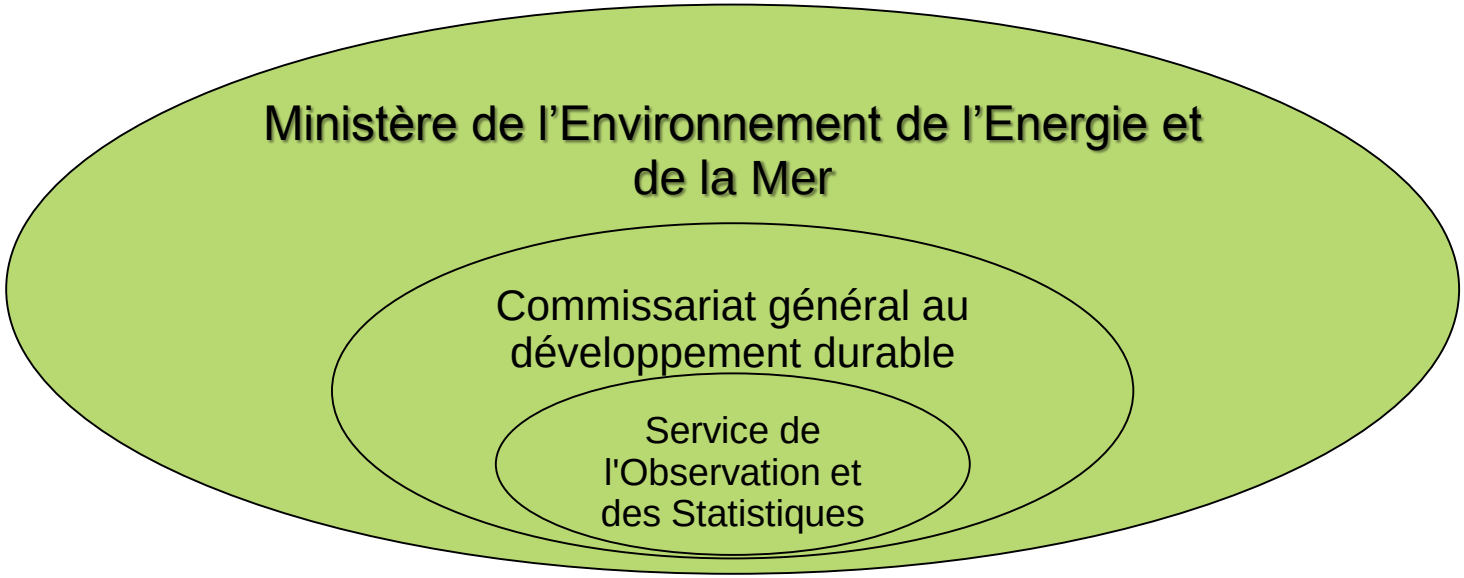
Tendance nationale de l'Indice Poisson Rivière estimée par un Modèle Additif Généralisé

Pascal Irz

Michaël Levi-Valensin

Commissariat Général au Développement Durable, Orléans

Cadre institutionnel



Missions du SOeS

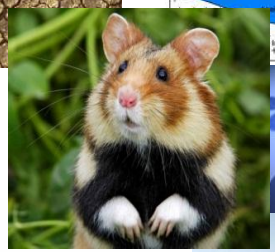
- Systèmes d'informations sur la biodiversité et l'eau
- Production d'indicateurs nationaux et/ou régionaux
- Synthèses thématiques
- Information du public

Travaux engagés

- Expérimentation des méthodes d'estimation de tendances (modélisation statistique)
- Précurseurs à la fois sur les types de modèles testés et sur l'utilisation de R

L'information environnementale : un champ vaste et complexe

















Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Ministère de l'Environnement, de l'Énergie et de la Mer

Les acteurs sont nombreux

Objectif

Evaluer si la qualité des cours d'eau, mesurée par l'état des communautés de poissons, tend à s'améliorer, ou à se dégrader sur le « long terme »

→ Recours à l'IPR

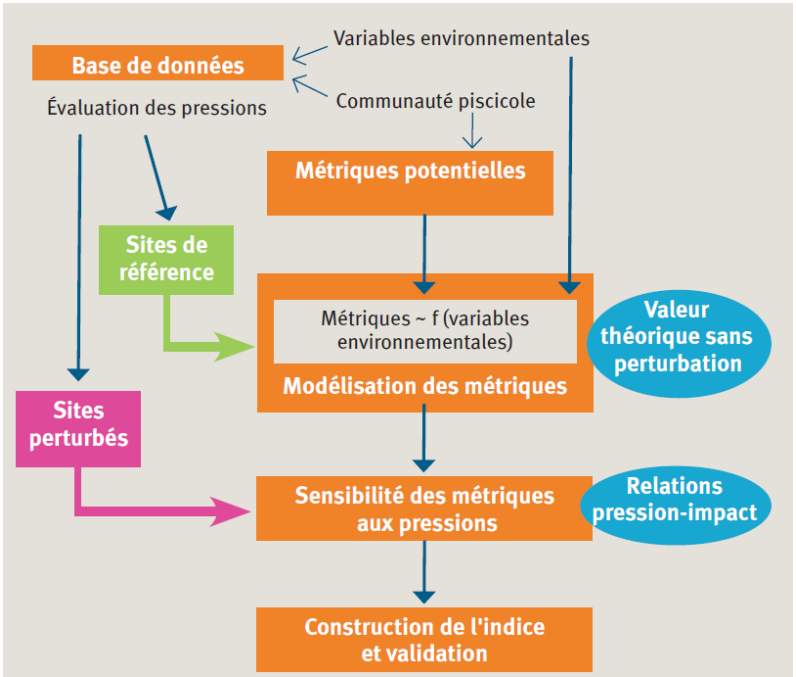
Qu'est-ce que l'IPR ?

- Outil d'évaluation de la qualité des cours d'eau[1] produit par l'Onema[2]
- La situation de référence dépend de 9 variables environnementales
- Pour chacune des 7 métriques

  Score = métrique prédite en situation de référence – métrique observée

- $Note\ IPR = \sum_{i=1}^{7} Score$
- IPR varie de 0 (conforme à la référence) à l'infini
- 5 classes de qualité

Note de l'IPR	Classe de qualité
<7	Excellente
]7–16]	Bonne
]16-25]	Médiocre
]25-36]	Mauvaise
>36	Très mauvaise



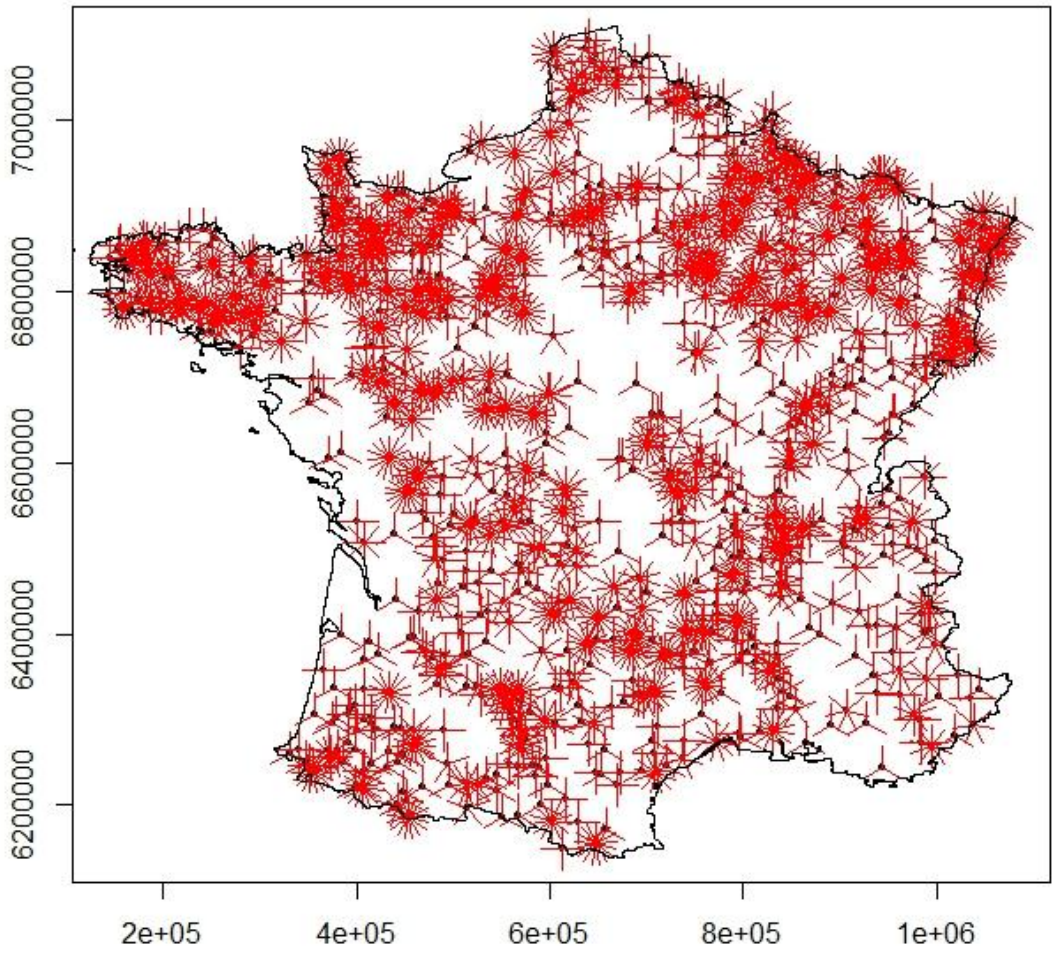
[1] Oberdorff *et al.* 2002 Development and validation of a fish-based index for the assessment of "river health" in France. Freshw. Biol. 47, 1720–1734.

[2] Onema, 2006. L'indice poissons rivière (IPR) - Notice de présentation - Edition avril 2006. Office National de l'Eau et des Milieux Aquatiques, Vincennes, France.

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Ministère de l'Environnement, de l'Énergie et de la Mer

Variabilité du réseau de stations

N ~ 2500 pêches sur 19 ans



Comment estimer une tendance d'évolution par modélisation statistique ?

Difficultés : Nombreux biais possibles

- ✓ Variabilité du réseau de stations
- ✓ Variabilité des saisons d'échantillonnage

Modèle additif

- linéaire si la forme de la relation est linéaire : GLM, ANOVA
- généralisé si la relation est non linéaire mais indéterminée *a priori* (fonction spline ou smooth) : GAM

Modèle pour données répétées => effet aléatoire

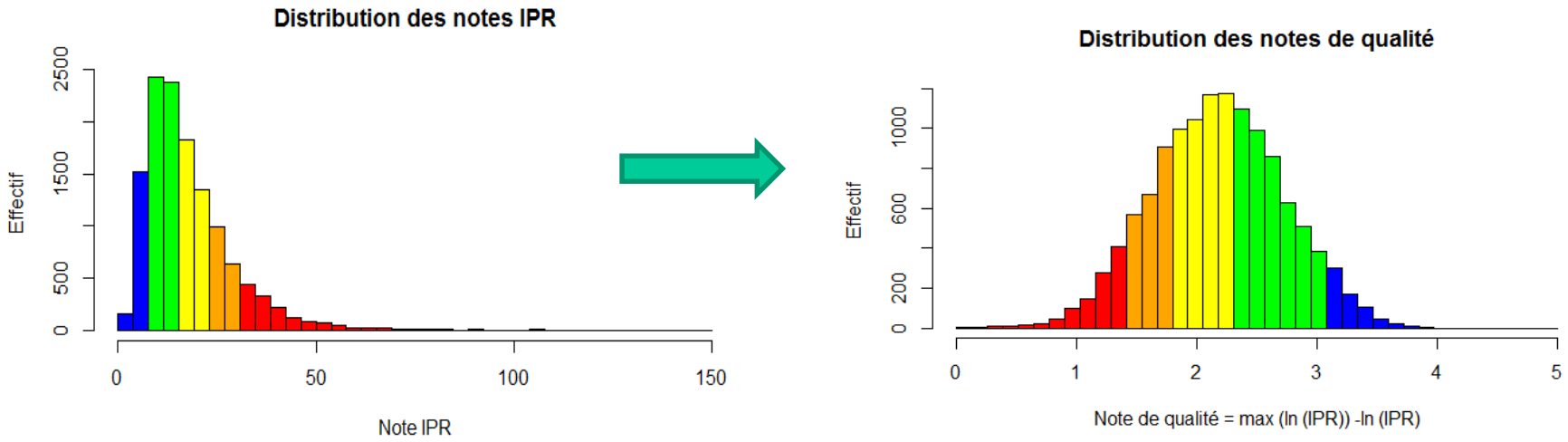
$$y_{i,t} = \eta_t + \gamma_i + \varepsilon_{i,t}$$

avec $t \in \{1, T\}$ l'année d'observation et $i \in \{1, \ldots, N\}$ l'index de station.

Prétraitement de l'indice

- *Distribution asymétrique de l'IPR -> normalisation et inversion*

IPR -> max(ln(IPR))-lnIPR



Choix du modèle

Modélisation NDQ ~ année + covariables

- ***Covariables***
  - *Mesures répétées -> code station*
  - *Données spatialisées -> coordonnées, bassin*
  - *Saisonnalité -> jour de pêche codé de 1 à 365*

- ***Spécifications***
  - *Bassin en effet aléatoire*
  - *Code.station en effet aléatoire « emboîté » dans le bassin*
  - *Spline (longitude, latitude)*
  - *Année -> soit qualitative pour indice annuel, soit quantitative (comme dans les régressions de Poisson, logiciel TRIM*)*

*TRends and Indices for Monitoring data

Package mgcv

- *Mixed GAM Computation Vehicle* : nombreux modèles linéaires et additifs avec des fonctions splines et smooth
- Fonction **gam**
- « *Backfitting algorithm* » pour les termes de lissage

et validation croisée généralisée (GCV) pour ajuster le paramètre de lissage

- Pour des modèles mixtes (avec effets aléatoires),en grande dimension, fonction **bam** (*Big Additive Models*)
  - moins d'espace mémoire
  - scinde les données en morceaux (« *chunks* »)

L'année en variable qualitative

Mod <- bam (NDQ ~ as.factor(annee) + s(jour.annee, bs = "cc") + s(code.station, bs = "re") + s(xlambert93, ylambert93) + s(bassin, bs = "re"))

- Extraction des éléments du modèle summary(mod)$p.table

coefficients annuels entre 1996 et 2013, écarts à la modalité de référence en 1995

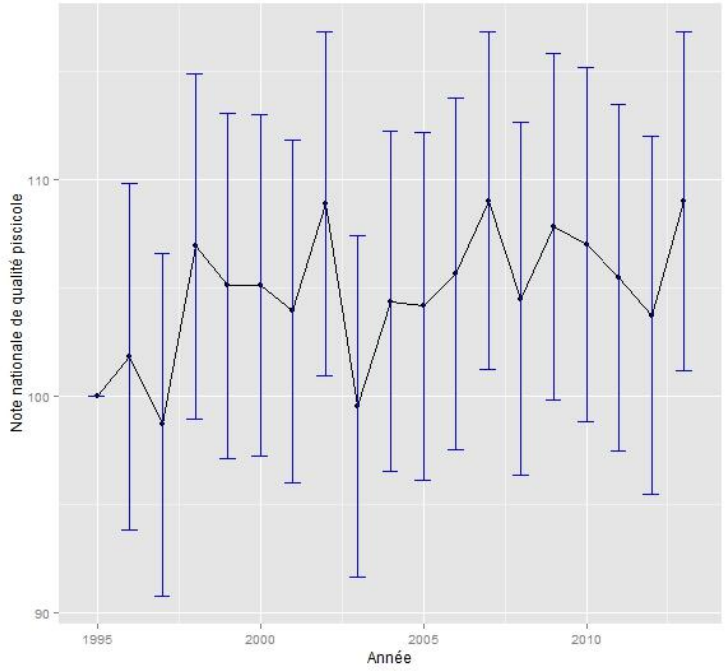
Std.Error : erreur-type

- Représentation des tendances annuelles (ggplot2)

Indices en base 100 :
ind100=100*(1+Estimate)

Intervalles de confiance :

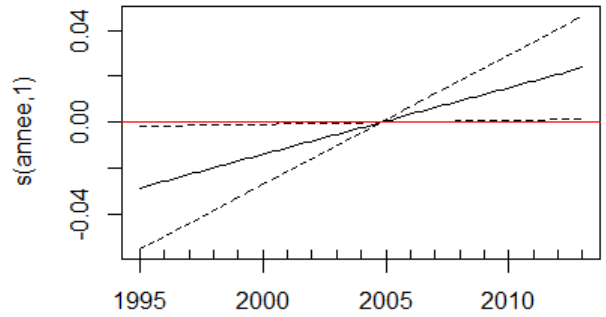
ind100 +/- 1.96*(100* Std.Error)



L'année en variable quantitative

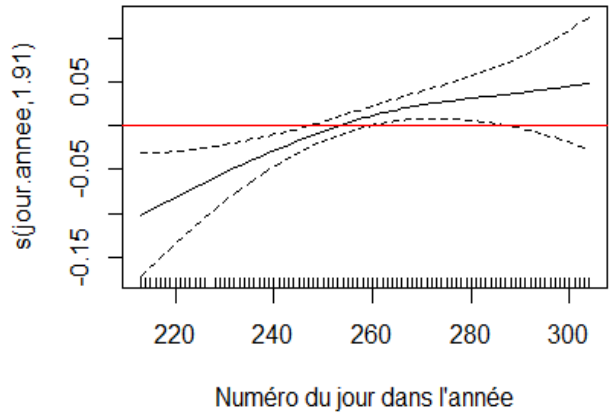
Mod <- bam (logipr ~ s(annee) + s(jour.annee, bs = "cc") + s(code.station, bs = "re") + s(xlambert93, ylambert93) + s(bassin, bs = "re"))

plot.gam(mod) : courbes splines des effets aléatoires année et jour



Notes de qualité

- *inférieures à la moyenne jusqu'à début sept. (jour.annee <250)*
- *- supérieures à la moyenne de mi-sept. à mi oct. (260<jour.annee<290)*



Des prolongements de l'étude

Sur l'IPR

- Tests de robustesse par bootstrap des estimations et des intervalles de confiance
- Cartographie et mesure de l'autocorrélation spatiale des résidus (package **ape**, fonction **I.Moran**)
- Document de travail

Tests pour

- d'autres éléments de faune et flore
- concentrations de macropolluants des eaux superficielles sur des bassins

⇒ Besoin de collaborations avec le monde académique

Véritables développements spatio-temporels

Et des travaux qui se poursuivent autour de R

- Au SOeS, utilisation de SAS pour la plupart des traitements statistiques
- Réflexion engagée sur l'utilisation progressive de R : comparatif des fonctionnalités, coût des licences
  En ce qui concerne la modélisation, R s'avère plus adapté que SAS
- Constitution d'un Groupe de Référents R : mise en place de formations


Merci de votre attention