'espace des observations en sous-domaines les plus homogènes possible 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 contenant le test de la variable est alors construit. 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.

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 présentera 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 un arbre 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 (son nombre de nœuds). L'arbre retenu est alors celui qui réalise le meilleur compromis entre erreur et taille.

> *L'**algorithme de construction des arbres de décision** (**Boët** et al. 2001) se déroule en deux phases successives :*
>
> *- À chaque pas de construction de l'arbre, l'algorithme détermine la meilleure séparation linéaire des observations de l'échantillon d'apprentissage, puis récursivement sur chaque sous-ensemble de l'échantillon ainsi formé. 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. En d'autres termes, cela consiste à effectuer une partition de l'espace X des observations vers l'ensemble C des différentes classes du problème. Cependant, la distribution des données dans l'espace des X induit fréquemment des chevauchements. Il n'existe donc pas de partition décrivant complètement les classes. La méthode utilisée pour effectuer la construction de l'arbre fait appel à une procédure « pas à pas ». Cela signifie qu'elle effectue la séparation suivante de manière optimale sans tenter d'optimiser les performances à l'échelle de l'arbre entier.*
>
> *- 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 fini 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, T la taille de l'arbre, et Rα(T) le critère pénalisé obtenu par Rα(T) = R(T) + α * Nombre de nœuds (T). Il est immédiat que si nous faisons augmenter α, la taille de l'arbre le pénalisera de plus en plus. Il existe 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, il 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 Nième 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 Rα(T) pour tout α. Le paramètre α correspondant au minimum est le paramètre déterminant finalement le meilleur arbre élagué.

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 qu'à une boîte noire!

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.

Exemple d'arbre de régression : le Titanic

Quels passagers du Titanic avaient le plus de chances de survivre au naufrage ? Dans ce cas, la variable d'intérêt est la survie. Et nous pouvons diviser les passagers en groupes basés sur l'âge, le sexe et la classe pour chercher ensuite la proportion de survivants dans chaque groupe. L'algorithme de construction de l'arbre choisit automatiquement les ensembles qui donnent des groupes homogènes ayant la plus grande différence de taux de survie.

Dans cet exemple, la méthode divise d'abord les observations en deux groupes : hommes et femmes. Chaque groupe est à son tour subdivisé et permet au final de connaître la proportion de survivants d'une catégorie donnée de passagers. L'arbre de la figure 6 donne ces résultats de façon synthétique.

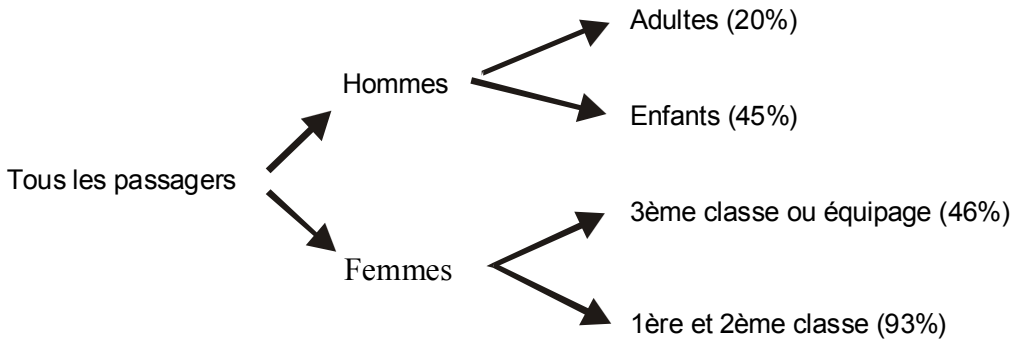


figure 6. Arbre de survie des passagers du Titanic.

3.3.2 Evaluation de l'efficacité des modèles (courbes ROC)

L'arbre de classification prédit en sortie de modèle une classe de la variable qualitative à prédire. L'évaluation de l'efficacité du modèle se fait à partir de la matrice de confusion. Cette matrice est le résultat d'un tableau croisé entre les valeurs observées de la variable qualitative et les valeurs prédites par le modèle pour cette même variable (tableau 1).

		Valeurs observées	
		ME	B
Valeurs Prédites	ME	Vrais positifs (a)	Faux positifs (b)
	B	Faux négatifs (c)	Vrais négatifs (d)

tableau 1. Exemple de matrice de confusion

Dans un premier temps, cette matrice permet de calculer le taux de mauvaise classification du modèle. Ce taux d'erreur est une première estimation de l'erreur du modèle.

On définit ensuite à partir de la matrice de confusion, la *sensibilité* (Se) et la *spécificité* (Sp) du modèle. La sensibilité est la proportion des vrais positifs parmi les stations en mauvais état. La spécificité est la proportion des vrais négatifs parmi les stations en bon état.

$$Se = \frac{a}{a+c} \qquad Sp = \frac{d}{b+d}$$

Les courbes ROC (Receiver Operating Characteristic) (**Hanley** et al. 1982) permettent d'étudier les variations de la spécificité et de la sensibilité d'un test pour différentes

valeurs du seuil de discrimination. Le terme de courbe ROC peut être envisagé comme une *"courbe de caractéristiques d'efficacité"*.

On porte sur l'axe des abscisses, la variable '1-spécificité', cette variable est égale à l'effectif de faux positifs parmi les stations en bon état. Sur l'axe des ordonnées, on retrouve la sensibilité, égale à l'effectif de vrais positifs parmi les stations en mauvais état. La courbe se construit de façon empirique en calculant la sensibilité puis la spécificité d'un test pour différents niveaux de seuils de discrimination (figure 7).

L'*aire sous la courbe ROC* est un estimateur de *l'efficacité globale* du test ; si le test n'est pas informatif, l'aire est de 0.5. Si le test est parfaitement discriminatif, l'aire sera de 1.

(**Swets** 1988) a définit une échelle d'interprétation de l'efficacité d'un test en fonction de la valeur de l'AUC (Area Under the Curve) (tableau 2)

AUC	Qualité du modèle
0.5-0.7	faible
0.7-0.9	satisfaisante
>0.9	excellente

tableau 2. Echelle d'efficacité d'un modèle en fonction de l'aire sous la courbe ROC (AUC)

Il existe différentes méthodes d'approximation de cette aire dont celle de Hanley (1982) qui n'est pas paramétrique c'est à dire qu'elle ne nécessite pas d'approximation statistique sur la courbe. Elle est fondée sur un calcul de rang.

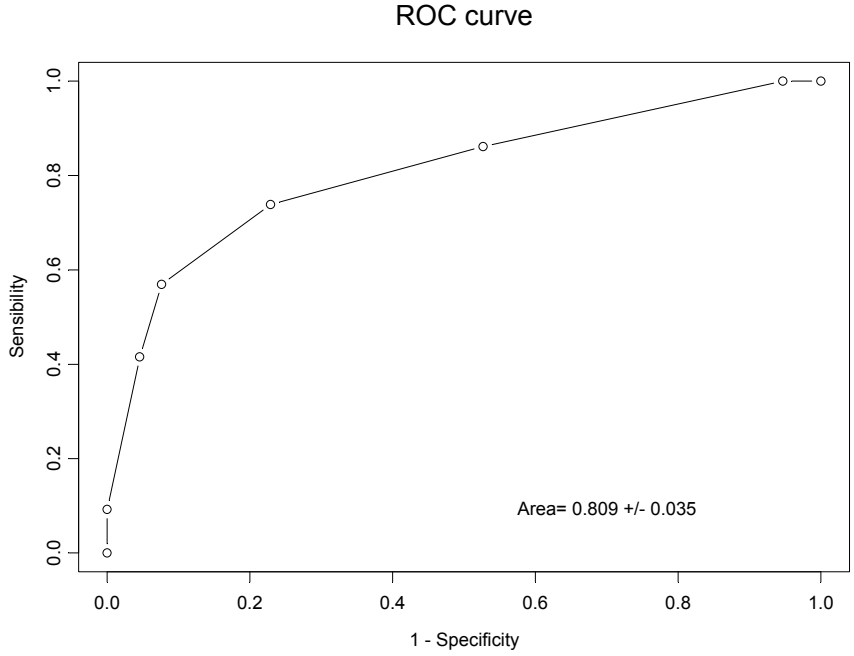


figure 7. Exemple de courbe ROC

4 Sélection des variables de pressions

4.1 Pressions d'occupation du sol

Dans un premier temps, nous avons éliminé les postes CORINE Land Cover sous représentés (tableau 3). Plus concrètement, les postes présentant plus de 90% de valeurs nulles dans les bassins versants des stations de notre base n'ont pas été utilisé pour l'analyse (figure 8) mais ont été pris en compte tout de même dans la constitution des groupes d'occupation du sol.

Codes CLC conservés	Codes CLC éliminés
1.1.1 – Tissu urbain continu	1.2.3 – Zones portuaires
1.1.2 – Tissu urbain discontinu	1.3.2 – Décharges
1.2.1 – Zones industrielles et commerciales	1.3.3 – Chantiers
1.2.2 – Réseaux routier et ferroviaires	1.4.1 – Espaces verts urbains
1.2.4 – Aéroports	2.1.2 – Périmètres irrigués
1.3.1 – Extraction de matériaux	2.1.3 – Rizières
1.4.2 – Equipements sportifs et de loisirs	2.2.3 – Oliveraies
2.1.1 – Terres arables hors irrigation	2.4.1 – Cultures an. Associées aux cult. perm.
2.2.1 – Vignobles	2.4.4 – Territoires agro-forestiers
2.2.2 – Vergers et petits fruits	3.3.1 – Plages, dunes, sable
2.3.1 – Prairies	3.3.4 – Zones incendiées
2.4.2 – Systèmes culturaux et parcellaires	3.3.5 – Glaciers, neiges éternelles
2.4.3 – Agriculture et veg. naturelle	4.1.2 – Tourbières
3.1.1 – Forêts de feuillus	4.2.1 – Marais maritimes
3.1.2 – Forêts de conifères	4.2.2 – Marais salants
3.1.3 – Forêts mélangées	4.2.3 – Zones intertidales
3.2.1 – Pelouses et pâturages naturels	5.1.1 – Cours et voies d'eau
3.2.2 – Landes et broussailles	5.2.1 – Lagunes littorales
3.2.3 – Végétation sclérophylle	5.2.2 – Estuaires
3.2.4 – Forêt et veg. arbustive	5.2.3 – Mers et océans
3.3.2 – Roches nues	
3.3.3 – Végétation clairsemée	
4.1.1 – Marais intérieurs	
5.1.2 – Plans d'eau	

tableau 3. Postes CLC conservés et éliminés pour l'analyse et descriptif rapide

En effet, une trop forte proportion de zéros dans les tableaux utilisés pour l'analyse de données entraînait une incapacité technique à réaliser une validation complète des modèles. Nous avons donc choisi d'éliminer ces postes en partant du postulat que si ces variables pouvaient avoir un effet significatif à une échelle d'étude locale, les chances de leur voir attribuer un impact important à une échelle nationale étaient quasi nulles. D'autre part, les principales catégories d'occupation du sol restent représentées de manière satisfaisante.

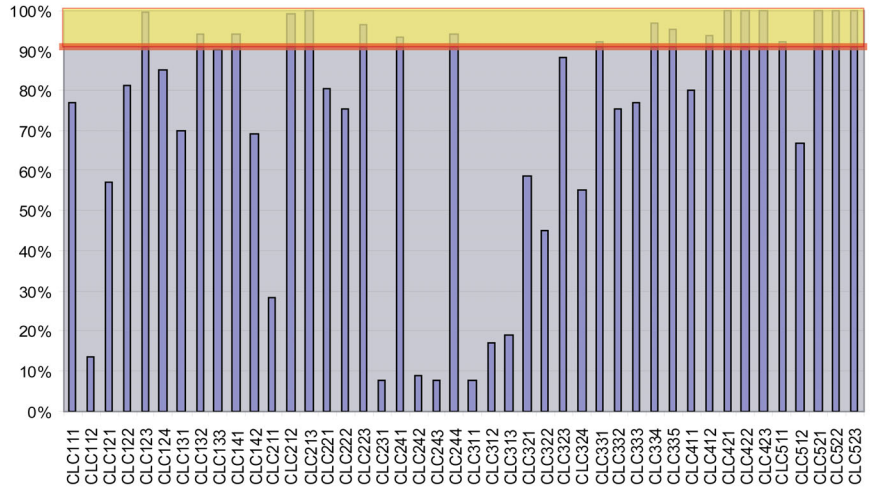


Figure 8. Pourcentage de valeurs nulles dans les bassins versants des stations de la base de données CEMAGREF pour les 44 postes CORINE land cover. Les variables ayant plus de 90% de valeurs nulles ont été éliminées de l'analyse

Ensuite, nous avons réalisé une régression PLS de la variable EQR-IBGN sur les 24 postes CLC retenus (tableau 3), pour tester le lien entre ces variables d'occupation du sol et l'EQR des stations.

Les coefficients de régression PLS représentent la contribution relative de chaque variable à la variation de l'EQR-IBGN ; ils sont visualisés figure 9. Parallèlement ont été représentés les coefficients de corrélation simple, qui servent à quantifier la relation entre l'EQR-IBGN et chacune des variables indépendamment des autres. On constate sur cette figure que pour une variable donnée, ces deux coefficients varient toujours dans le même sens et dans des proportions équivalentes, ce qui permet de confirmer la validité des coefficients de la régression PLS.

L'étude des coefficients de la régression PLS permet de différencier 6 groupes d'occupation du sol correspondant à des intensités de pressions différentes sur l'IBGN :

groupe 1 : effet fortement négatif - CLC 111, 112,121
groupe 2 : effet moyennement négatif - CLC 122, 124, 131, 142
groupe 3 : effet faiblement négatif - CLC 211, 221, 222, 231, 242
groupe 4 : effet nul - CLC 243, 311, 312, 313
groupe 5 : effet faiblement positif - CLC 321, 322, 323, 324, 332, 333
groupe 6 : effet moyennement positif - CLC 411, 512

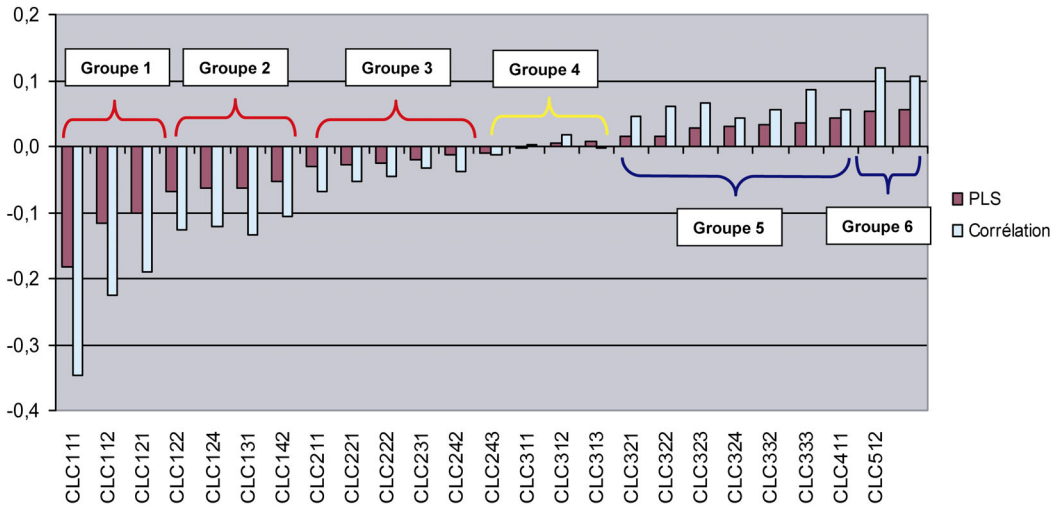


figure 9. Coefficients de la régression PLS de l'EQR en fonction des 24 postes CLC sélectionnés et corrélations linéaires correspondantes et identification de 6 groupes d'intensité de pression.

Nous avons ensuite retravaillé ces groupes en essayant de conserver la nature des différentes variables dans le classement final. Les groupes 1 et 2 correspondent à des espaces artificialisés, le groupe 3 à des espaces d'agriculture intensive et les groupes 4, 5 et 6 représentent l'agriculture de « faible intensité » (notamment prairies) ou les espaces naturels (y compris les divers types de forêts).

Sur la base de ces résultats, 4 grands types d'occupation du sol ont été retenus pour la modélisation :

- **Territoires artificialisés** (variable ARTIF) : zones urbanisées, zones industrielles ou commerciales et réseaux de communication, mines, décharges et chantiers, espaces verts artificialisés, non agricoles - codes CLC : 1.1.1, 1.1.2, 1.2, 1.3, 1.4.1, 1.4.2.
- **Territoires agricoles de forte intensité** (variable AGRI_I) : terres arables, cultures permanentes, vergers, vignes, oliveraies, cultures

annuelles associées aux cultures permanentes, systèmes culturaux et parcellaires complexes - codes CLC : 2.1, 2.2, 2.4.1, 2.4.2.

- **Territoires agricoles de faible intensité** (variable AGRI_F) : prairies, territoires principalement occupés par l'agriculture avec présence de végétation naturelle importante, territoires agro-forestiers - codes CLC : 2.3.1, 2.4.3, 2.4.4.
- **Territoires à faible anthropisation** (variable ESP_NAT) : forêts et milieux semi-naturels, zones humides, surfaces en eau - codes CLC : 3.1.1, 3.1.2, 3.1.3, 3.2, 3.3, 4 et 5.

4.2 Corrélation entre occupation du sol et pressions polluantes

Dans une analyse des relations entre l'IBGN et l'occupation des sols, on admet implicitement que les zones urbanisées correspondent à des sources de rejets directs, et que l'intensité de ces rejets est proportionnelle en première approximation à la superficie urbanisée. Pour vérifier cette hypothèse, nous avons étudié la corrélation entre les **rejets connus des Agences de l'Eau** et les 4 variables définies précédemment pour synthétiser les grands types d'occupation du sol.

Les données concernant les rejets sont issues de l'étude **SIEE-*Cemagref*** « Définition d'un réseau national de stations ou de tronçons de référence » (SIEE et al. 2002). Dans cette étude, les rejets ont été évalués pour chacune des 6200 zones hydrographiques (ZH), en cumulant l'ensemble des rejets connus à l'amont de l'exutoire de chaque ZH. En l'état actuel, il n'est pas possible d'affecter ces rejets aux bassins versants réels des stations biologiques, mais seulement à partir d'une approximation basée sur le cumul des zones hydrographiques amont de chaque station ; ceci pourrait entraîner une augmentation de l'erreur du modèle liée à la mauvaise adéquation pression-station si elles étaient introduites dans les analyses au même titre que les pressions d'occupation du sol.

Les *variables de rejets* sont les suivantes :
- DBO5
- NH4+
- Pt (phosphore total)
- MI (matières inhibitrices)
- METOX (métaux lourds)

Par ailleurs, la pression polluante provenant des élevages ne peut être évaluée à partir de la simple superficie des prairies, car une partie importante des élevages, dans certaines régions, sont réalisées hors sol. Dans l'étude SIEE ***Cemagref***, la densité du cheptel a été calculée pour chaque ZH et la variable correspondante (densCHEP) a également été introduite dans l'analyse. Il n'est pas encore possible d'affecter ces rejets aux bassins versants réels des stations.

Les corrélations simples entre ces variables de pressions et les variables d'occupation du sol sont fournies dans le tableau suivant (tableau 4). Nous avons également procédé à une analyse en composantes principales (ACP) de l'ensemble des variables (occupation du sol et rejets) pour mieux observer leur organisation relative.

	ARTIF	AGRI_I	AGRI_F	ESP_NAT
DBO5	***0.36***	***0.18***	-0.11	***-0.14***
Nh4	***0.45***	0.06	-0.08	-0.08
Pt	***0.43***	0.10	-0.08	-0.11
MI	***0.18***	0.05	-0.04	-0.05
METOX	***0.13***	0.08	0.00	-0.09
densCHEP	-0.05	***0.31***	***0.51***	***-0.57***

tableau 4. Corrélations entre les variables de rejet et les groupes de variables d'occupation du sol

L'examen de la matrice de corrélation (tableau 4) montre que les variables DBO5, NH4, Pt sont positivement corrélées de façon conséquente (coefficient > 0.30) au pourcentage de territoires artificialisés (ARTIF) ; les rejets toxiques (MI et METOX) sont également partiellement corrélées positivement à ARTIF.

La densité de cheptel (densCHEP) est fortement corrélée négativement avec les espaces naturels, et positivement avec les territoires agricoles ; cette corrélation est logiquement plus forte avec les prairies (AGRI_F), mais la corrélation avec les cultures (AGRI_I), montre bien l'importance des élevages hors-sol.

Ces résultats sont illustrés par le cercle des corrélations des variables représentant plus de 50% de la variabilité totale des données de pressions (figure 10).

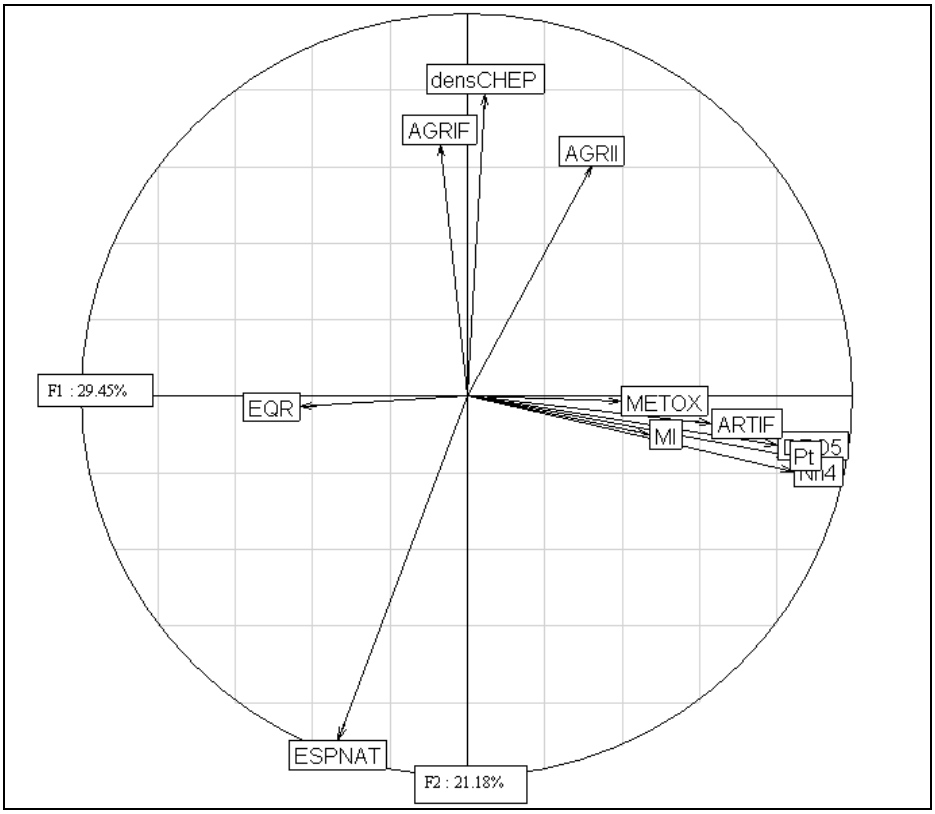


figure 10. Cercle des corrélations de l'ACP du tableau de données « pressions » (plan 1, 2).

Deux conclusions importantes ressortent de cette analyse :

- les territoires artificialisés (ARTIF) représentent correctement les pressions polluantes des rejets domestiques (DBO5, NH4+, Pt), mais assez mal les rejets toxiques (MI, METOX).
- Les pressions liées aux élevages sont principalement représentées par les prairies (AGRI_F), mais ne sauraient être exclues des zones de cultures (AGRI_I).

4.3 Discussion : intérêt et limites des variables d'occupation du sol

On peut en conclure en première analyse que *les variables d'occupation du sol représentent bien les principales sources de pressions polluantes*.

Le principal intérêt, à l'heure actuelle, de l'utilisation de l'occupation du sol est de permettre une affectation plus précise des pressions aux bassins versants réels et aux corridors rivulaires des stations biologiques. En effet, les sources de données existantes à l'échelle nationale ne permettent d'affecter les rejets polluants qu'aux zones hydrographiques pour les rejets urbains, et encore avec une approximation certaine, et à l'échelle des cantons pour les pollutions dues aux élevages en raison de la confidentialité des données du recensement agricole (RGA).

Pour éviter l'erreur topologique dans l'affectation des rejets, ainsi qu'une redondance dans le jeu de données de pressions pouvant altérer la qualité des analyses, nous avons choisi de ne conserver que les variables d'occupation du sol.

Toutefois, cette approche a des limites évidentes :

- L'occupation du sol par les « territoires artificialisés » ne tient compte ni de la densité de population des zones urbanisées, ni du taux de traitement des rejets urbains, et représente assez mal les sources ponctuelles de rejets toxiques.
- Aux zones de grandes cultures (« agriculture intensive ») peuvent être associées les sources de pollutions diffuses (nutriments, pesticides…), mais à l'exception des nitrates, la relation quantitative entre superficie cultivée et rejets polluants est loin d'être simple ; les effets de proximité, qui sont analysés à travers l'occupation des corridors rivulaires, peuvent avoir une grande importance.
- Les pollutions dues aux élevages ne sauraient être directement reliées aux superficies de prairies (classées en « agriculture faible »), car l'intensification des pratiques d'élevage est très différente selon les régions.
- Enfin, l'occupation du sol ne révèle qu'indirectement les pressions hydro-morphologiques ; certaines altérations physiques comme la chenalisation peuvent être systématiquement suspectées dans les zones urbanisées où d'agriculture intensive, mais nombre de pressions importantes comme les barrages, les dérivations, la navigation, l'endiguement, ne sont pas du tout prises en compte par les variables d'occupation du sol.

On devra donc garder à l'esprit ces limitations pour l'interprétation des résultats.

5 Modèles d'extrapolation spatiale

Les ***modèles d'extrapolation spatiale*** permettent de représenter à l'échelle nationale *l'état écologique actuel probable* des cours d'eau, avec l'IBGN, et en fonction de différentes hypothèses de limites de bon état.

5.1 Construction des modèles d'extrapolation spatiale

Le principe de l'extrapolation spatiale du modèle va se dérouler en trois temps comme décrit dans la figure 11.

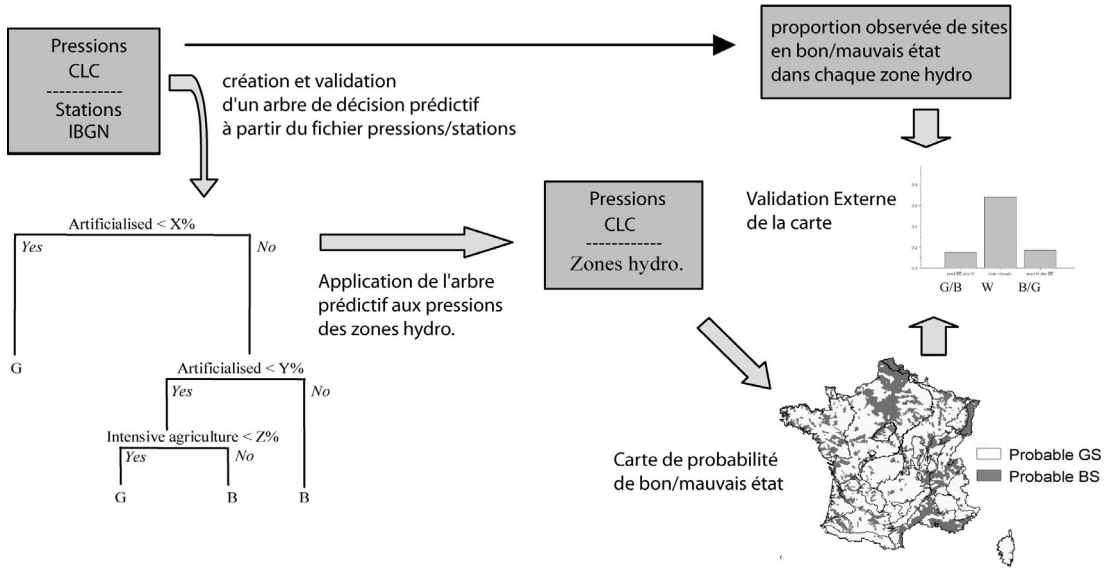


figure 11. Description du processus de modélisation et d'extrapolation spatiale de l'arbre de décision validé.

1 – En premier lieu, un arbre de décision est développé et validé sur la base des variables d'occupation du sol calculées pour les bassins versants réels des stations IBGN.

2 – Cet arbre de décision est ensuite utilisé pour prédire l'état écologique actuel probable (« bon » ou « mauvais ») des zones hydrographiques, d'après l'occupation du sol du bassin versant cumulé à l'amont de chaque ZH.

3 – On utilise les sorties prédictives du modèle pour représenter sur une carte la probabilité actuelle de non atteinte du bon état pour le drain principal de chaque ZH.

4 – On procède ensuite à une validation externe du modèle en comparant l'état des ZH « prédit » par le modèle et l'état « observé » sur des stations situées dans les mêmes ZH.

Dans la mesure où il s'agit d'un modèle que nous considérons comme fiable uniquement sur les drains principaux, l'état « observé » des zones hydrographiques est déterminé en utilisant les stations situées en aval du drain principal pour chaque ZH. Cette validation externe est donc effectuée sur 1802 ZH.

5.2 Choix des modèles d'extrapolation spatiale

Les arbres correspondant aux modèles d'extrapolation sont présentés en détail dans l'annexe 2. Ces arbres ont été développés et validés dans un but prédictif et non explicatif. De ce fait, ils ne répondent pas toujours à une représentation théorique acceptable de l'information. Mais ils répondent en revanche à la question posée : prédire le plus efficacement possible l'état d'une station.

Une analyse explicative plus poussée sera réalisée à l'aide de modèles de diagnostic (chapitre 7).

Dans la mesure où nous utiliserons les numéros des hydro-écorégions tout au long de ce chapitre, le tableau 5 permettra de faire la correspondance avec leur nom.

n°	HER1	n°	HER1
1	PYRENEES	12	ARMORICAIN
2	ALPES INTERNES	13	LANDES
3	MASSIF CENTRAL	14	COTEAUX AQUITAINS
4	VOSGES	15	PLAINE SAONE
5	JURA-PREALPES DU NORD	16	CORSE
6	MEDITERRANEEN	17	DEPRESSIONS SEDIMENTAIRES
7	PREALPES DU SUD	18	ALSACE
8	CEVENNES	19	GRANDS CAUSSES
9	TABLES CALCAIRES	20	DEPOTS ARGILO SABLEUX
10	COTES CALCAIRES EST	21	MASSIF CENTRAL NORD
11	CAUSSES AQUITAINS	22	ARDENNES

tableau 5. Identification des hydro-écorégions de niveau 1

5.3 Modèle type « France entière »

Dans un premier temps, nous avons procédé à la mise en place d'un arbre de décision prédisant l'état des stations sur la base de l'occupation du sol pour la France entière (figure 12).

La validation croisée du modèle général a été répétée 100 fois afin de déterminer de manière plus précise le nombre de nœuds à retenir pour l'élagage. L'arbre à 36 nœuds est celui qui minimisait l'erreur de classement du modèle de façon récurrente sur l'ensemble des répétitions. Nous l'avons donc utilisé comme arbre prédictif.

L'erreur de classement du modèle est de 26% ce qui signifie qu'il prédit l'état écologique d'une station avec moins d'une chance sur quatre de se tromper. L'aire sous la courbe ROC (AUC) est de 0.75 ; le modèle a donc une capacité prédictive moyenne mais satisfaisante.

Toutefois, on constate sur la figure 12 A que l'essentiel de l'explication du modèle est donné par les premières branches. La séquence de l'arbre à 4 nœuds n'a que 32% d'erreur, contre 26% pour l'arbre optimum. Cette séquence été représentée dans un but illustratif (figure 12 B), l'arbre a 36 nœuds étant trop complexe pour être représenté.

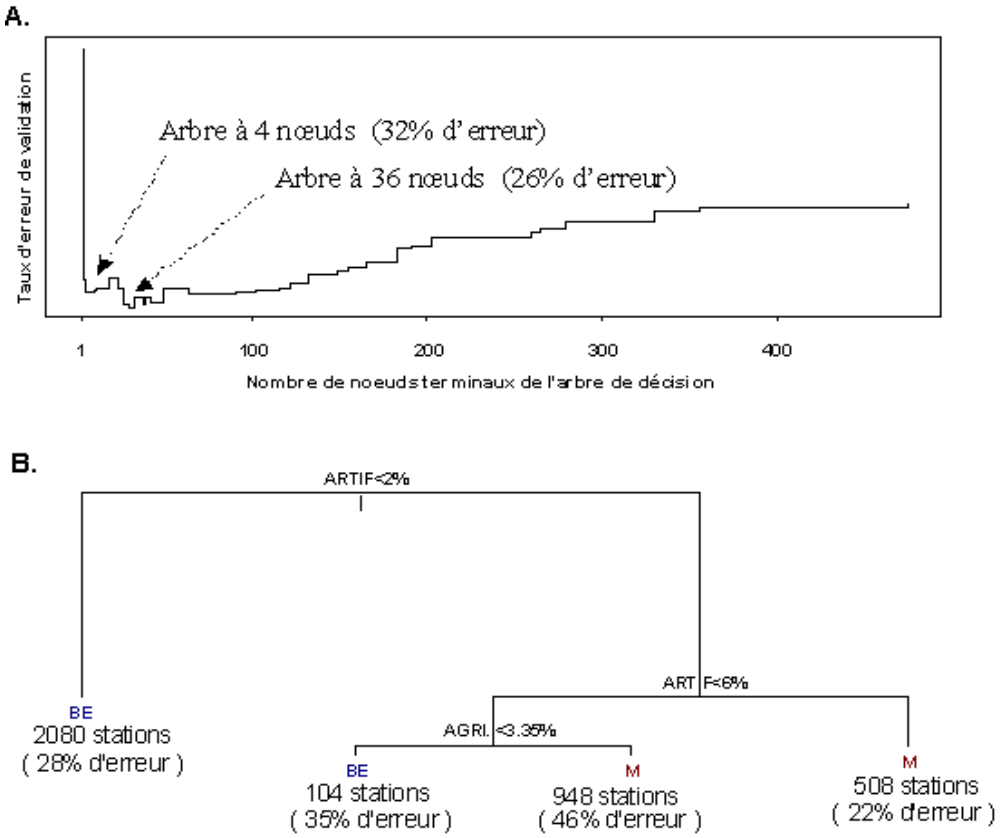


figure 12. *A. Graphique de validation croisée représentant l'erreur totale de validation en fonction du nombre de nœuds terminaux de l'arbre. B. Arbre de décision à 4 nœuds élagué d'après la validation croisée. Chaque nœud terminal donne la classe d' « état » prédite par le modèle, le nombre de stations associées à ce nœud et entre parenthèses, l'erreur de prédiction associée à chaque nœud.*

En suivant cet arbre, la probabilité de « bon état » basée sur l'IBGN est principalement liée à la proportion de territoires artificialisés dans le bassin versant, avec un premier seuil à 2% en dessous duquel les stations sont majoritairement (à 72%) en bon état, et un second seuil à 6 % au dessus duquel les stations sont majoritairement (à 78%) en mauvais état. L'agriculture intensive intervient secondairement dans le modèle ; entre les deux seuils d'artificialisation ci-dessus, les stations aux bassins non cultivés apparaissent plutôt en bon état.

Ces résultats semblent cohérents si l'on se réfère aux distributions des variables d'occupation du sol en fonction de l'état des stations présentées dans la figure 13. Sur cette figure, seuls les territoires artificialisés révèlent une différence nette de distribution entre les bassins des stations en « bon » et en « mauvais » état.

Toutefois, comme nous l'avons signalé, ces modèles n'ont pas un but explicatif mais prédictif, et leur interprétation doit être complétée par d'autres analyses.

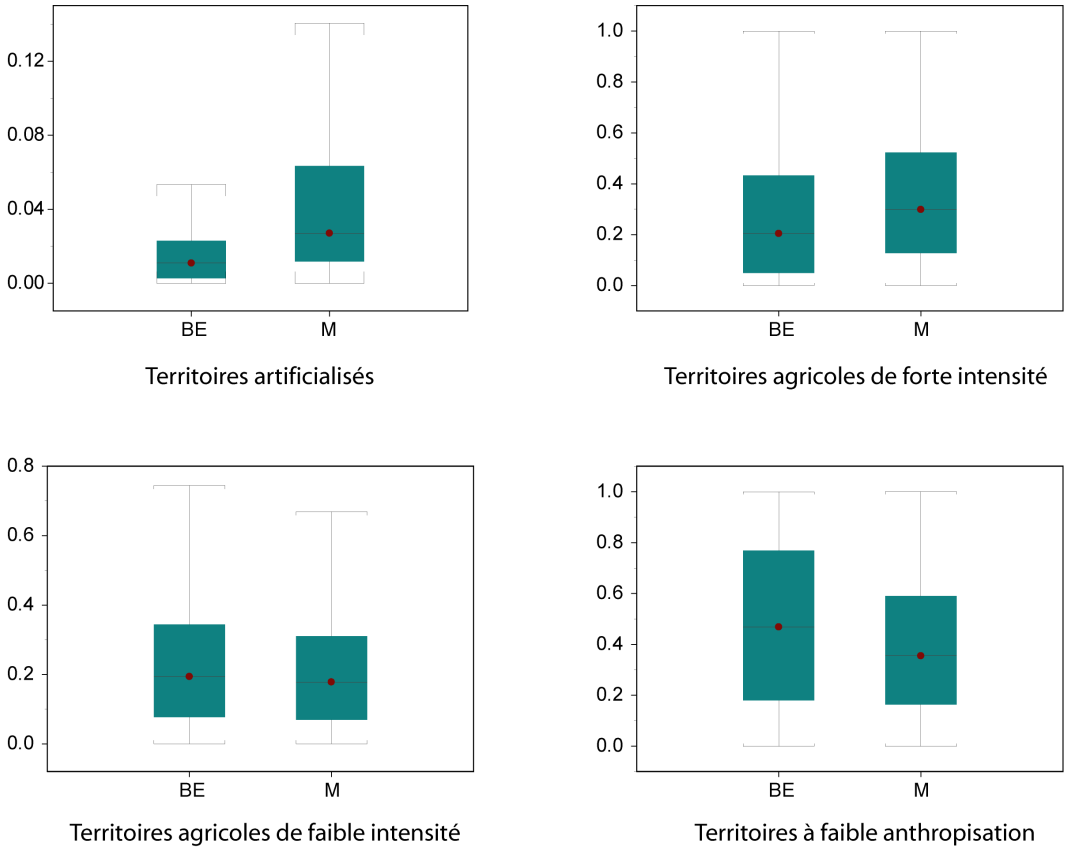


figure 13. Boites à moustaches représentant les distributions des variables de pression en fonction de l'état des 3640 stations de la France entière.

Lors de la validation externe, réalisée à partir de l'état des stations situées dans la partie aval des 1802 zones hydrographiques du jeu de validation, nous avons observé un taux de mauvais classement de 32%, cohérent avec l'erreur du modèle (figure 14). La répartition quasi symétrique de l'erreur de classement montre que le modèle n'est pas biaisé par une erreur systématique.

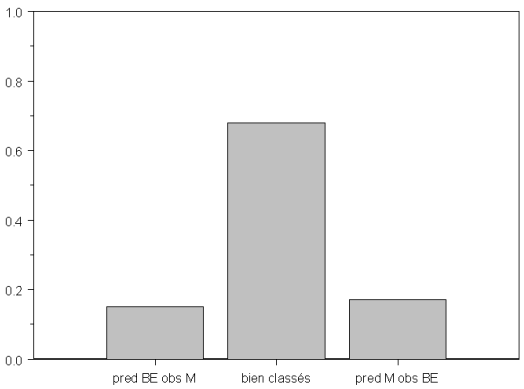


figure 14. Comparaison de l'état des stations « prédit » en sortie du modèle « France entière » et de l'état « observé » sur 1802 stations situées en partie aval du drain principal des zones hydrographiques.

Le résultat de l'extrapolation spatiale de ce modèle est représenté sur la figure 15.

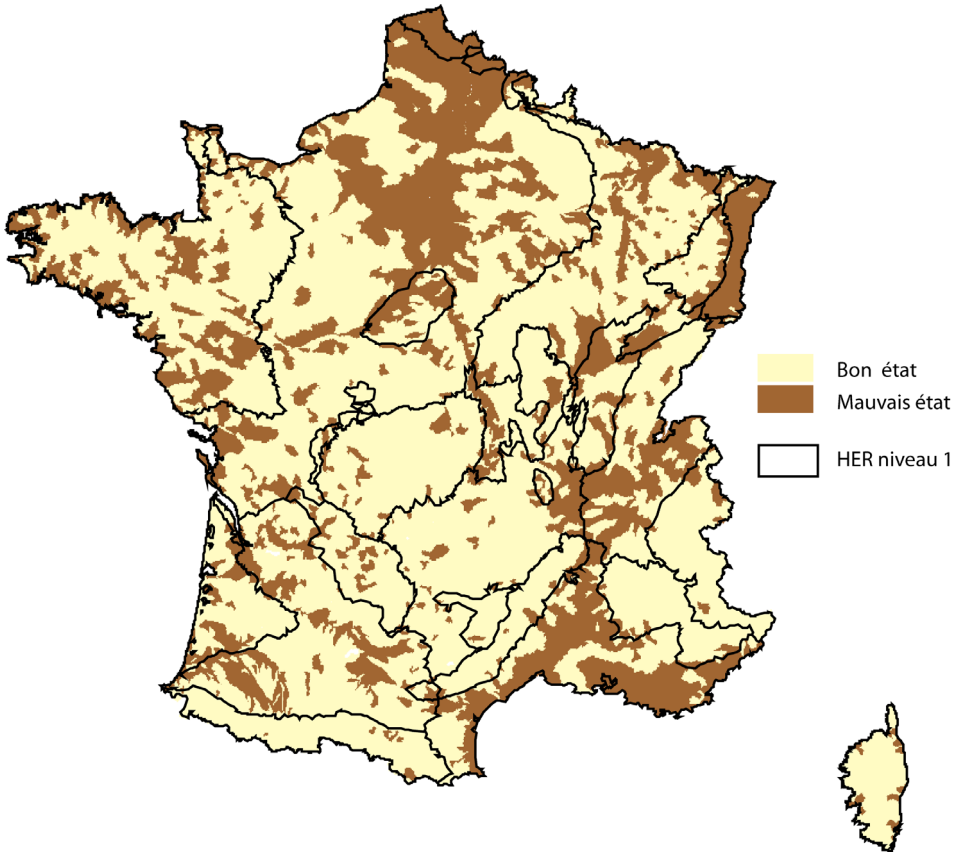


figure 15. Extrapolation spatiale du modèle « France entière »

5.4 Modèles régionalisés

5.4.1 Modèle basé sur 3 groupes d'HER

Dans le but d'améliorer les capacités d'extrapolation de ce modèle, nous avons essayé de mettre en place de nouveaux modèles en créant une partition fonctionnelle du territoire. L'hypothèse de base est que les relations « pressions / impacts » pourraient s'exprimer de manière différente en fonction du fonctionnement physique des cours d'eau.

Nous avons donc regroupé les hydro-écorégions en trois grands types :

- un type de « plaine » (HER 9, 12, 14, 15, 18, 10, 20, 22, 13 et 11) ;
- un type « montagne » (HER 3, 4, 17, 19, 21, 1, 2 et 5) ;
- un type « méditerranée » (HER 6, 7, 8 et 16).

Si les taux d'erreur des modèles restent satisfaisants (de 23% à 36%), l'aire sous la courbe ROC (0.65 à 0.67) montre que ces modèles sont peu informatifs (tableau 6).

Type de modèle	Erreur du modèle	AUC courbe ROC	Erreur de validation externe
Modèle « plaine »	36%	0.67	41%
Modèle « montagne »	29%	0.65	29%
Modèle « méditerranée »	23%	0.65	20%

tableau 6. Taux d'erreur des arbres de décision de types « plaine », « montagne » et méditerranée » et taux d'erreur de validation correspondants.

De plus, ce regroupement n'est pas satisfaisant dans la mesure où le taux d'erreur de validation externe (tableau 6) n'est pas significativement diminué et que l'erreur n'est pas équilibrée (figure 16) : la proportion de stations prédites en bon état et observées en mauvais état est supérieure à l'erreur inverse. Ces trois modèles ne sont donc pas performants pour prédire le mauvais état, ce qui semble délicat compte tenu du but de notre étude.

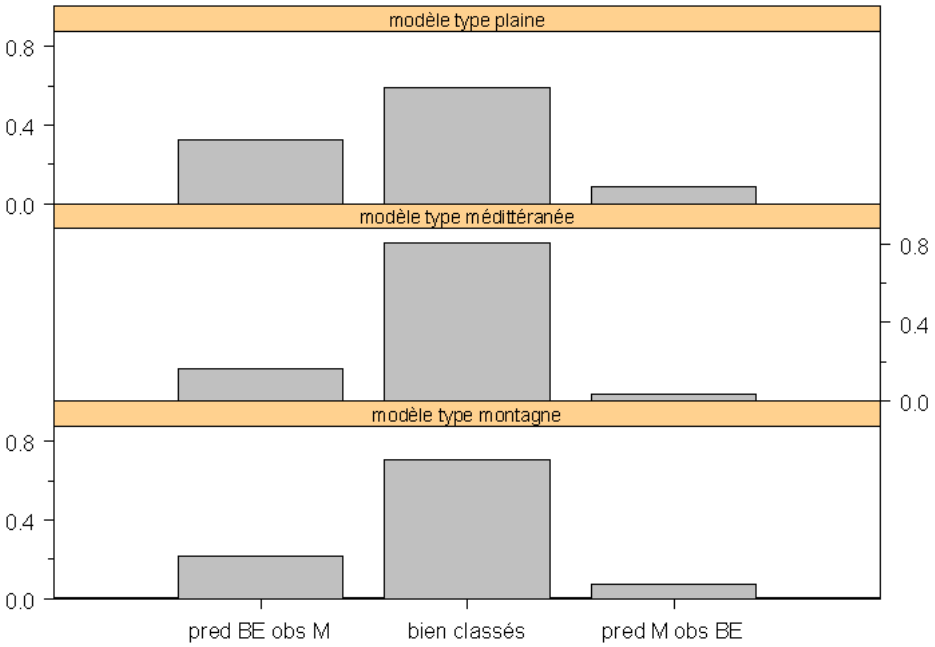


figure 16. Comparaison de l'état des stations en sortie des modèles par types fonctionnels et de l'état des stations déterminé d'après la base de données.

L'extrapolation spatiale du modèle est représentée sur la figure 17. Les arbres de décision correspondants sont détaillés dans l'annexe 2.

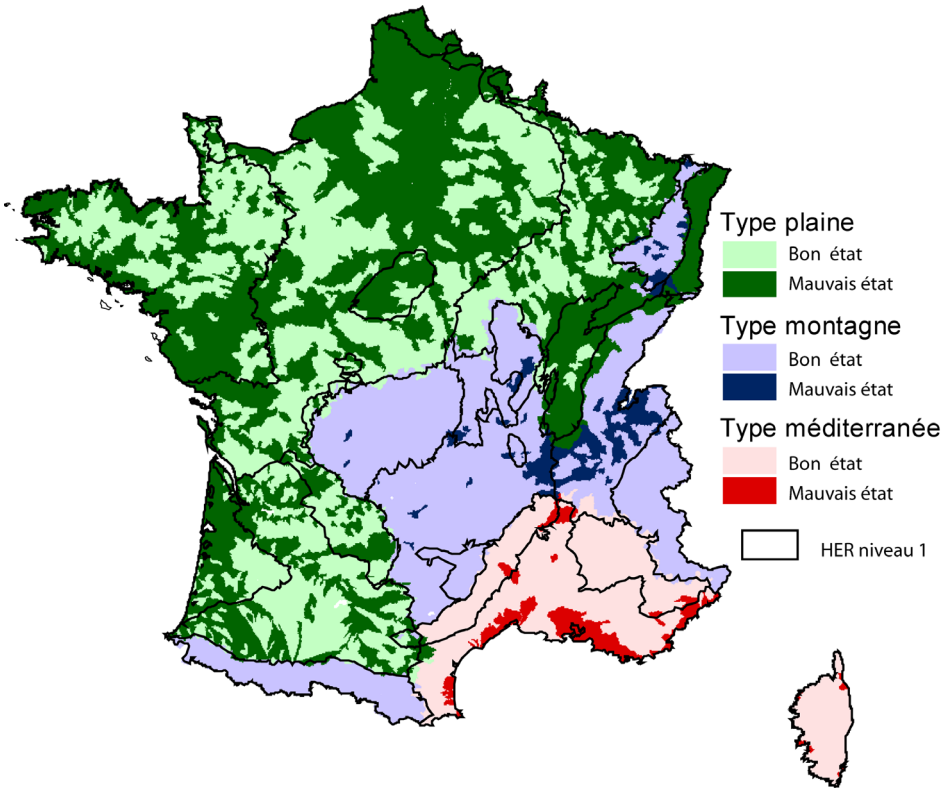


figure 17. Extrapolation spatiale des modèles par types fonctionnels (plaine, montagne, méditerranée)

5.4.2 Modèles basés sur 6 groupes d'HER

Nous avons donc revu cette typologie pour les types « plaine » et « montagne » en tenant compte des principales HER les composant. Le modèle Méditerranée, avec des taux d'erreur plus faibles que le modèle France entière, semble assez satisfaisant.

Aucun modèle n'a pu être validé pour les HER 10, 11, 13, 14, 15, 20, 22. L'erreur était trop forte lors de la validation externe et l'image d'extrapolation n'était pas réaliste. Ceci provient probablement d'une hétérogénéité trop forte des petites régions, et d'une qualité douteuse des données pour les Coteaux Aquitains. Le type « plaine » a donc été réduit aux Tables calcaires, Dépôts argilo-sableux et Massif Armoricain (HER 9, 20, 12).

Le type montagne a été séparé en trois types correspondant aux hautes montagnes (Alpes et Pyrénées, HER 1 et 2), aux moyennes montagnes (Massif Central et Vosges, HER 3, 4, 17, 19 , 21) et à l'ensemble Jura PréAlpes du Nord (HER 5).

Le modèle méditerranéen reste inchangé.

Les performances de ces modèles sont résumées dans le tableau 7, en comparaison avec celles du modèle France entière ; le taux d'erreur est cette fois beaucoup plus satisfaisant (de 16% à 26%), et l'aire sous la courbe ROC (0.71 à 0.80) montre que les modèles régionalisés présentent une bonne efficacité. On observe également un meilleur équilibre des erreurs lors de la validation externe (figure 18).

Ce regroupement en cinq groupes d'HER apportera donc de meilleurs résultats prédictifs que la typologie plaine-montagne-mediterranée.

Type de modèle	Erreur du modèle	AUC courbe ROC	Erreur de validation externe
Plaine (HER 9, 20, 12)	26%	0.75	32%
Moyenne Montagne (HER 3, 4, 21, 17, 19)	26%	0.72	21%
Haute Montagne (HER 1, 2)	16%	0.80	28%
Jura PréAlpes du Nord (HER 5)	26%	0.71	36%
Méditerranée (HER 6, 7, 8, 16)	23%	0.65	20%
France entière	***26%***	***0,75***	***32%***

tableau 7. Taux d'erreur des arbres de décision pour les modèles issus du redécoupage des types plaines et montagnes selon les HER les constituant et taux d'erreur de validation correspondants.

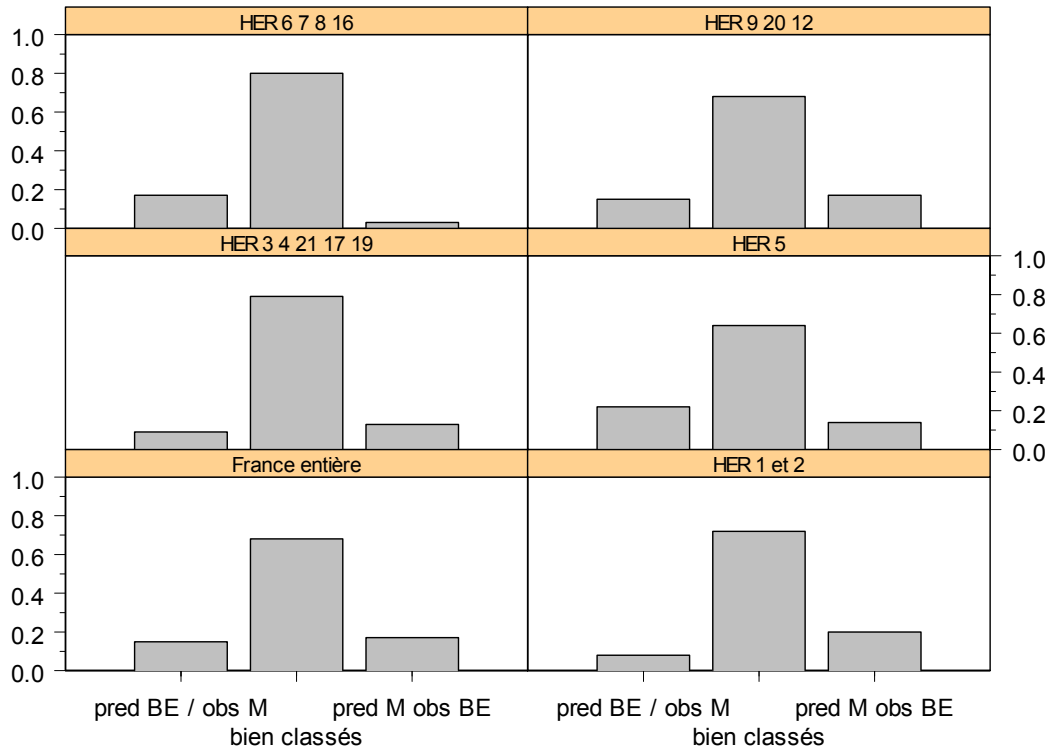


figure 18. Comparaison de l'état des stations en sortie des modèles par groupes d'HER et de l'état des stations déterminé d'après la base de données.

Nous avons donc réalisé une extrapolation spatiale sur la base de modèles (figure 19), en appliquant le modèle « France entière » pour les HER non représentées dans les 5 groupes d'HER ci-dessus.

Les arbres de décision correspondants sont détaillés dans l'annexe 2.

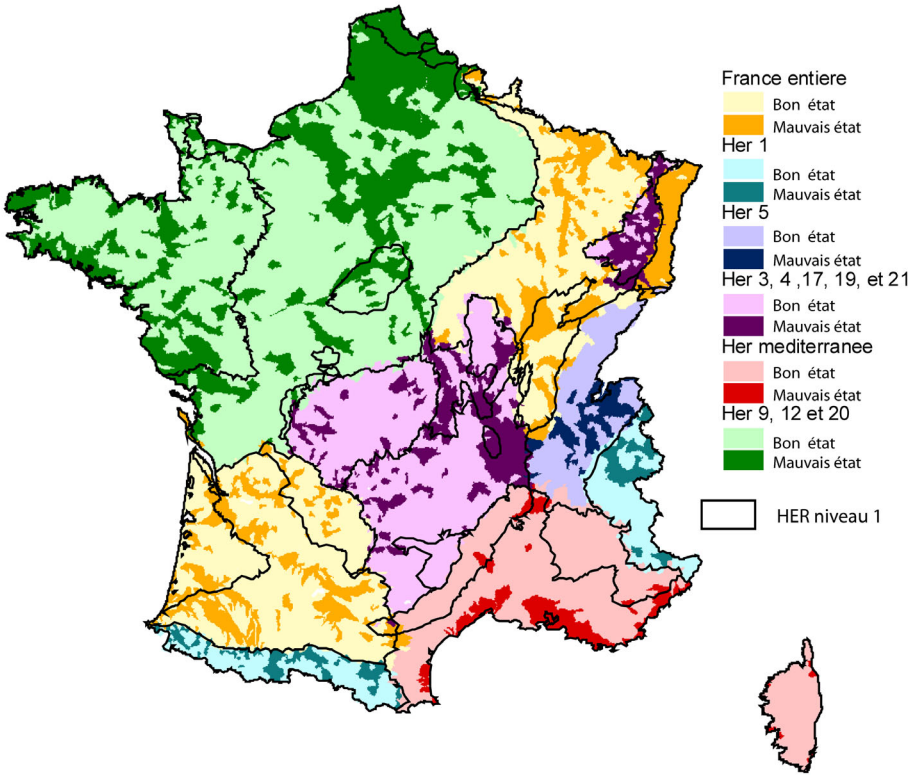


figure 19. Extrapolation spatiale des modèles par types d'HER.

5.5 Choix du modèle le plus performant

Il semble évident que nous ne conserverons pas le modèle à 3 groupes d'HER (plaine-montagne-méditerranée). Il n'est pas assez performant et ne donne pas une image d'extrapolation suffisamment fiable pour être utilisable.

Par contre il est difficile de choisir entre le modèle France entière et le modèle à 6 groupes d'HER sans éléments discriminatoires supplémentaires.

A partir du modèle France entière, nous avons calculé l'aire sous la courbe ROC des sous-groupes correspondants aux modèles par groupes d'HER, et nous avons ensuite comparé ces AUC aux AUC des modèles par groupes d'HER correspondants comme résumé dans le tableau 8.

	modèle 6 groupes d'HER		**modèle France entière**	
Groupes d'HER	*erreur*	*AUC*	*erreur*	*AUC*
Plaine (HER 9, 20, 12)	0.26	0.75	**0.25**	**0.76**
Moyenne Montagne (HER 3, 4, 21, 17, 19)	0.26	0.72	**0.23**	**0.77**
Haute Montagne (HER 1, 2)	**0.16**	**0.80**	0.29	0.68
Jura PréAlpes du Nord (HER 5)	0.26	0.71	**0.26**	**0.74**
Méditerranée (HER 6, 7, 8, 16)	0.23	0.65	**0.21**	**0.75**
autres HER	-	-	**0.33**	**0.72**

tableau 8. Erreur de classement et AUC des courbes ROC pour les modèles France entière (calculées par sous-groupes d'HER) et 6 types.

Pour chaque groupe de région, à l'exception des hautes montagnes (HER 1 et 2), les erreurs de classement sont plus faibles et l'AUC plus élevée avec le modèle France entière qu'avec les modèle 6 types correspondant.

Regardons enfin à titre de comparaison, le diagramme de validation externe correspondant au modèle France entière original distribué par sous groupes d'HER (figure 20). On constate que la validation est satisfaisante pour tous les groupes, excepté pour les hautes montagnes (HER 1 et 2).

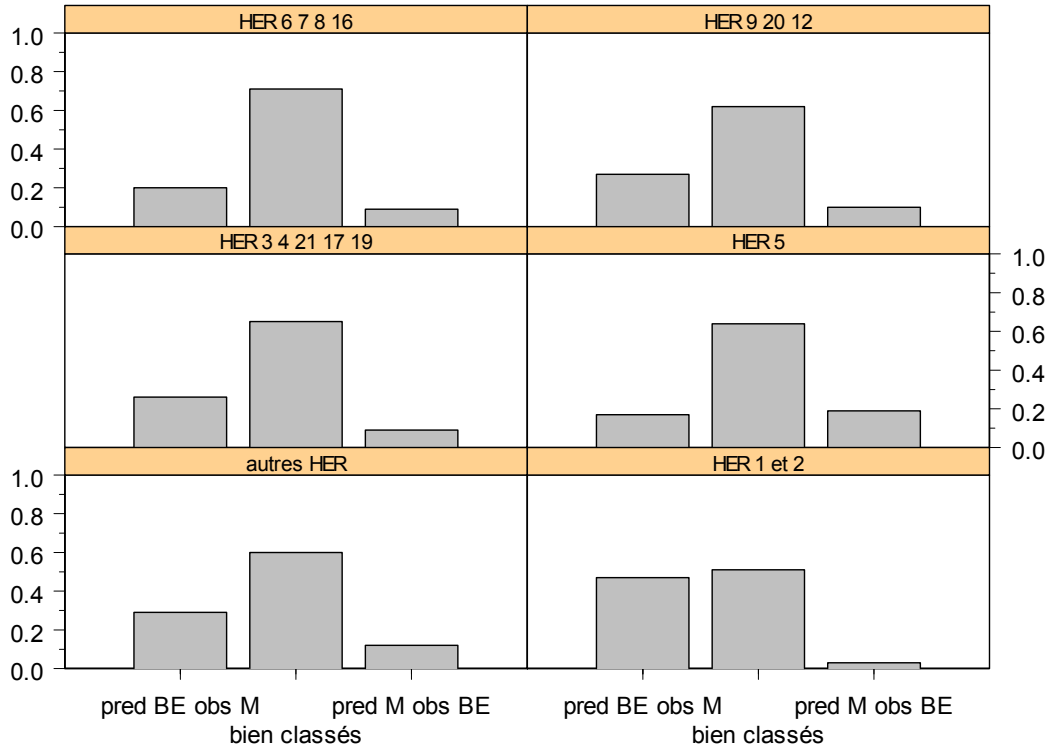


figure 20. Comparaison de l'état des stations en sortie du modèle « France entière » et de l'état des stations déterminé d'après la base de données dans différents groupes d'HER

Nous pouvons par conséquent dire que le **modèle France entière** est légèrement plus performant sur la plus grande partie du territoire. Restent les zones de haute montagne (HER 1 et 2) qui nécessiteraient apparemment un modèle spécifique pour optimiser les capacités prédictives.

5.6 Discussion : intérêt et limites des modèles d'extrapolation

Les modèles développés dans ce chapitre permettent donc de visualiser à l'échelle nationale une extrapolation spatiale basée sur un « apprentissage » par un arbre de décision. Les différentes techniques de validation permettent de choisir les modèles les plus performants en apprentissage et les moins biaisés en extrapolation.

L'analyse des performances des modèles réalisés à différentes échelles (nationale, par types fonctionnels ou par groupes d'HER) montre que le modèle « France entière » est globalement plus performant. Cela vient du fait que l'apprentissage sur le modèle global se fait sur un beaucoup plus grand nombre de situations.

En revanche, les HER de hautes montagnes (Alpes internes et Pyrénées) nécessitent un modèle particulier. La relation entre l'occupation du sol des bassins et les pressions

qui s'exercent sur les milieux est probablement différente du fait de la spécificité des altérations hydrologiques et morphologiques (barrages, endiguements..).

L'intérêt d'une représentation globale d'un « état écologique probable » est évident ; en revanche, la principale limitation de ces modèles vient du fait que seules des grandes catégories d'occupation du sol à l'échelle des bassins versants peuvent être utilisées en extrapolation. Les modèles sont donc *robustes mais pas très discriminants*, et leur valeur explicative est limitée.

Pour l'interprétation des résultats, il faut cependant garder en mémoire plusieurs points importants.

- *L'extrapolation est limitée aux drains principaux* des zones hydrographiques (ZH) ; l'état écologique probable représenté est celui du drain principal, qui ne présume en rien de celui des petits affluents de la ZH.
- *L'erreur locale des modèles est importante, mais non biaisée* ; cela signifie qu'il ne faut pas chercher à interpréter l'état d'une ZH particulière (le risque d'erreur est alors de 20 à 30%), mais que l'image globale représente assez bien la proportion des ZH en « bon » ou « mauvais » état probable.

Ce dernier point permet donc de comparer la proportion de stations classées en « bon » ou « mauvais » état dans les réseaux de suivi, avec le résultat du modèle d'extrapolation. L'image brute fournie par le réseau est en fait biaisée par le fait que les stations sont généralement placées à des points de contrôle de la pollution. L'image extrapolée par le modèle corrige ce biais en appliquant les mêmes critères à l'ensemble des drains principaux des ZH.

La figure 21 visualise l'état écologique des stations IBGN des réseaux (base GIRAFE), en regard de l'état des drains principaux prédit par le modèle d'extrapolation. On constate que si 42% des stations sont observées en « mauvais état », le modèle d'extrapolation ne prédit que 32% de drains principaux en « mauvais état probable ». Ceci montre bien que l'image fournie par les réseaux est un peu pessimiste, et *la réalité se situe probablement dans une fourchette de l'ordre de* ***35 à 40% de masses d'eau en « mauvais état »*** *pour l'IBGN, en fonction de notre hypothèse de départ*.

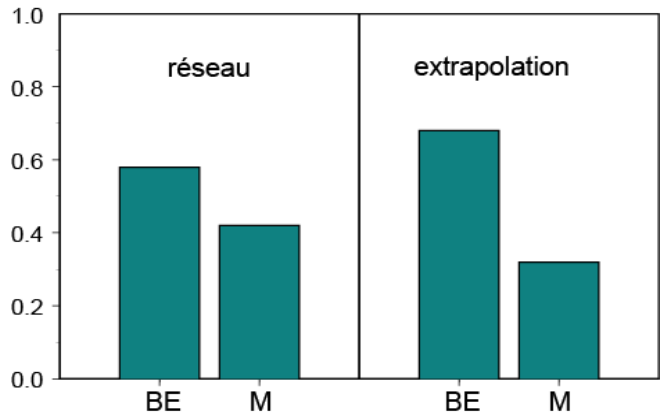


figure 21. Etat écologique des stations IBGN des réseaux (base GIRAFE), en regard de l'état des drains principaux prédit par le modèle d'extrapolation

6 Test de la sensibilité de la limite de bon état

6.1 Objectif

L'intérêt de ce type de modèle est aussi de pouvoir tester différentes hypothèses de classifications, et d'en visualiser le résultat à l'échelle de la France entière.

L'hypothèse de départ testée dans les chapitres précédents est celle décrite au § 2.1. La limite de « Bon état » est fixée en première approximation en divisant en 4 parties égales l'étendue des valeurs de l'IBGN entre la limite du « Très bon état », qui découle de la distribution observée sur les sites de référence, et une valeur minimum réaliste fixée à IBGN = 1 (**Wasson** et al. 2003).

Cependant, cette limite est arbitraire, car rien n'indique dans la DCE que les classes d'état écologique doivent avoir une étendue égale. D'autres hypothèses peuvent être testées, les résultats ci-dessus ayant montré que la classification de départ n'est pas exagérément « pessimiste ».

Nous avons donc testé une deuxième hypothèse de classification, que nous appellerons « Bon état + 1 » (ou BE+1), *en relevant uniformément de 1 point IBGN pour tous les types la limite de « Bon état » déterminée en première hypothèse*. L'objectif est de visualiser par le même type de modèle les conséquences en termes de probabilité actuelle de non atteinte du bon état.

6.2 Modèle d'extrapolation « Bon état + 1 »

Nous avons pour cela réalisé des modèles suivant la typologie validée précédemment sur l'hypothèse de départ (modèle France entière).

Le modèle « BE+1 » donne des résultats moyens. Son taux d'erreur est de 32% avec une AUC de 0.69 (tableau 9). C'est un peu moins satisfaisant que pour le modèle de départ mais cela reste acceptable.

La validation externe est elle aussi acceptable avec un taux d'erreur de 35% et une bonne répartition de l'erreur entre bon et mauvais état comme on peut le voir sur la figure 22A.

Type de modèle	AUC courbe ROC	Erreur du modèle	Erreur de validation externe
France entière BE+1	0.69	32%	35%

tableau 9. Aire sous la courbe ROC, taux d'erreur des arbres de décision et taux d'erreur de validation externe correspondants pour le modèle France entière BE+1 (limite de bon état relevée de 1 point IBGN).

Cependant, les résultats de la validation réalisée par groupes d'HER (tableau 10) montrent que pour le groupe de Haute Montagne (HER 1 et 2), l'erreur du modèle est très forte (51% avec une AUC de 0.53), et la prédiction est fortement « optimiste» car de nombreuses zones hydrographiques prédites en « bon état » sont observées en « mauvais état ». *Le modèle n'est donc pas du tout applicable aux régions de Haute Montagne*.

A l'inverse, pour l'HER 5 (Jura PréAlpes du Nord), de nombreuses ZH prédites « mauvaises » sont observées « bonnes ». Le modèle est donc nettement « pessimiste » pour cette HER.

	France		France « BE +1 »	
Groupes d'HER	*erreur*	*AUC*	*erreur*	*AUC*
Plaine (HER 9, 20, 12)	0.25	0.76	0.32	0.69
Moyenne Montagne (HER 3, 4, 21, 17, 19)	0.23	0.77	0.28	0.74
Haute Montagne (HER 1, 2)	0.29	0.68	***0.51***	***0.53***
Jura PréAlpes du Nord (HER 5)	0.26	0.74	0.34	0.65
Méditerranée (HER 6, 7, 8, 16)	0.21	0.75	0.27	0.71
autres HER	0.33	0.72	0.33	0.68

tableau 10. Erreur de classement et AUC des courbes ROC pour les modèles France entière et France entière BE+1 (calculées par sous-groupes d'HER).

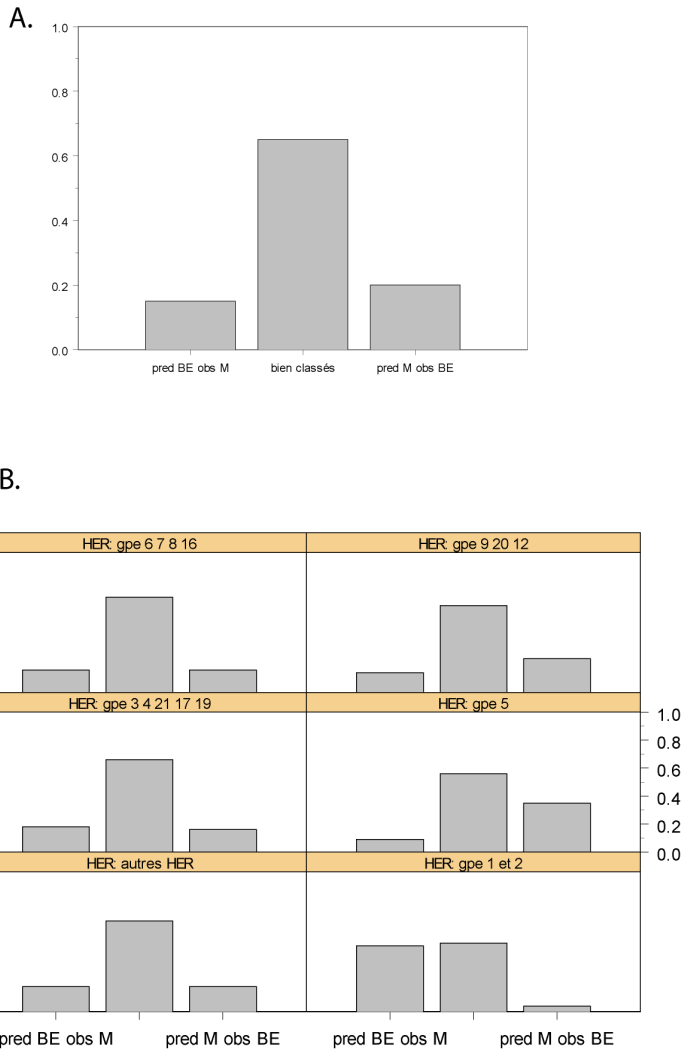


figure 22. Comparaison de l'état des stations en sortie du modèle « France entière BE +1» et de l'état des stations déterminé d'après la base de données (A) et dans différents groupes d'HER en particulier (B).

La validation externe par groupes d'HER a été réalisée en parallèle pour essayer de comprendre si le niveau de la limite de bon état avait la même importance quelque soit la zone du territoire concernée. La figure 22B montre que l'extrapolation des modèles pose un problème pour le groupe HER 1 et 2 avec une erreur très forte et déséquilibrée.

Le diagramme de validation externe correspondant au modèle 6 types BE+1 (figure 23) montre que la même expérience sur le modèle 6 types donne des résultats similaires.

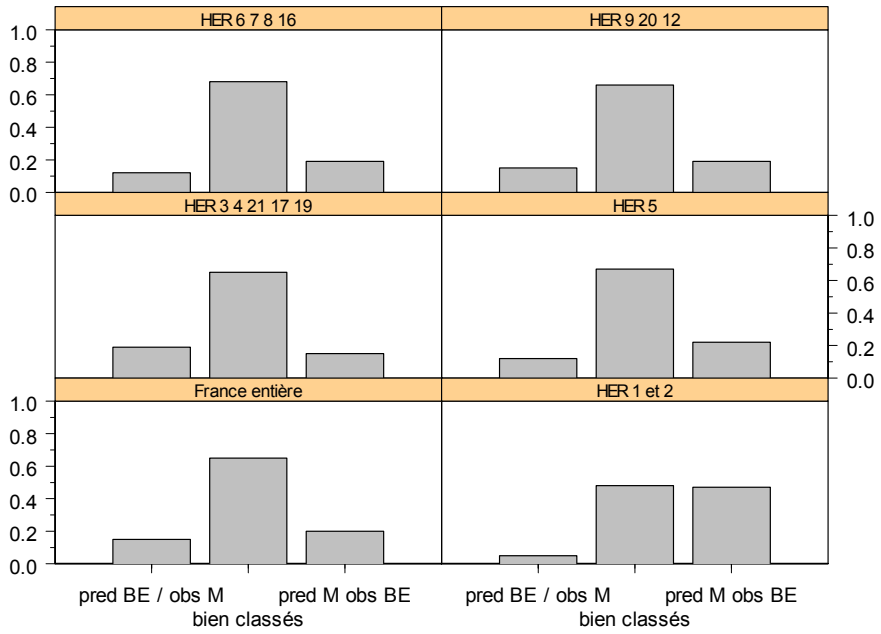


figure 23. Comparaison de l'état des stations en sortie du modèle 6 types BE+1et de l'état des stations déterminé d'après la base de données

6.3 Modèle « Bon état + 1 » : visualisation des situations limites

A l'exception des Alpes Internes et des Pyrénées, les modèles sont donc validés et peuvent faire l'objet d'une extrapolation spatiale suffisamment fiable pour être utilisée.

On constate logiquement que l'extrapolation cartographique du modèle « BE+1 » donne une image plus « pessimiste », puisque dans cette hypothèse un certain nombre de ZH basculent du côté du « mauvais état ». Il nous a semblé intéressant de visualiser sur une même carte les conséquences probables d'un changement de limite. Les résultats sont présentés sur la figure 24.

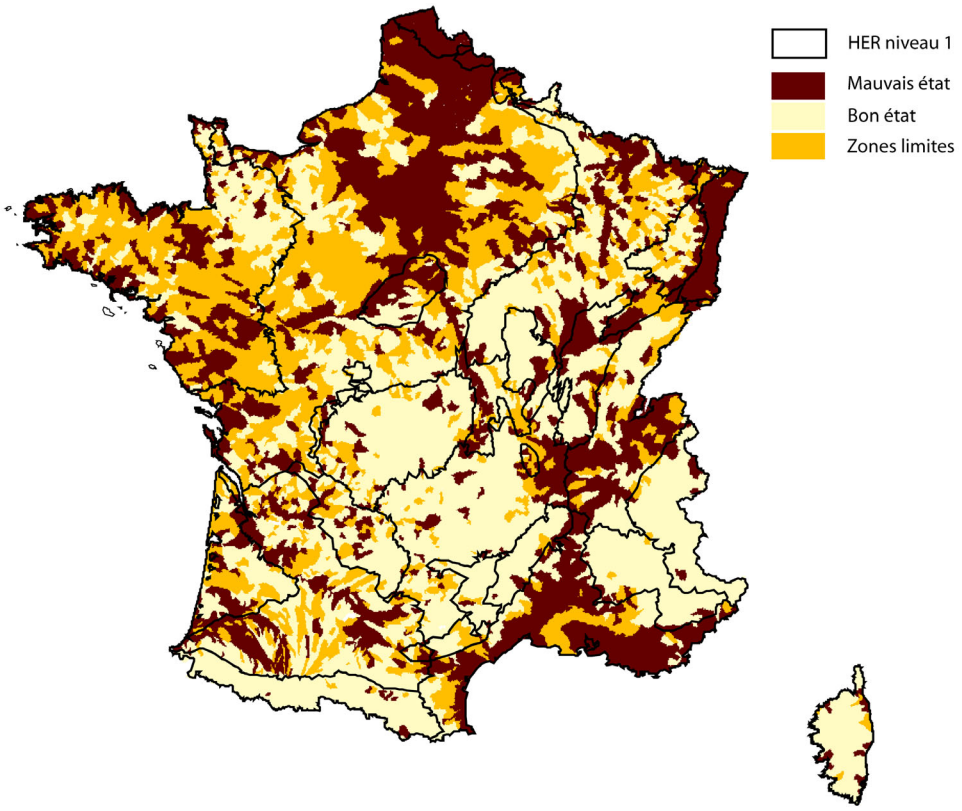


figure 24. Représentation de l'évolution de la probabilité de bon état lors de l'augmentation de la limite de bon état de un point

Cette carte correspond à la représentation simultanée de deux modèles :

- le modèle « France » avec l'hypothèse de départ pour la limite de bon état,
- le modèle « France BE+1 » avec la limite de bon état relevée d'un point IBGN.

Les zones hydrographiques (ZH) qui présentent dans les deux hypothèses une forte probabilité de « mauvais état » sont figurées en sombre (marron) ; celles qui présentent dans les deux cas une forte probabilité de « bon état » sont en clair (blanc) ; enfin les ZH qui « basculent » du bon vers le mauvais état sont figurées en teinte intermédiaire (jaune). Cette représentation fait donc apparaître les zones hydrographiques ***en situation limite***, qui se situent en moyenne à 1 point IBGN de la limite du « Bon état » considérée en première hypothèse.

6.4 Discussion : simuler différentes hypothèses de « Bon état » ?

La première question concerne la fiabilité de la simulation pour les deux hypothèses testées. Les modèles sont un peu moins performants pour l'hypothèse « Bon Etat +1 », ***ce qui laisse penser que les descripteurs d'occupation du sol utilisés sont moins pertinents dans la gamme des faibles pressions***. Néanmoins les performances des modèles sont suffisamment proches et le biais d'extrapolation spatiale suffisamment faible pour qu'on puisse comparer les résultats des deux hypothèses.

Comme pour l'hypothèse de départ, la figure 25 visualise le pourcentage de stations des réseaux (base GIRAFE) classées en « mauvais état », avec le pourcentage de drains principaux prédits en « mauvais état probable » en sortie du modèle d'extrapolation dans l'hypothèse « BE+1 ». Cette figure est à comparer à la 21(§5.6).

Cette comparaison permet de tirer deux enseignements :

- dans les réseaux, le pourcentage de stations observées en « mauvais état » passe de 42% à 53% dans l'hypothèse « BE+1 », soit une augmentation de 11% par rapport à l'hypothèse de départ ;
- en sortie du modèle d'extrapolation, le pourcentage de drains principaux prédits en « mauvais état probable » (52%) est du même ordre que celui observé dans les réseaux, mais l'augmentation est de 20% par rapport à l'hypothèse de départ (32%).

Nous avons signalé d'une part que les performances du modèle « BE+1 » sont un peu moins bonnes avec des prédictions légèrement pessimistes (figure 22A), d'autre part que l'image fournie par les réseaux est elle aussi probablement pessimiste. On peut donc supposer que la réalité se situe un peu en dessous de ces valeurs, avec *un pourcentage de masses d'eau en « mauvais état » pour l'IBGN* ***de l'ordre de 50%*** *dans l'hypothèse « BE+1 »*.

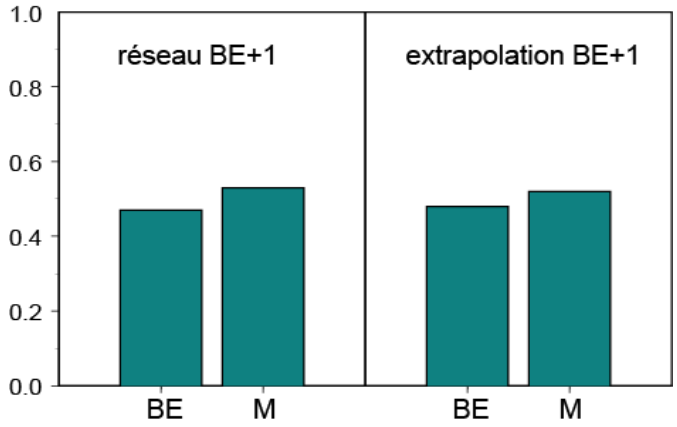


figure 25. pourcentage de stations des réseaux (base GIRAFE) classées en « mauvais état », et pourcentage de drains principaux prédits en « mauvais état probable » en sortie du modèle d'extrapolation dans l'hypothèse « BE+1 »

On constate que les changements liés à la remontée d'un point de la limite de « Bon état » sont nettement plus importants dans les zones fortement agricoles (Tables Calcaires, Massif Armoricain…) et dans la région Méditerranée. L'image ainsi obtenue est conforme à la représentation attendue d'après connaissances de terrain des gestionnaires et opérationnels.

Zones en situation limite : quels facteurs pénalisant ?

La comparaison des arbres de décision des modèles dans les deux hypothèses donne des informations intéressantes sur les facteurs qui font basculer des zones hydrographiques vers le mauvais état dans l'hypothèse d'une remontée d'un point IBGN de la limite de « Bon Etat ».

Dans le modèle « France entière » pour l'hypothèse de départ (figure 12B, §5.3), les stations qui ont moins de 2% de territoires urbanisés dans leur bassin sont majoritairement en bon état ; en revanche, pour le modèle « France entière BE+1 » (annexe 2.4), un critère [agriculture intensive < 13%] se rajoute au critère [artificialisation < 2%] pour discriminer les stations majoritairement en bon état. On peut en conclure qu'à l'échelle nationale, un grand nombre de sites dont le bassin est occupé par des terres labourées se trouvent en situation limite.

Si l'on regarde maintenant à l'échelle des groupes d'HER, on retrouve des résultats comparables.

Pour la région méditerranéenne (HER 6, 7, 8 et 16), le modèle de départ, très simple, donne majoritairement en bon état les stations ayant moins de 4,5% de territoires artificialisés dans leur bassin (annexe 2.2) ; dans l'hypothèse BE+1 (annexe 2.5.1), le seuil d'artificialisation descend à 2,5% et il se rajoute un critère [agriculture intensive < 13%]. Ceci laisse penser que dans cette région de nombreux sites sont en situation limite à la fois du fait des impacts urbains et agricoles.

Pour les Tables Calcaires et le Massif Armoricain (HER 9, 20, 12), régions de plaine à forte occupation agricole, on retrouve un schéma similaire au niveau des premières branches. Avec l'hypothèse de départ (annexe 2.2), les stations ayant un bassin avec moins de 2,1% d'artificialisation sont majoritairement en bon état. Pour le modèle BE+1 (annexe 2.5.3), il se rajoute aussi un seuil [agriculture intensive < 58%] pour discriminer les stations majoritairement en bon état.

En revanche, *pour les régions de moyenne montagne* (Massif Central et Vosges, HER 3, 21, 4, 17, 19), les modèles ne diffèrent pratiquement pas, avec un seuil autour de 2% de territoires artificialisés pour discriminer les stations en bon ou mauvais état dans les deux hypothèses (annexe 2.5.2). On peut en conclure que dans ces régions les pressions dominantes sont liées à l'urbanisation, et les que activités agricoles à dominante d'élevage peu intensif ont un impact globalement limité.

Nous n'irons pas plus dans le détail dans l'interprétation de ces modèles dont on a déjà dit qu'ils sont conçus pour l'extrapolation et non pour l'explication des phénomènes. Néanmoins, les premières branches des arbres de décision, qui correspondent aux structures fortes des jeux de données, sont suffisamment fiables et concordantes pour autoriser certaines conclusions. Dans l'hypothèse d'une remontée de un point de la limite de Bon état, un grand nombre de sites qui basculent vers le mauvais état se trouvent en « situation limite» du fait de pressions liées principalement à l'agriculture intensive (terres labourées), et probablement aussi à l'urbanisation dans les régions méditerranéennes.

Mais il faut rappeler que ces *résultats sont encore provisoires* et demandent à être confirmés avec les valeurs révisées des limites de classes sur l'IBGN.

7 Modèles de Diagnostic : Influence de l'occupation du sol sur l'état écologique mesuré par l'IBGN

On a signalé précédemment que les arbres de décision utilisés dans les modèles d'extrapolation n'ont pas forcément un caractère explicatif. D'autre part, pour ces modèles, le jeu de variables utilisé est limité aux grandes catégories d'occupation du sol des bassins versants, afin de permettre l'extrapolation spatiale à l'échelle des zones hydrographiques ; les variables qui, à l'heure actuelle, ne peuvent être extrapolées à l'ensemble du réseau - en particulier l'occupation des corridors rivulaires - ne sont donc pas utilisées dans ces modèles, ce qui limite d'autant leur capacité d'explication.

7.1 Objectifs des modèles de Diagnostic

L'objectif des modèles de diagnostic est donc d'évaluer l'impact relatif des pressions anthropiques sur l'IBGN transformé en EQR, révélateur de l'état écologique actuel des stations. Ces modèles ne permettent pas en l'état de représentation cartographique, mais cherchent à *mettre en évidence les causes majeures d'altération des milieux à des échelles nationales et régionales*.

Pour les raisons de disponibilité des données sur les pressions signalées au §1.4, *nous avons utilisé comme indicateurs de pressions l'occupation du sol des bassins versants et des corridors rivulaires*, à partir des données de CORINE Land Cover, selon les méthodologies décrites au §2.2.

Les modèles mathématiques utilisés sont des *régressions PLS entre l'EQR-IBGN et les variables d'occupation du sol,* couplées à des analyses multivariées, (cf. chapitre 3). Il est important de noter que dans cette approche, on ne cherche plus à prédire une classification binaire (« bon » ou « mauvais » état), mais à analyser l'influence des variables de pressions *sur l'ensemble du gradient d'état écologique* révélé par l'EQR-IBGN.

Un objectif important de ce travail est d'analyser des ***effets d'échelles***, à deux niveaux différents :

- En comparant les résultats de modèles développés *à l'échelle nationale et à l'échelle des hydro-écorégions*, l'HER Massif Armoricain étant prise comme région test.
- En analysant un « effet de proximité » à travers l'impact de l'occupation du sol à *l'échelle des bassins versants et des corridors rivulaires*.

Enfin, nous avons cherché à évaluer, sur l'exemple du Massif Armoricain, le gain d'information apporté par l'utilisation d'un *niveau détaillé des catégories d'occupation du sol*.

7.2 Modèle de Diagnostic « France entière » à partir des grandes catégories d'occupation du sol (bassin et corridor)

L'objectif de cette première série d'analyses est de déterminer à l'échelle du territoire métropolitain l'influence des grandes catégories d'occupation du sol sur l'état écologique des stations mesuré par l'EQR-IBGN.

Les quatre grandes catégories d'occupation du sol, rappelées ci-dessous, sont les mêmes que celles utilisées pour les modèles d'extrapolation spatiale (cf. § 4.1), mais elle sont quantifiées à l'échelle du bassin versant réel et du corridor rivulaire de chaque station IBGN.

Territoires artificialisés (variable **ARTIF**) : zones urbanisées, zones industrielles ou commerciales et réseaux de communication, mines, décharges et chantiers, espaces verts artificialisés, non agricoles - codes CLC : 1.1.1, 1.1.2, 1.2, 1.3, 1.4.1, 1.4.2.

Territoires agricoles de forte intensité (variable **AGRI_I**) : terres arables, cultures permanentes, vergers, vignes, oliveraies, cultures annuelles associées aux cultures permanentes, systèmes culturaux et parcellaires complexes - codes CLC : 2.1, 2.2, 2.4.1, 2.4.2

Territoires agricoles de faible intensité (variable **AGRI_F**) : prairies, territoires principalement occupés par l'agriculture avec présence de végétation naturelle importante, territoires agro-forestiers - codes CLC : 2.3.1, 2.4.3, 2.4.4

Territoires à faible anthropisation (variable **ESP_NAT**) : forêts et milieux semi-naturels, zones humides, surfaces en eau - codes CLC : 3.1.1, 3.1.2, 3.1.3, 3.2, 3.3, 4 et 5.

7.2.1 Régression PLS : IBGN f (occupation du sol).

Une régression PLS a été réalisée entre l'EQR-IBGN et les variables globales d'occupation du sol des bassins versants et des corridors rivulaires.

Le R^2 du modèle est de 16.3%.

Les coefficients de régression normés sont présentés dans la figure 26.

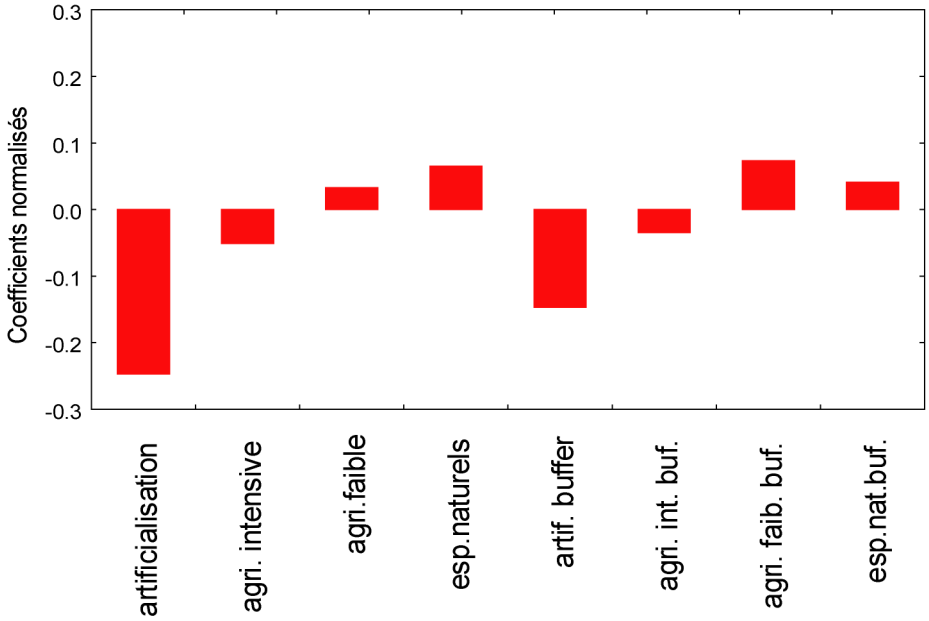


figure 26. Coefficients normés de régression PLS
EQR= f(groupes d'occupation du sol) pour la France entière

L'examen des coefficients de régression montre que ce sont les territoires artificialisés, au niveau du bassin comme du corridor, qui ont la plus forte contribution dans le modèle, avec un effet négatif sur l'EQR-IBGN. Les autres catégories ont des contributions moins importantes, négatives pour l'agriculture intensive, et positives pour l'agriculture faible et les espaces naturels.

On remarque que les prédicteurs agissent dans le même sens au niveau du bassin et du corridor rivulaire.

Cette analyse apporte une bonne identification des facteurs impactants et de l'ampleur relative des impacts reliés aux différentes catégories d'occupation du sol. Les impacts relatifs de l'urbanisation et de l'agriculture intensive (terres labourées) correspondent à ce qui avait été observé dans la hiérarchie d'apparition de ces facteurs dans les arbres de régression pour le modèle « France entière » (cf. § 5.3, figure 12).

Cependant, cette analyse soulève des interrogations quant à l'impact de l'agriculture intensive qui apparaît ici relativement faible par rapport à l'urbanisation, alors qu'on l'attendait plutôt un effet important, sinon majeur.

7.2.2 Typologie des pressions d'occupation du sol.

Pour essayer de mieux comprendre l'impact de l'agriculture, nous avons analysé les corrélations entre les variables d'occupation du sol à partir d'une ACP normée suivie d'une régression sur composante principale. Cette analyse fournit une représentation de la typologie des pressions d'occupation du sol qui s'exercent à l'échelle nationale.

On retrouve 58 % de la variance dans le plan F1 x F2. La figure 27 représente la projection des variables dans ce plan.

- L'axe F1 oppose les espaces naturels à l'agriculture (intensive et faible qui sont pourtant orthogonales dans le plan). L'artificialisation est mal représentée sur cet axe.
- L'axe F2 oppose l'artificialisation et l'agriculture intensive (corrélées entre elles) à l'agriculture faible (prairies). Les espaces naturels ne sont pas représentés sur cet axe.

On remarque en même temps que les variables mesurées au niveau des corridors (ou buffer) sont corrélées aux mêmes variables mesurées au niveau du bassin. Toutefois, ces corrélations sont nettement plus marquées pour les espaces naturels et les prairies (agri. faible) que pour les cultures (agri. intensive) et les zones urbanisées (artificialisation).

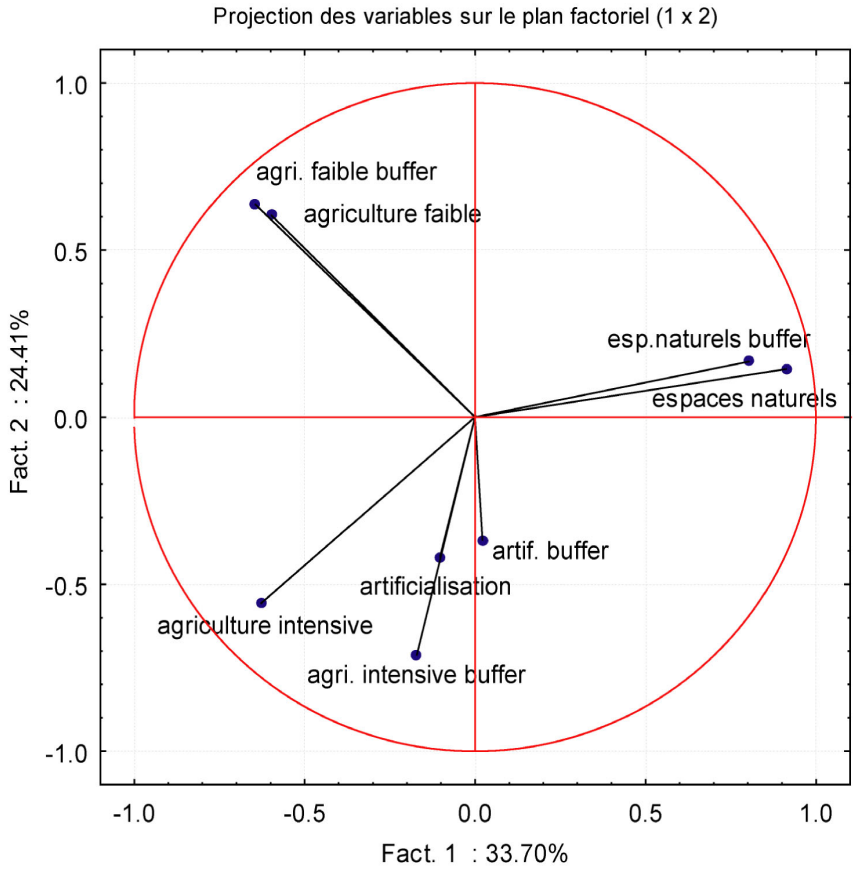


figure 27. Projection des variables sur le plan (1,2) de l'ACP normée des prédicteurs

L'examen de l'espace des individus (figure 28) confirme cette structure en montrant que les 3640 stations s'organisent suivant trois gradients pour former une structure triangulaire. Ces gradients correspondent respectivement à :

- 1 : artificialisation et agriculture intensive,
- 2 : agriculture faible
- 3 : espaces naturels.

Cette structure permet de tirer les conclusions suivantes quant à l'organisation de l'espace, à l'échelle de la France entière :

- les zones urbanisées et les zones d'agriculture intensives (terres labourées et cultures permanentes) sont associées spatialement sur les mêmes territoires ; autrement dit, les bassins impactés le sont généralement à la fois par l'urbanisation

et l'agriculture intensive. Cependant, dans ces zones, l'occupation du corridor rivulaire peut être sensiblement différente de celle du bassin versant.

- Le reste du territoire se partage entre des espaces naturels et des zones agricoles de faible intensité, où dominent les prairies. Dans ces zones, l'occupation du corridor est en général la même que celle du bassin.
- Evidemment, toutes les situations intermédiaires entre ces trois pôles d'occupation du sol peuvent être rencontrées.

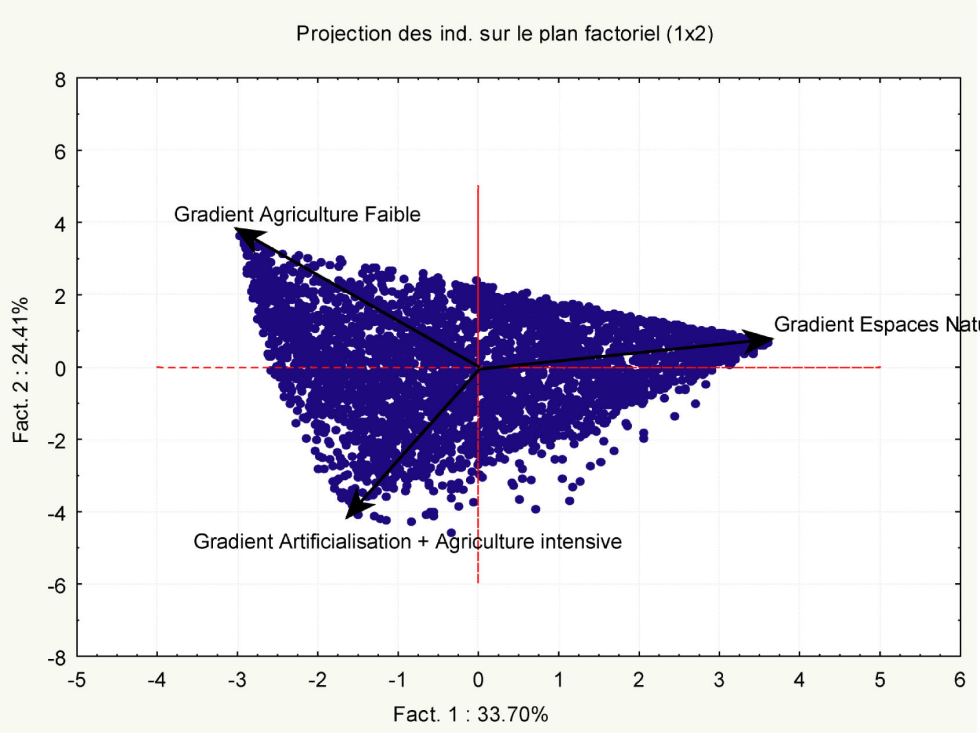


figure 28. Projection des individus sur le plan (1,2) de l'ACP normée des prédicteurs

7.2.3 Relation entre variables synthétiques de pressions et IBGN

Les axes F1 et F2 de l'ACP précédente représentent des variables synthétiques exprimant l'intensité des pressions liées à l'occupation du sol ; nous avons donc cherché les corrélations entre ces axes et l'état écologique exprimé par l'EQR-IBGN.

Les régressions linéaires des axes F1 et F2 sur l'EQR-IBGN sont significatives, et leurs coefficients nous permettent d'affirmer que chaque axe représente indirectement un gradient d'impact sur l'EQR-IBGN.

On peut donc interpréter ces résultats de la manière suivante :

- l'axe 1 de l'ACP montre que les espaces naturels ont un effet positif sur l'état écologique d'une station (évalué avec l'IBGN), alors que les autres catégories d'occupation du sol (y compris l'agriculture faible) ont plutôt un impact négatif.
- L'axe 2, qui a une contribution plus importante que l'axe 1 à la variabilité de l'EQR-IBGN, montre que l'artificialisation et l'agriculture intensive ont un impact toujours négatif sur l'état écologique, et s'opposent à l'agriculture faible (prairies) qui est corrélée positivement à l'EQR-IBGN sur cet axe.

7.2.4 Enseignements tirés du modèle de diagnostic « France entière »

Ce premier modèle très global, basé sur des grandes catégories d'occupation du sol à l'échelle de la France entière permet de tirer quelques enseignements généraux importants.

L'analyse des pressions révèle une *association spatiale des zones urbanisées et de l'agriculture intensive ; cette association rend plus difficile la discrimination des impacts* dus à ces deux causes importantes de dégradation de l'état écologique des rivières.

Malgré cela, ces premières analyses nous permettent d'affirmer qu'à l'échelle nationale *les impacts liés à **l'urbanisation** constituent le principal facteur d'altération de l'état écologique* évalué par l'IBGN.

*Les impacts liés à **l'agriculture intensive*** (terres labourées et cultures permanentes) *semblent intervenir secondairement*. L'impact sur l'IBGN est toujours négatif, mais de plus faible amplitude que celui associé aux territoires artificialisés.

L'agriculture de faible intensité (principalement les prairies) joue un rôle qui peut s'avérer négatif ou positif selon le contexte.

Les *espaces naturels* ont, par contre, *un effet toujours positif sur l'IBGN*.

Donc on a d'une part la pression négative forte de l'urbanisation et d'autre part l'effet positif des espaces naturels. Entre ces deux facteurs, l'agriculture intensive a un impact négatif moins marqué, mais d'une manière générale les impacts agricoles sont moins évidents à appréhender à l'échelle nationale.

Il semble à ce stade nécessaire de changer d'échelle d'étude, et donc de réaliser une analyse similaire à une échelle régionale correspondant à des types de pressions plus homogènes.

7.3 Modèle de Diagnostic « Régional » à partir des grandes catégories d'occupation du sol : le Massif Armoricain

Nous présentons à titre d'exemple une étude similaire à la précédente, mais portant uniquement sur l'hydro-écorégion « Massif Armoricain » (HER 12). Une étude plus complète sur l'ensemble des HER fera l'objet d'un prochain rapport.

Le tableau de données utilisé pour ces analyses est constitué de 276 stations pour lesquelles on dispose de l'EQR-IBGN. L'occupation du sol est renseignée d'après CORINE Land Cover au niveau du bassin versant et du corridor rivulaire. Les groupes d'occupation du sol utilisés sont identiques à ceux de l'étude précédente au niveau national (§ 7.2).

7.3.1 Régression PLS : IBGN *f* (occupation du sol)

Une régression PLS a été réalisée pour modéliser l'EQR-IBGN en fonction des pressions d'occupation du sol des bassins versants et des corridors rivulaires. Le R^2 du modèle est de 11.9%.

L'analyse des contributions des variables au modèle est présentée dans la figure 29. Elle montre que c'est l'artificialisation du bassin et du corridor qui a la plus forte contribution négative, tandis que les espaces naturels au niveau du bassin et du corridor ont la plus forte contributions positives.

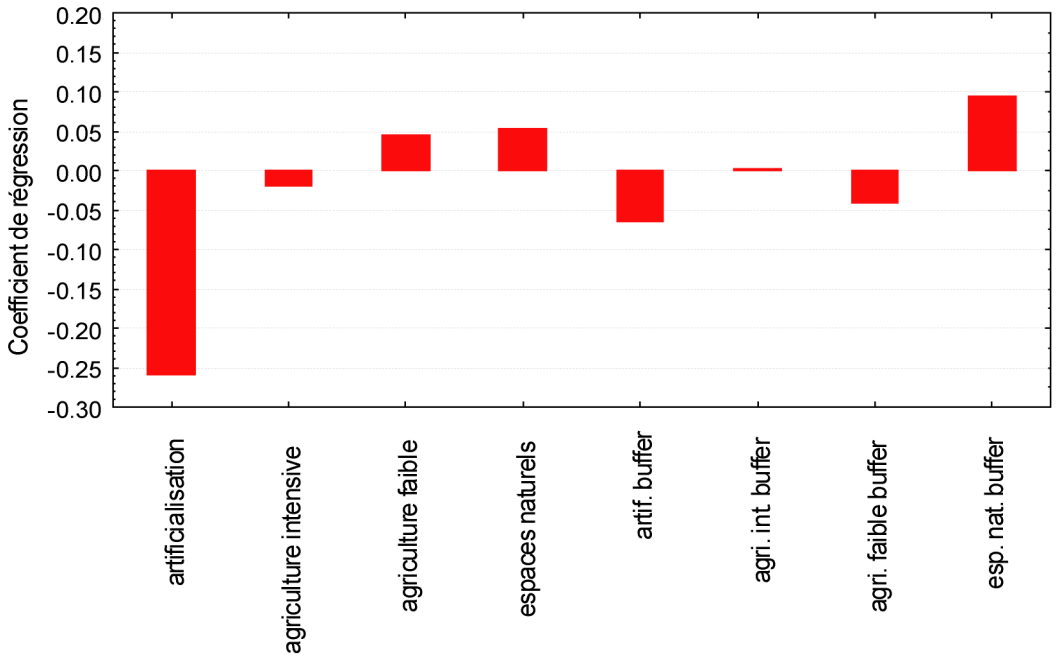


figure 29. Coefficients normés de régression PLS EQR=f(groupes d'occupation du sol) pour l'HER 12.

Toutefois, cette figure est sensiblement différente de celle du modèle France entière (figure 26, § 7.2.1).

- D'une part, la contribution de l'agriculture intensive du bassin est très faible et pratiquement négligeable au niveau du corridor (ou buffer) ; pour une région fortement agricole, cela ne laisse pas d'étonner.
- D'autre part, alors que dans le modèle France entière les prédicteurs agissent dans le même sens au niveau du bassin et du corridor, on constate dans le Massif Armoricain que l'agriculture faible (prairies) a un effet positif au niveau du bassin mais négatif au niveau du corridor.

Autrement dit, les impacts agricoles sont ici plus liés aux prairies, ce qui apparaît logique dans une région d'élevage intensif. Mais ces impacts n'apparaissent que lorsque les prairies sont situées en bordure des cours d'eau, car un paysage général de prairies au niveau du bassin versant constitue ici comme ailleurs un environnement plutôt protecteur. Néanmoins, cette situation nécessite une analyse plus poussée.

7.3.2 Typologie des pressions d'occupation du sol, et relation avec l'IBGN

Pour essayer de mieux comprendre quel est le rôle de l'agriculture, nous avons réalisé comme précédemment une ACP normée des variables d'occupation du sol, suivie d'une régression des composantes principales avec l'EQR-IBGN.

On retrouve 56 % de la variance dans le plan (F1 x F2). La figure 30 représente la projection des variables dans ce plan, et la figure 27 celle des stations.

- L'axe F1 oppose l'agriculture faible (prairies) à l'agriculture intensive (terres labourées et cultures permanentes). L'artificialisation et les espaces naturels ont une faible contribution à cet axe.
- L'axe F2 oppose l'artificialisation et les espaces naturels. L'agriculture (faible et intensive) contribue peu à cet axe.

Cette disposition rappelle celle du modèle France entière ; l'agriculture intensive et l'urbanisation sont partiellement corrélées. On remarque en même temps que les variables mesurées au niveau des corridors (buffer) sont corrélées aux mêmes variables mesurées au niveau du bassin.

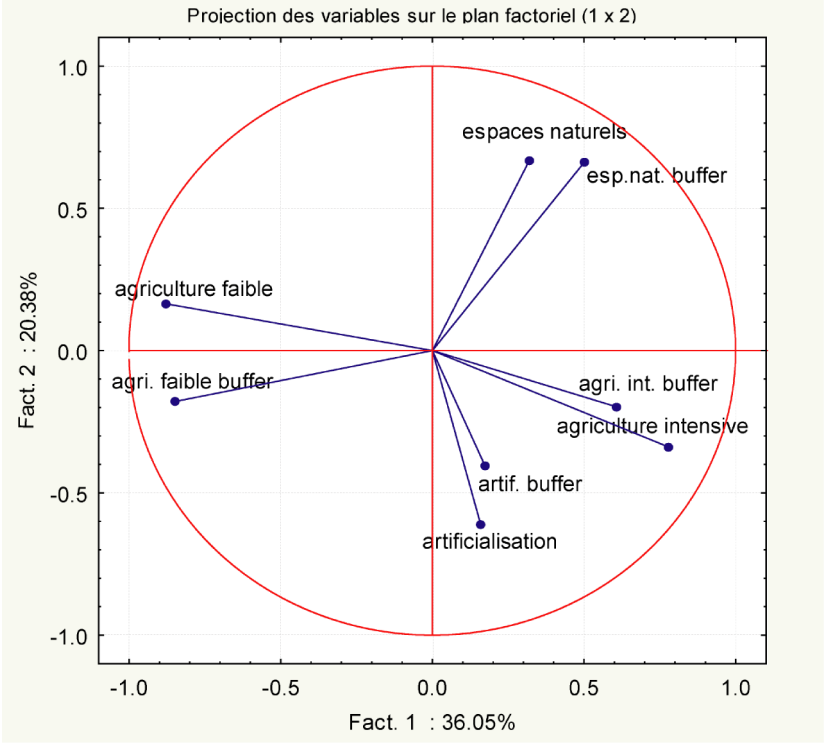


figure 30. Projection des variables sur le plan (1,2) de l'ACP normée des prédicteurs

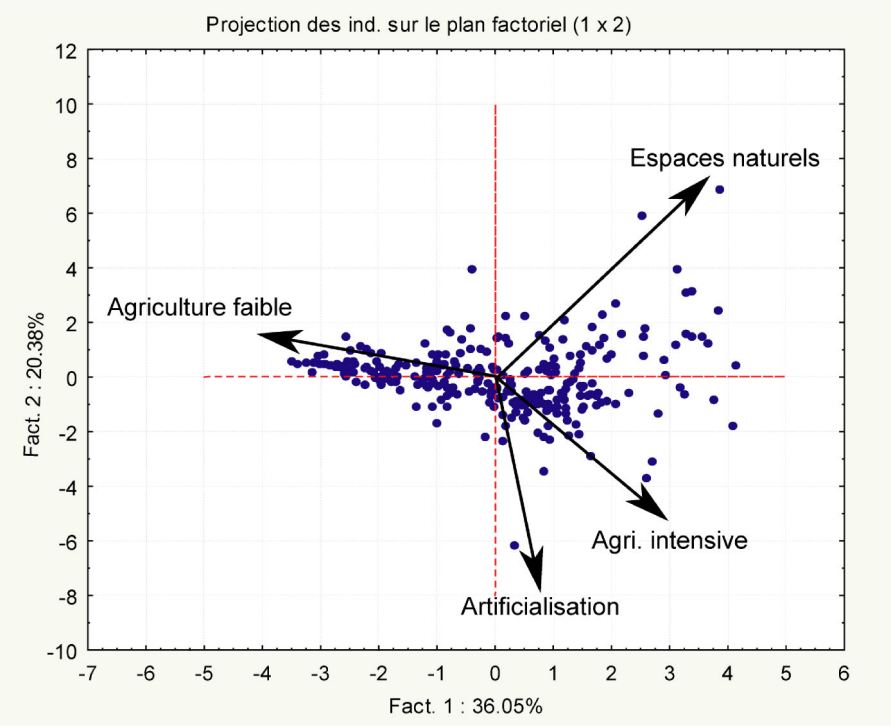


figure 31. Projection des individus sur le plan (1,2) de l'ACP normée des prédicteurs

Les 276 stations s'organisent principalement le long de l'axe F1 (figure 31) qui correspond plutôt à un gradient opposant les prairies aux terres labourées.

Les régressions linéaires des deux axes sur l'EQR-IBGN montrent que seul l'axe 2 est corrélé à l'IBGN avec un coefficient positif ; donc l'axe 2 représente un gradient croissant d'EQR.

On peut donc interpréter ces résultats de la manière suivante :

- L'axe 1 de l'ACP (qui n'est pas du tout corrélé avec l'EQR-IBGN) révèle que l'agriculture n'a pas d'impact direct sur l'état écologique d'une station (évaluée avec l'IBGN), même si on peut voir une forte opposition entre agriculture faible et agriculture intensive correspondant à une structuration de l'espace agricole.
- L'axe 2 (qui a contribution significative à la variabilité de l'EQR-IBGN) montre que l'artificialisation a un impact négatif sur la qualité biologique d'une station et s'oppose aux espaces naturels qui ont un effet positif tant au niveau du bassin que du corridor.

7.3.3 Enseignements tirés du modèle de diagnostic « Régional »

On retrouve donc les mêmes résultats que précédemment.

A ce stade, on peut s'interroger sur la nature des variables qui entrent dans le modèle ; il est possible en effet que le regroupement en 4 catégories d'occupation du sol, justifié pour des modèles à l'échelle nationale, ne soit plus pertinent à l'échelle d'une hydro-écorégion qui présente son identité propre autant, par construction, en termes de caractéristiques géographiques naturelles qu'en termes d'utilisation du sol, en raison justement de ses particularités.

7.4 Modèle de Diagnostic « Régional détaillé » à partir de toutes les catégories d'occupation du sol : le Massif Armoricain

Le Massif Armoricain présente une physionomie particulière, d'une part en termes de répartition de l'urbanisation avec des villes situées le long des côtes et quelques rares agglomérations à l'intérieur des terres, d'autre part par son agriculture décrite dans la terminologie CORINE comme « systèmes culturaux et parcellaires complexes » (code CLC 2.4.2). Cette agriculture est principalement tournée vers l'élevage intensif, et la densité d'animaux y est généralement supérieure à 1,25 UGB / ha de bassin versant, ce qui constitue une particularité remarquable à l'échelle nationale (UGB = Unités de Gros Bétail). Une analyse plus détaillée de l'impact des différentes catégories d'occupation du sol est donc justifiée.

7.4.1 Régression PLS : IBGN f (occupation détaillée du sol)

Dans ce but, un nouveau modèle a été construit à partir des 44 catégories détaillées de CORINE Land Cover (cf. § 4.1 et annexe 1). Une régression PLS a été réalisée pour modéliser l'EQR-IBGN en fonction de l'occupation du sol détaillée au niveau des bassins versants et des corridors rivulaires.

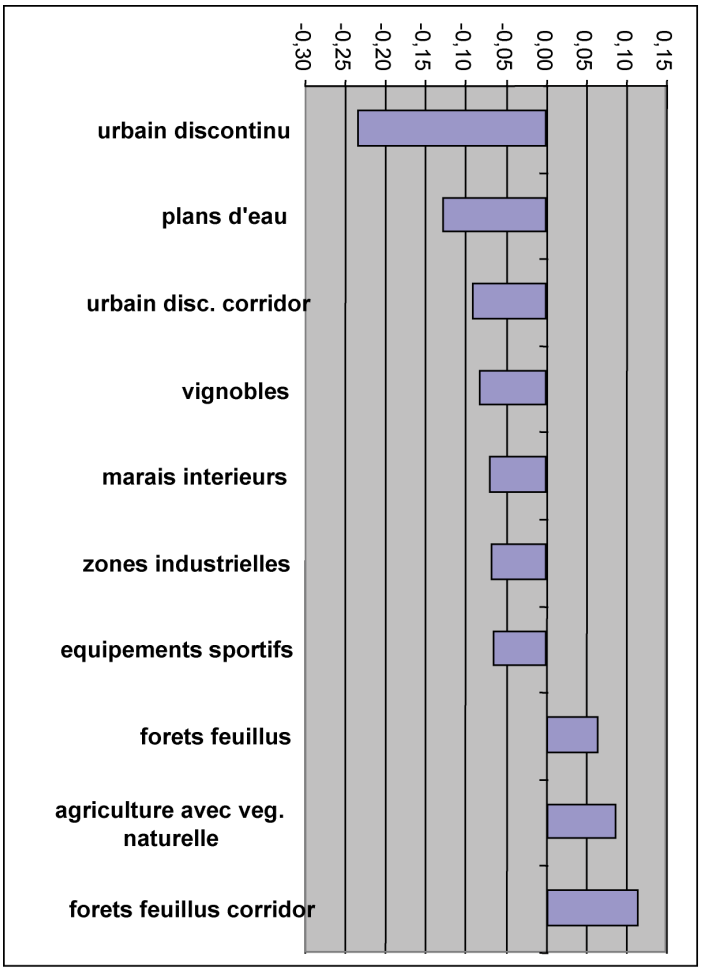


figure 32. Coefficients normés les plus forts de régression PLS (EQR-IBGN / Codes CLC détaillés)

Le R^2 du modèle est de 22%.

Les coefficients de régression normés sont présentés dans la figure 32 ; leur analyse est cette fois très instructive. Le tissu urbain discontinu du bassin est toujours le principal facteur d'impact négatif, mais il est immédiatement suivi des plans d'eau. Viennent ensuite, l'artificialisation du corridor, et un peu plus loin les zones industrielles et équipements sportifs à rattacher au tissu urbain discontinu.

En matière d'impact agricole, seuls les vignobles, pourtant très localisés, apparaissent avec un effet clairement négatif !

Enfin, les principaux facteurs positifs sont les forêts de feuillus au niveau du bassin, mais surtout du corridor, et les zones agricoles avec une forte proportion de végétation naturelle (code CLC 2.4.3).

7.4.2 Enseignements tirés du modèle de diagnostic « Régional détaillé »

Le tissu urbain discontinu du bassin reste donc le principal facteur d'impact sur l'état écologique révélé par l'IBGN.

L'apparition des plans d'eau, en grande majorité artificiels, comme second facteur d'impact confirme bien la perception des acteurs locaux sur l'importance des altérations hydro-morphologiques dans cette région. On notera aussi l'effet négatif des marais intérieurs, dont

il reste à préciser s'il s'agit d'une perturbation ou d'un phénomène naturel (à corriger dans ce cas en créant un type particulier pour les cours d'eau qui en sortent).

Au total, dans le Massif Armoricain, l'état écologique mesuré par l'IBGN serait dégradé principalement par le tissu urbain discontinu, les plans d'eau révélateurs d'altérations hydro-morphologiques, et localement par les vignobles. Les zones naturelles constituent comme partout un environnement protecteur, mais l'effet le plus positif sur l'IBGN est lié à la présence de forêts dans le corridor rivulaire : ce dernier résultat ouvre des perspectives particulièrement encourageantes pour la restauration des milieux !

7.5 Discussion : intérêt et limites des modèles de diagnostic

Ces modèles sont basés sur une méthode de régression PLS qui assure l'indépendance des prédicteurs, même lorsque les variables d'occupation du sol sont spatialement corrélées entre elles. D'autre part, ils mettent en évidence l'influence de ces prédicteurs sur l'ensemble du gradient de variation de la variable de réponse, ici l'EQR-IBGN, et non sur une discrimination en classes pré-établies. Ils ont donc une véritable capacité explicative permettant d'établir l'influence relative (positive ou négative) des variables d'occupation du sol sur la réponse de l'IBGN. Ils permettent donc de hiérarchiser les facteurs de dégradation de l'état écologique et de faire ressortir les facteurs « protecteurs ». Ils sont en cela très complémentaires des modèles d'extrapolation spatiale basés sur des arbres de décision.

D'un point de vue méthodologique, on retiendra plusieurs points importants.

- Les modèles globaux à l'échelle France entière (basés sur les grandes catégories d'occupation du sol) sont utiles pour valider et expliquer les résultats des modèles d'extrapolation spatiale. Mais les mêmes modèles à l'échelle régionale n'apportent pas d'information déterminante.
- En revanche, les modèles « régionaux détaillés », à l'échelle de grandes HER et basés sur des catégories détaillées d'occupation du sol, apportent une information très intéressante permettant d'approcher les véritables facteurs d'impact, en fonction des spécificités régionales des caractéristiques naturelles et des activités humaines.
- Ces modèles permettent également de faire ressortir des effets de proximité, en comparant les effets des même catégories d'occupation du sol à l'échelle des bassins versants et au niveau du corridor rivulaire.

Dans l'état actuel des données, les impacts hydro-morphologiques peuvent seulement être suspectés à partir des catégories d'occupation du sol génératrices d'impact. Mais rien ne s'oppose, dans l'avenir, à l'intégration de nouvelles variables de pressions dans les modèles.

En contrepartie, ces modèles ne peuvent servir à l'extrapolation spatiale ; mais ils pourraient être utilisés dans l'avenir pour prédire l'état écologique de sites sur lesquels on disposerait d'informations cohérentes sur les pressions.

8 Conclusions opérationnelles et Perspectives

8.1 Rappel des objectifs

Le présent rapport avait pour premier objectif de développer une méthodologie pour l'analyse des relations entre l'état écologique des cours d'eau et les pressions anthropiques génératrices d'impact. Nous avons donc exploré en introduction les aspects conceptuels de la problématique pour nous frayer un chemin dans l'écheveau complexe de ces relations « pressions / impacts ». En effet, les méthodes biologiques utilisées pour définir l'état écologique d'une rivière ne permettent pas à elles seules d'indiquer les causes d'altération.

Ces modèles ont pour objectif d'apporter des éléments de réponses à des questions très opérationnelles :

- à quel niveau de pression correspondrait un seuil de « bon état écologique », ambitieux et réaliste, compatible avec les exigences de la DCE ?
- où se rencontrent les problèmes majeurs, quelles sont les régions les plus impactées, et quelles en sont les causes dominantes ?
- quelles sont les grandes structures socio-économiques responsables des impacts, et sur quels facteurs peut-on agir ?

Ces questions concernent un niveau plus décisionnel qu'opérationnel, ce qui nous a orienté vers des modèles à large échelle et spatialisés, permettant de visualiser des résultats à des échelles nationales et régionales. L'approche par hydro-écorégions, cohérente avec la typologie des milieux et la spatialisation des pressions, fournit un cadre fonctionnel adéquat pour l'analyse des problèmes et la définition de politiques d'action.

L'extrapolation spatiale permettant la simulation de différentes hypothèses, le *diagnostic* des causes d'altération et la recherche de facteurs gérables pour la *restauration* sont donc les objectifs communs à ces modèles pressions / impacts.

8.2 Limites de l'exercice

En fonction à la fois de ces objectifs et de l'état des données disponibles au début de notre travail (fin 2003), nous avons limité notre approche à l'indice IBGN (méthode standardisée basée sur les invertébrés benthiques) pour définir l'état écologique des cours d'eau, et à l'occupation du sol (à partir de la base européenne CORINE Land Cover) pour l'évaluation des pressions. Ces choix sont explicités en introduction.

Les invertébrés benthiques sont les bioindicateurs les plus utilisés en eau courante au niveau européen, et seront au centre de l'exercice d'intercalibration. Les variations naturelles de l'IBGN ont été normalisées par rapport aux valeurs de référence établies par type de milieu, la variable biologique utilisée étant l'EQR-IBGN (*Ecological Quality Ratio*, ou écart à la référence pour chaque type). La corrélation entre l'occupation du sol et les principales pressions polluantes a également été testée.

Rappelons cependant deux **limitations importantes** du présent travail qui doivent être gardées à l'esprit pour l'interprétation des résultats :

- *les valeurs de référence et les limites de bon état pour l'IBGN utilisées ici sont des valeurs **provisoires**,* qui ont été affinées fin 2004 et sont encore susceptibles d'évoluer en fonction des données qui seront recueillies sur le nouveau réseau de sites de référence, et des résultats de l'intercalibration ;
- *les pressions hydro-morphologiques sont **mal** représentées* par les variables d'occupation du sol et les impacts correspondant sont donc insuffisamment pris en compte dans nos modèles.

Malgré cela, on soulignera que l'IBGN est un indice robuste, que les ajustements attendus sont limités, et que le seuil de bon état testé ici est cohérent avec celui de nos partenaires européens. Les résultats présentés ici paraissent donc suffisamment fiables pour autoriser quelques conclusions préliminaires.

8.3 Acquis méthodologiques

Rappelons ici qu'il n'existait pas au départ de notre recherche un corpus de méthodes validées pour le développement de modèles répondant aux objectifs que nous nous sommes fixés. Concernant l'objectif premier de développement d'outils, les résultats sont conséquents sur de nombreux points résumés ci-dessous.

1. *Développement d'une plate-forme informatique* couplant, à l'échelle nationale, des bases de données et un système d'information géographique (SIG) qui permet d'associer à toute station disposant de données biologiques les informations disponibles sur les caractéristiques naturelles et les pressions anthropiques à l'échelle du bassin versant et du corridor rivulaire. Cette plate-forme inclut notamment :
    a. la quasi-totalité des données existantes sur les bioindicateurs utilisés pour la définition de l'état écologique pour les invertébrés (IBGN), les diatomées (IBD), et les poissons (IP) ;
    b. la totalité du réseau hydrographique (BD CARTHAGE), avec la topologie simplifiée des drains principaux des zones hydrographiques, et le réseau ordonné (rangs de Strahler) ;
    c. la typologie des milieux et les caractéristiques géographiques naturelles (géologie, lithologie, relief, climat) des bassins versants, les caractéristiques hydrologiques (modules) et hydro-chimiques de base, ainsi que les entités administratives ;
    d. l'occupation des sols (CORINE Land Cover), et des données relatives aux pressions agricoles (risque d'érosion, recensement général agricole) ;
    e. des données chimiques du réseau national (RNDE), et certaines pressions hydro-morphologiques (grands barrages).

    Cette plate-forme gérée par le ***Cemagref***, qui constitue actuellement un outil unique en France, est continuellement perfectionnée et enrichie de nouvelles données, en particulier sur les pressions anthropiques.
2. *Développement d'un corpus de modèles* permettant d'analyser les relations pressions / impact entre toute variable biologique utilisée pour la définition de l'état écologique et des indicateurs de pressions anthropique de toute nature. Ces modèles permettent notamment :
    a. *d'extrapoler*, à des échelles nationales et régionales, un état écologique probable pour une variable biologique donnée à partir d'indicateurs de pressions spatialement homogènes ;
    b. *de simuler*, à partir des mêmes variables, le résultat de différentes hypothèses de limite du « bon état » ;
    c. d'identifier des *seuils de pressions* correspondant à ces limites de « bon état » ;
    d. *de diagnostiquer et de hiérarchiser* les facteurs les plus pénalisants ;
    e. de dégager l'impact respectif des pressions s'exerçant au niveau des *bassins versants* et *des corridors rivulaires.*

Ces modèles ont été appliqués à l'IBGN, à partir des variables d'occupation du sol ; d'importantes conclusions méthodologiques en ressortent.

- *Pour **l'extrapolation*** spatiale, la simulation de différentes hypothèses, et l'identification de seuils de pressions, ***les arbres de décisions** sont les modèles les plus adaptés* ; mais avec les données disponibles, seules les grandes catégories d'occupation du sol des bassins permettent l'extrapolation. Ces modèles n'ont donc qu'une faible valeur explicative.
- Ces modèles d'extrapolation sont plus fiables appliqués à *l'échelle nationale*, mais il ne faut pas en attendre une bonne capacité prédictive locale.
- *Pour le **diagnostic*** et la hiérarchisation des causes, et l'analyse des effets de proximité spatiale, *les modèles adéquats sont les **régressions PLS*** ;
- Les modèles de diagnostic sont beaucoup plus performants et informatifs à *l'échelle des hydro-écorégions*, et avec des *catégories détaillées* d'occupation du sol ; ils sont applicables à des variables de pression de toute nature et permettent d'étudier des *effets de proximité* par l'analyse des corridors rivulaires.

Les arbres de décisions pourront néanmoins être appliqués à des variables de pressions détaillées, pour en améliorer la valeur explicative et rechercher des effets de seuils, mais sans possibilité d'extrapolation spatiale.

8.4 Premières conclusions opérationnelles

Ces conclusions ne concernent *que l'état écologique évalué d'après les peuplements d'invertébrés* avec l'indice IBGN, dans la situation actuelle (i.e. sur la base des données observées entre 1992 et 2003) ; elles devront être complétées par la prise en compte des autres peuplements (poissons, diatomées, macrophytes), et bien évidemment la physico-chimie.

On se reportera aux chapitres 5 (§ 5.3 et 5.6), 6 (§ 6.4) et 7 (§ 7.2.4 et 7.4.2) pour le détail des analyses .

Une situation générale relativement favorable…

Pour l'ensemble du territoire métropolitain, et avec l'hypothèse de base pour la limite du bon état sur l'IBGN, *le pourcentage de tronçons de cours d'eau qui n'atteignent pas le « bon état écologique » serait **de l'ordre de 35% à 40%.***

*La probabilité actuelle de ne pas atteindre le « bon état » sur l'IBGN est principalement liée à **l'urbanisation*** ; les pressions qui en découlent concernent toutes les formes de rejets directs (pollution organique et toxique due aux rejets domestiques et industriels), mais aussi l'artificialisation des cours d'eau liée à l'occupation des lits majeurs.

*L'occupation du sol par **l'agriculture intensive***, (évaluée en pourcentage de terres labourées dans les bassins) *intervient secondairement*, et ne semble pas représenter à elle seule une cause essentielle de non atteinte du bon état sur l'IBGN. Néanmoins, *des impacts réels existent, mais qui ne semblent pas liés directement à la superficie occupée par l'agriculture* : ce point nécessite une analyse plus approfondie à des échelles appropriées.

***Les prairies** constituent en général un environnement plutôt favorable* à l'échelle nationale, mais cette conclusion doit être fortement nuancée en fonction du contexte régional.

Les espaces « naturels » (ou faiblement anthropisés) *ont par contre un effet toujours positif*.

… mais de nombreuses zones en situation limite.

Néanmoins, cette situation semble fragile. Nous avons simulé grâce aux modèles d'extrapolation spatiale le résultat d'une augmentation de la limite du bon état de un point IBGN (hypothèse BE+1), soit environ 6% pour l'EQR-IBGN.

Dans cette hypothèse, *le pourcentage de tronçons de cours d'eau qui n'atteignent pas le « bon état écologique » serait* ***de l'ordre de 50%,*** soit une augmentation comprise entre 10% et 20% selon que l'on se fie aux observations des réseaux ou aux résultats des modèles. *Cela signifie que de nombreuses zones se trouvent en « situation limite »,* et basculent vers le mauvais état dans l'hypothèse BE+1.

Des différences régionales très marquées, liées principalement aux impacts agricoles.

On constate que *ces zones en situation limite sont essentiellement situées dans les régions très agricoles* : Tables Calcaires, Massif Armoricain, Coteaux Aquitains. A l'inverse, la situation semble moins fragile dans les massifs montagneux et forestiers : Alpes et Pyrénées, PréAlpes, Massif Central, Vosges, Landes, Cévennes, Corse. Dans la région Méditerranéenne (HER 6), le basculement semble lié à la fois à des impacts agricoles et urbains.

Cette influence des impacts liés à l'agriculture intensive est confirmée par l'analyse des arbres de décision dans l'hypothèse BE+1.

On peut émettre l'hypothèse au vu de ces résultats qu'il existe une certaine ***synergie*** entre les impacts liés à l'agriculture intensive et ceux liés à l'urbanisation : *les rivières soumises à des impacts agricoles seraient plus sensibles aux impacts de l'urbanisation,* en particulier à la pollution. Néanmoins, cette hypothèse demande à être confirmée par d'autres analyses.

Des diagnostics révélateurs de « pathologies » régionales …

Cette vision générale est confirmée par les modèles de diagnostic, qui permettent de hiérarchiser les principales causes d'altération de l'état écologique révélé par l'IBGN, sur l'ensemble du gradient de variation de cet indice. On retrouve au niveau national la même hiérarchie entre urbanisation, terres labourées, prairies et espaces naturels.

Mais ces modèles permettent surtout de préciser le diagnostic à des échelles beaucoup plus pertinentes pour l'analyse des relations pressions / impact. L'exemple du Massif Armoricain (HER 12), région très particulière par sa géographie et ses activités agricoles, est à ce titre particulièrement instructif.

L'urbanisation constitue comme ailleurs le principal facteur négatif. L'impact général des terres agricoles, objet d'une polyculture complexe, reste flou ; mais dans cette région d'élevage intensif, les prairies qui ont un effet positif au niveau du bassin semblent avoir un impact sur l'IBGN lorsqu'elles sont situées en bordure des rivières.

A un niveau plus détaillé, on voit apparaître les plans d'eau, révélateurs d'altérations hydro-morphologiques, comme second facteur pénalisant.

… et du rôle essentiel des corridors rivulaires.

Mais le plus intéressant concerne le rôle majeur des corridors rivulaires, car le facteur le plus positif sur l'IBGN est la présence de forêts de feuillus au bord du cours d'eau, en opposition à l'effet négatif de l'urbanisation diffuse des fonds de vallées.

Cette *mise en évidence à large échelle du rôle* ***protecteur*** *des corridors rivulaires* constitue une information nouvelle : si de nombreuses études de cas ont démontré l'importance fonctionnelle de ces corridors, il n'existait pas jusqu'alors de démonstration d'une influence positive sur l'état écologique à l'échelle d'un réseau de suivi. Ce résultat constitue un

élément essentiel pour la validation des effets attendus d'une politique de restauration portant sur ces espaces « gérables ».

Quelles actions prioritaires pour un objectif « réaliste et ambitieux » ?

Ces premières observations débouchent sur des questions centrales : quelle limite de bon état correspondrait à un objectif « réaliste et ambitieux » pour l'état écologique des rivières ? Comment atteindre cet objectif ?

On remarquera tout d'abord que *la situation actuelle* correspondant à l'hypothèse de base *n'est pas exagérément alarmante*. *Une légère augmentation de la limite de bon état (6% en EQR-IBGN) ne semble donc pas hors de portée*, et elle aurait pour effet de faire porter un effort significatif sur une large partie du territoire, principalement dans les zones agricoles.

Or si l'agriculture n'apparaît pas comme la cause majeure d'altération pour les invertébrés, cela signifie en contrepartie ***que l'objectif de « bon état » ne semble pas irréaliste, même dans les régions très agricoles***. En effet, on peut trouver des cours d'eau en « bon état » même dans des bassins largement cultivés. Il semble ici que le type de culture, les pratiques agricoles, et un corridor rivulaire protecteur aient plus d'effet sur l'état écologique que la seule superficie cultivée dans le bassin.

Il semble donc possible, sans remettre en cause les activités agricoles, mais avec des actions volontaristes et généralisées sur l'ensemble des territoires concernés, d'améliorer d'un ou deux points IBGN l'état des rivières dans les régions dédiées à l'agriculture intensive.

Des actions significatives devront être conduites à la fois :

- au niveau des parcelles cultivées, pour réduire l'érosion des sols, la perte de nutriments et la diffusion des pesticides ;
- au niveau des corridors rivulaires, pour limiter les transferts de polluants, permettre la reconstitution des habitats et la restauration des processus écologiques fondamentaux.

Néanmoins, il faudra garder à l'esprit *les délais de réponse des systèmes écologiques* dans les actions de restauration de ce genre ; en particulier, la restructuration physique sous l'action des processus morpho-dynamiques naturels peut s'avérer très lente pour les rivières à faible énergie dans les régions de plaine.

Il n'en reste pas moins que les situations les plus dégradées apparaissent généralement liées à l'urbanisation. ***La réduction de la pollution et des impacts urbains en général reste donc absolument nécessaire***.

De ce point de vue, on peut s'interroger sur la marge de manœuvre restante pour récupérer les milieux les plus dégradés par la pollution. En effet, en 2001, plus de 95% des logements étaient raccordés à un système épuratoire collectif ou autonome, et le taux d'abattement de la pollution organique était de 92% en sortie des stations de plus de 2000 équivalent-habitant (**IFEN** 2004). Là encore, il semble possible de gagner quelques points d'IBGN au prix d'un effort soutenu, mais il est probable que des procédures de report d'objectifs, voire d'objectifs moins stricts, soient à envisager dans les zones très fortement urbanisées.

8.5 Perspectives

Les bases méthodologiques étant maintenant fixées et validées, les développements des modèles pressions / impact se feront dans trois directions :

1. l'application des modèles d'extrapolation aux autres variables biologiques (poissons et diatomées) qui entrent dans l'évaluation de l'état écologique ;
2. l'extension des modèles de diagnostic avec variables détaillées à l'ensemble des grandes hydro-écorégions, avec un regard particulier sur le rôle des corridors rivulaires ;
3. l'intégration de nouvelles variables de pression, notamment
    a. pour caractériser les activités agricoles,
    b. pour intégrer certaines pressions hydro-morphologiques (selon les données disponibles).

Enfin, l'intégration à la base de données des variables chimiques permettra de rechercher les seuils de paramètres correspondant à la limite de bon état. Ce point semble prioritaire dans l'agenda, mais nécessitera sans doute de nouveaux développements méthodologiques.

Références bibliographiques

Billen, G. & Garnier, J. (1999). Nitrogen transfers through the Seine drainage network: a budget based on the application of the 'Riverstrahler model'. *Hydrobiologia,* **410**: 139-150.

Boët, P., Fuhs, T., Gorges, G. & Toupotte, L. (2001). Modélisation prédictive des peuplements piscicoles, à l'échelle du bassin de la Seine. Rapport d'activité PIREN-Seine 2000.

Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984). Classification and Regression Trees. Wadsworth International Group. Belmont, California.

Garcia, A., Villeneuve, B. & Wasson, J.G. (in press). Combined pressures and geographical context. **In:** *REBECCA WP4 Rivers. Deliverable 6: Report on existing methods and relationships linking pressures, chemistry and biology in rivers*. Andersen, J.M., Dunbar, M. & Friberg, N. (Eds.). Contract No: SSPI-CT-2003-502158, European Commission.

Garcia, A. & Wasson, J.G. (2005). Combined pressures. **In:** *Relationships between ecological and chemical status of surface waters: Analysis of the current knowledge gaps for the implementation of the Water Framework Directive*. Heiskanen, A.S. & Solimini, A.G. (Eds.). Rapport EUR 21497 EN, European Commission. p 45-47.

Hanley, J.A. & McNeil, J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology,* **143**(1): 29:36.

IFEN. (2004). L'épuration des eaux usées urbaines. *Les données de l'environnement,* **98**.

Oberdorff, T., Pont, D., Hugueny, B. & Chessel, D. (2001). A probabilistic model characterizing fish assemblages of French rivers: a framework for environmental assessment. *Freshwater Biology,* **46**(3): 399-415.

Perez-Correa, M. (2004). Développement d'une méthode de cartographie de l'occupation du sol le long des cours d'eau à partir de données de télédétection, SILAT.

SIEE, STRATEGIS & **Cemagref.** (2002). Définition d'un réseau national de stations ou tronçons de référence. Tome 1 : Méthodologie. Rapport final, Ministère de l'Aménagement du Territoire et de l'Environnement, Agences de l'Eau Adour-Garonne. 94 p.

Souchon, Y., Andriamahefa, H., Cohen, P., Breil, P., Pella, H., Lamouroux, N., Malavoi, J.R. & Wasson, J.G. (2000). Régionalisation de l'habitat aquatique dans le bassin de la Loire, Agence de l'eau Loire Bretagne, Cemagref Lyon BEA/LHQ. 291p.

Swets, J.A. (1988). Measuring accuracy of diagnostic systems. *Science,* **240**(4857): 1285-1293.

Tenenhaus, M. (1998). La régression PLS. Technip, Paris.

Usseglio-Polatera, P., Richoux, P., Bournaud, M. & Tachet, H. (2001). A functional classification of benthic macroinvertebrates based on biological and ecological traits: application to river condition assessment and stream management. *Archiv für Hydrobiologie, Suppl*. **139**(1): 53-83.

Wasson, J.G., Chandesris, A., Pella, H. & Blanc, L. (2002). Définition des Hydro-écorégions françaises métropolitaines. Approche régionale de la typologie des eaux courantes et éléments pour la définition des peuplements de référence d'invertébrés. Ministère de l'Aménagement du Territoire et de l'Environnement, Cemagref Lyon BEA/LHQ. 190 p.

Wasson, J.G., Chandesris, A., Pella, H., Blanc, L., Villeneuve, B. & Mengin, N. (2003). Détermination des valeurs de référence de l'IBGN et propositions de valeurs limites du "Bon Etat". Cemagref Lyon BEA/LHQ, Valorez, ZABR. 74 p + annexes.

ANNEXE 1 : Nomenclature CORINE Land Cover

Le programme CORINE land cover repose sur une nomenclature standard hiérarchisée à **3 niveaux** et **44 postes** répartis selon **5 grands types d'occupation du territoire** :

1. Territoires artificialisés

1.1. Zones urbanisées

1.1.1. Tissu urbain continu
1.1.2. Tissu urbain discontinu

1.2. Zones industrielles ou commerciales et réseaux de communication

1.2.1. Zones industrielles et commerciales
1.2.2. Réseaux routier et ferroviaire et espaces associés
1.2.3. Zones portuaires
1.2.4. Aéroports

1.3. Mines, décharges et chantiers

1.3.1. Extraction de matériaux
1.3.2. Décharges
1.3.3. Chantiers

1.4. Espaces verts artificialisés, non agricoles

1.4.1. Espaces verts urbains
1.4.2. Equipements sportifs et de loisirs

2. Territoires agricoles

2.1. Terres arables

2.1.1. Terres arables hors périmètres d'irrigation
2.1.2. Périmètres irrigués en permanence
2.1.3. Rizières

2.2. Cultures permanentes

2.2.1. Vignobles
2.2.2. Vergers et petits fruits
2.2.3. Oliveraies

2.3. Prairies

2.3.1. Prairies

2.4. Zones agricoles hétérogènes

2.4.1. Cultures annuelles associées aux cultures permanentes
2.4.2. Systèmes culturaux et parcellaires complexes
2.4.3. Territoires principalement occupés par l'agriculture,avec présence de végétation
naturelle importante
2.4.4. Territoires agro-forestiers

3. Forêts et milieux semi-naturels

3.1. Forêts

3.1.1. Forêts de feuillus
3.1.2. Forêts de conifères
3.1.3. Forêts mélangées

3.2. Milieux à végétation arbustive et/ou herbacée

3.2.1. Pelouses et pâturages naturels
3.2.2. Landes et broussailles
3.2.3. Végétation sclérophylle
3.2.4. Forêt et végétation arbustive en mutation

3.3. Espaces ouverts, sans ou avec peu de végétation

3.3.1. Plages, dunes et sable
3.3.2. Roches nues
3.3.3. Végétation clairsemée
3.3.4. Zones incendiées
3.3.5. Glaciers et neiges éternelles

4. Zones humides

4.1. Zones humides intérieures

4.1.1. Marais intérieurs
4.1.2. Tourbières

4.2. Zones humides maritimes

4.2.1. Marais maritimes
4.2.2. Marais salants
4.2.3. Zones intertidales

5. Surfaces en eau

5.1. Eaux continentales

5.1.1. Cours et voies d'eau
5.1.2. Plans d'eau

5.2. Eaux maritimes

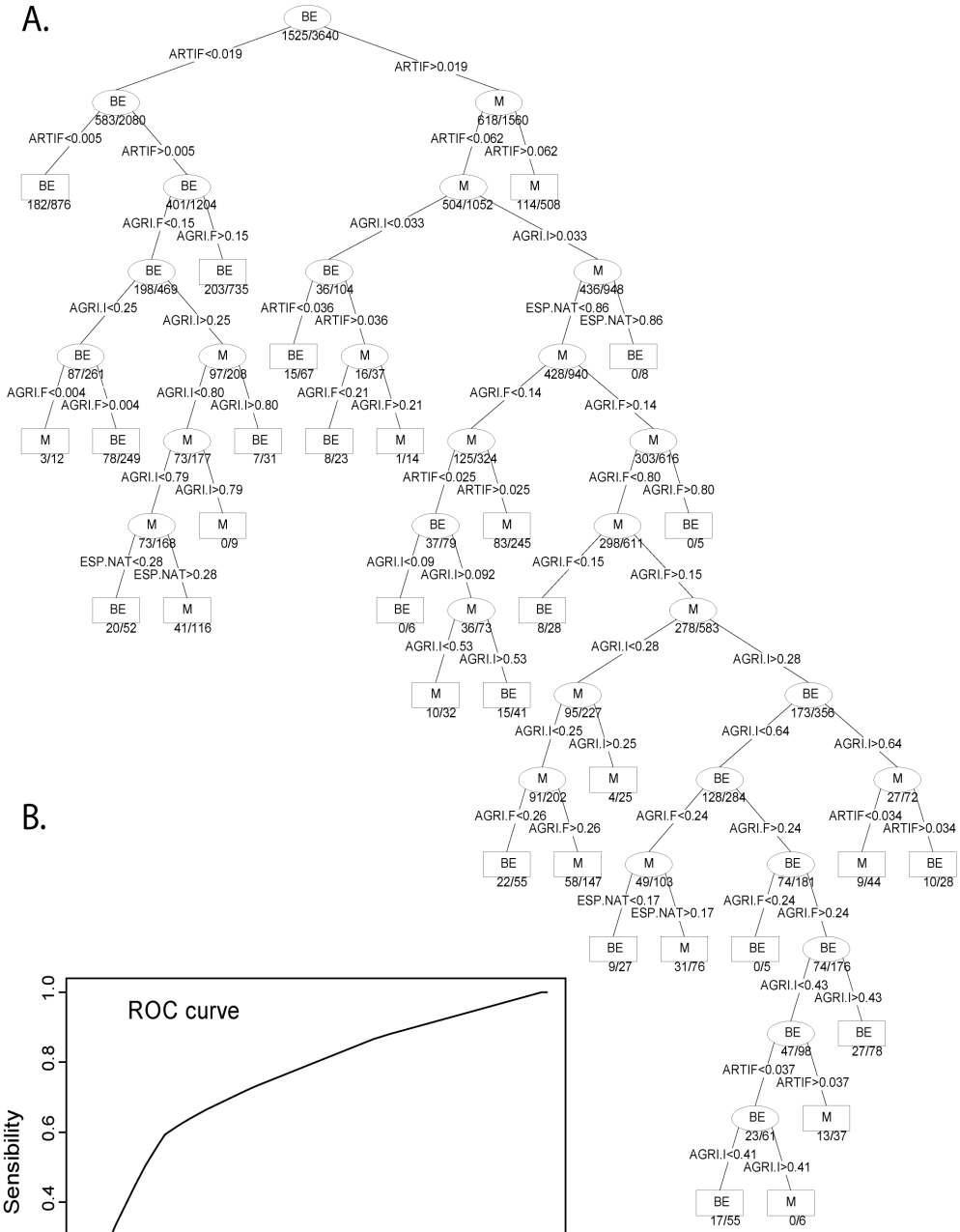
5.2.1. Lagunes littorales
5.2.2. Estuaires
5.2.3. Mers et océans

ANNEXE 2 : arbres de décision des modèles d'extrapolation spatiale

Annexe 2.1 - Modèle France entière

A. schéma de l'arbre de décision, B. courbe ROC du modèle

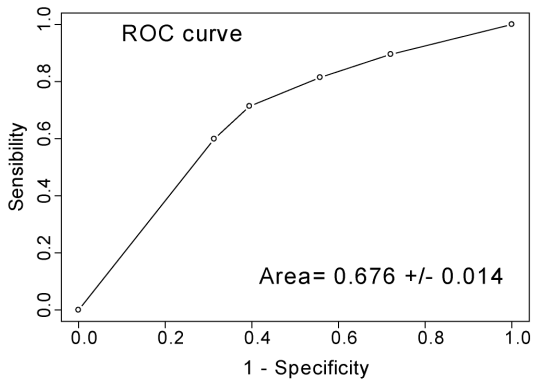
A.

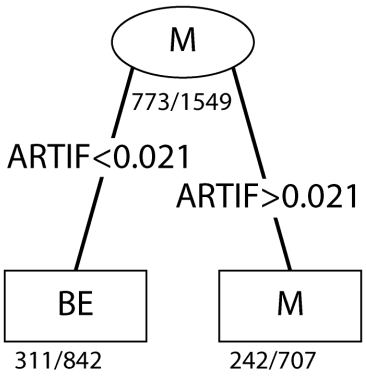


B.

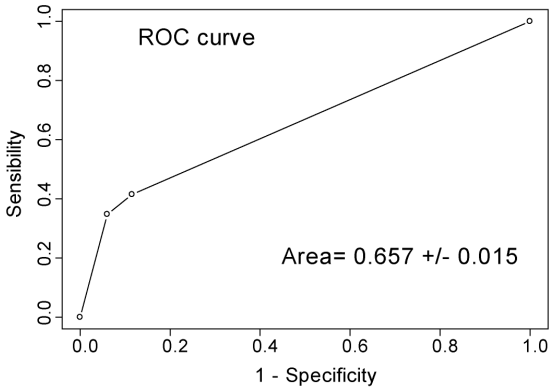
Annexe 2.2 - Modèles régionalisés (3groupes d'HER)

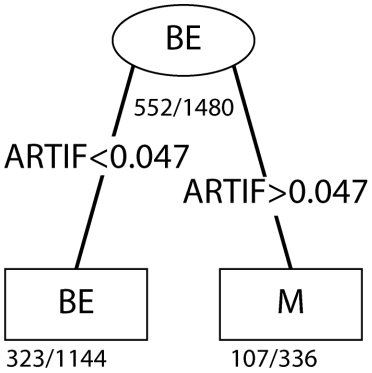
type plaine : courbe ROC et arbre de décision



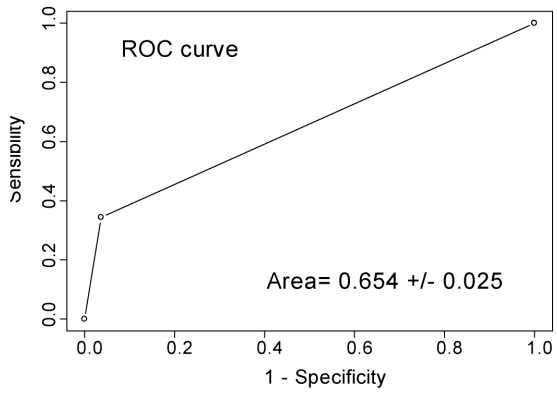


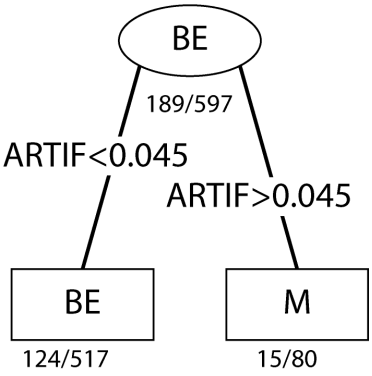
type montagne : courbe ROC et arbre de décision





type mediterranée : courbe ROC et arbre de décision

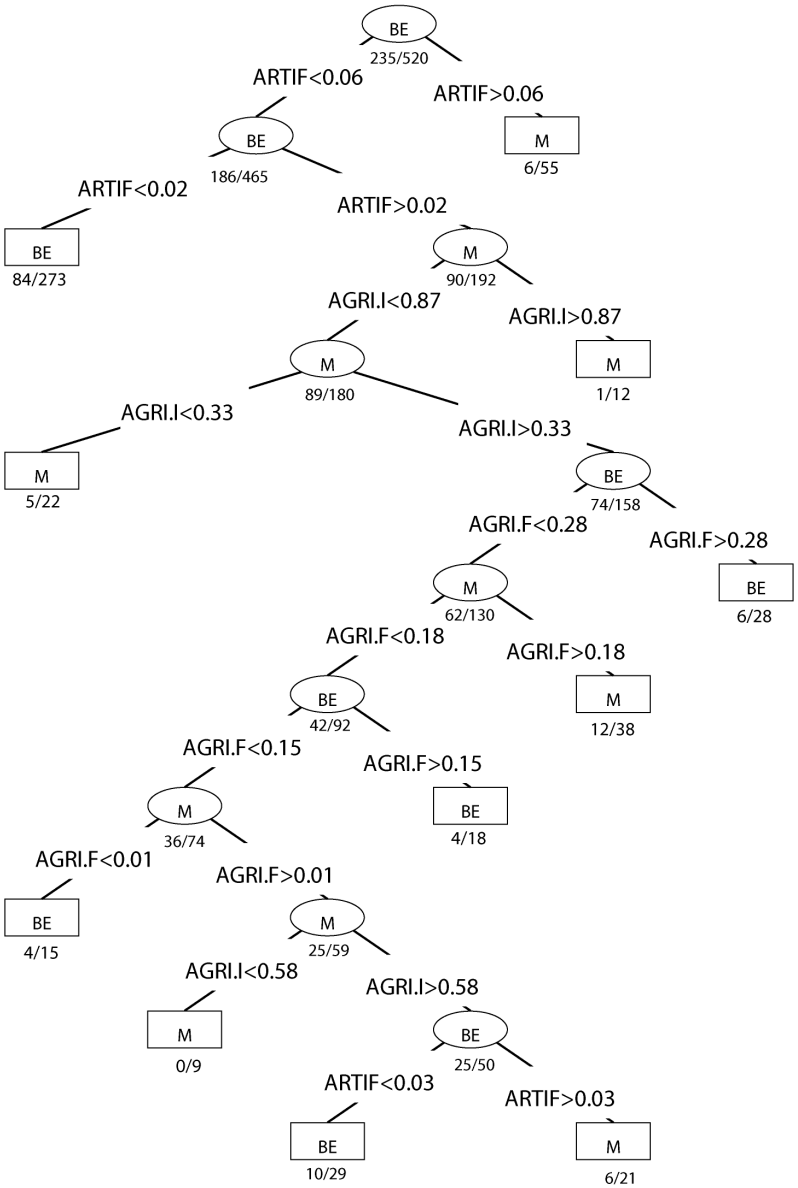




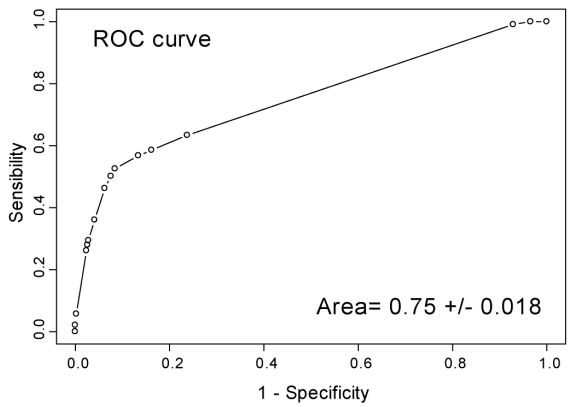
Annexe 2.3.1 - Modèle HER 9, 12 et 20

A. schéma de l'arbre de décision, B. courbe ROC du modèle

A.



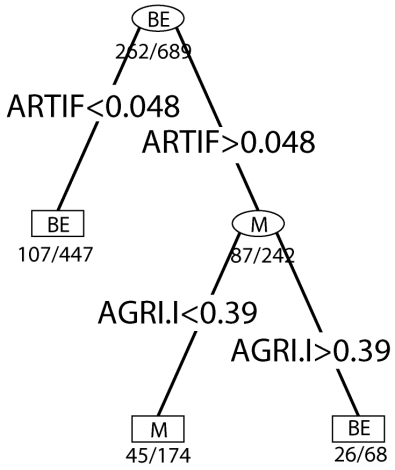
B.



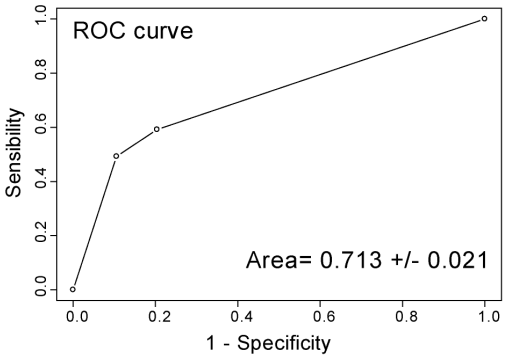
Annexe 2.3.2 - Modèle HER 5

A. schéma de l'arbre de décision, B. courbe ROC du modèle

A.



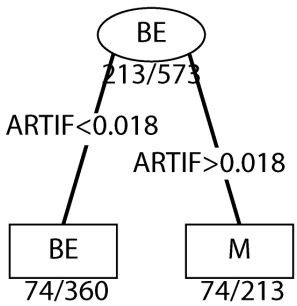
B.



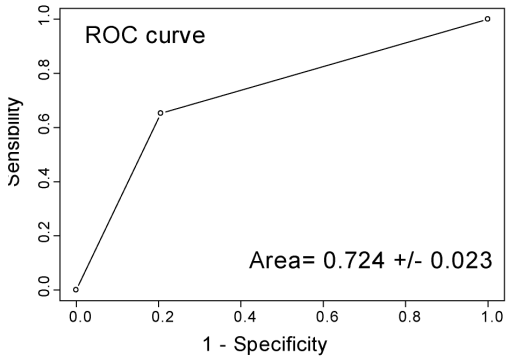
Annexe 2.3.3 - Modèle HER 3, 4, 17, 19, 21

A. schéma de l'arbre de décision, B. courbe ROC du modèle

A.



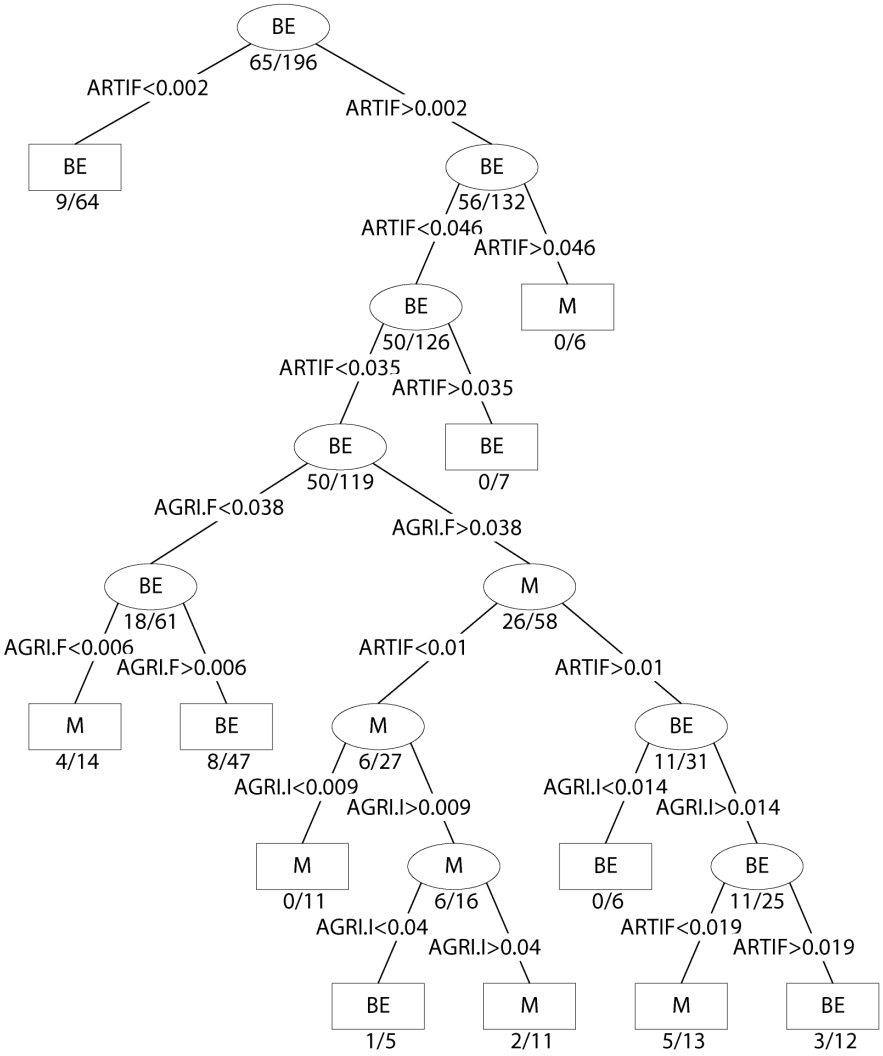
B.



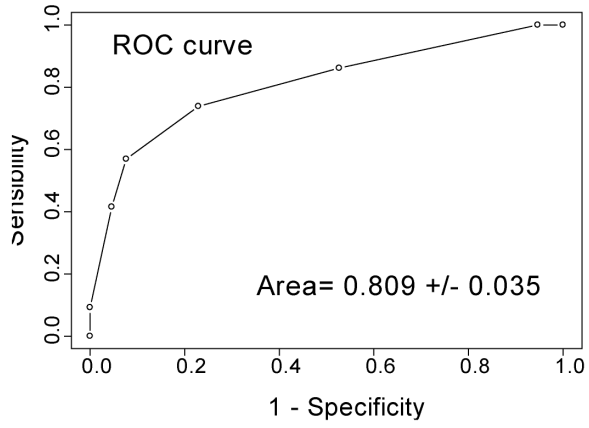
Annexe 2.3.4 - Modèle HER 1 ET 2

A. schéma de l'arbre de décision, B. courbe ROC du modèle

A.



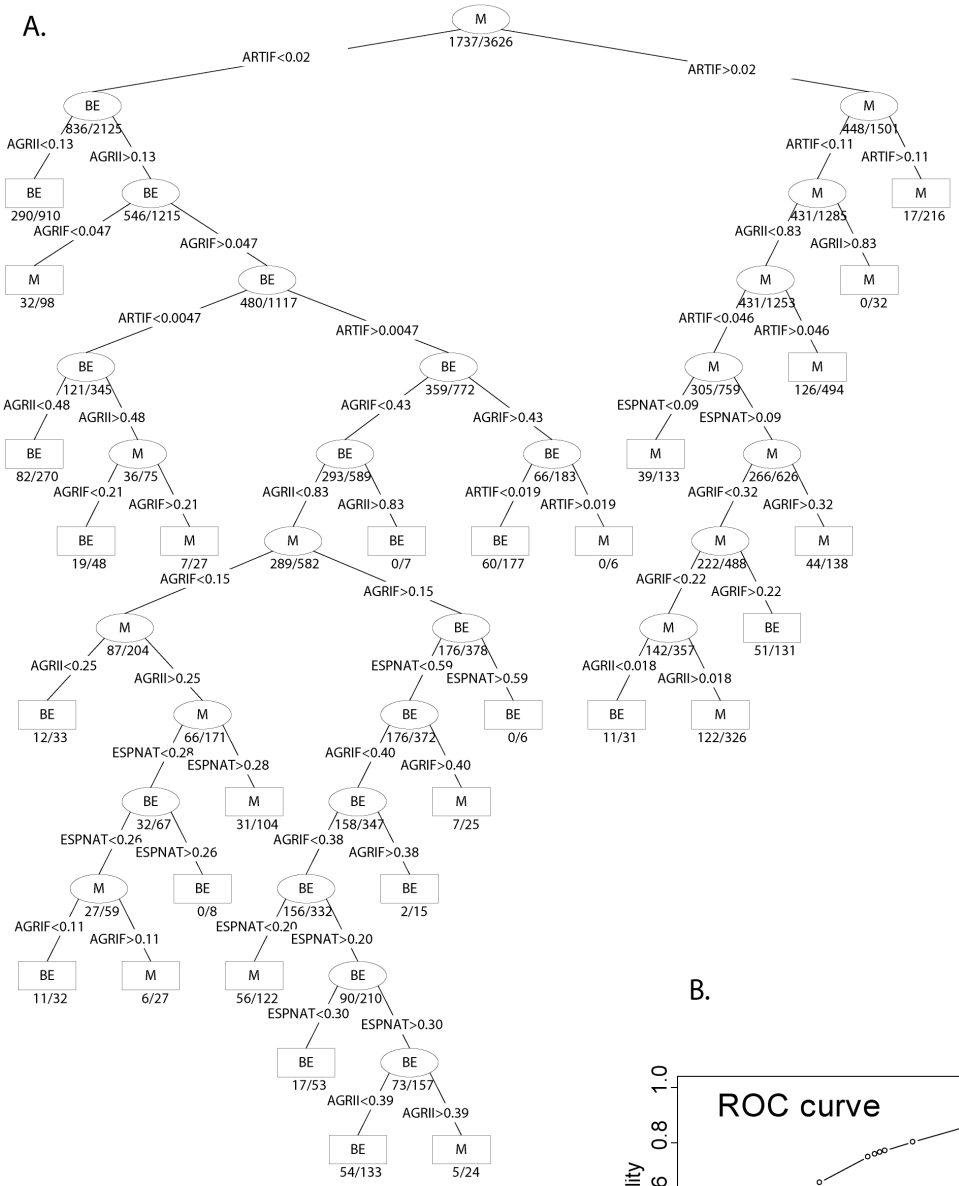
B.



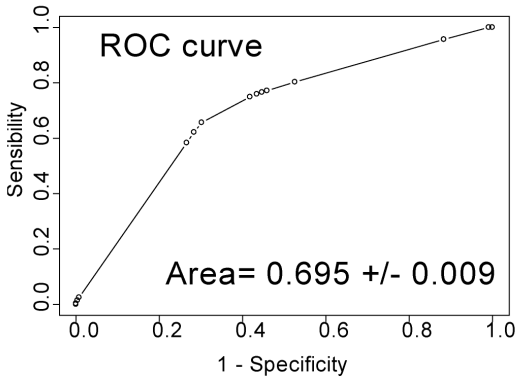
Annexe 2.4 - Modèle France entière (limite BE+1)

A. schéma de l'arbre de décision, B. courbe ROC du modèle

A.

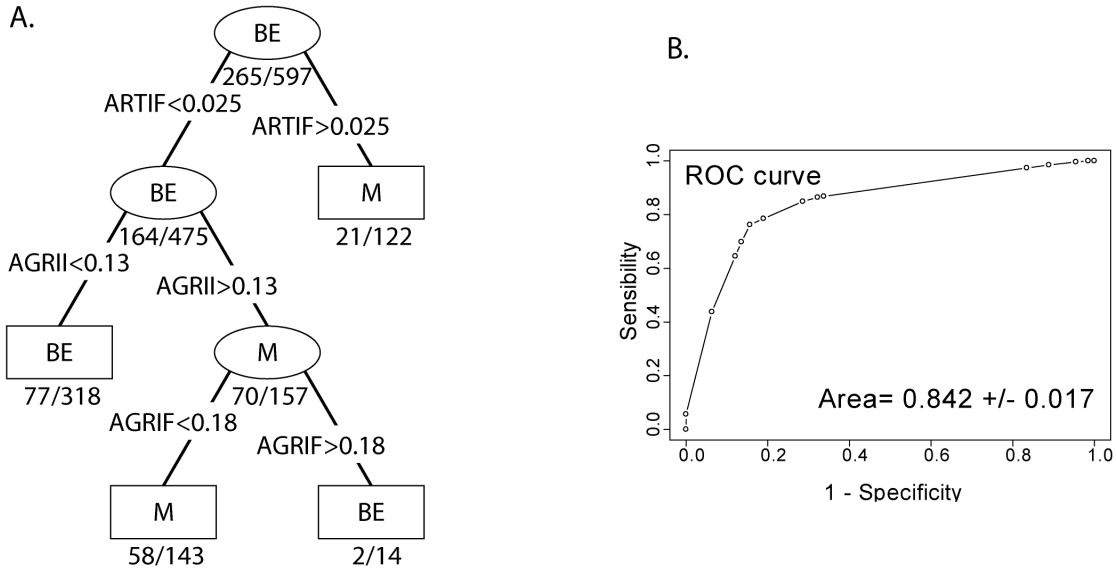


B.



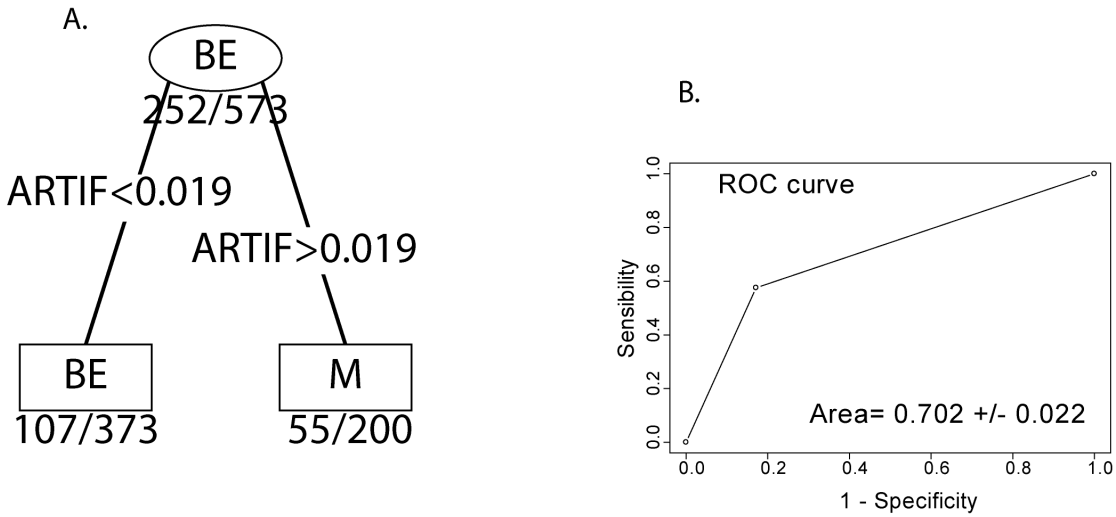
Annexe 2.5.1 - Modèle méditerranée (limite BE+1)

A. schéma de l'arbre de décision, B. courbe ROC du modèle



Annexe 2.5.2 - Modèle HER 3, 21, 4, 17, 19 (limite BE+1)

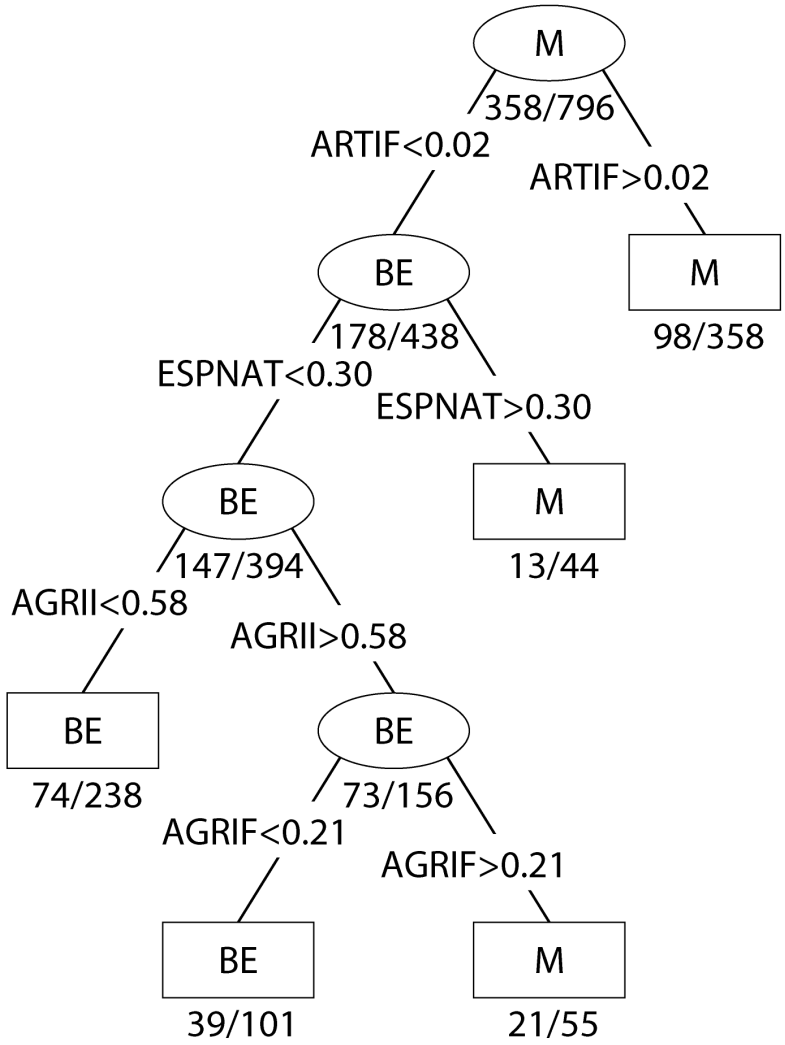
A. schéma de l'arbre de décision, B. courbe ROC du modèle



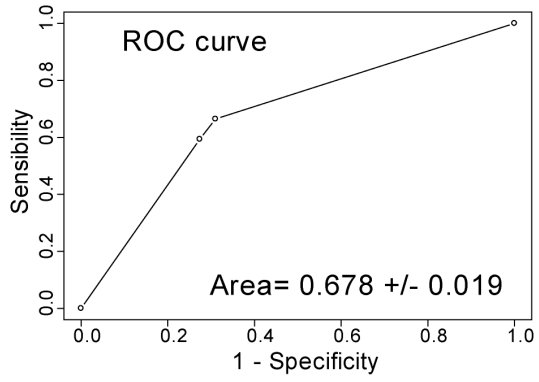
Annexe 2.5.3 - Modèle HER 9, 12 et 20 (limite BE+1)

A. schéma de l'arbre de décision, B. courbe ROC du modèle

A.



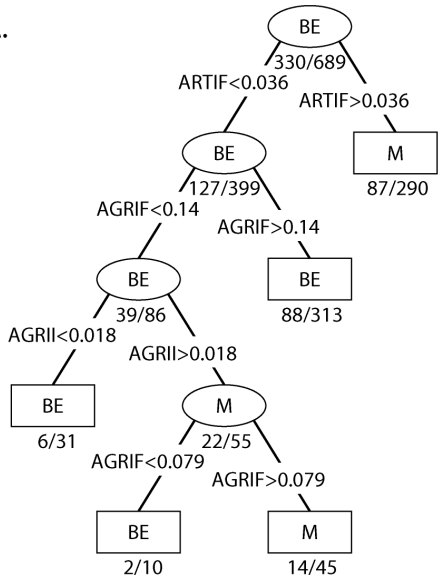
B.



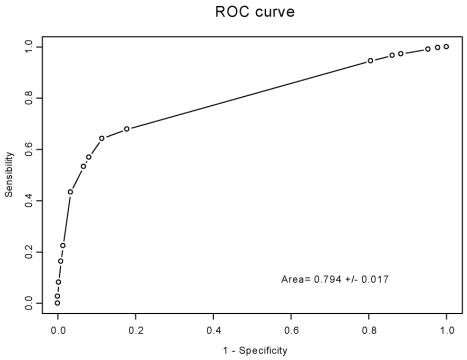
Annexe 2.5.4 - Modèle HER 5 (limite BE+1)

A. schéma de l'arbre de décision, B. courbe ROC du modèle

A.



B.



Annexe 2.5.5 - Modèle HER 1 et 2 (limite BE+1)

schéma de l'arbre de décision