COMMISSARIAT
GÉNÉRAL AU
DÉVELOPPEMENT
DURABLE

SERVICE
DE L'OBSERVATION
ET DES STATISTIQUES

Juin
2016

Document de travail n° 22

Environnement

Évaluation de tendances nationales à partir de données ponctuelles issues de réseaux d'observation

Application à l'indice poisson rivière (IPR)

Service de l'observation et des statistiques
www.statistiques.developpement-durable.gouv.fr

Auteurs : Pascal Irz, Michaël Levi-Valensin, Marlène Kraszewski (*)

(*) Les auteurs tiennent à remercier Nicolas Poulet (Onema) pour ses avis et la fourniture de données, ainsi que Thierry Oberdorff (IRD) et Christophe Barbraud (CNRS) pour leurs éclairages méthodologiques.

Ce document annule et remplace la version mise en ligne en avril 2016.

Sommaire

1 Introduction

1.1 Cadre général

Les réseaux de suivi environnemental se sont largement développés au cours des dernières décennies. Leur mise en place a généralement répondu à des exigences réglementaires. À titre d'exemples, le réseau EMEP (The European Monitoring and Evaluation Programme) assure la surveillance des pollutions atmosphériques au niveau européen. Au niveau national, le ROCCH (Réseau d'observation de la contamination chimique du littoral) et le RMQS (Réseau de mesures de la qualité des sols) suivent respectivement la contamination chimique du littoral et celle des sols. La faune et la flore font également l'objet de suivis.

Les données produites par ces réseaux varient selon les objectifs poursuivis, les phénomènes étudiés, l'échelle spatiale et temporelle considérée et les moyens alloués à leur acquisition. Des compromis sont nécessairement arbitrés entre la finesse du maillage spatial, la fréquence des observations, le nombre de paramètres mesurés et la précision de la mesure.

Malgré leur diversité, ces réseaux partagent des caractéristiques communes. Les mesures sont localisées ponctuellement et généralement répliquées dans le temps. Si un réseau a été bien conçu et correctement doté en moyens, il doit produire des données permettant de répondre aux objectifs ayant présidé à sa mise en place. Les variations spatio-temporelles des paramètres mesurés, analysées par un traitement statistique approprié, doivent permettre de répondre à diverses questions dont les plus courantes sont :

- La tendance est-elle favorable ?
- La situation est-elle spatialement homogène ?
- Quelle partie de la variabilité spatio-temporelle est attribuable aux activités anthropiques ?

Les données brutes ne permettent généralement pas de répondre directement à ces questions. Diverses interrogations sont forcément soulevées sur la manière d'agréger, de caractériser les tendances centrales et la dispersion, de traiter les valeurs extrêmes et les données manquantes, de contrôler les effets des biais connus.

De même que les réseaux de collecte évoluent au fil du temps, les questions sociétales changent et les outils d'analyse progressent considérablement. Il est donc utile d'interroger périodiquement nos pratiques en termes d'adéquation entre les données mobilisées, les méthodes employées et les questions considérées.

1.2 Objectifs de l'étude

La présente étude a pour objectif de tester des méthodes de modélisation statistique pour traiter les données d'indice poisson rivière (IPR) produites par l'Onema. Une attention particulière sera portée sur l'évaluation de la tendance pluriannuelle.

La démarche est présentée de manière la plus générique possible pour que certains de ses éléments puissent être valorisés dans le traitement d'autres jeux de données. Dans un souci de traçabilité, l'ensemble des étapes est décrit, depuis les données brutes jusqu'à la mise en forme finale. Les scripts sont fournis en annexes.

2 Données et méthodes

2.1 L'indice poisson rivière (IPR)

2.1.1 Qu'est-ce que l'IPR ?

L'IPR est un outil d'évaluation de la qualité des cours d'eau, sur la base de leurs communautés de poissons. Une notice en précise les conditions d'applications normalisées (Onema, 2006).

Pour chaque station, une situation de référence est estimée pour un jeu de 7 métriques décrivant le peuplement :

- nombre total d'espèces ;
- nombre d'espèces rhéophiles ;
- nombre d'espèces lithophiles ;
- densité d'individus tolérants ;
- densité d'individus invertivores ;
- densité d'individus omnivores ;
- densité totale d'individus.

Définition des traits des espèces de poissons utilisés pour obtenir les métriques

Invertivore :	qui se nourrit d'invertébrés. Pour les poissons, il s'agit de maillons intermédiaires dans les chaînes alimentaires.
Lithophile :	qui est dépendant de substrats minéraux (sable, graviers, galets) pour sa reproduction.
Omnivores :	qui peut se nourrir sur une large gamme d'aliments, animaux et végétaux.
Rhéophile :	qui affectionne les eaux vives.
Tolérant :	se dit d'une espèce dont la présence et l'abondance sont peu influencées par la dégradation du milieu en conséquence des activités anthropiques.

La station considérée est par ailleurs décrite par 9 variables environnementales :

- surface du bassin versant ;
- distance à la source ;
- largeur et profondeur de la station ;
- pente du cours d'eau ;
- altitude ;
- régime de température (2 variables) ;
- unité hydrographique (bassin versant).

À partir de ces 9 variables, les valeurs de référence des métriques sont estimées pour chacun des sites. Les références sont donc différentes selon la localisation et les caractéristiques du site et de son environnement, en particulier amont. Ainsi, l'IPR est conçu pour permettre de distinguer, dans la variabilité observée des communautés de poissons, ce qui est de l'ordre de la variabilité naturelle de ce qui est attribuable aux perturbations anthropiques.

Pour chaque métrique, un score est calculé comme la différence entre la valeur de référence et la valeur observée lors de la pêche (cf. *§2.1.2*). La note IPR de la station considérée au moment de la mesure est la somme de sept scores :

$$IPR = \sum_{i=1}^{7} Score_i$$

La note IPR peut théoriquement varier de zéro (situation de référence parfaite) à l'infini.

2.1.2 Champ d'application

La méthodologie de construction de l'IPR a été présentée en deux articles.

Le premier est consacré à la modélisation des données de présence/absence de 34 espèces de poissons, en fonction des caractéristiques des stations (Oberdorff, Pont, Hugueny, & Chessel, 2001). Les inventaires piscicoles utilisés pour caler les modèles logistiques sont réalisés par pêche électrique. Les 650 stations échantillonnées sont réputées « de référence », donc subissant des pressions anthropiques minimales. Elles sont situées en France continentale et les pêches sont réalisées en période de basses eaux, d'août à octobre.

Le second article a proposé un jeu de métriques « candidates », décrivant divers aspects des communautés piscicoles (Oberdorff, Pont, Hugueny, & Porcher, 2002). Les métriques basées sur l'occurrence des espèces (nombre d'espèces) ont été obtenues à partir des modèles du précédent article. Celles basées sur les abondances des espèces (densités) ont été modélisées par régression multiple, en fonction des mêmes variables caractérisant les stations. La sélection des métriques a été réalisée en testant leur capacité à distinguer des sites réputés de référence de sites réputés perturbés (Oberdorff et al., 2002).

L'usage de l'IPR doit donc être réservé aux conditions d'application des modèles sous-jacents. L'outil IPR ayant été ajusté pour des stations situées en France continentale, échantillonnées par pêche électrique entre août et octobre, en toute rigueur il ne doit pas être employé en dehors de ce cadre. Toutefois, l'extension de l'étude pour inclure les pêches réalisées quelle que soit

leur date nous a semblé avoir des vertus démonstratives, sur une méthode de prise en compte de la cyclicité annuelle du phénomène étudié, aussi est-elle présentée en *annexe 6*.

2.2 Les données IPR et leur prétraitement

2.2.1 Description des données

Les données analysées proviennent de l'Onema (direction de la connaissance et de l'information sur l'eau) et sont extraites de son entrepôt de données du système d'évaluation de l'état des eaux (SEEE). Cette table comprend les notes IPR par station et par date. Elle couvre la période 1995 – 2013.

La table des données IPR a été complétée par des informations sur les méthodes de pêche électrique utilisées sur le terrain. Ce travail a également été réalisé par l'Onema (direction de l'action scientifique et technique).

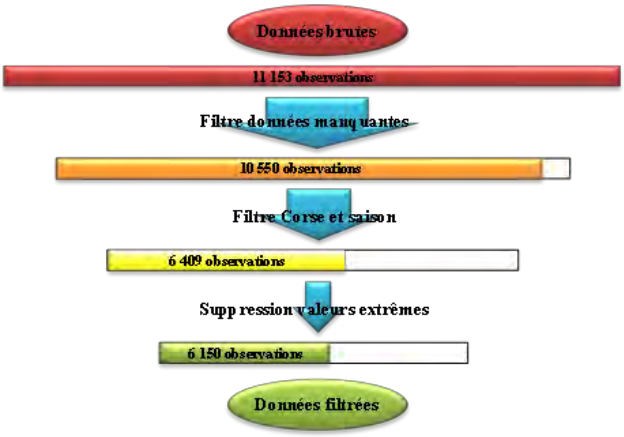
Figure 1 : cartographie des données IPR 2011-2012 (n = 1 388 stations échantillonnées)

Note : les notes IPR sont discrétisées et transformées en classes de qualité (Onema, 2006). Quand une même station a été échantillonnée à plusieurs reprises, elle se voit attribuer la moyenne géométrique des notes successives.
Source : Onema, 2014. Traitements : SOeS

2.2.2 Prétraitements des données

Avant de calculer l'indice annuel, des prétraitements doivent être opérés sur le jeu de données. Les premiers sont détaillés en *annexe 1* et les scripts correspondants sont fournis en *annexes 2 et 3*, respectivement pour *SAS* et *R*. Ils sont schématisés en Figure 2.

Figure 2 : schématisation des étapes du prétraitement des données



À partir de 2007, le nombre des stations du réseau a augmenté, mais chacune d'entre elles n'est plus prospectée qu'une année sur deux. Jusqu'à 2006, un échantillonnage complet était réalisé chaque année. Ensuite, l'échantillonnage est complet avec les couples d'années consécutives (2007-2008, ...).

Dans un premier temps, des prétraitements sont appliqués (*annexe 1* et *Figure 2)*.

Ensuite, le jeu de données est filtré pour ne conserver que la méthode de pêche complète et le moyen de pêche à pied. Il est plus rigoureux de se limiter à ce cadre car seules les pêches complètes ont été utilisées de manière uniforme tout au long de la période (*Figure 4*). Ce filtrage a pour conséquence d'exclure de l'analyse les cours d'eau qui n'ont été échantillonnés qu'au moyen d'autres méthodes. C'est le cas en particulier de ceux qui sont trop profonds pour être prospectés autrement qu'en bateau. Les interprétations formulées ici ne concerneront donc que les petits cours d'eau, typiquement de largeur inférieure à 10 m.

La table de données qui contenait 6 150 observations n'en contient plus que 2 415.

Les stations présentant moins de 3 observations sur l'ensemble de la période d'étude (1995-2013) sont supprimées. Ce filtre a pour objectif de ne conserver que des stations qui ont une certaine représentativité sur la période. Selon des essais réalisés, un filtrage plus strict (5 et 8 occurrences) n'altère pas les résultats des modèles.

La table de données ne contient désormais plus que 2 221 observations.

2.3 Proposition de méthode

Le suivi environnemental produit des mesures répétées sur un réseau de stations de mesures (ter Braak, van Strien, Meijer, & Verstrael, 1994).

L'analyse de ces données longitudinales autorise potentiellement :

- une estimation annuelle de la situation sur une zone donnée ;
- une analyse des changements entre plusieurs années ;
- une estimation de la tendance d'évolution sur une période.

Les données de suivi sont très généralement incomplètes. En effet, la probabilité de n'avoir aucune donnée manquante diminue dès lors que la période couverte s'allonge et que le nombre de points de mesures augmente.

2.3.1 Méthodes d'analyse des données de surveillance des milieux

La modélisation statistique des mesures (brutes ou transformées) en fonction d'une variable qualitative identifiant les stations, de l'année de collecte et éventuellement de covariables est envisagée. Elle est d'usage courant dans l'analyse des indices biologiques d'abondance comme *l'indice planète vivante* (IPV, cf. Collen et al., 2009) ou pour le *suivi temporel des oiseaux* (STOC, cf. Gregory et al., 2005). Elle a pour principaux avantages de permettre la prise en compte de covariables et de fournir

des statistiques de diagnostic. Outre le suivi temporel des populations animales et végétales, ces méthodes sont communes en épidémiologie environnementale (cf. par exemple Dominici, McDermott, Zeger, & Samet, 2002), en génétique (Foll & Gaggiotti, 2006) et également employées en géostatistiques (Diggle, Tawn, & Moyeed, 1998).

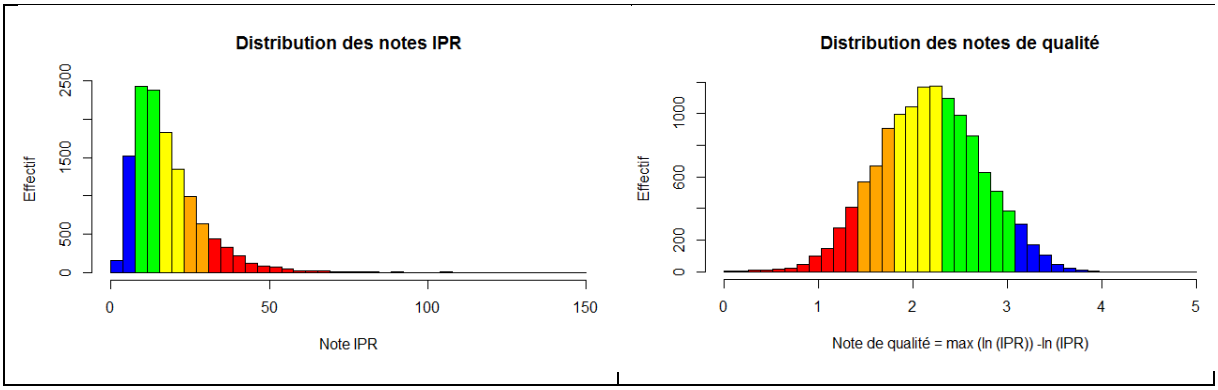
2.3.2 Classes de modèles envisageables

La note de qualité (*NDQ*), dérivée de l'IPR, tient le rôle de variable dépendante. Celle-ci est obtenue par transformation logarithmique de l'IPR afin de normaliser la distribution de la variable dépendante :

$$NDQ = max\ (log(IPR+1)) - log(IPR+1) \qquad \text{Équation 5}$$

La valeur de la NDQ est d'autant plus élevée que la qualité l'est, et sa distribution est approximativement gaussienne (Figure 3).

Figure 3 : transformation initiale de la variable IPR



Notes : à gauche, la distribution des notes IPR et à droite, celle des notes de qualité. Toutes les années sont confondues. Les couleurs représentent les classes de qualité, depuis le bleu pour la classe « Excellente », jusqu'au rouge pour la classe « Mauvaise ».

Les prédicteurs sont l'identifiant de la station, l'année et une ou plusieurs covariables :

$$NDQ \sim code.station + f\ (annee) + g_1\ (covariable_1) + g_2\ (covariable_2) + \ldots \qquad \text{Équation 6}$$

Différents types de modèles peuvent être envisagés. Le modèle le plus simple consisterait à régresser la variable dépendante sur une tendance linéaire, mais l'estimation de ce modèle conduirait vraisemblablement à des valeurs biaisées de la tendance. En effet, il y a tout lieu de penser que d'autres causes de variabilité temporelle des mesures existent et ce type de modèle les renvoie dans le terme d'erreur engendrant ainsi des phénomènes de pseudoreplication tels que décrits par Hurlbert (1984) ou d'endogénéité (Wooldridge, 2002). Des modélisations plus abouties sont donc requises.

On peut par exemple considérer un modèle linéaire généralisé (GLM - Generalized Linear Model) si l'on s'attend à des relations de proportionnalité entre la variable prédite et des prédicteurs ou un modèle additif généralisé (ou *GAM - Generalized Additive Model*) si la relation est non linéaire et indéterminée *a priori* (exprimée par une fonction *spline*). Le principe des *splines* est d'approcher au plus près des courbes complexes par une fonction polynomiale définie par intervalles. Cela est très utile pour décrire les relations entre deux variables sans faire d'hypothèse préalable sur la forme de cette relation (qui est néanmoins supposée continue).

Enfin, un modèle peut être à effets fixes (avec des coefficients estimés pour chaque modalité des variables nominales) ou fixes et aléatoires. Dans ce dernier cas, le modèle est dit mixte (GLMM, GAMM cf. Zuur, Ieno, Walker, Saveliev, & Smith, 2009).

2.4 Application : modélisation statistique des données IPR

2.4.1 Identification des sources de variation de l'IPR

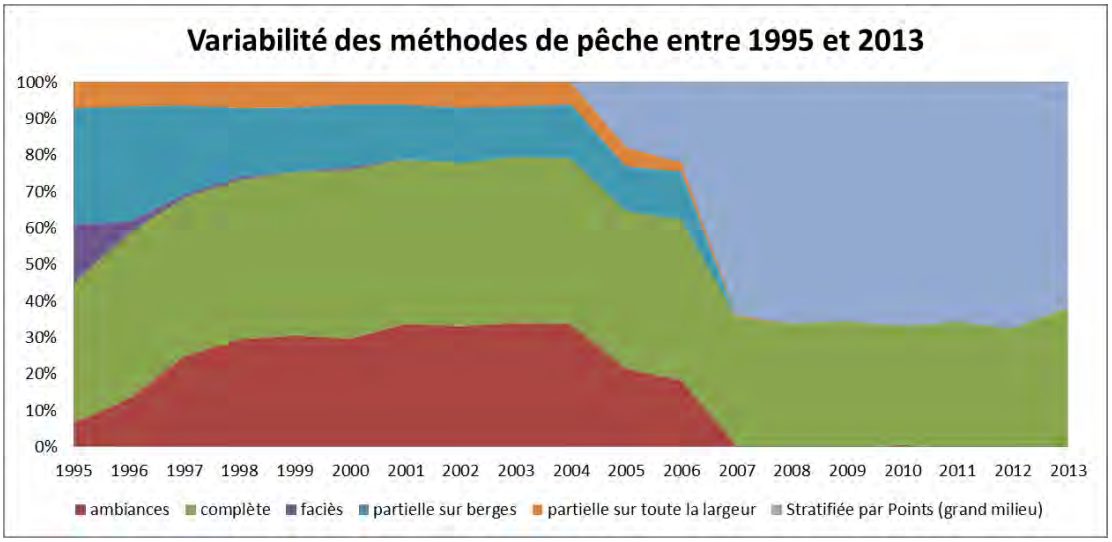
Les modèles statistiques sont nécessairement très dépendants du choix des variables explicatives. Il est donc essentiel de déterminer les phénomènes susceptibles de biaiser les observations (Stewart-Oaten, Murdoch, & Parker, 1986), de collecter les données permettant de les caractériser et éventuellement de les transformer en covariables pertinentes.

Pour la présente étude, l'IPR étant calculé à partir d'inventaires piscicoles, tout phénomène susceptible d'engendrer une variation dans les caractéristiques des communautés de poissons ou de modifier l'efficacité de la pêche peut être considéré comme un biais potentiel.

Les communautés de poissons connaissent de fortes variations saisonnières. La reproduction est en général annuelle. Des phénomènes de déplacements sur de longues distances concernent les migrateurs amphihalins (entre la mer et les eaux douces), mais aussi des espèces potamodromes, qui occupent des habitats différant selon les périodes de l'année, comme le Barbeau commun *Barbus barbus* ou le Chevesne *Squalius cephalus* (Benitez, Matondo, Dierckx, & Ovidio, 2015). Les communautés observées en un point donné du réseau hydrographique varient donc selon le cycle saisonnier.

Autre facteur susceptible de biaiser les observations, les méthodes de prospection sur le terrain ont varié au cours du temps et entre régions. Sur le jeu de données, pas moins de six de ces méthodes ont été déployées. Les pêches « par ambiance » et « partielle sur berge » ont été abandonnées en 2007, tandis que le protocole « stratifié par points » a été largement déployé sur les grands cours d'eau à partir de 2004 (Figure 4). De telles modifications des modes d'acquisition des données sont susceptibles d'induire des biais. À titre d'exemple, sur le Danube, il a été montré que la prospection continue des berges était moins efficace que la pêche par point pour capturer les juvéniles (Janáč& Jurajda, 2007).

Figure 4 : variabilité interannuelle de la part de chaque méthode de prospection par pêche électrique, sur le jeu de données brutes fourni par l'Onema



La localisation des sites échantillonnés est aussi potentiellement un facteur de biais. Cet effet est pris en compte par l'inclusion dans les modèles de l'identifiant du site (variable *code.station*). Cependant, cette inclusion d'un effet fixe revient à ajuster un paramètre par station, ce qui est consommateur de degrés de liberté. Aussi, il peut être utile de réduire cette consommation en introduisant une modélisation des proximités. Des sites proches tendent à partager des caractéristiques communes (conditions climatiques, chimie de l'eau, pressions liées à l'occupation du sol). Des mécanismes biologiques, migration, dispersion passive ou transmission de pathogènes, génèrent des dynamiques à l'échelle des bassins versants. À l'inverse, deux sites peuvent être proches en distance euclidienne, mais quasiment indépendants car n'échangeant pas d'individus. C'est le cas au niveau des lignes de partages des eaux. Par exemple, les sites échantillonnés en Lozère peuvent appartenir au bassin de la Loire, de la Garonne ou du Rhône. Bien que proches, ils fonctionnent en métapopulations indépendantes. Il apparaît donc également nécessaire de tenir compte du découpage des bassins versants.

2.4.2 Description des variables

L'affectation des stations dans les bassins versants (sous-unités DCE) a été réalisée à partir de la couche SIG disponible sur le site eaufrance (*annexe 1*). Le script *R* correspondant est fourni en *annexe 3*.

La liste des variables qui ont été obtenues et intégrées, à différents stades des analyses, est la suivante :

- l'année (*variable annee*). C'est la variable dont le lien avec l'IPR décrit la tendance. Elle a pu être utilisée en numérique ou en variable qualitative ;
- l'identifiant de la station (variable nominale *code.station*). Il est indispensable pour distinguer d'une part, les effets relatifs de la tendance en elle-même et d'autre part, ceux dus à la variation dans le réseau de stations ;
- les méthodes d'intervention sur le terrain. La stratégie de prospection, stratifiée ou non par habitat, avec ou sans filets de barrage du cours d'eau, peut influencer les résultats d'inventaires (Bohlin, Hamrin, Heggberget, Rasmussen, & Saltveit, 1989), donc potentiellement la note IPR. Ces méthodes ont été décrites par la variable *intit.method*, codée en 6 modalités (*Figure 4*). Il en est de même pour le moyen de prospection (pêche à pieds ou en bateau), qui a été caractérisé par la variable nominale *intit.moyen* à 3 modalités ;
- le nombre de jours écoulés, au jour de la pêche, depuis le 1er janvier précédent (variable *jour.annee*). Cette façon de coder la date permet de décrire la cyclicité saisonnière sans avoir à choisir arbitrairement le nombre des intervalles ni leurs bornes comme dans le cas d'un codage discret. Elle est communément employée pour caractériser la phénologie végétale (Myneni, Keeling, Tucker, Asrar, & Nemani, 1997) ou pour désaisonnaliser des tendances (Cleveland, Cleveland, McRae, & Terpenning, 1990) ;
- le bassin versant dans lequel se trouve la station échantillonnée (variable *bassin*). Certains phénomènes de dispersion entre populations (métapopulations, cf. Hanski, 1998) de poissons connectées par le réseau hydrographique tendent à générer un synchronisme ;
- la localisation géographique « pure » indiquée par les coordonnées (variables *xlambert93* et *ylambert93*).

2.4.3 Description des modèles

Les modèles ont été ajustés au moyen du package *mgcv* pour *R* (S. Wood, 2015). S'agissant de *GAM*s, plusieurs options de lissage sont envisageables. Des fonctions *splines* ont été employées. Il s'agit d'un lissage polynomial cubique défini par intervalles. Le lissage a été optimisé par *Generalized Cross Validation* (GCV, cf. Craven & Wahba, 1978 ; Golub, Heath, & Wahba, 1979). La fonction *bam*, plutôt que la fonction *gam* du même package, a été employée car elle est nettement plus rapide pour les gros jeux de données (S. N. Wood, Goude, & Shaw, 2015). Cette fonction accélère certaines étapes du calage du modèle (conditions initiales et procédures itératives) en ne les réalisant que sur un sous-échantillon du jeu de données complet (S. N. Wood et al., 2015). Une vérification a été réalisée, selon laquelle sur le modèle final retenu, les fonctions *bam* et *gam* donnaient des résultats équivalents.

L'objectif de notre étude est l'estimation d'une tendance nationale, à partir de mesures répétées sur des stations. Par ailleurs, les mesures pour une station donnée ne peuvent pas être considérées comme indépendantes les unes des autres. Les modèles mixtes sont particulièrement adaptés à ce type de situation (Zuur et al., 2009). Ils comprennent à la fois des prédicteurs à effet fixe (dont le nombre de modalités est fini et connu) et des prédicteurs à effet aléatoire (dont toutes les modalités ne sont pas forcément représentées dans le jeu de données). Les premiers sont communs à l'ensemble des observations alors que les seconds varient selon les groupes d'observations (les sites d'observations par exemple). Un des avantages majeurs des modèles mixtes est qu'ils spécifient explicitement les mesures répétées, avec une grande souplesse. Contrairement aux traitements classiques des séries chronologiques, les données manquantes ne sont pas rédhibitoires, et les pas de temps entre observations peuvent varier. Autre avantage, les effets n'étant pas estimés pour chacune des modalités, mais globalement pour le prédicteur, le nombre de degrés de liberté associé est nettement réduit dans le modèle mixte. Il en résulte que celui-ci requiert un moins grand nombre d'observations par variable explicative que le modèle à effets fixes correspondant, ou que des effets ténus ont plus de chances d'être mis en évidence par un modèle mixte que par un modèle à effets uniquement fixes qui est par construction plus paramétré. Enfin, inclure un effet aléatoire permet d'extrapoler les résultats à des unités statistiques (par exemple des sites de pêche) qui n'ont pas contribué à ajuster le modèle (Bolker et al., 2009).

Les modèles suivants, incluant tous l'identifiant des sites de pêche (variable *code.station*) en effet aléatoire gaussien, ont donc été ajustés. Ils sont présentés du plus simple au plus compliqué. Pour les modèles 1 à 5, l'année de la pêche est entrée comme variable numérique. Pour le modèle 6, elle est entrée comme variable qualitative.

Le modèle 1 est un GLM mixte simple. Un effet aléatoire de type gaussien est introduit sur le code station (dans le code *R*, spline avec option bs = « re ») comme si la station était issue d'un tirage aléatoire simple sans remise parmi toutes les stations potentielles.

Mod1 : NDQ~glm(logipr~annee+s(code.station,bs="re"))

Le modèle 2 est un modèle additif généralisé (GAM mixte) avec une fonction spline cubique de l'année.

Mod2 : NDQ~gam(logipr~s(annee)+s(code.station,bs="re"))

Le modèle 3 est toujours un GAM mixte mais avec deux effets aléatoires imbriqués de la station dans le bassin.

Mod3 : NDQ~gam(s(annee)+ s(code.station,bs="re")+ s(bassin,bs="re"))

Le modèle 4 introduit en effet fixe une spline cubique du jour de la mesure (numéro du jour dans l'année).

Mod4 : NDQ~gam(s(annee)+ s(code.station,bs="re")+ s(bassin,bs="re")+s(jour.annee))

Le modèle 5 ajoute une *spline* (bi)cubique des coordonnées Lambert de la station pour prendre en compte une éventuelle autocorrélation spatiale.

Mod5 : NDQ~gam(s(annee)+ s(code.station,bs="re")+s(bassin,bs="re")+s(jour.annee)+s(xlambert93,ylambert93))

Le modèle 6 est équivalent au modèle 5, sauf que l'année y est entrée en variable qualitative à 19 modalités.

Mod6 : NDQ~gam(as.factor(annee)+s(code.station,bs="re")+s(bassin,bs="re")+s(jour.annee)+s(xlambert93,ylambert93))

Les effets estimés sont alors annuels et s'interprètent comme des écarts à l'effet moyen de l'année de référence. Ils peuvent se traduire en base 100 en 1995.

3 Résultats et interprétation

3.1 Comparaison des modèles

Les différents modèles ont été comparés sur la base du critère d'information d'Akaïke (AIC, cf. Akaike, 1974). Cette statistique est utile pour effectuer des choix parmi différents modèles selon d'une part, leur maximum de vraisemblance et d'autre part, leur nombre de degrés de liberté (ddl). Plus l'*AIC* est faible, plus le modèle est performant en termes de compromis entre qualité de l'ajustement et parcimonie de la spécification. Le modèle 5 est, de ce point de vue, le meilleur (*Tableau 1*).

Tableau 1 : comparaison des modèles GAM sur la base du critère d'Akaïke

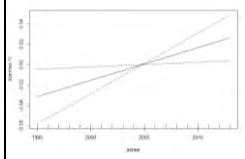
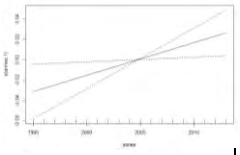
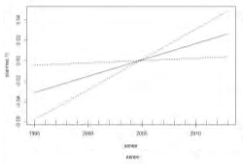
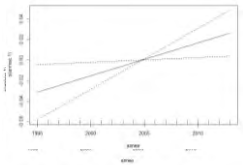
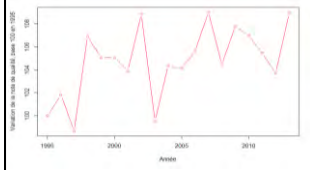
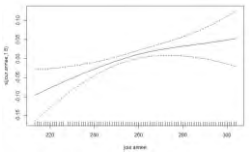
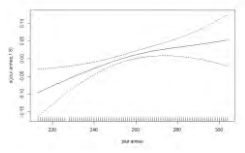
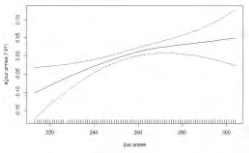
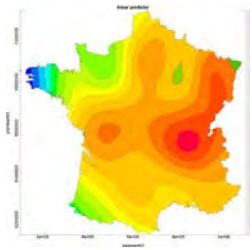
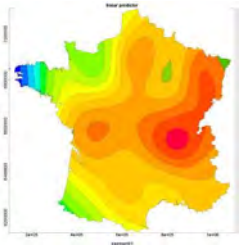
Modèle	mod2	mod3	mod4	mod5	mod6
Degrés de liberté (ddl)	305.5336	302.2440	305.0469	301.7353	300.5018
AIC	707.5026	704.9700	693.7906	691.4858	697.4358

Note : le modèle 5a est identique au modèle 5, à l'exception de la variable annee qui a été retirée.

Bien que l'AIC nous montre que l'ajustement du modèle 5 est le meilleur, il est intéressant de comparer les modèles successifs obtenus. En effet, l'objectif n'est pas ici de caler le modèle ayant la meilleure capacité prédictive, mais de tester une tendance temporelle non biaisée par des effets indésirables. Il est donc préférable que le modèle comprenne des prédicteurs dont la contribution est non significative, ou n'améliore pas significativement le modèle, sans toutefois altérer la qualité de l'estimation des paramètres, plutôt que d'exclure de tels prédicteurs au risque d'aboutir à une interprétation biaisée.

Les différents modèles qui ont été successivement ajustés sont synthétisés au *Tableau 2*.

Tableau 2 : résumé des étapes de modélisation aboutissant aux modèles finaux avec l'année en variable numérique (mod5) ou en variable nominale (mod6). Chaque colonne correspond à un modèle. Chaque ligne correspond à un prédicteur, à l'exception des deux dernières.

Modèle	Mod1 (GLM mixte)	Mod2 (GAM mixte)	Mod3	Mod4	Mod5	Mod6
annee	Estimation =0,0032 P=0,0189	 ddl=1 P=0,0195	 ddl=1 P=0,0204	 ddl=1 P=0,025	 ddl=1 P=0,033	
code.station (ddl de référence=320)	ddl=302,4 p<2e-16	ddl=302,4 p<2e-16	ddl=279,82 p=4,18e-08	ddl=280,2 p=8,65e-08	ddl=263,2 p=1,51e-11	ddl=263 p=5,64e-13
bassin (ddl de référence=31)			ddl=19,23 p=1.70e-07	ddl=18,9 p=1.73e-06	ddl=15,84 p=0,00132	ddl=15,8 p=0,0013
s(jour.annee)				 ddl=1,8 p=0,0047	 ddl=1,91 p=0,0042	 ddl=1,91 p=0,0052
s(latitude,longitude)					 ddl=16,7 p=3.36e-05	 ddl=16,7 p=3.36e-05
R^2 ajusté	0,775	0,775	0,775	0,777	0,777	0,777
Déviation expliquée	80,6%	80,6%	80,6%	80,7%	80,7%	80,3%

3.2 Présentation des résultats du modèle retenu (mod5)

Le coefficient de détermination (r^2 ajusté) est de 0,78 et la déviance expliquée de 80,7 %. L'ensemble des prédicteurs contribue significativement au modèle (*Tableau 3)*. L'effet année est donc significatif. Il ne prend qu'un degré de liberté car la fonction *spline* cubique ne permet pas un meilleur ajustement qu'une simple droite. La distribution des résidus semble approximativement normale. On ne note pas de dépendance entre résidus en valeurs prédites (*Figure 5*). Les effets aléatoires gaussiens de *code.station* et de *bassin* sont à peu près conformes à la normalité, malgré certaines discontinuités pour le *code.station*.

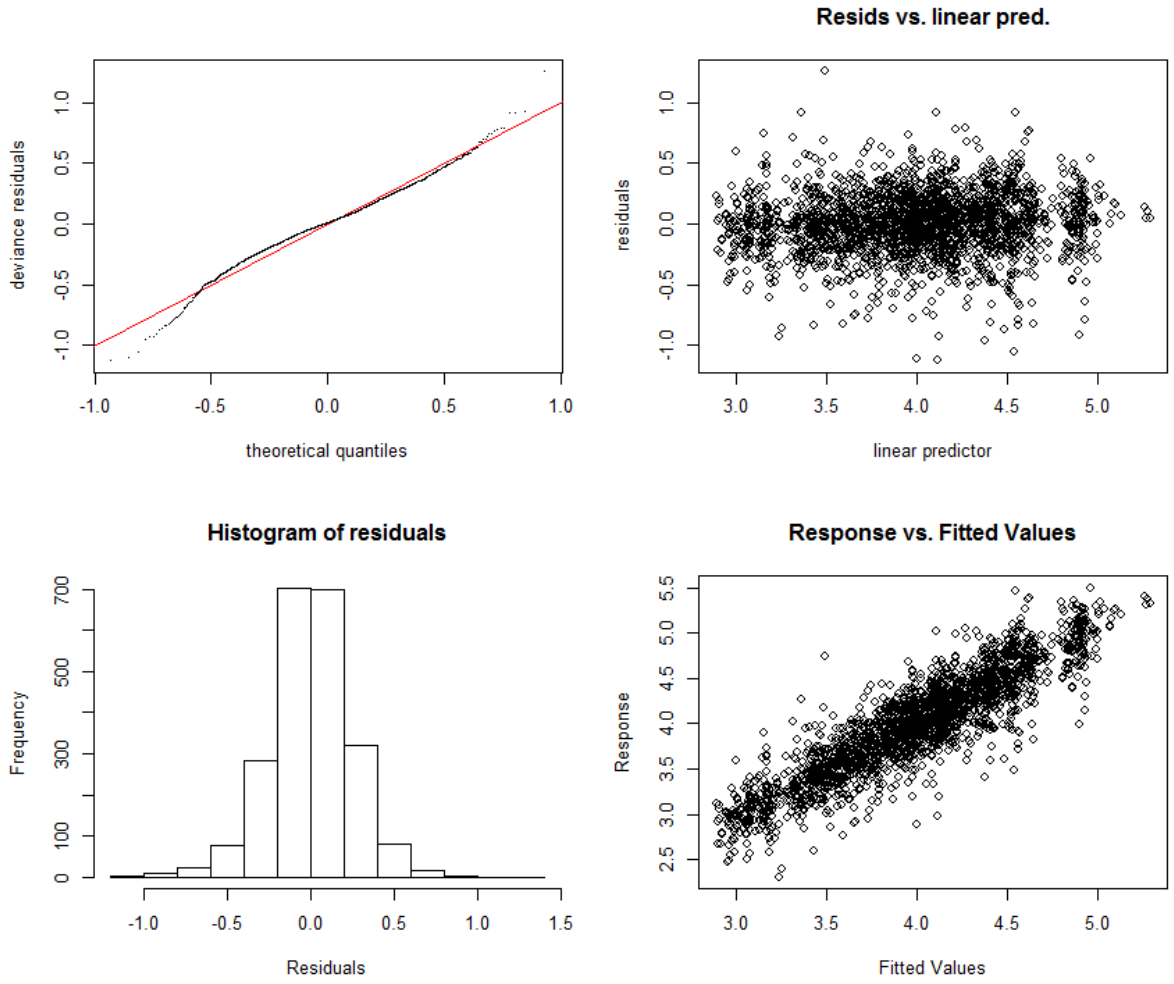
Tableau 3 : résultats du modèle 5 ; effets aléatoires des prédicteurs nominaux code.station et bassin, et effet des prédicteurs numériques

Variable	*ddl*	*Ref.ddl*	*F*	*p-value*	*Sig.*
s(annee)	*1,00*	*1,000*	*4,522*	*0,03359*	*
s(code.station)	*263,18*	*318,000*	*16,207*	*1,51e-11*	***
s(bassin)	*15,84*	*31,000*	*177,723*	*0,00132*	**
s(jour.annee)	*1,91*	*2,447*	*5,013*	*0,00421*	**
s(xlambert93,ylambert93)	*16,74*	*17,093*	*2,890*	*6,49e-05*	***

Note : une étoile indique une p-value comprise entre 0,01 et 0,05, deux étoiles entre 0,001 et 0,01 et trois étoiles inférieure à 0,001.

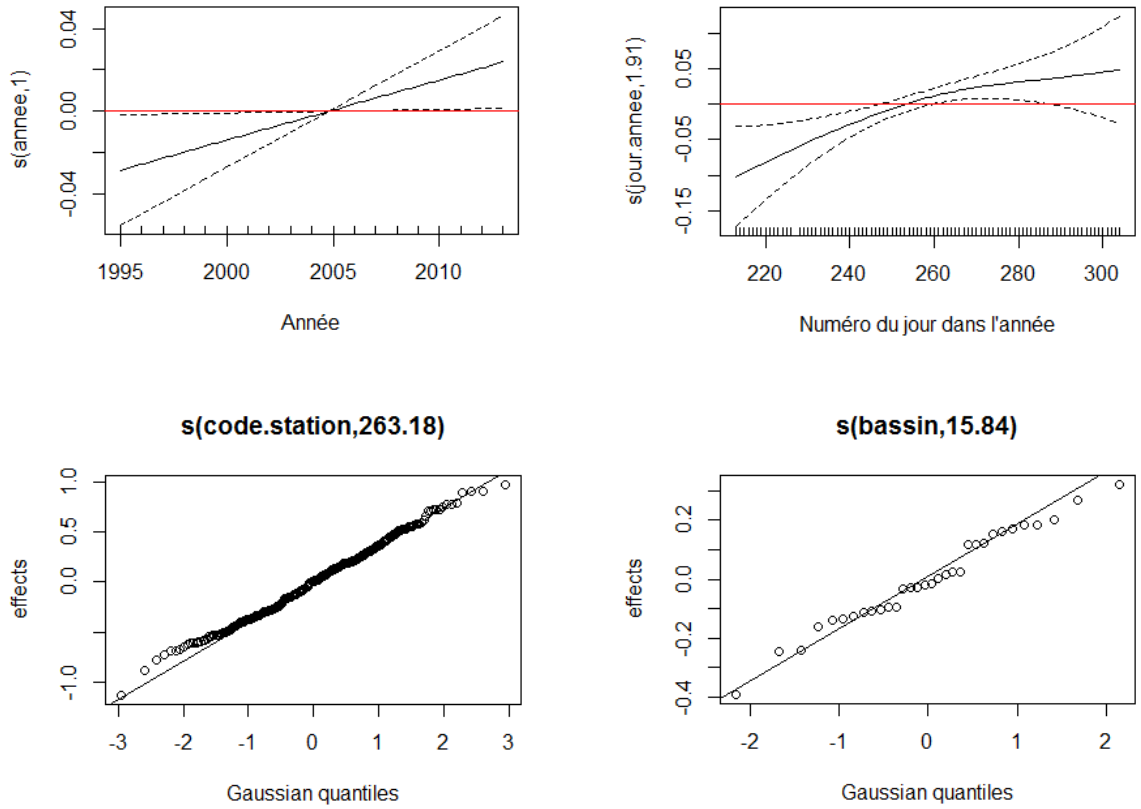
Pour s'assurer de la significativité de l'effet année, un modèle équivalent au modèle 5, mais sans la variable année a été construit. Son *AIC* est de 697,44, plus élevé que pour le modèle 5 (*Tableau 1)*. Une *ANOVA* comparant ces deux modèles est possible car ils sont « emboîtés », les variables explicatives du dernier modèle étant un sous-ensemble de celles du modèle 5. Cette *ANOVA* indique que le modèle 5 est significativement supérieur (*F*-test, *p*=00011).

Figure 5 : graphiques de diagnostic du modèle 5



L'essentiel de la variance expliquée l'est par l'identifiant de la station, ce qui montre que la variance inter-station est importante par rapport à la variance intra-station. C'est assez logique dans la mesure où le jeu de données comprend des stations couvrant une large gamme d'intensité de pressions anthropiques.

Figure 6 : effets des variables explicatives du modèle 5



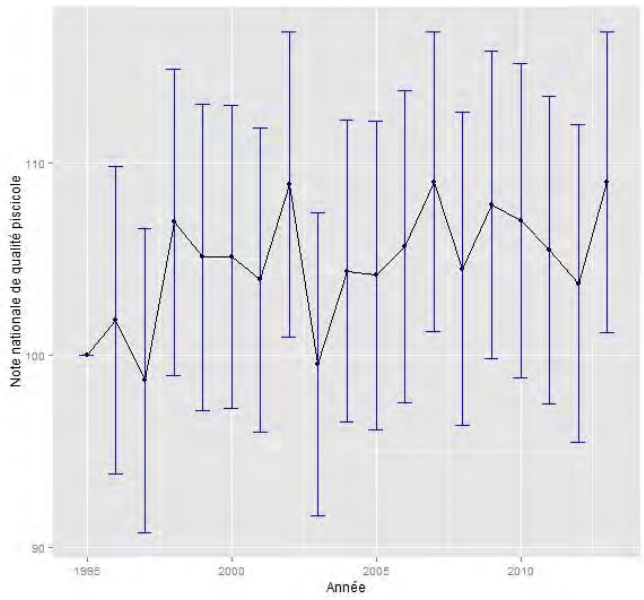
Notes : les effets de code.station et de bassin sont aléatoires. Les fonctions de transformation de annee et de jour.annee sont des splines cubiques.

L'effet du *bassin* est modéré. L'inclusion de cette variable explicative n'améliore pas les statistiques du modèle, mais il a été choisi de la conserver car elle correspond à la description de mécanismes potentiellement à l'œuvre à l'échelle des bassins. En comparant les modèles 2 et 3, il apparaît que la variance expliquée par la variable *bassin* vient en déduction de celle expliquée par *code.station*.

La fonction de lissage des coordonnées Lambert *s(xlambert93,ylambert93)* entre très significativement dans le modèle. Même si, comme le *bassin*, elle ne contribue pas à améliorer la variance expliquée par le modèle, elle permet de prendre en compte l'autocorrélation entre les observations, et donc d'éviter que les statistiques de diagnostic ne soient biaisées.

Le modèle 6 est quasiment identique au modèle 5, à l'exception de l'interprétation de l'effet de la variable *annee*. Celle-ci est entrée en prédicteur qualitatif, donc les valeurs annuelles ne sont pas lissées (Figure 10). On observe une stabilité de 1995 à 1997, suivie d'une forte amélioration en 1998 qui se maintient jusqu'en 2002. L'année 2003 est marquée par une forte dégradation. En 2004, les notes IPR sont de nouveau du même ordre de grandeur qu'avant la canicule de 2003. Elles le restent jusqu'à la fin de la série de données.

Figure 7 : variabilité interannuelle de l'indice de qualité (effet fixe annuel), obtenue par le modèle 6



Notes : la variable annee *est entrée dans le modèle comme variable catégorielle. Les intervalles de confiance représentent les indices annuels ± 1,96 erreur standard.*

3.3 Interprétation

3.3.1 Tendance générale (interprétation du modèle 5)

Comme précisé au paragraphe 2.2.2, seuls les petits cours d'eau entrent dans le champ d'analyse développé ici. La principale conclusion est que la qualité de ces cours d'eau, mesurée par l'IPR, tend significativement à s'améliorer sur la période 1995-2013. La forme de la relation entre l'année et la note de qualité ne diffère pas significativement d'une relation linéaire. La pente de la droite est malgré tout évaluée avec une incertitude conséquente (*Figure 6 Figure 5)*. L'effet année est cohérent au fil de l'inclusion des prédicteurs (*Tableau 2*). Il est notable que sur le jeu de données complet (avant restriction aux pêches complètes, donc aux petits cours d'eau), l'inclusion ou l'exclusion des variables caractérisant les méthodes de terrain modifie fortement la tendance indiquée sur la période par le modèle (comparaison entre le modèle 5 et celui présenté en *annexe 6*).

L'effet de la variable *jour.annee*, caractérisant la période à laquelle la station a été échantillonnée, est significatif malgré l'exclusion des pêches réalisées hors de la période de référence août-octobre. Cet effet est nettement plus marqué en amplitude quand l'année complète est considérée (*annexe 6*), mais il reste utile de conserver cette covariable.

3.3.2 Variabilité interannuelle (interprétation du modèle 6)

Seuls les effets fixes annuels des années 2002, 2007 et 2013 diffèrent significativement de la référence en 1995. En effet, les intervalles de confiance sont larges, ce qui ne permet pas de détecter avec certitude de faibles variations entre les années. Une vérification a été réalisée en estimant indépendamment les intervalles de confiance non paramétriques par *bootstrap* (n=1 000 tirages). Celle-ci a confirmé l'amplitude des IC déterminés à l'ajustement du modèle (*Figure 7*).

L'élément saillant du modèle 6 est la détection de ce qui semble être une anomalie en 2003 (*Figure 7*). Celle-ci peut être rapprochée de la canicule qui, entre le 1er et le 20 août 2003, a causé une surmortalité dans la population humaine estimée à 15 000 décès (Fouillet et al., 2006). Couplées à un épisode de sécheresse rare en Europe (García-Herrera, Díaz, Trigo, Luterbacher, & Fischer, 2010), ces températures (plus de 10 °C au-dessus des normales saisonnières) ont manifestement contribué à dégrader la qualité des habitats aquatiques. Après la chute de la qualité des cours d'eau mesurée par l'IPR en 2003, ces milieux ont, dès 2004, pratiquement repris leur qualité antérieure (*Figure 7*).

Les informations fournies par le modèle 6 sont donc complémentaires de celles fournies par le modèle 5, au niveau de l'interprétation de la trajectoire des milieux. L'inclusion de la variable *annee* en codage qualitatif permet de mettre en évidence des événements brutaux, comme l'impact de la canicule de 2003, qui tendent à être gommés par le lissage effectué dans le modèle 5.

4 Conclusion

Nous avons testé, dans la présente étude, des méthodes de modélisation statistique en les appliquant au traitement des données d'indice poisson rivière sur les petits cours d'eau. Cette démarche a permis de mettre en évidence les points suivants :

- **les questions auxquelles sont censées répondre les indicateurs évoluent** : elles changent au fil du temps. Il est donc inévitable de réévaluer périodiquement lesquelles de ces questions méritent un investissement, compte tenu de l'enjeu, et lesquelles sont de nature conjoncturelle, ou de moindre portée ;
- **le patrimoine des données mobilisées pour construire les indicateurs évolue lui aussi, qualitativement et quantitativement**. D'un côté, bon nombre de réseaux de suivi ont été allégés et rationalisés. De l'autre, les technologies d'acquisition actuelles (télédétection, collecte automatique, sciences participatives etc.) ouvrent des perspectives nouvelles, que ce soit pour traiter de nouveaux sujets, ou pour obtenir des variables venant expliquer nos observations ;
- **les moyens de traitement de l'information sont de plus en plus performants**. La capacité de calcul croît en continu. Les moyens de décentraliser ce calcul se démocratisent, autorisant le traitement en parallèle sur de nombreux processeurs. Les méthodes et outils disponibles pour manipuler et analyser d'importantes volumétries de données sont aussi en plein progrès.

Il est donc logique que l'adéquation questions/données/méthodes soit périodiquement interrogée. L'application de la modélisation statistique au cas particulier de l'IPR a mis en évidence une réelle plus-value. Pour être mise en œuvre de manière satisfaisante et efficace, sans jamais perdre de vue les objectifs, cette démarche nécessite d'associer étroitement spécialistes des traitements et spécialistes thématiques : pour identifier d'une part, les facteurs susceptibles de venir perturber le signal étudié (les biais), les sources de données mobilisables et les acteurs du domaine susceptibles d'être associés et d'autre part, tester et valider différentes approches et modèles statistiques.

Les méthodes de modélisation statistique se sont considérablement diversifiées dans les dernières décennies. Elles permettent désormais de traiter des données dont les distributions sont variées, et liées par des relations non linéaires. Les modèles mixtes permettent de faire face aux mesures répétées à pas de temps irrégulier avec des données manquantes. La famille des modèles statistiques permet donc d'expliquer la variabilité de descripteurs de l'environnement par le temps (*annee*) et des covariables visant à corriger certains biais. Cela s'applique ainsi à l'agrégation des données ponctuelles collectées par les réseaux de suivi environnemental en vue de dégager des tendances temporelles.

5 Références bibliographiques

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Benitez, J.-P., Matondo, B. N., Dierckx, A., & Ovidio, M. (2015). An overview of potamodromous fish upstream movements in medium-sized rivers, by means of fish passes monitoring. *Aquatic Ecology*, 1–17.

Bohlin, T., Hamrin, S., Heggberget, T. G., Rasmussen, G., & Saltveit, S. J. (1989). Electrofishing—theory and practice with special emphasis on salmonids. *Hydrobiologia, 173*(1), 9–43.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution, 24*(3), 127–135.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics, 6*(1), 3–73.

Collen, B., Loh, J., Whitmee, S., McRae, L., Amin, R., & Baillie, J. E. M. (2009). Monitoring change in vertebrate abundance: the Living Planet Index. *Conservation Biology, 23*(2), 317‑327.

Craven, P., & Wahba, G. (1978). Smoothing noisy data with spline functions. *Numerische Mathematik, 31*(4), 377–403.

Diggle, P. J., Tawn, J. A., & Moyeed, R. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 47*(3), 299–350.

Dominici, F., McDermott, A., Zeger, S. L., & Samet, J. M. (2002). On the use of generalized additive models in time-series studies of air pollution and health. *American journal of epidemiology, 156*(3), 193–203.

Foll, M., & Gaggiotti, O. (2006). Identifying the environmental factors that determine the genetic structure of populations. *Genetics, 174*(2), 875‑891.

Fouillet, A., Rey, G., Laurent, F., Pavillon, G., Bellec, S., Guihenneuc-Jouyaux, C., … Hémon, D. (2006). Excess mortality related to the August 2003 heat wave in France. *International Archives of Occupational and Environmental Health, 80*(1), 16‑24.

García-Herrera, R., Díaz, J., Trigo, R. M., Luterbacher, J., & Fischer, E. M. (2010). A review of the European summer heat wave of 2003. *Critical Reviews in Environmental Science and Technology, 40*(4), 267‑306.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics, 21*(2), 215–223.

Gregory, R. D., van Strien, A., Vorisek, P., Gmelig Meyling, A. W., Noble, D. G., Foppen, R. P. B., & Gibbons, D. W. (2005). Developing indicators for European birds. *Philosophical Transactions of the Royal Society B: Biological Sciences, 360*(1454), 269‑288. http://doi.org/10.1098/rstb.2004.1602

Hanski, I. (1998). Metapopulation dynamics. *Nature, 396*(6706), 41–49.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological monographs, 54*(2), 187–211.

Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of computational and graphical statistics, 5*(3), 299–314.

Janáč, M., & Jurajda, P. (2007). A comparison of point abundance and continuous sampling by electrofishing for age-0 fish in a channelized lowland river. *North American Journal of Fisheries Management, 27*(4), 1119–1125.

Lofland, C. L., & Ottesen, R. (2013). The SAS Versus R debate in industry and academia. *SAS Global Forum*, Paper 348 - 2013.

Myneni, R. B., Keeling, C. D., Tucker, C. J., Asrar, G., & Nemani, R. R. (1997). Increased plant growth in the northern high latitudes from 1981 to 1991. *Nature, 386*, 698‑702.

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.

Oberdorff, T., Pont, D., Hugueny, B., & Porcher, J.-P. (2002). Development and validation of a fish-based index for the assessment of « river health » in France. *Freshwater Biology, 47*(9), 1720‑1734.

Ohri, A. (2014). *R for cloud computing - An approach for data scientists*. Springer.

Onema. (2006). *L'indice poissons rivière (IPR) - Notice de présentation - édition avril 2006* (p. 24). Vincennes, France : Office national de l'eau et des milieux aquatiques.

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing (Version 3.1.1). Vienna, Austria. URL http://www.R-project.org/.

Ringuedé, S. (2014). *SAS: Introduction au décisionnel: du data management au reporting* (3e édition). Pearson Education France.

SAS Institute. (2013). *SAS® 9.4 Online documentation*. Cary, NC: SAS Institute Inc. Consulté à l'adresse https://support.sas.com/documentation/94/index.html

Schmidberger, M., Morgan, M., Eddelbuettel, D., Yu, H., Tierney, L., & Mansmann, U. (2009). State-of-the-art in parallel computing with R. *Journal of Statistical Software, 47*(1).

Stewart-Oaten, A., Murdoch, W. W., & Parker, K. R. (1986). Environmental impact assessment: « Pseudoreplication » in time? *Ecology, 67*(4), 929–940.

ter Braak, C. J. F., van Strien, A. J., Meijer, R., & Verstrael, T. J. (1994). Analysis of monitoring data with many missing values: which method? *Bird*, (1992), 663–673.

Wood, S. (2015). *Package 'mgcv' version 1.8-7*.

Wood, S. N., Goude, Y., & Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 64*(1), 139‑155.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge and London: MIT press.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (Éd.). (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer.

6 Aides à la lecture

6.1 Glossaire

Les définitions données ci-dessous sont appropriées dans le contexte particulier de la présente étude. Elles ne donnent pas le sens général des termes définis.

Amphihalin	Espèce dont une partie du cycle biologique s'effectue en mer et une autre partie en rivière.
Invertivore	Qui se nourrit d'invertébrés. Pour les poissons, il s'agit de maillons intermédiaires dans les chaînes alimentaires.
Lithophile	Qui est dépendant de substrats minéraux (sable, graviers, galets) pour sa reproduction.
Métapopulation	Ensemble de populations liées par la dispersion des individus.
Potamodrome	Qui effectue des migrations (obligatoires ou facultatives) en eau douce pour assurer ses fonctions de reproduction, nourrissage ou pour trouver abris.
Rhéophile	Qui affectionne les eaux vives.
Spline	Fonction de lissage polynomial définie par intervalles.
Tolérant	Se dit d'une espèce dont la présence et l'abondance sont peu influencées par la dégradation du milieu en conséquence des activités anthropiques.
Reporting	Désigne une famille d'outils d'analyse décisionnelle destinés à assurer la réalisation, la publication et la diffusion de tableaux de bord selon un format prédéterminé.

6.2 Abréviations et sigles

ANCOVA	Analyse de covariance
ANOVA	Analyse de variance
DEB	Direction de l'eau et de la biodiversité (ministère de l'Environnement, de l'Énergie et de la Mer)
ETL	Les processus ETL (Extraction-Transformation-Loading) servent à l'alimentation d'un système décisionnel : extraction de données des applications et des bases de production, transformation, chargement des données résultantes dans les différentes applications.
GCV	Generalized Cross Validation (validation croisée généralisée)
GLM	Generalized Linear Model (modèle linéaire généralisé)
GAM	Generalized Additive Model (modèle additif généralisé)
IC	Intervalle de confiance
IPR	Indice poisson rivière
Irstea	Institut national de recherche en sciences et technologies pour l'environnement et l'agriculture
MOOC	Massive Online Open-access Course (cours de masse gratuit en ligne)
ONB	Observatoire national de la biodiversité
Onema	Office national de l'eau et des milieux aquatiques
SEEE	Système d'évaluation de l'état des eaux
SQL	SQL (Structured Query Language) est un langage de définition de données, de manipulation de données et de contrôle de données pour les bases de données relationnelles.
SOeS	Service de l'observation et des statistiques (ministère de l'Environnement, de l'Énergie et de la Mer)

7 Annexes

Annexe n° 1
Prétraitement des données – étapes communes à l'ensemble des analyses

Cette section offre une description des étapes communes à la constitution de toutes tables analysées.

Import et assemblage des données

Import des fichiers de données (logiciel SAS)

Les fichiers de données IPR sont au format texte « *csv* ». Il en existe un par agence de l'eau (bassin) et par année :

- 6 agences de l'eau :
  - AEAG : agence de l'eau Adour-Garonne ;
  - AEAP : agence de l'eau Artois-Picardie ;
  - AELB : agence de l'eau Loire-Bretagne ;
  - AERM : agence de l'eau Rhin-Meuse ;
  - AERMC : agence de l'eau Rhône-Méditerranée Corse ;
  - AESN : agence de l'eau Seine-Normandie ;
- 19 années (1995 à 2013).

28 variables sont communes à l'ensemble des fichiers. D'autres apparaissent ponctuellement dans certains fichiers et ne sont pas retenues. Le tableau ci-dessous liste et décrit les 28 variables, précise si elles sont conservées ou supprimées. Si elles sont conservées, elles sont renommées.

Tableau 4 : variables conservées à partir du fichier de données brutes

Code de la variable dans le fichier de données	Description	Suppression/nouveau code de la variable	
		R	SAS
Code masse	<Vide>	Suppression	
Code site d'évaluation	Code SANDRE du point de prélèvement	code.sandre	CODE_SANDRE
Code point de contrôle	Code de la station de prélèvement	code.station	CODE_STATION
Date début	Date de début d'évaluation	date1	DATE_DEBUT
IPR - Fusion - 0.005, IPR : VB	Indice poissons rivière – valeur brute	ipr	IPR_VB

Les tables sont toutes importées, empilées et les variables sont renommées comme indiqué. Les codes station et SANDRE sont éventuellement préfixés par des « 0 » pour atteindre la longueur de 8 caractères.

À cette étape, la table de données nommée *donn* contient **11 183 observations**.

Ajout des coordonnées (logiciel)

Le fichier ESRI shapefile associant des coordonnées à chaque station est téléchargé sur :

http://services.sandre.eaufrance.fr/telechargement/geo/HYD/StationHydro/FXX/StationHydro_FXX-shp.zip

Il est transformé en fichier CSV et nommé *StationMesureEauxSurface.csv* avant d'être importé dans une table nommée *sites.*

Le codage correspond à un système de coordonnées **Lambert 93.**

Seules les variables d'intérêt sont conservées et renommées.

Tableau 5 : variables de localisation conservées à partir du fichier de données brutes

Code de la variable dans le fichier de données	Description	Suppression/nouveau code de la variable	
		R	SAS
CdStationM.C.254	Code SANDRE du point de prélèvement	code.sandre	CODE_SANDRE
CoordXStat.N.24.15	Abscisse des coordonnées Lambert 93	xlambert93	X_L93
CoordYStat.N.24.15	Ordonnée des coordonnées Lambert 93	ylambert93	T_L93
CdMasseDEa.C.254	Code de la masse d'eau	code.masse.eau	CODE_MASSE
CdCommune.C.254	Code de la commune	code.commune	CODE_COM

Un zéro est ajouté devant le code sandre lorsqu'il ne compte que 7 caractères. Cette variable est déclarée en classe *factor* (variable nominale).

Une jointure asymétrique est effectuée sur le code sandre entre la table de données (dataframe) *donn* et la table *sites,* de manière à conserver l'intégralité des observations contenues dans la table et pour ajouter les quatre nouvelles variables.

Ajout des indications sur la méthode et le moyen de pêche (logiciel SAS)

Le fichier *IPR_Pascal - Tableau final - Méthodes pêche.xlsx* fourni par l'Onema contient les méthodes et moyens de pêches associés à chaque mesure du fichier de données.

Le fichier est importé dans une table nommée *methode_moyen.* Seules les variables d'intérêt sont conservées et renommées.

Tableau 6 : variables utilisées pour caractériser le moyen et le mode de pêche

Code de la variable dans le fichier de données	Description	Suppression/nouveau code de la variable	
		R	SAS
CODE_SANDRE	Code SANDRE du point de prélèvement	code.sandre	CODE_SANDRE
CODE_STATION	Code de la station de prélèvement	code.station	CODE_STATION
DATE_DEBUT	Date de début d'évaluation	date1	DATE_DEBUT
methode_prospection	Libellé associé à la méthode de prospection	init.method	Methode_prospection
moyen_prospection	Libellé associé au moyen de prospection	init.moyen	Moyen_prospection

Des corrections sont apportées aux informations contenues dans la table :

- le moyen de prospection codé 852 (« non renseigné ») ne fournit aucune information, il est modifié : le code et le libellé sont remplacés par des valeurs manquantes ;
- des prélèvements sont en doublon (avec 2 méthodes et/ou moyens de pêche différents), ils sont gérés à la main, les modalités les moins pertinentes étant supprimées au profit des plus pertinentes (la pertinence étant jugée à partir des méthodes et moyens des prélèvements antérieurs et postérieurs au prélèvement qui pose problème) :
    - la station « 01800020 » est évaluée avec la méthode « ambiances » et le moyen « mixte » une fois par an de 1995 à 2005. Se pose le problème de l'année 1997 pour laquelle il existe un deuxième couple méthode-moyen « complète » – « à pied ». La ligne associée à cette deuxième modalité est supprimée ;
    - pour la station « 03500004 » et le prélèvement du 1er octobre 1998, la ligne associée à la méthode « partielle sur berges » est supprimée ;
    - pour la station « 03610179 » et les prélèvements du 29 septembre 1998 et du 12 octobre 1999, les lignes associées à la méthode « partielle sur berges » sont supprimées ;
    - pour la station « 04350052» et les prélèvements du 3 octobre 1995, du 19 juin 1996 et du 24 septembre 1996, les lignes associées à la méthode « EPA » sont supprimées ;
    - pour la station « 04350052 » et le prélèvement du 27 septembre 2000, la ligne associée à la méthode « faciès » est supprimée ;
    - pour la station « 04560087 » et les prélèvements du 16 juin 1995 et du 14 juin 1996, les lignes associées à la méthode « faciès » et au moyen « à pied » sont supprimées ;
    - pour la station « 04560087 » et le prélèvement du 23 septembre 1996, la ligne associée à la méthode « complète » et au moyen « à pied » est supprimée ;
    - pour la station « 04580004 » et le prélèvement du 6 juin 2001, la ligne associée à la méthode « partielle sur berges » est supprimée ;

- pour la station « 04636705 » et le prélèvement du 25 août 2008, la ligne associée à la méthode « EPA » est supprimée ;
- pour la station « 06040082 » et les prélèvements du 4 juillet 2006, du 17 juillet 2008, du 26 mai 2009, du 13 octobre 2010, du 24 mai 2011 et du 20 juin 2012, les lignes associées à la méthode « autres » sont supprimées ;
- pour la station « 06050061 » et le prélèvement du 31 mai 2007, la ligne associée à la méthode « stratifiée par points (grand milieu) » est supprimée ;
- pour la station « 06110041 » et le prélèvement du 19 juin 1995, la ligne associée à la méthode « ambiances » est supprimée ;
- pour la station « 06380236 » et le prélèvement du 17 juin 2009, la ligne associée à la méthode « complète » est supprimée ;
- pour la station « 04420107 » et le prélèvement du 28 septembre 1995, les 2 lignes sont transformées en une seule en conservant la méthode « ambiances » commune aux 2 lignes et en regroupant les moyens « à pied » et « en bateau » en « mixte » (modalité utilisée les autres années) ;
- pour la station « 04710110 » et les prélèvements du 12 août 1998 et du 17 août 1999, les lignes associées au moyen « à pied » sont supprimées ;

- des prélèvements ne sont pas associés à une méthode ou un moyen de pêche (893 couples moyen-méthode manquants sur les 11 183 prélèvements du fichier de données dont les 794 prélèvements effectués en 2013 ; 10 moyens supplémentaires manquants à la suite de la transformation de la modalité « non renseigné »). Selon l'Onema, on peut considérer qu'au-delà de 2008, il n'y a pas eu de changement. Sur cette hypothèse, les méthodes et moyens de pêches sont complétés en reprenant la dernière information connue.
  Après cette étape, il reste 116 observations avec méthode et moyen de pêche manquants, dont :
  - 108 sont les observations effectuées en Corse (qui ne seront pas retenues). Note : ces stations n'ont pas de valeurs renseignées ;
  - 4 concernent 4 stations qui apparaissent en 2013 (qui ne seront pas retenues car une seule observation). Note : une seule de ces stations a des valeurs renseignées ;
  - 1 concerne 1 station qui n'existe qu'en 2000 (qui ne sera pas retenue car une seule observation).
  - au final, il ne reste que la station 01590033 (code SANDRE 01008000) qui présente des mesures en 2008, 2010 et 2012 mais n'a jamais d'indication sur la méthode ou le moyen de pêche. Cette station sera supprimée.

Les quatre nouvelles variables sont ajoutées dans la table de données par fusion des deux tables sur le code station et sur la date de début.

Ajout des indications sur le bassin (logiciel)

Le fichier des limites des sous-unités DCE est téléchargé sur le site eaufrance. Le format est ESRI Arcgis, le système de coordonnées WGS84. Le fichier est téléchargé et nommé *SsBassinDCEAdmin*. Ses coordonnées sont converties en Lambert 93 (epsg : 2154). Les DOM sont supprimés.

Ensuite, la table de données *donn* est spatialisée en Lambert 93 (en utilisant ses variables *xlambert93* et *ylambert93*). Elle était de classe *DataFrame* (table de données contenant des variables de natures variées) et devient de classe *SpatialPointsDataFrame* (table attributaire d'objets spatiaux ponctuels).

Une requête spatiale est exécutée pour affecter chaque station dans un des polygones représentant les sous-unités DCE.

Au moment de la vérification, quatre stations frontalières ne sont pas affectées. Ces stations sont gérées par une affectation manuelle.

Tableau 7 : affectation "manuelle" des stations frontalières

Code SANDRE du point de prélèvement	Code de la sous-unité DCE
01059000	FRA_ESCA
01590127	FRA_ESCA
05311011	FRF_GARO
06660103	FRD_COLR

Un code bassin est alors associé à chaque station selon son code SANDRE. Ces informations sont ajoutées dans la table de données par fusion asymétrique des deux tables sur le code SANDRE.

Le code de la sous-unité DCE est nommé *bassin* quel que soit le logiciel utilisé.

Un fond de carte de France est créé à partir de la cartothèque du package « *maps* » pour *R*. Il est transformé en Lambert 93. Il sera par la suite utilisé pour visualiser les effets spatiaux sur l'IPR et les résidus.

Transformation et filtrage des données (Logiciels *SAS* et R en parallèle)

Création de nouvelles variables

De nouvelles variables sont créées pour finaliser la table de données complète.

Tableau 8 : variables créées à partir des données brutes en vue des traitements

Description	Suppression/nouveau code de la variable Description	
	R	SAS
Note de qualité	logipr	NOTE_QUALI
Année du prélèvement	annee	ANNEE
Mois du prélèvement	mois	MOIS
Jour dans l'année du prélèvement	jour.annee	JOUR

La note de l'indice IPR est une note qui fonctionne de la façon suivante : plus elle est élevée, plus faible est la qualité.

L'objectif est de créer une nouvelle variable qui donne une information sur la qualité et qui fonctionnerait de la façon suivante : plus elle est élevée, plus élevée est la qualité.

De plus, les valeurs extrêmes perturbent l'analyse. Un passage au logarithme est utilisé de manière à écraser les valeurs extrêmes et à rendre la distribution de la note gaussienne.

La note de qualité est créée telle que Note de qualité = max(ln(IPR_VB +1))-ln(IPR_VB +1).

À cette étape, la table de données contient toujours **11 183 observations**.

Suppression des valeurs manquantes

Les observations avec un indice poisson non calculé (valeur manquante) sont supprimées.

À cette étape, la table de données ne contient désormais plus que **10 550 observations**.

Filtre sur les stations de Corse et la période de novembre à juillet (étape non réalisée pour l'annexe n° 1)

L'indice IPR est censé n'être valable que sur la France métropolitaine hors Corse et sur les mois d'août, septembre et octobre, les prélèvements effectués en Corse (code station commençant par « 062A » ou « 062B ») ou ne répondant pas à la seconde contrainte sont donc supprimés[1].

La suppression des stations de Corse n'a, en réalité, aucun effet puisque les prélèvements effectués en Corse sont tous associés à des valeurs manquantes dans les données et ont donc été supprimés à l'étape précédente.

À cette étape, la table de données contient **6 409 observations**.

Suppression des doublons stricts

S'il y a des doublons stricts, ils sont supprimés.

Aucun doublon strict n'étant repéré, la table de données contient toujours 6 409 observations.

Suppression des valeurs extrêmes au niveau global

La moyenne et l'écart-type de l'ensemble des 6 409 notes de qualité sont calculés.

Les observations pour lesquelles la note de qualité n'est pas comprise dans l'intervalle de ± 3 fois l'écart-type autour de la moyenne sont supprimées.

À cette étape, la table de données contient **6 403 observations**.

Suppression des valeurs extrêmes au niveau local

La moyenne et l'écart-type de l'ensemble des notes de qualité sont calculés pour chaque station.

Pour une station donnée, les observations pour lesquelles la note de qualité n'est pas comprise dans l'intervalle de ± 3 fois l'écart-type autour de la moyenne sont supprimées.

À cette étape, la table de données contient **6 150 observations**.

Suppression des doublons en réalisant des moyennes par station et date

Les stations présentant plusieurs prélèvements (différents) le même jour sont traitées : la note de qualité moyenne est utilisée.

---

[1] Oberdorff, T., Pont, D., Hugueny, B., Chessel, D., 2001. A probabilistic model characterizing fish assemblages of French rivers: a framework for environmental assessment. Freshw. Biol. 46, 399–415. doi:10.1046/j.1365-2427.2001.00669.x

Oberdorff, T., Pont, D., Hugueny, B., Porcher, J.-P., 2002. Development and validation of a fish-based index for the assessment of "river health" in France. Freshw. Biol. 47, 1720–1734. doi:10.1046/j.1365-2427.2002.00884.x

Cette étape n'a dans les faits pas d'effet, il n'y a pas ce genre de cas dans la table de données.

À cette étape, la table de données contient toujours **6 150 observations** et est prête à être utilisée.

Annexe n° 2
Scripts *SAS* de prétraitement des données

```
/*************************************************************************************************/

/* Indice Poisson Rivière                                                                        */
/*-----------------------------------------------------------------------------------------------*/
/* Prétraitement des données                                                                     */
/*************************************************************************************************/

%LET fichdep=T:\B_travaux\Biodiv\201505 - Indice IPR\Fichdep; /* Répertoire d'entrée */
%LET fichfin=T:\B_travaux\Biodiv\201505 - Indice IPR\Fichfin; /* Répertoire de sortie */
LIBNAME fichfin "T:\B_travaux\Biodiv\201505 - Indice IPR\Fichfin"; /* Librairie de sortie */

/*-----------------------------------------------------------------------------------------------*/
/* Import des fichiers de données                                                                */
/*-----------------------------------------------------------------------------------------------*/

%Macro import;

        /* Liste des dossiers présents dans le répertoire SEEE_IPR_1995-2013 */

        Data liste_doss (DROP=rc did nbfiles i);
        LENGTH dossier $500.;
        rc=filename("mydir","&fichdep.\SEEE_IPR_1995-2013");
        did=dopen("mydir");
        IF did=<0 THEN STOP;
        nbfiles=dnum(did);
        IF nbfiles ne 0 THEN DO i=1 TO nbfiles;
                dossier=dread(did,i);
                OUTPUT;
        END;
        did=dclose(did);
        Run;

        /* Création de macro-variables : nbdoss = nombre de dossiers ; doss1, ..., doss(i) = nom des i dossiers
*/

        Data _NULL_;
        SET liste_doss;
        CALL SYMPUT("nbdoss",_N_);
        CALL SYMPUT(compress("doss"!!_N_),strip(dossier));
        Run;

        Proc datasets NOLIST; DELETE liste_doss; Quit; /* Suppression de la table liste_doss */

        %DO i=1 %TO &nbdoss.; /* Boucle sur les i dossiers */

                /* Liste des sous-dossiers présents dans le dossier i */

                Data liste_ssdoss (DROP=rc did nbfiles i);
                LENGTH sous_dossier $500.;
                rc=filename("mydir","&fichdep.\SEEE_IPR_1995-2013\&&doss&i..");
                did=dopen("mydir");
                IF did=<0 THEN STOP;
                nbfiles=dnum(did);
                IF nbfiles ne 0 THEN DO i=1 TO nbfiles;
                        sous_dossier=dread(did,i);
                        OUTPUT;
                END;
                did=dclose(did);
                Run;

                /* Création de macro-variables : nbssdoss = nombre de sous-dossiers dans le dossier i ;
ssdoss1, ..., ssdoss(j) = nom des j sous-dossiers */

                Data _NULL_;
                SET liste_ssdoss (WHERE=(sous_dossier ne "sites.txt"));
                CALL SYMPUT("nbssdoss",_N_);
                CALL SYMPUT(compress("ssdoss"!!_N_),strip(sous_dossier));
                Run;

                Proc datasets NOLIST; DELETE liste_ssdoss; Quit; /* Suppression de la table liste_ssdoss */

                %DO j=1 %TO &nbssdoss.; /* Boucle sur les j sous-dossiers */

                        /* Liste des fichiers présents dans le sous-dossier j */

                        Data liste_fich (DROP=rc did nbfiles i);
                        LENGTH fichier $500.;
                        rc=filename("mydir","&fichdep.\SEEE_IPR_1995-2013\&&doss&i..\&&ssdoss&j..");
                        did=dopen("mydir");
                        IF did=<0 THEN STOP;
```

```
                        nbfiles=dnum(did);
                        IF nbfiles ne 0 THEN DO i=1 TO nbfiles;
                                fichier=dread(did,i);
                                OUTPUT;
                        END;
                        did=dclose(did);
                        Run;

                        /* Création de macro-variables : nbfich = nombre de fichiers dans le sous-dossier j ;
fich1, ..., fich(k) = nom des k fichiers */

                        Data _NULL_;
                        SET liste_fich;
                        CALL SYMPUT("nbfich",_N_);
                        CALL SYMPUT(compress("fich"!!_N_),strip(fichier));
                        Run;

                        Proc datasets NOLIST; DELETE liste_fich; Quit; /* Suppression de la table liste_fich
*/

                        %DO k=1 %TO &nbfich.; /* Boucle sur les k fichiers */

                                /* Import du kième fichier */

                                Proc                import                DATAFILE="&fichdep.\SEEE_IPR_1995-
2013\&&doss&i..\&&ssdoss&j..\&&fich&k.."  OUT=data&i.&j.&k.  DBMS=DLM  REPLACE;  DELIMITER=";";  GUESSINGROWS=1000;
Run;

                                /* Mise en place de nouveaux noms, codes stations et SANDRE préfixés par des
"0" */

                                Data data&i.&j.&k.  (KEEP=CODE_STATION CODE_SANDRE DATE_DEBUT DATE_FIN IPR_D:
IPR_N: IPR_VB IPR_CE);
                                ATTRIB CODE_STATION CODE_SANDRE DATE_DEBUT DATE_FIN LABEL="";
                                SET data&i.&j.&k.;
                                FORMAT CODE_STATION CODE_SANDRE $8. DATE_DEBUT DATE_FIN date9. IPR_D: IPR_N:
IPR_VB IPR_CE best12.;
                                IF substr(VAR3,1,1) ne "0" THEN CODE_STATION=compress("0"!!VAR3);
                                ELSE CODE_STATION=compress(VAR3);
                                CODE_SANDRE=compress("0"!!Code_Site_d_Evaluation);
                                DATE_DEBUT=VAR4;
                                IPR_DII_VB=IPR___Fusion___0_005__DII___VB*1;
                                IPR_DIIa_VB=IPR___Fusion___0_005__DII_attend*1;
                                IPR_DIIo_VB=IPR___Fusion___0_005__DII_observ*1;
                                IPR_DIO_VB=IPR___Fusion___0_005__DIO___VB*1;
                                IPR_DIOa_VB=IPR___Fusion___0_005__DIO_attend*1;
                                IPR_DIOo_VB=IPR___Fusion___0_005__DIO_observ*1;
                                IPR_DIT_VB=IPR___Fusion___0_005__DIT___VB*1;
                                IPR_DITa_VB=IPR___Fusion___0_005__DIT_attend*1;
                                IPR_DITo_VB=IPR___Fusion___0_005__DIT_observ*1;
                                IPR_DTI_VB=IPR___Fusion___0_005__DTI___VB*1;
                                IPR_DTIa_VB=IPR___Fusion___0_005__DTI_attend*1;
                                IPR_DTIo_VB=IPR___Fusion___0_005__DTI_observ*1;
                                IPR_NEL_VB=IPR___Fusion___0_005__NEL___VB*1;
                                IPR_NELa_VB=IPR___Fusion___0_005__NEL_attend*1;
                                IPR_NELo_VB=IPR___Fusion___0_005__NEL_observ*1;
                                IPR_NER_VB=IPR___Fusion___0_005__NER___VB*1;
                                IPR_NERa_VB=IPR___Fusion___0_005__NER_attend*1;
                                IPR_NERo_VB=IPR___Fusion___0_005__NER_observ*1;
                                IPR_NTE_VB=IPR___Fusion___0_005__NTE___VB*1;
                                IPR_NTEa_VB=IPR___Fusion___0_005__NTE_attend*1;
                                IPR_NTEo_VB=IPR___Fusion___0_005__NTE_observ*1;
                                IPR_VB=IPR___Fusion___0_005__IPR___VB*1;
                                IPR_CE=IPR___Fusion___0_005__IPR___CE*1;
                                Run;

                        %END;

                %END;

        %END;

        /* Regroupement de toutes les tables en une seule */

        Data donn;
        SET %DO i=1 %TO &nbdoss.; %DO j=1 %TO &nbssdoss.; %DO k=1 %TO &nbfich.; data&i.&j.&k. %END; %END; %END;;
        Run;

        /* Suppression des tables intermédiaires */

        Proc datasets NOLIST; DELETE %DO i=1 %TO &nbdoss.; %DO j=1 %TO &nbssdoss.; %DO k=1 %TO &nbfich.;
data&i.&j.&k. %END; %END; %END;; Quit;

%Mend;

%import;
```

```
/*-------------------------------------------------------------------------------------------------------*/
/* Ajout des coordonnées                                                                                 */
/* (à partir du fichier StationMesureEauxSurface.xlsx créé depuis R)                                     */
/*-------------------------------------------------------------------------------------------------------*/

/* Import */

Proc import OUT=coordonneesXY DATAFILE="U:\pirz\Indicateurs\Indicateurs SOeS\Indicateurs et
Indices\081_Qualite_piscicole_cours_deau_IPR\Traitements\Donnees\StationMesureEauxSurfaceFXX-
shp\StationMesureEauxSurface.xlsx" DBMS=EXCEL REPLACE; Run;

/* Fusion des 2 tables */

Proc sort DATA=donn; BY CODE_SANDRE; Run;
Proc sort DATA=coordonneesXY (KEEP=CdStationM_C_254 CoordXStat_N_24_15 CoordYStat_N_24_15) OUT=coordonneesXY
(RENAME=(CdStationM_C_254=CODE_SANDRE CoordXStat_N_24_15=X CoordYStat_N_24_15=Y)) NODUPKEY; BY CdStationM_C_254
CoordXStat_N_24_15 CoordYStat_N_24_15; Run;

Data donn;
MERGE donn (IN=a) coordonneesXY (IN=b);
BY CODE_SANDRE;
IF a;
Run;
Proc sort DATA=donn; BY CODE_STATION DATE_DEBUT; Run;

/* Suppression des tables intermédiaires */

Proc datasets NOLIST; DELETE coordonneesxy; Quit;

/*-------------------------------------------------------------------------------------------------------*/
/* Ajout des indications sur la méthode et le moyen de pêche                                             */
/* (à partir du fichier IPR_Pascal - Tableau final - Méthodes pêche.xlsx fourni par l'Onema)             */
/*-------------------------------------------------------------------------------------------------------*/

/* Import */

Proc import DATAFILE="T:\B_travaux\Biodiv\201505 - Indice IPR\Fichdep\IPR_Pascal - Tableau final - Méthodes
pêche.xlsx" OUT=methode_moyen (KEEP=CODE_STATION DATE_DEBUT op_cd_methodeprospection Methode_prospection
op_cd_moyenprospection Moyen_prospection) DBMS=EXCEL REPLACE; Run;

Data methode_moyen;
SET methode_moyen;
IF op_cd_moyenprospection=852 THEN DO;
        op_cd_moyenprospection=.;
        Moyen_prospection="";
END;
Run;
Proc sort DATA=methode_moyen NODUPKEY; BY CODE_STATION DATE_DEBUT op_cd_methodeprospection Methode_prospection
op_cd_moyenprospection Moyen_prospection; Run;

* Méthode de pêche non manquante - Doublons ;

Data nonmanq_methode (KEEP=CODE_STATION DATE_DEBUT op_cd_methodeprospection Methode_prospection);
SET methode_moyen (WHERE=(op_cd_methodeprospection ne .));
Run;
Proc sort DATA=nonmanq_methode NODUPKEY; BY CODE_STATION DATE_DEBUT op_cd_methodeprospection Methode_prospection;
Run;
Proc sort DATA=nonmanq_methode; BY CODE_STATION DATE_DEBUT; Run;

Data test_nonmanq_methode; * Doublons ;
SET nonmanq_methode;
BY CODE_STATION DATE_DEBUT;
IF first.DATE_DEBUT=0 or last.DATE_DEBUT=0;
Run;

Proc datasets NOLIST; DELETE test_nonmanq_methode; Quit; /* Suppression de la table test_nonmanq_methode */

* Moyen de pêche non manquant - Doublons ;

Data nonmanq_moyen (KEEP=CODE_STATION DATE_DEBUT op_cd_moyenprospection Moyen_prospection);
SET methode_moyen (WHERE=(op_cd_moyenprospection ne .));
Run;
Proc sort DATA=nonmanq_moyen NODUPKEY; BY CODE_STATION DATE_DEBUT op_cd_moyenprospection Moyen_prospection; Run;
Proc sort DATA=nonmanq_moyen; BY CODE_STATION DATE_DEBUT; Run;

Data test_nonmanq_moyen; * Doublons ;
SET nonmanq_moyen;
BY CODE_STATION DATE_DEBUT;
IF first.DATE_DEBUT=0 or last.DATE_DEBUT=0;
Run;

Proc datasets NOLIST; DELETE test_nonmanq_moyen; Quit; /* Suppression de la table test_nonmanq_moyen */

* Rectificatifs doublons ;

Data methode_moyen;
```

```
SET methode_moyen;
IF CODE_STATION="01800020" and DATE_DEBUT="01JUL1997"d and Methode_prospection="complète" THEN DELETE;
IF CODE_STATION="03500004" and DATE_DEBUT="01OCT1998"d and Methode_prospection="partielle sur berges" THEN
DELETE;
IF CODE_STATION="03610179" and DATE_DEBUT="29SEP1998"d and Methode_prospection="partielle sur berges" THEN
DELETE;
IF CODE_STATION="03610179" and DATE_DEBUT="12OCT1999"d and Methode_prospection="partielle sur berges" THEN
DELETE;
IF CODE_STATION="04350052" and DATE_DEBUT="03OCT1995"d and Methode_prospection="EPA" THEN DELETE;
IF CODE_STATION="04350052" and DATE_DEBUT="19JUN1996"d and Methode_prospection="EPA" THEN DELETE;
IF CODE_STATION="04350052" and DATE_DEBUT="24SEP1996"d and Methode_prospection="EPA" THEN DELETE;
IF CODE_STATION="04350052" and DATE_DEBUT="27SEP2000"d and Methode_prospection="faciès" THEN DELETE;
IF CODE_STATION="04560087" and DATE_DEBUT="16JUN1995"d and Methode_prospection="faciès" THEN DELETE;
IF CODE_STATION="04560087" and DATE_DEBUT="14JUN1996"d and Methode_prospection="faciès" THEN DELETE;
IF CODE_STATION="04560087" and DATE_DEBUT="23SEP1996"d and Methode_prospection="complète" THEN DELETE;
IF CODE_STATION="04580004" and DATE_DEBUT="06JUN2001"d and Methode_prospection="partielle sur berges" THEN
DELETE;
IF CODE_STATION="04636705" and DATE_DEBUT="25AUG2008"d and Methode_prospection="EPA" THEN DELETE;
IF CODE_STATION="06040082" and DATE_DEBUT="04JUL2006"d and Methode_prospection="autres" THEN DELETE;
IF CODE_STATION="06040082" and DATE_DEBUT="17JUL2008"d and Methode_prospection="autres" THEN DELETE;
IF CODE_STATION="06040082" and DATE_DEBUT="26MAY2009"d and Methode_prospection="autres" THEN DELETE;
IF CODE_STATION="06040082" and DATE_DEBUT="13OCT2010"d and Methode_prospection="autres" THEN DELETE;
IF CODE_STATION="06040082" and DATE_DEBUT="24MAY2011"d and Methode_prospection="autres" THEN DELETE;
IF CODE_STATION="06040082" and DATE_DEBUT="20JUN2012"d and Methode_prospection="autres" THEN DELETE;
IF CODE_STATION="06050061" and DATE_DEBUT="31MAY2007"d and Methode_prospection="Stratifiée par Points (grand
milieu)" THEN DELETE;
IF CODE_STATION="06110041" and DATE_DEBUT="19JUN1995"d and Methode_prospection="ambiances" THEN DELETE;
IF CODE_STATION="06380236" and DATE_DEBUT="17JUN2009"d and Methode_prospection="complète" THEN DELETE;
IF CODE_STATION="04420107" and DATE_DEBUT="28SEP1998"d and Moyen_prospection="A pied" THEN DELETE;
IF CODE_STATION="04420107" and DATE_DEBUT="28SEP1998"d and Moyen_prospection="En bateau" THEN DO;
op_cd_moyenprospection=855; Moyen_prospection="Mixte"; END;
IF CODE_STATION="04710110" and DATE_DEBUT="12AUG1998"d and Moyen_prospection="A pied" THEN DELETE;
IF CODE_STATION="04710110" and DATE_DEBUT="17AUG1999"d and Moyen_prospection="A pied" THEN DELETE;
Run;
Proc sort DATA=methode_moyen NODUPKEY; BY CODE_STATION DATE_DEBUT; Run;

* Méthode de pêche non manquante ;

Data nonmanq_methode (KEEP=CODE_STATION DATE_DEBUT op_cd_methodeprospection Methode_prospection);
SET methode_moyen (WHERE=(op_cd_methodeprospection ne .));
Run;
Proc sort DATA=nonmanq_methode; BY CODE_STATION DESCENDING DATE_DEBUT; Run;

* Moyen de pêche non manquant ;

Data nonmanq_moyen (KEEP=CODE_STATION DATE_DEBUT op_cd_moyenprospection Moyen_prospection);
SET methode_moyen (WHERE=(op_cd_moyenprospection ne .));
Run;
Proc sort DATA=nonmanq_moyen; BY CODE_STATION DESCENDING DATE_DEBUT; Run;

* Méthode de pêche manquante ;

Data manq_methode (KEEP=CODE_STATION DATE_DEBUT op_cd_methodeprospection Methode_prospection);
SET methode_moyen (WHERE=(op_cd_methodeprospection=.));
Run;
Proc sort DATA=manq_methode; BY CODE_STATION DATE_DEBUT; Run;

* Moyen de pêche manquant ;

Data manq_moyen (KEEP=CODE_STATION DATE_DEBUT op_cd_moyenprospection Moyen_prospection);
SET methode_moyen (WHERE=(op_cd_moyenprospection=.));
Run;
Proc sort DATA=manq_moyen; BY CODE_STATION DATE_DEBUT; Run;

* Complétude des méthodes de pêche manquantes ;

Proc sql NOPRINT;
CREATE TABLE manq_nonmanq_methode AS
SELECT T1.CODE_STATION, T1.DATE_DEBUT, T2.DATE_DEBUT AS DATE_REF, T2.op_cd_methodeprospection,
T2.Methode_prospection
FROM manq_methode AS T1, nonmanq_methode AS T2 HAVING T1.CODE_STATION=T2.CODE_STATION AND
T2.DATE_DEBUT<=T1.DATE_DEBUT
ORDER BY T1.CODE_STATION, T1.DATE_DEBUT, T2.DATE_DEBUT DESC;
Quit;

Data manq_nonmanq_methode (DROP=DATE_REF);
SET manq_nonmanq_methode;
BY CODE_STATION DATE_DEBUT;
IF first.DATE_DEBUT=1;
Run;

* Complétude des moyens de pêche manquants ;

Proc sql NOPRINT;
CREATE TABLE manq_nonmanq_moyen AS
SELECT T1.CODE_STATION, T1.DATE_DEBUT, T2.DATE_DEBUT AS DATE_REF, T2.op_cd_moyenprospection, T2.Moyen_prospection
```

```
FROM manq_moyen AS T1, nonmanq_moyen AS T2 HAVING T1.CODE_STATION=T2.CODE_STATION AND
T2.DATE_DEBUT<=T1.DATE_DEBUT
ORDER BY T1.CODE_STATION, T1.DATE_DEBUT, T2.DATE_DEBUT DESC;
Quit;

Data manq_nonmanq_moyen (DROP=DATE_REF);
SET manq_nonmanq_moyen;
BY CODE_STATION DATE_DEBUT;
IF first.DATE_DEBUT=1;
Run;

* Complétude de la table methode_moyen ;

Data methode_moyen;
UPDATE methode_moyen manq_nonmanq_methode;
BY CODE_STATION DATE_DEBUT;
Run;

Data methode_moyen;
UPDATE methode_moyen manq_nonmanq_moyen;
BY CODE_STATION DATE_DEBUT;
Run;

* Ajout dans la table des données ;

Data donn;
ATTRIB CODE_STATION CODE_SANDRE DATE_DEBUT DATE_FIN LABEL="";
MERGE methode_moyen donn;
BY CODE_STATION DATE_DEBUT;
Run;

/* Suppression des tables intermédiaires */

Proc datasets NOLIST; DELETE methode_moyen nonmanq_methode nonmanq_moyen manq_methode manq_moyen
manq_nonmanq_methode manq_nonmanq_moyen; Quit;

/*-------------------------------------------------------------------------------------------------*/
/* Ajout des indications sur le bassin                                                             */
/* (à partir du fichier Nouveau_jeu_de_donnees_avec_methodes_peche_complete_et_BV.csv créé avec R) */
/*-------------------------------------------------------------------------------------------------*/

/* Import */

Proc import DATAFILE="T:\B_travaux\Biodiv\201505 - Indice
IPR\Fichdep\Nouveau_jeu_de_donnees_avec_methodes_peche_complete_et_BV.csv" OUT=bassin (KEEP=code_sandre
code_masse_eau code_troncon bassin) DBMS=DLM REPLACE; DELIMITER=","; Run;

/* Fusion des tables */

Proc sql NOPRINT;
CREATE TABLE donn AS
SELECT T1.*, T2.code_masse_eau, T2.code_troncon, T2.bassin
FROM donn AS T1 LEFT JOIN bassin AS T2 ON T1.code_sandre=T2.code_sandre;
Quit;

Proc datasets NOLIST; DELETE bassin; Quit; /* Suppression de la table bassin */

/*-------------------------------------------------------------------------------------------------*/
/* Création de nouvelles variables                                                                 */
/*-------------------------------------------------------------------------------------------------*/

/* Calcul de la note de qualité */

/* La note de l'indice IPR est une note qui fonctionne de la façon suivante : plus elle est élevée, plus faible
est la qualité.
L'objectif est de créer une nouvelle note qui donne une information sur la qualité et qui fonctionnerait de la
façon suivante : plus elle est élevée, plus élevée est la qualité.
De plus, les valeurs extrêmes perturbent l'analyse. Un passage au logarithme est utilisé de manière à écraser les
valeurs extrêmes.
La note NOTE_QUALI est créée telle que NOTE_QUALI=max(log(Note+1))-log(Note+1) */

Proc sql NOPRINT;
CREATE TABLE donn AS
SELECT *, max(log(1+IPR_VB))-log(1+IPR_VB) AS NOTE_QUALI
FROM donn;
Quit;

/* Création des variables ANNEE, MOIS et JOUR */

Data donn;
SET donn;
ANNEE=year(DATE_DEBUT);
MOIS=month(DATE_DEBUT);
JOUR=INTCK('day',mdy(01,01,year(DATE_DEBUT)),DATE_DEBUT)+1;
Run;
```

```
/*-------------------------------------------------------------------------------------------------------*/
/* Suppression des valeurs manquantes                                                                    */
/*-------------------------------------------------------------------------------------------------------*/

Data donn;
SET donn (WHERE=(IPR_VB ne .));
Run;

/*-------------------------------------------------------------------------------------------------------*/
/* OPTION : Filtre sur les stations de Corse et la période de novembre à juillet                         */
/*-------------------------------------------------------------------------------------------------------*/

Data donn;
SET donn (WHERE=(substr(CODE_STATION,1,4) not in ("062A","062B") and month(DATE_DEBUT) not in
(11,12,01,02,03,04,05,06,07)));
Run;

/*-------------------------------------------------------------------------------------------------------*/
/* Suppression des doublons stricts                                                                      */
/*-------------------------------------------------------------------------------------------------------*/

Proc sort DATA=donn NODUPKEY; BY CODE_STATION DATE_DEBUT NOTE_QUALI; Run;

/*-------------------------------------------------------------------------------------------------------*/
/* Suppression des valeurs extremes au niveau global                                                     */
/*-------------------------------------------------------------------------------------------------------*/

/* +/- 3 fois l'écart-type */

Proc sql NOPRINT;
CREATE TABLE donn AS
SELECT *, MEAN(NOTE_QUALI)-3*STD(NOTE_QUALI) AS VEG_INF, MEAN(NOTE_QUALI)+3*STD(NOTE_QUALI) AS VEG_SUP
FROM donn
ORDER BY NOTE_QUALI;
Quit;

/* Numéro d'ordre de chaque note de qualité */

Data donn;
SET donn;
RETAIN ORDRE 0;
ORDRE+1;
Run;

/* Pour visualisation */

SYMBOL1 COLOR=CX99CC00 INTERPOL=nojoin VALUE=dot HEIGHT=0.5;
SYMBOL2 COLOR=CX008080 INTERPOL=join VALUE=none HEIGHT=0.5;
LEGEND1 LABEL=NONE VALUE=(TICK=1 "Notes de qualité" TICK=2 "Limite valeurs extrêmes");

Proc gplot DATA=donn;
WHERE NOTE_QUALI<3;
PLOT (NOTE_QUALI VEG_INF)*ORDRE / OVERLAY LEGEND=legend1;
Run;
Quit;

Proc gplot DATA=donn;
WHERE NOTE_QUALI>3;
PLOT (NOTE_QUALI VEG_SUP)*ORDRE / OVERLAY LEGEND=legend1;
Run;
Quit;

/* Suppression des valeurs extrêmes */

Data donn (DROP=VEG_INF VEG_SUP ORDRE);
SET donn (WHERE=(VEG_INF<NOTE_QUALI<VEG_SUP));
Run;

/*-------------------------------------------------------------------------------------------------------*/
/* Suppression des valeurs extrêmes au niveau local                                                      */
/*-------------------------------------------------------------------------------------------------------*/

/* +/- 3 fois l'écart-type par station */

Proc sql NOPRINT;
CREATE TABLE donn AS
SELECT *, MEAN(NOTE_QUALI)-3*STD(NOTE_QUALI) AS VEG_INF, MEAN(NOTE_QUALI)+3*STD(NOTE_QUALI) AS VEG_SUP, CASE WHEN
MIN(NOTE_QUALI)<MEAN(NOTE_QUALI)-3*STD(NOTE_QUALI) or MAX(NOTE_QUALI)>MEAN(NOTE_QUALI)+3*STD(NOTE_QUALI) THEN 1
ELSE 0 END AS A_VOIR
FROM donn
GROUP BY CODE_STATION
ORDER BY CODE_STATION, NOTE_QUALI;
Quit;

/* Numéro d'ordre de chaque note de qualité */

Data donn;
```

```
SET donn;
BY CODE_STATION;
IF first.CODE_STATION=1 THEN ORDRE=1; ELSE ORDRE+1;
Run;

/* Pour visualisation */

SYMBOL1 COLOR=CX99CC00 INTERPOL=nojoin VALUE=dot HEIGHT=1;
SYMBOL2 COLOR=CX008080 INTERPOL=join VALUE=none HEIGHT=1;
SYMBOL3 COLOR=CX008080 INTERPOL=join VALUE=none HEIGHT=1;
LEGEND1 LABEL=NONE VALUE=(TICK=1 "Notes de qualité" TICK=2 "Limites valeurs extrêmes");

Proc gplot DATA=donn;
BY CODE_STATION;
WHERE A_VOIR=1;
PLOT (NOTE_QUALI VEG_INF VEG_SUP)*ORDRE / OVERLAY LEGEND=legend1;
Run;
Quit;

/* Suppression des valeurs extrêmes */

Data donn (DROP=VEG_INF VEG_SUP ORDRE A_VOIR);
SET donn (WHERE=(VEG_INF<NOTE_QUALI<VEG_SUP));
Run;

/*-------------------------------------------------------------------------------------------------------*/
/* Suppression des doublons en réalisant des moyennes par station et date                                */
/*-------------------------------------------------------------------------------------------------------*/

Proc sort DATA=donn OUT=test; BY CODE_STATION DATE_DEBUT NOTE_QUALI; Run;

Data test;
SET test;
BY CODE_STATION DATE_DEBUT NOTE_QUALI;
IF first.DATE_DEBUT=0 or last.DATE_DEBUT=0;
Run;

/* Pas de doublons */

Proc datasets NOLIST; DELETE test; Quit; /* Suppression de la table test */

/*-------------------------------------------------------------------------------------------------------*/
/* Pré-traitement pour calcul de l'indice chaîné                                                         */
/*-------------------------------------------------------------------------------------------------------*/

/* Regroupement des années 2 à 2 à partir de 2007 */

Data donn_indice (DROP=_ANNEE_);
SET donn (RENAME=(ANNEE=_ANNEE_));
IF _ANNEE_ in (2007,2008) THEN ANNEE="2007-2008";
ELSE IF _ANNEE_ in (2009,2010) THEN ANNEE="2009-2010";
ELSE IF _ANNEE_ in (2011,2012) THEN ANNEE="2011-2012";
ELSE IF _ANNEE_=2013 THEN DELETE;
ELSE ANNEE=compress(_ANNEE_);
Run;

/* Notes moyennes par année (ou couple d'années à partir de 2007) */

Proc means DATA=donn_indice NWAY NOPRINT;
CLASS CODE_STATION ANNEE;
VAR NOTE_QUALI;
OUTPUT OUT=donn_indice (DROP=_TYPE_ _FREQ_) MEAN=;
Run;

/* Création d'une table biannuelle */

Data donn_indice_1;
SET donn_indice;
Run;
Proc sort DATA=donn_indice_1; BY CODE_STATION ANNEE; Run;

Data donn_indice_2 (DROP=NOTE_QUALI);
SET donn_indice (WHERE=(ANNEE ne "2011-2012"));
LABEL ANNEER="Année (N-1)" NOTE_QUALIR="Note de qualité (N-1)";
ANNEER=ANNEE;
IF ANNEER="2009-2010" THEN ANNEE="2011-2012";
ELSE IF ANNEER="2007-2008" THEN ANNEE="2009-2010";
ELSE IF ANNEER="2006" THEN ANNEE="2007-2008";
ELSE ANNEE=strip(ANNEER+1);
NOTE_QUALIR=NOTE_QUALI;
Run;
Proc sort DATA=donn_indice_2; BY CODE_STATION ANNEE; Run;

Data biann_indice;
ATTRIB CODE_STATION LABEL="Code de la station" ANNEE LABEL="Année" NOTE_QUALI LABEL="Note de qualité" ANNEER
LABEL="Année (N-1)" NOTE_QUALIR LABEL="Note de qualité (N-1)";
```

```
MERGE donn_indice_1 donn_indice_2;
BY CODE_STATION ANNEE;
Run;

/* Suppression des tables intermédiaires */

Proc datasets NOLIST; DELETE donn_indice_1 donn_indice_2; Quit;

/*-------------------------------------------------------------------------------------------------*/
/* Pré-traitement pour modélisation statistique                                                    */
/*-------------------------------------------------------------------------------------------------*/

/* Filtre sur le jeu de données pour ne conserver que la méthode de pêche complète et le moyen de pêche à pied */

Data donn_modeles;
SET donn (WHERE=(Methode_prospection="complète" and Moyen_prospection="A pied"));
Run;

/* Suppression des stations présentant moins de 3 observations sur l'ensemble de la période d'étude */

Proc sql NOPRINT;
CREATE TABLE donn_modeles AS
SELECT * FROM donn_modeles
GROUP BY CODE_STATION
HAVING count(NOTE_QUALI)>2
ORDER BY CODE_STATION, DATE_DEBUT;
Quit;
```

Annexe n° 3
Scripts de prétraitement des données

```
##################################################################################################
################## Preparation du fichier de donnees ##############################################
##################################################################################################

################## Importation des donnees IPR et métriques
# le fichier de données Nouveau_jeu_de_donnees_avec_methodes_peche_complete.csv a été agrégé sous SAS par Marlène
rm(list=ls())

setwd("T:/Indicateurs/Indicateurs SOeS/Indicateurs et Indices/081_Qualite_piscicole_cours_deau_IPR/Traitements/Donnees/IPR_issu_SEEE")

data <- read.csv ("Nouveau_jeu_de_donnees_avec_methodes_peche_complete.csv", sep = ";", dec = ",")


# suppression des variables et observations inutiles et renommage des variables conservées
data <- data [c("CODE_STATION", "CODE_SANDRE", "DATE_DEBUT", "op_cd_methodeprospection",
                "Methode_prospection", "op_cd_moyenprospection", "Moyen_prospection",
                "IPR_VB")]

colnames(data) <- c("code.station", "code.sandre", "date1", "code.method", "intit.method", "code.moyen",
"intit.moyen", "ipr")

# rem : la date est nommée "date1" et non "date" pour éviter des confusions avec la fonction "date"


# suppression des stations en Corse et des notes IPR vides
data <- subset (data, ! substr(data$code.station,1,4) == "062A" & ! substr(data$code.station,1,4) == "062B" )

data <- data [ ! is.na(data$ipr) == T,]


# ajout d'un zéro devant le code.sandre s'il ne compte que 7 caractères
data$code.sandre <- as.character (data$code.sandre)

data$code.sandre <- ifelse (nchar(data$code.sandre) == 7, paste ("0", data$code.sandre, sep=""), data$code.sandre)

data$code.sandre <- as.factor (data$code.sandre)


# gestion des dates
data$date1 <- as.Date (data$date1, format = "%d/%m/%Y")

library(lubridate)

data$annee <- year (data$date1)

data$mois <- month (data$date1)

data$jour.annee <- yday (data$date1)

summary(data$code.method)


# recodage méthodes de pêche
data$code.method <- as.factor(data$code.method)

data$code.moyen <- as.factor(data$code.moyen)

table (data$annee, data$code.method)

table (data$code.moyen, data$code.method)


# transformation log de l'IPR
data$logipr <- max (log (1 + data$ipr)) - log (1 + data$ipr)


######################## importation fichier avec coordonnées Lambert des sites
# C'est du Lambert 93 :
#http://www.sandre.eaufrance.fr/squelettes/consulter_fiche_attribut.php?attribut=ProjStationMesureEauxSurface&dictionnaire=/db/sandre/Schemas/stq/2.2/sandre_fmt_xml_stq.xsd
# lecture du fichier
file <- "U:/pirz/Indicateurs/Indicateurs SOeS/Indicateurs et Indices/081_Qualite_piscicole_cours_deau_IPR/Traitements/Donnees/StationMesureEauxSurfaceFXX-shp/StationMesureEauxSurface.csv"

sites <- read.csv(file, sep=";", dec=",")

colnames(sites)

# sélection et renommage des variables d'intérêt
sites <- sites [, c("CdStationM.C.254", "CoordXStat.N.24.15", "CoordYStat.N.24.15", "CdMasseDEa.C.254",
                    "CdTronconH.C.254", "CdCommune.C.254")]

colnames(sites) <- c("code.sandre", "xlambert93", "ylambert93", "code.masse.eau",
                     "code.troncon", "code.commune")
```

```
# ajout d'un zéro devant le code sandre quand il ne compte que 7 caractères et passage en classe "factor"
sites$code.sandre <- as.character (sites$code.sandre)
sites$code.sandre <- ifelse (nchar(sites$code.sandre) == 7, paste ("0", sites$code.sandre, sep=""), sites$code.sandre)
sites$code.sandre <- as.factor (sites$code.sandre)


# ajout de ces variables au dataframe "data"
data <- merge (x=data, y=sites, by="code.sandre", all.x = T)


# exportation pour Marlène
write.csv(data, "StationMesureEauxSurface.csv")


######################## sélection des observations
# sélection des échantillonnages entre août et octobre
data <- data [data$mois > 7 & data$mois < 11 ,]
# sélection des échantillonnages réalisés à pied selon la méthode "complète"
data <- data [data$intit.method == "complète" ,]
data <- data [data$intit.moyen == "A pied" ,]


data <- subset(data, ! is.na (code.method == T))


# suppression des doublons stricts (vérification)
data <- data[!duplicated(data),]


# suppression des valeurs au-delà de moyenne + ou - 3 écarts-type sur la distribution générale
moy <- mean (data$logipr)
et <- sd (data$logipr)
data <- subset (data, abs (data$logipr - moy) < 3 * et)
rm(moy, et)


# ajout de colonnes pour les nb de valeurs, moyennes et écarts-types des notes de qualité
tab2 <- aggregate(data[, "logipr"], by=list(data$code.station), FUN=length, simplify=T)
colnames(tab2) <- c("code.station", "nb.notes.qualite")
tab3 <- aggregate(data[, "logipr"], by=list(data$code.station), FUN=mean, simplify=T)
colnames(tab3) <- c("code.station", "moy.notes.qualite")
tab4 <- aggregate(data[, "logipr"], by=list(data$code.station), FUN=sd, simplify=T)
colnames(tab4) <- c("code.station", "et.notes.qualite")
data <- merge (x=data, y=tab2, by="code.station", all.x=T)
data <- merge (x=data, y=tab3, by="code.station", all.x=T)
data <- merge (x=data, y=tab4, by="code.station", all.x=T)
rm(tab3, tab4)


# sélection des stations avec 3 données IPR ou plus
sub3 <- subset (data, nb.notes.qualite > 2)
sub3$code.station <- droplevels (sub3$code.station)


# élimination des notes au-delà de 3 écarts-types pour chaque station
# calcul de l'écart à la moyenne de la station pour chaque note de qualité et suppression au-delà du seuil
sub3$ecart.moy <- abs ((sub3$logipr-sub3$moy.notes.qualite)/sub3$et.notes.qualite)
sub3 <- subset (sub3, sub3$ecart.moy <= 3)


# re-sélection des stations avec plus de 3 données IPR
sub3 <- subset (sub3, nb.notes.qualite > 2)
sub3$code.station <- droplevels (sub3$code.station)
n3 <- nrow (tab2[ which (tab2$nb.notes.qualite > 3),])


# recherche des pseudo-doublons - 2 notes IPR sur un même site à même date - moyenne si plusieurs IPR
vars <- as.list(names(data))
vars <- vars [vars !="logipr"]
sub3b <- aggregate(sub3["logipr"], by=list(sub3$code.station, sub3$date), FUN=mean, simplify=T)
colnames(sub3b) [1:2] <- c("code.station", "date1")
```

```
# visualisation des faux doublons éventuels
sub4 <- merge (x=sub3b, y=sub3,  by=c("code.station", "date1"), all.y=T)

sub4$diff <- sub4$logipr.x - sub4$logipr.y

min(sub4$diff)

max(sub4$diff)

sub4b <- subset (sub4, !sub4$logipr.x==sub4$logipr.y)

rm(sub4b, sub4)


# suppression des faux doublons éventuels
sub3 <- merge (x=sub3b, y=sub3,  all.x=T)

sub3 <- droplevels(sub3)

rm(tab2, sub3b, file, n3)

#############################################################################################################
# manipulation des fichiers SIG pour affecter les stations dans les bassins et préparer les
# représentations cartographiques
#############################################################################################################

######################### récupération des limites des sous-unités DCE
#(http://www.sandre.eaufrance.fr/atlascatalogue/?mode=ModeMeta&uuid=bd7c0d56-5c93-49ab-91fe-2f677796b178#)
wd1 <- getwd()

wd2               <-               "T:/Indicateurs/Indicateurs               SOeS/Indicateurs               et
Indices/081_Qualite_piscicole_cours_deau_IPR/Traitements/Donnees/Affectation_stations_sous_unites_dce"

setwd(wd2)

library(rgdal)

ogrInfo(".", "SsBassinDCEAdmin")

ssudce <- readOGR(".", "SsBassinDCEAdmin")

names(ssudce@data)

print(proj4string(ssudce))

# passage en Lambert 93
ssudcel93 <- spTransform (ssudce, CRS ("+init=epsg:2154") )

par (mfrow = c( 1,1))


# élimination de la Corse et des DOM
metr  <-  subset  (ssudcel93,  !  ssudcel93@data$CdEuSsBass  %in%  c("FRL_REU",  "FRK_GUY",  "FRJ_MAR",  "FRI_GUA",
"FRE_CORS"))

plot(metr)

summary(metr)


# spatialisation de sub3
sub3$x <- sub3$xlambert93

sub3$y <- sub3$ylambert93

coordinates(sub3) <- c("x", "y")

proj4string(sub3) <- CRS ("+init=epsg:2154")

print(proj4string(sub3))

plot(sub3)

plot(metr, add=T)


# affectation des stations dans des sous-unités DCE
require(sp)

require(maps)

sub3$bassin <- over(sub3, metr)$CdEuSsBass

sub3[is.na(sub3$bassin) == T,]

# affectation manuelle pour quelques stations frontalières
sub3$bassin[sub3$code.station == "01059000"] <- "FRA_ESCA"

sub3$bassin[sub3$code.station == "01590127"] <- "FRA_ESCA"

sub3$bassin[sub3$code.station == "05311011"] <- "FRF_GARO"

sub3$bassin[sub3$code.station == "06660103"] <- "FRD_COLR"

sub3$bassin <- droplevels (sub3$bassin)


rm(wd1, wd2, ssudce, vars, sites, data)


# visualisation
# création du fond de carte
library(maps)
```

```
library(rgdal)
par (mfrow = c( 1,1))


################# Définition de la fonction SpacializedMap()
# Fonction équivalente à la fonction map() mais qui crée des cartes géoréférencées de classe SpatialPolygons.
SpacializedMap=function(database="world",regions=".",...){
  #Create objects of class SpatialPolygons from geographical maps from map() function
  #BBORGY 08/20/2014
  require(maps)
  require(sp)
  require(mapdata)
  k=map(database=database,regions=regions,plot=F,fill=T,...)
  pol=vector("list",length(unique(k$names)))
  w.start=c(1,which(is.na(k$x))+1)
  w.end=c(which(is.na(k$x))-1,length(k$x));w=1
  for(j in 1:length(unique(k$names))){
    pol[[j]]=vector("list",sum(k$names==unique(k$names)[j]))
    for(i in 1:sum(k$names==unique(k$names)[j])){
      pol[[j]][[i]]=Polygon(cbind(k$x[c(w.start[w]:w.end[w],w.start[w])],k$y[c(w.start[w]:w.end[w],w.start[w])]))
      w=w+1}
    pol[[j]]=Polygons(pol[[j]],unique(k$names)[j])}
  pol=SpatialPolygons(pol,proj4string=CRS("+init=epsg:4326"))
  return(pol)}
################ fin de définition de la fonction SpacializedMap()

library (rgdal)
par (mfrow =c (1,1))
pol_WGS84 <- SpacializedMap('france')
plot (pol_WGS84, axes = T) ; mtext ("France continentale / Latlong", cex=.8)


fond_WGS84 <- SpacializedMap ('worldHires','france', xlim = range (-10,8.4), ylim = range (42.5,51), interior =
F)
plot (pol_WGS84, axes = T)
fond <- spTransform (fond_WGS84, CRS ("+init=epsg:2154") )
plot (fond, axes = T)
mar.orig <- par("mar")
# superposition du fond de carte, des stations et des limites de bassins
plot(sub3, pch=20)
plot(fond, add=T)
plot(metr, add=T)
```

Annexe n° 4
Scripts pour la modélisation statistique et les représentations graphiques

```
#################################################################################################
############ modélisation IPR
# calage de différents modèles ; les n° 2 à 5 & 5a sont avec l'année en variable numérique

library(mgcv)


# mod2
mod2 <- bam (logipr ~ s(annee) +
               s(code.station, bs="re"),
               family=gaussian, data = sub3)
# mod3
mod3 <- bam (logipr ~ s(annee) +
               s(code.station, bs="re") +
               s(bassin, bs="re"),
               family=gaussian, data = sub3)
# mod4
mod4 <- bam (logipr ~ s(annee) +
               s(code.station, bs="re") +
               s(bassin, bs="re") +
               s(jour.annee),
               family=gaussian, data = sub3)


# mod5
mod5 <- bam (logipr ~ s(annee) +
               s(code.station, bs="re") +
               s(bassin, bs="re") +
               s(jour.annee) +
               s(xlambert93, ylambert93) ,
               family=gaussian, data = sub3)


# mod5a : sans annee
mod5a <- bam (logipr ~ s(code.station, bs="re") +
               s(bassin, bs="re") +
               s(jour.annee) +
               s(xlambert93, ylambert93),
               family=gaussian, data = sub3)


# comparaison des modèles
AIC (mod2, mod3, mod4, mod5, mod5a)
Anova (mod5, mod5a, test = "F")


# examen de chaque modèle. Commencer par nommer "mod" le modèle choisi puis faire tourner la suite
mod <- mod5
summary(mod)
# graphiques de diagnostic
par(par.def)
gam.check (mod)
par(mfrow=c(1,1))
library(stats)
scatter.smooth(mod$residuals~mod$fitted.values, cex=0.5, col="red")
# relation note qualité ~ prédicteurs !!!! l'argument "select=" correspond à l'ordre d'entrée des variables
# dans le modèle ; le script a été calé pour mod5 mais doit être adapté pour les autres
plot.gam(mod, select=1, scale=0, xlab="Année")
abline (h=0, col="red", lty=1)
plot.gam(mod, select=2, se=F)
plot.gam(mod, select=3)
plot.gam(mod, select=4)
plot.gam(mod, select=5, se=F, main = "s(xlambert93, ylambert93)")
```

```
# ajout du fond de carte
plot(fond, add=T)

# carte avec de la couleur
vis.gam(mod, view=c("xlambert93", "ylambert93"), plot.type="contour", xlim=c(100000,1100000), color="terrain")

plot(fond, add=T)

# pour l'exporter
bmp(filename = "effet_geographique.bmp", width = 2000, height = 2000, units = "px", pointsize = 12, bg = "white")

par(mar = c(7, 8, 4.1, 2.1), mgp = c(6, 1, 0))

vis.gam(mod5, view=c("xlambert93", "ylambert93"), plot.type="contour", xlim=c(100000,1100000),
        color="terrain", cex.lab=3, cex.axis=3, cex.main=3)

plot(fond, add=T)

library(graphics)

grid()

dev.off()


# récupération du graphique de l'effet année en HD
bmp(filename = "effet_annee.bmp", width = 1200, height = 1000, units = "px", pointsize = 12, bg = "white")

par(mar = c(7, 8, 4.1, 2.1), mgp = c(6, 1, 0))

plot.gam(mod, select=1, scale=0, ylab = "Effet sur la note de qualité", xlab="Année", cex.lab=2.5, cex.axis=2,
las=1,
         lwd=2, rug=F)

abline (h=0, col="red", lty=2, lwd=1.5)

# text (2012, -0.08, "(a)", cex=4)
library(graphics)

grid()

dev.off()


#############################################################################################################
# même modèle que mod5 mais avec l'année en facteur pour obtenir un indice annuel

mod5b <- bam (logipr ~ as.factor(annee) +
                        s(code.station, bs="re") +
                        s(bassin, bs="re") +
                        s(jour.annee) +
                        s(xlambert93, ylambert93),
                        family=gaussian, data = sub3)


mod <- mod5b

summary(mod)

gam.check(mod)

# probabilité de l'absence d'effet année
summary(mod)$pTerms.pv

# récupération des indices annuels
indice.annuel <- mod$coefficients[1:length(levels(as.factor(sub3$annee)))]

# mise à zéro pour l'année de référence au lieu de l'intercept
indice.annuel[1] <- 0


# graphique avec les barres d'erreurs et indice 100 en première année ; exportation en bmp HD
library(Hmisc)

bmp(filename = "effet_annee_facteur.bmp", width = 1200, height = 1000, units = "px", pointsize = 12, bg =
"white")

par(mar = c(5.1, 6, 4.1, 2.1), mgp = c(4, 1, 0))

coef.table <- as.data.frame(summary(mod)$p.table)

colnames(coef.table)[2] <- "SE"

coef.table$annee <- as.numeric(substr( row.names(coef.table), nchar( row.names(coef.table))-3, nchar(
row.names(coef.table))))

coef.table$annee[1] <- min(na.omit(sub3$annee))

coef.table$Estimate[1] <- 0

coef.table$ind100 <- 100*(1+coef.table$Estimate)

coef.table$SE100 <- 100*coef.table$SE

coef.table$SE100 [1]<- 0

errbar(coef.table$annee, coef.table$ind100, yplus=coef.table$ind100+1.96*coef.table$SE100, las=1,
       yminus=coef.table$ind100-1.96*coef.table$SE100, type = "b", xlab="Année",
```

```
       ylab = "Indice annuel de qualité", lwd=2, cex=2, cex.lab=2.5, cex.axis=2, ylim=c(90,120))
# text (2012, 93.5, "(b)", cex=4)
abline (h = 100, col="red", lty=2)
library(graphics)
grid()
dev.off()


#############################################################################################
################## Estimation par bootstrap des Intervalles de confiance pour mod6 ########
#############################################################################################

mat <- matrix (ncol =19, nrow =1000)
for (i in 1:1000)
{
  modboot <- bam (logipr ~ as.factor (annee) + s (code.station, bs="re") + s (bassin, bs="re") + s (jour.annee) +
             s (xlambert93, ylambert93), data = sub3[sample (x = nrow(sub3), size = nrow(sub3), replace=T) , ])
  coefboot <- modboot$coefficients[1:19]
  mat[i,] <- as.numeric(coefboot)
}
colnames(mat) <- paste("A", levels(as.factor(sub3$annee)), sep="")


# agrégation pour obtenir les statistiques par année dans le data frame "synth"
# création des fonctions pour les quantiles en vue d'obtenir l'IC à 95%
quantile2.5 <- function (var) { quantile (var, probs=0.025) }
quantile97.5 <- function (var) { quantile (var, probs=0.975) }


synth1 <- apply (mat,2, median) # applique la fonction 'median' à la seconde dimension (les colonnes)
synth2 <- apply (mat,2, quantile2.5)
synth3 <- apply (mat,2, quantile97.5)
# empilage des trois listes et transformation en data frame
synth <- as.data.frame (rbind (synth1, synth2, synth3))
rm (synth1, synth2, synth3)


# ajout de l'année de référence 1995, mise des colonnes dans l'ordre chronologique et transposition
synth$A1995 <- c(0, 0, 0)
synth <- synth [c(length (levels (data$annee)), 1: length (levels (data$annee))-1)]
synth <- as.data.frame (t(synth))
# renommage des variables et création d'une variable numérique "annee"
colnames (synth) <- c("median", "quantile_2.5", "quantile_97.5")
synth$annee <- c(1995:2013)


# visualisation
# graphique avec les barres d'erreurs et indice 100 en première année ; exportation en bmp HD
library(Hmisc)
bmp(filename = "effet_annee_facteur_IC_bootstrap.bmp", width = 1200, height = 1000, units = "px", pointsize = 12,
bg = "white")
par(mar = c (5.1, 6, 4.1, 2.1), mgp = c (4, 1, 0))
errbar(synth$annee, synth$median, yplus=synth$quantile_97.5, las=1,
       yminus=synth$quantile_2.5, type = "b", xlab="Année",
       ylab = "Indice annuel de qualité", lwd=2, cex=2, cex.lab=2.5, cex.axis=2)
# text (2012, 93.5, "(b)", cex=4)
abline (h = 0, col="red", lty=2)
library(graphics)
grid()
dev.off()
```

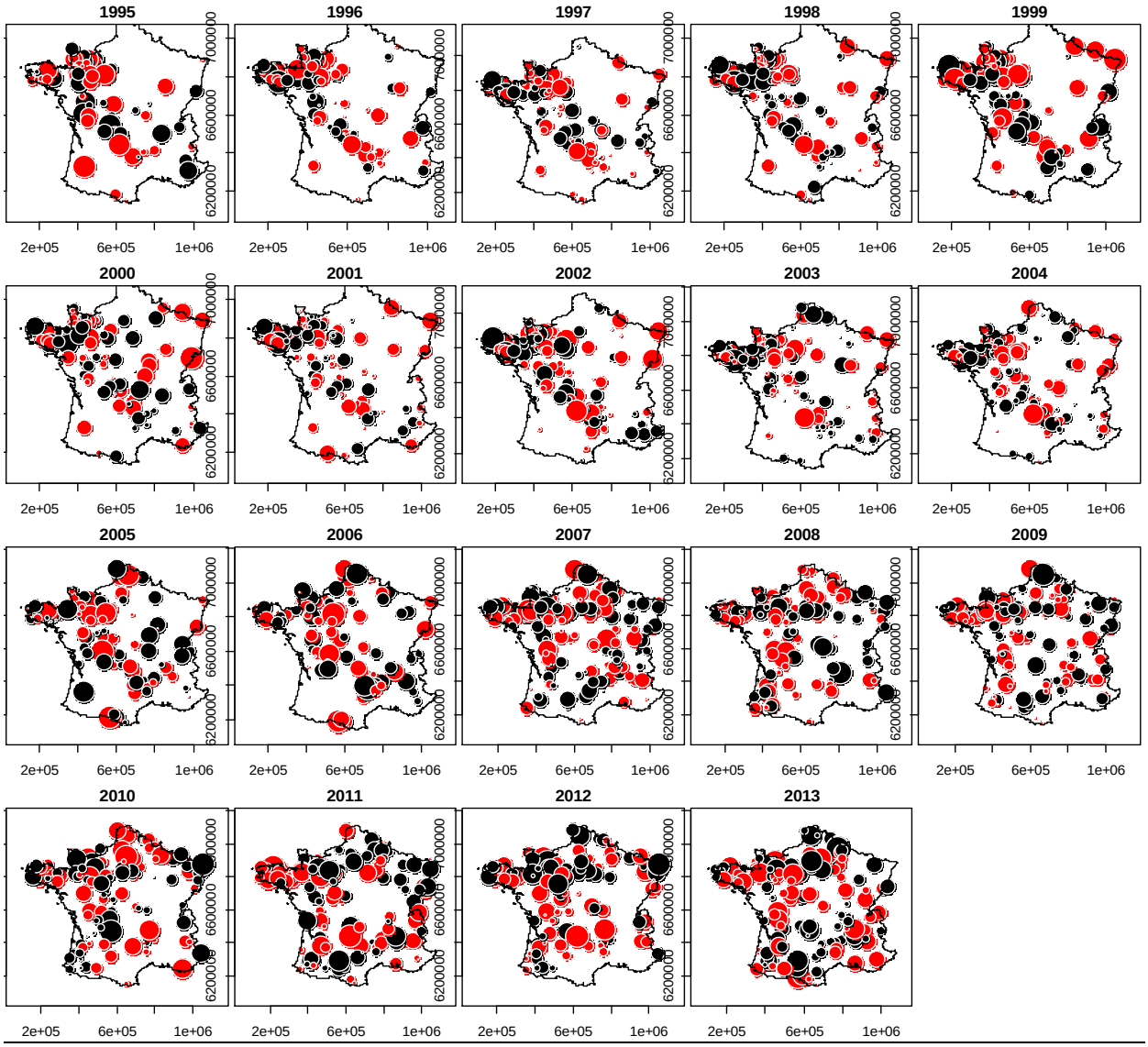
Annexe n° 5
Autocorrélation spatiale des résidus

Script R de cartographie des résidus du modèle

```
#1. Fonction SpacializedMap()
# comme la fonction map(),  crée des objets géoréférencés de classe SpatialPolygons.
SpacializedMap=function(database="world",regions=".",...){
  # crée des objets à partir des fonds géographiques de la fonction map()
  require(maps)
  require(sp)
  require(mapdata)
  k=map(database=database,regions=regions,plot=F,fill=T,...)
  pol=vector("list",length(unique(k$names)))
  w.start=c(1,which(is.na(k$x))+1)
  w.end=c(which(is.na(k$x))-1,length(k$x));w=1
  for(j in 1:length(unique(k$names))){
    pol[[j]]=vector("list",sum(k$names==unique(k$names)[j]))
    for(i in 1:sum(k$names==unique(k$names)[j])){
      pol[[j]][[i]]=Polygon(cbind(k$x[c(w.start[w]:w.end[w],w.start[w])],k$y[c(w.start[w]:w.end[w],w.start[w])]))
      w=w+1}
    pol[[j]]=Polygons(pol[[j]],unique(k$names)[j])}
  pol=SpatialPolygons(pol,proj4string=CRS("+init=epsg:4326"))
  return(pol)}
library(rgdal)
par(mfrow=c(1,1))
pol_WGS84=SpacializedMap('france')
plot(pol_WGS84,axes=T);mtext("France continentale / Latlong",cex=.8) # Affiche le contour de la France et des départements
# 2. Création du fond de carte
library(maps)
 library(rgdal)
par (mfrow = c( 1,1))
# si ce n'est pas déjà fait, faire tourner le script de la fonction SpacializedMap, ci-dessus
fond_WGS84 <- SpacializedMap ('worldHires','france', xlim = range (-10,8.4), ylim = range (42.5,51), interior = F)
plot (pol_WGS84, axes = T)
fond <- spTransform (fond_WGS84, CRS ("+init=epsg:2154") )
# 3. Cartographie des résidus : graphique composé par année, couleurs différentes selon résidus positifs ou négatifs, surface bulle proportionnelle à la valeur absolue du résidu - avec boucle sur l'année
don$radius <- sqrt( abs( mod6$residuals ) / pi )
don$coul <- as.factor (sign (mod6$residuals))
levels (don$coul) <-  c("red", "green")
# partitionnement de la fenêtre graphique
par(mfrow=c(4, 5), mar = c(2,0.2,2,0.2))
for (an in levels (as.factor (don$annee)))
{  # sélection de l'année
  dat <- subset (don, don$annee == an)
  symbols (dat$xlambert93 , dat$ylambert93, circles = dat$radius, inches=0.1, fg="white",
           bg=dat$coul, xlab="Longitude (Lambert 93)", ylab="Latitude (Lambert 93)", main = an)
  # ajout du fond de carte
  plot(fond, add=T) }
rm (dat.neg, dat.pos)
par(par.def)
```

Présentation des résultats

Avec le précédent script, les résidus du modèle 6 sont cartographiés par année d'après leurs coordonnées Lambert.



Les phénomènes d'autocorrélation spatiale positive des résidus peuvent être visualisés si certaines zones affichent une surreprésentation de résidus de même signe. Mais il existe des indicateurs d'autocorrélation spatiale à partir des données X_i et X_j des stations repérées par leurs coordonnées (i,j).

Pour cela, il faut construire une matrice de poids décroissants w_{ij} en fonction de la distance d_{ij} entre deux stations.

L'indice I de Moran est une mesure de ce phénomène. $$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}$$

Une valeur fortement positive traduit une autocorrélation spatiale positive : les valeurs proches (positives ou négatives) ont tendance à se regrouper.

En cas d'absence d'autocorrélation spatiale, la valeur attendue de I est $$E(I) = \frac{-1}{N-1}$$

Le test associé donne la significativité de l'autocorrélation.

```
tabres=cbind(mod6$residuals,don)  # récupération des résidus du modèle et ajout dans la table de données
res=tabres[tabres$annee==1998,] # sélection des résidus sur l'année 1998
dists<-  as.matrix(dist(cbind(res$xlambert93,  res$ylambert93)))   # matrice de distances euclidiennes entre les coordonnées des stations
dists.inv <- 1/(dists+1) # matrice de poids
diag(dists.inv) <- 0 # la diagonale de la matrice est nulle par définition
```

```
library(ape)
Moran.I(res[,1],dists.inv) # statistique de Moran et test de significativité
```

```
$observed
[1] -0.02991323     # I de Moran

$expected
[1] -0.009803922   # valeur attendue sous l'hypothèse d'absence d'autocorrélation

$sd
[1] 0.02007716 # statistique du test d'égalité des deux grandeurs

$p.value
[1] 0.3165363 # p-value du test
```

Ainsi, en 1998, ce test traduit une absence d'autocorrélation spatiale au seuil de 5 %.

En sélectionnant les résidus pour chaque année, l'hypothèse d'autocorrélation spatiale est rejetée (avec des p-values comprises entre 0,1 en 2001 et 0,89 en 2007) sur toute la période.

Annexe n° 6
Modèle comprenant le jour de l'année et les méthodes de pêche en prédicteurs

Données

Le modèle 7 est calé sur le jeu de données non filtré par les méthodes de pêche ni la période de l'année (n = 10 279 observations).

Les variables explicatives sont les mêmes que pour le modèle 5 qui est le modèle retenu pour la détermination de la tendance, auxquelles sont ajoutées :

- la variable *jour.annee*, sous la forme d'une fonction spline cubique cyclique. La cyclicité contraint les extrémités (1 pour le 1er janvier et 365 pour le 31 décembre) à se raccorder en valeur, mais aussi en dérivées première (pente) et seconde (inflexion de la pente) ;
- la variable *intit.method* (8 modalités) qui caractérisent la stratégie de prospection en pêche électrique. Les modalités sont les suivantes : « ambiances », « autres », « complète », « EPA », « faciès », « partielle sur berges », « partielle sur toute la largeur » et « stratifiée par points (grand milieu) » ;
- la variable *intit.moyen* (3 modalités) qui caractérise le moyen de prospection en pêche électrique (en bateau, mixte ou à pied). Les modalités sont les suivantes : « à pied », « en bateau » et « mixte ». Les effets fixes de la stratégie et du moyen de pêche ne sont pas croisés avec l'année : ils sont supposés indépendants du temps.

Résultats

Les sorties du modèle sont ci-dessous.

Family: gaussian

Link function: identity

Formula:

logipr ~ s(annee) + s(jour.annee, bs = "cc") + s(code.station, bs = "re") + s(bassin, bs = "re") + s(xlambert93, ylambert93) + intit.method + intit.moyen

R-sq.(adj) = 0.675 Deviance explained = 71.6%

fREML = 4330.8 Scale est. = 0.097175 n = 10264

Tableau 9 : résultats du modèle 7 ; effets aléatoires des prédicteurs nominaux code.station et bassin, et effet des prédicteurs numériques

Variable	ddl	Ref.ddl	F	p-value	Sig.
s(annee)	5,108	6,232	2,560	0,0165	*
s(jour.annee)	4,227	8,000	14,123	0,0344	*
s(code.station)	1229,893	1442,000	10,589	< 2e-16	***
s(bassin)	19,323	31,000	324,205	7,48e-07	***
s(xlambert93,ylambert93)	22,815	23,224	5,726	< 2e-16	***

Tableau 10 : résultats du modèle 7 ; effets fixes des prédicteurs nominaux intit.method et intit.moyen (coefficients paramétriques)

Paramètre	Estimation	Erreur Std.	t	p-value	Sig.
(Intercept)	3,79156	0,03327	113,966	< 2e-16	***
intit.methodautres	0,22053	0,23728	0,929	0,352719	
intit.methodcomplète	0,08532	0,01978	4,312	1,63e-05	***
intit.methodEPA	0,03318	0,32219	0,103	0,917989	
intit.methodfaciès	0,00459	0,03899	0,118	0,906302	
intit.methodpartielle sur berges	0,11228	0,02335	4,809	1,54e-06	***
intit.methodpartielle sur toute la largeur	0,11448	0,03517	3,255	0,001139	**
intit.methodStratifiée par Points (grand milieu)	0,15569	0,01731	8,994	< 2e-16	***
intit.moyenEn bateau	-0,17237	0,02256	-7,639	2,41e-14	***
intit.moyenMixte	-0,08244	0,02472	-3,335	0,000858	***

Note : les modalités de références pour ces deux variables sont respectivement « ambiances » et « à pied ».

Figure 8 : graphiques de diagnostic du modèle 7

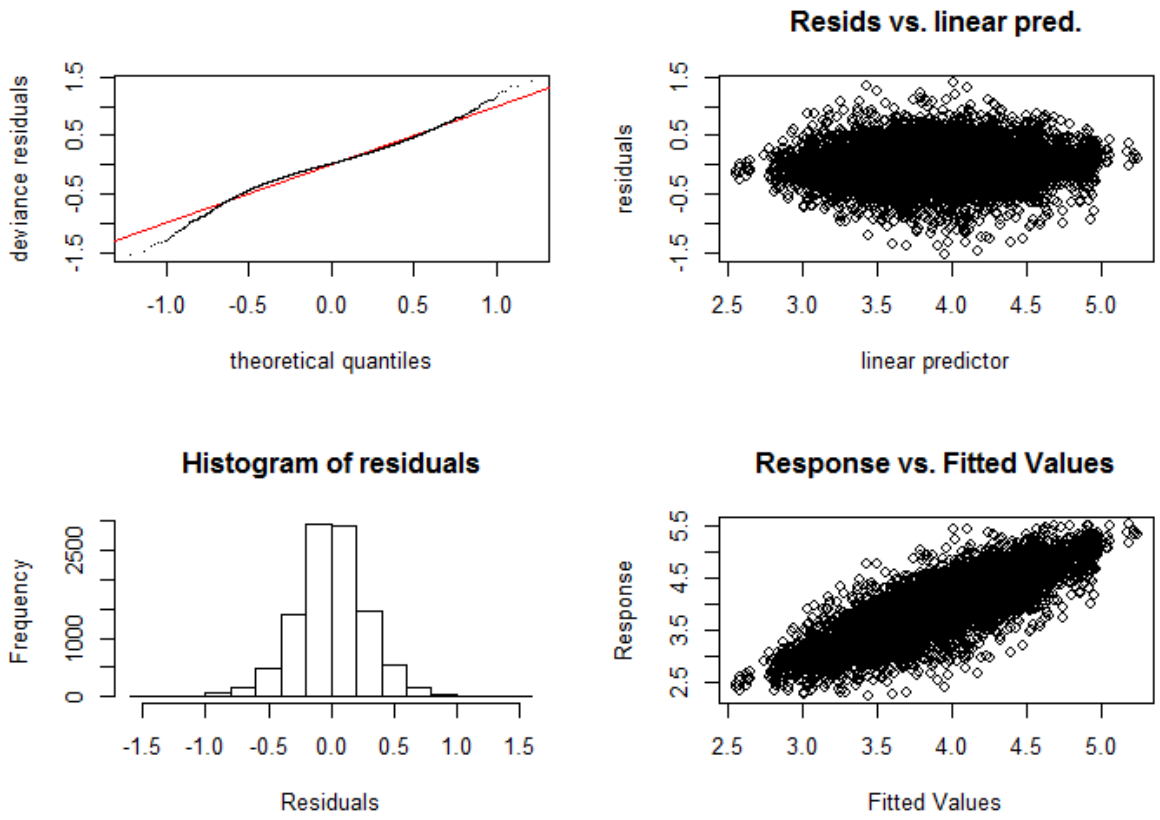
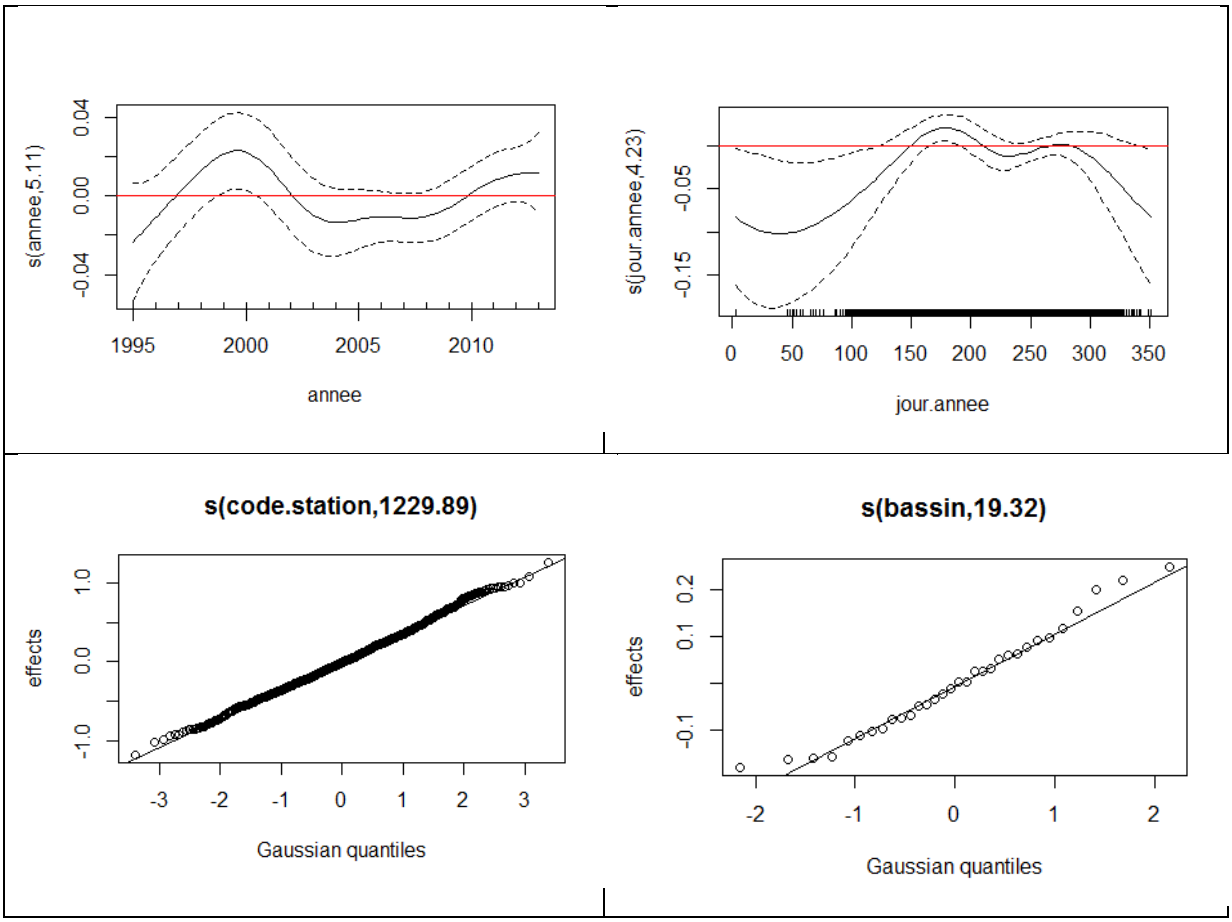
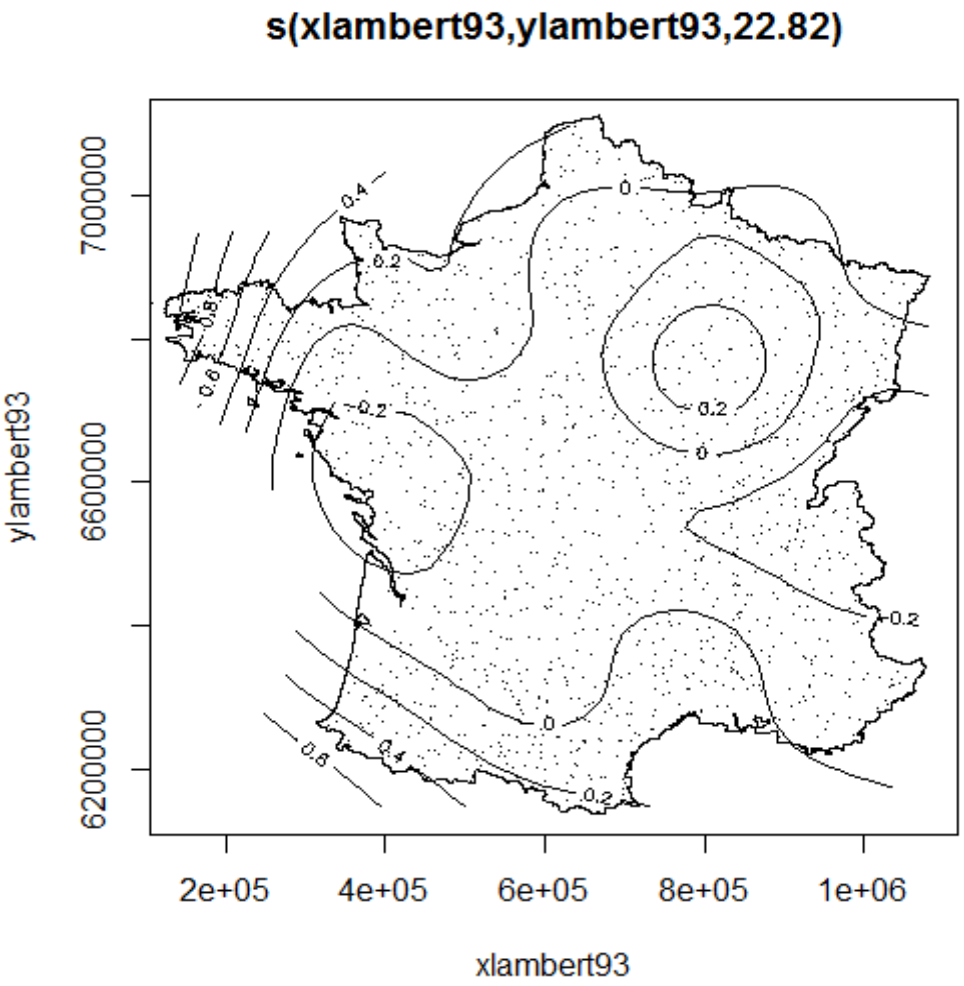


Figure 9 : effets des variables explicatives du modèle 7



Note : les effets de code.station *et de* bassin *sont aléatoires. Les fonctions de transformation de* annee *et de* jour.annee *sont des splines cubiques.*

Figure 10 : graphique de diagnostic du modèle 7



Note : la fonction de transformation des coordonnées Lambert est un spline cubique.

Interprétation

Ce modèle est présenté pour sa valeur démonstrative mais il est calé sur des données ne respectant pas rigoureusement le champ d'application de l'IPR. L'ensemble des variables explicatives contribue significativement à expliquer la variabilité de la note de qualité (*Tableau 9* et *Tableau 10*).

Outre les effets des prédicteurs qui sont également présents dans le modèle 5 (cf. §3.2), il apparaît que l'effet de la variable *jour.annee* s'interprète, comme attendu, comme une saisonnalité. Les notes de qualité sont en effet significativement inférieures à la moyenne quand les pêches ont été réalisées de janvier à avril, et supérieures à la moyenne de mi-juin (jour 165) à mi-juillet (jour 190).

L'effet de la méthode de pêche est aussi net. Par rapport à la modalité de référence (pêche par ambiance), les pêches "complète", "partielle sur berges", "partielle sur toute la largeur" et "stratifiée par points (grand milieu)" donnent des notes de qualité supérieures.

L'effet de la variable *intit.moyen* indique que les pêches réalisées en bateau ou mixtes donnent des notes de qualité inférieures à celles obtenues quand la pêche est réalisée à pied.

Conclusion

Ces résultats mettent l'accent sur l'importance des biais potentiels que représentent les méthodes de collecte de la donnée ainsi que les effets saisonniers. Face à ces difficultés, il est possible d'essayer d'exercer un contrôle statistique sur les biais (approche de la présente annexe). La limite du contrôle statistique réside dans le grand nombre des interactions potentielles entre variables explicatives, qu'il n'est pas toujours possible de tester faute d'observations en nombre suffisant. Dans le cas présenté, un exemple serait entre la méthode de prospection, l'année et le bassin, car les différentes directions interrégionales de l'Onema n'ont pas nécessairement harmonisé leurs protocoles de terrain de manière synchrone.

L'alternative « radicale » est d'écarter les observations qui ont été faites dans les conditions qui ne semblent pas comparables avec le reste des données.

Annexe n° 7
Considérations logicielles

Dans cette étude, deux logiciels ont été utilisés : le logiciel *SAS (SAS Institute, 2013)* et le logiciel *R* (Ihaka & Gentleman, 1996).

Qu'est-ce que  ?

R est un logiciel d'analyse de données, de représentation graphique ainsi qu'une plate-forme de programmation (R Core Team, 2014). Il est gratuit et *open source*. Ses créateurs sont des scientifiques de l'université d'Auckland (Ihaka & Gentleman, 1996). Très performant en graphiques, statistiques, traitement de données géographiques, R peut être interfacé avec tous les formats *open source*.

Le logiciel de base comprend les fonctions essentielles, mais pour des analyses plus poussées, il est nécessaire de recourir à des modules complémentaires. Ceux-ci, quand ils ont été validés (publication de référence dans la revue *Journal of Statistical Software*), sont téléchargeables sur le site R-CRAN. Ils sont au total 7 339 au 15 octobre 2015. Cette abondance a pour inconvénient que certains sont redondants, donc qu'il existe de nombreuses façons d'effectuer les mêmes analyses avec R. Elle a pour avantage que pratiquement toutes les méthodes connues d'analyse statistique, de modélisation ou de visualisation des données peuvent être mises en œuvre avec R.

S'agissant d'un logiciel libre, la communauté des utilisateurs est très active sur des forums. Ceux-ci se rencontrent également à l'occasion de colloques comme, en France, les Rencontres R, dont la 4e édition a eu lieu à Grenoble en juin 2015. Les demandes d'assistance reçoivent généralement des réponses rapides qui permettent de lever les blocages. Le monde académique a, dans sa majorité, adopté R. Le nombre des étudiants formés à ce logiciel est en constante croissance. Des formations en ligne se développent en langue française (ex : MOOC Introduction à la statistique avec R, piloté par l'université Paris Sud) et permettent d'accéder à des cours parmi les plus prestigieux, comme ceux dispensés à Princeton.

Le principal reproche adressé à R est sa limitation sur les gros jeux de données, car celles-ci sont chargées en mémoire vive. Des progrès ont toutefois été réalisés avec la version 64-*bits* de R et la mise en ligne de packages dédiés au *big data* (ex : *biglm, bigmemory, biganalytics*).

L'autre reproche couramment adressé à R est sa lenteur sur les boucles, car il s'agit d'un langage interprété.

Les évolutions récentes tendent à résoudre ces limitations en mettant à profit les opportunités de calcul décentralisé : en parallèle (Schmidberger et al., 2009), cloud computing (Ohri, 2014), etc.

Qu'est-ce que SAS ?

Statistical Analysis System, ou *SAS*, est une solution de *Business Intelligence* couvrant l'ensemble du processus d'informatique décisionnelle, de l'intégration de sources de données d'origines multiples, leur transformation, leur stockage et leur analyse, jusqu'à la diffusion de l'information et de la connaissance. Il est le fruit d'un projet de l'université d'état de Caroline du Nord mené à la fin des années 1960 par Anthony J. Barr et James H. Goodnight. Il a été commercialisé dès 1976 par *SAS Institute*.

SAS est un progiciel composé de logiciels intégrés appelés modules. Le module *Base SAS* constitue le cœur du système, il est nécessaire au fonctionnement des autres produits et permet la manipulation, l'extraction et le stockage des données, la génération de statistiques descriptives et de rapports, ainsi que la création de macros qui simplifient la programmation (Ringuedé, 2014). Pour des besoins plus précis, l'utilisateur acquiert les modules adaptés : l'offre de *SAS* est « à la carte ».

Les utilisateurs forment une communauté très active et il existe, essentiellement basés aux États-Unis, de nombreux groupes d'utilisateurs au premier rang desquels se trouve *SAS Global Forum*. Il organise chaque année une conférence internationale au cours de laquelle des articles d'utilisateurs illustrant les possibilités de telle ou telle procédure sont exposés avant d'être mis à la disposition des utilisateurs *SAS* du monde entier.

La manipulation et la transformation de données sont des phases essentielles puisque les données, dans leur forme brute, n'ont jamais la forme sur lesquelles les outils économétriques et statistiques peuvent directement être appliqués. Le logiciel *SAS* est l'un des plus puissants dans ce domaine grâce, en partie, à la très grande adaptabilité de son langage spécifique qui offre des solutions recouvrant l'ensemble des besoins qu'un utilisateur puisse formuler. *SAS* présente également le grand avantage d'être polyvalent et de savoir tout faire dans le domaine de la donnée (*reporting*, graphes, *ETL*, *SQL*, analyse statistique ...). Même s'il n'excelle pas dans chacune des fonctions qu'il possède et que des outils spécialisés peuvent mieux faire, il faudrait pour le remplacer complètement recourir à plusieurs logiciels spécifiques.

Le principal frein à l'utilisation de *SAS* est son prix, ce qui explique en partie qu'il soit boudé par les universités, bien qu'il existe des partenariats académiques en France offrant un environnement d'enseignement et d'apprentissage gratuit ou à faible coût pour accéder à ses logiciels.

Dans une logique commerciale, *SAS* a dernièrement beaucoup axé son évolution sur des solutions adaptées au monde du marketing, sa cible principale. D'importants investissements ont notamment été faits sur les solutions *Big Data* et un outil interactif très puissant dédié au *reporting* agile, à l'exploration visuelle et à l'analyse des données a été mis au point : *SAS Visual Analytics*. L'intégration de nouvelles méthodes statistiques a ainsi été relayée au second plan, ce qui explique un certain retard de *SAS* dans ce domaine, notamment par rapport à *R*.

Comparaison R - SAS

Ces deux logiciels représentent l'essentiel des usages en programmation statistique (Lofland & Ottesen, 2013). *R* est dominant dans le monde académique et *SAS* au sein des entreprises.

C'est dans le domaine de la manipulation des (gros) jeux de données que *SAS* est nettement supérieur. Ce logiciel a été conçu comme un langage de manipulation de données (*Data Manipulation Language, DML*, en anglais). Il comprend donc des commandes logiques permettant facilement, et de manière intuitive, la manipulation des données (lectures et écritures) dans une base de données. Et à la différence de *R* (hormis dans des packages spécialisés), les données ne sont pas chargées *in extenso* en mémoire vive de l'ordinateur, mais seulement au moment des appels des procédures de calcul, dans l'objectif d'optimiser l'usage de la mémoire.

R peut faire toutes les analyses que *SAS* peut faire, ce qui n'est pas réciproque. Par exemple des méthodes récemment développées ont fait l'objet de packages *R* alors qu'elles ne sont pas disponibles en *SAS* (ex : *Adaptive Boosting, Net Regularized Generalized Linear Models, Random Forests for Classification and Regression, Generalized Additive Mixed Models*, etc.). La moindre réactivité de *SAS* face à l'innovation s'explique par une approche commerciale qui donne la priorité aux méthodes les plus employées, et par un processus de certification rigoureux mais prenant du temps.

En conclusion, se limiter à l'utilisation de *SAS* interdit de nombreuses méthodes statistiques. La rapidité des évolutions méthodologiques dans les sciences de l'environnement, pour répondre aux progrès dans les connaissances et dans la disponibilité de la donnée, milite pour une utilisation de *R* en complément de *SAS* quand celui-ci s'avère limitant dans une thématique.

Ministère de l’Environnement, de l'Énergie et de la Mer
Commissariat général au développement durable
Service de l’observation et des statistiques
Tour Séquoia
92055 La Défense Cedex
Mél : diffusion.soes.cgdd@developpement-durable.gouv.fr