Stochastic Search Variable Selection:

application à l'influence de l'environnement sur les peuplements de poissons

Jérémy Piffady – Pôle Onema-Irstea Hydroécologie des Cours d'Eau
Eric Parent – Equipe MORSE, UMR 518 INRA - AgroParisTech

Pour mieux affirmer ses missions, le Cemagref devient Irstea

www.irstea.fr

Rencontres AppliBUGS – 26 juin 2012

La question du choix de modèle

- Problème crucial en régression multiple: quels régresseurs introduire?
- Comparer les 2^P modèles possibles
  - AIC, BIC, DIC, Facteur de Bayes?
  - Problèmes quand p augmente
- Méthodes heuristiques?
  - Stepwise
- Reversible Jump
  - Exploration aléatoire de l'espace des modèles
- Stochastic Search Variable Selection

Stochastic Search Variable Selection (George & McCulloch, 1993)

- Intégration du modèle de régression dans un modèle normal bayésien hiérarchique

$$Y|\beta,\sigma^2 \sim N_n(X\beta,\sigma^2 I)$$

- Utilisation de variables latentes pour identifier les régresseurs d'intérêt

$$p(\gamma_i = 1) = 1 - p(\gamma_i = 0) = p_i$$

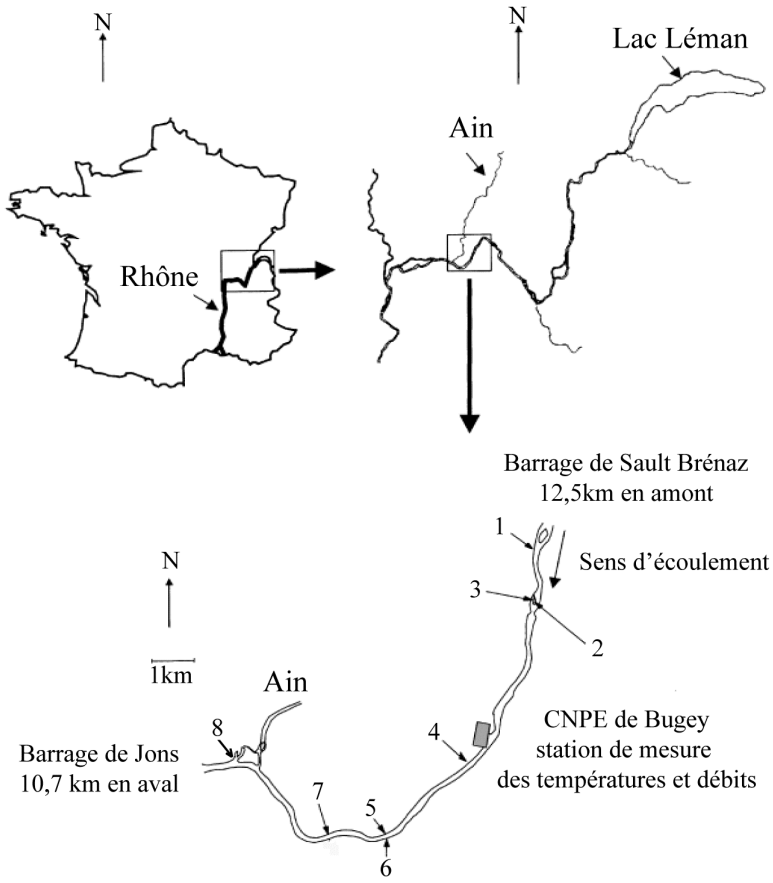
  - Probabilités a posteriori les plus élevées

- Prior de β : mélange de lois normales

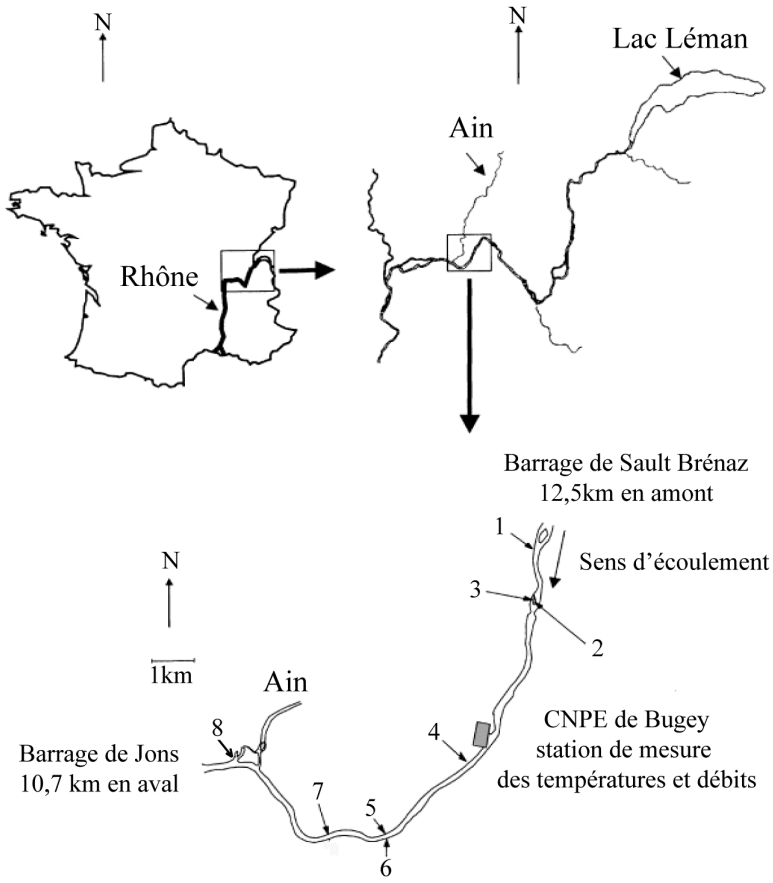
$$\beta_i|\gamma_i \sim (1-\gamma_i)N(0,\tau_i^2) + \gamma_i N(0,c_i\tau_i^2)$$

- Adaptation ici au cas GLM Poisson
  - Application aux assemblages de juvéniles de Cyprinidés du Haut-Rhône

Le site d'étude: le suivi piscicole de Bugey

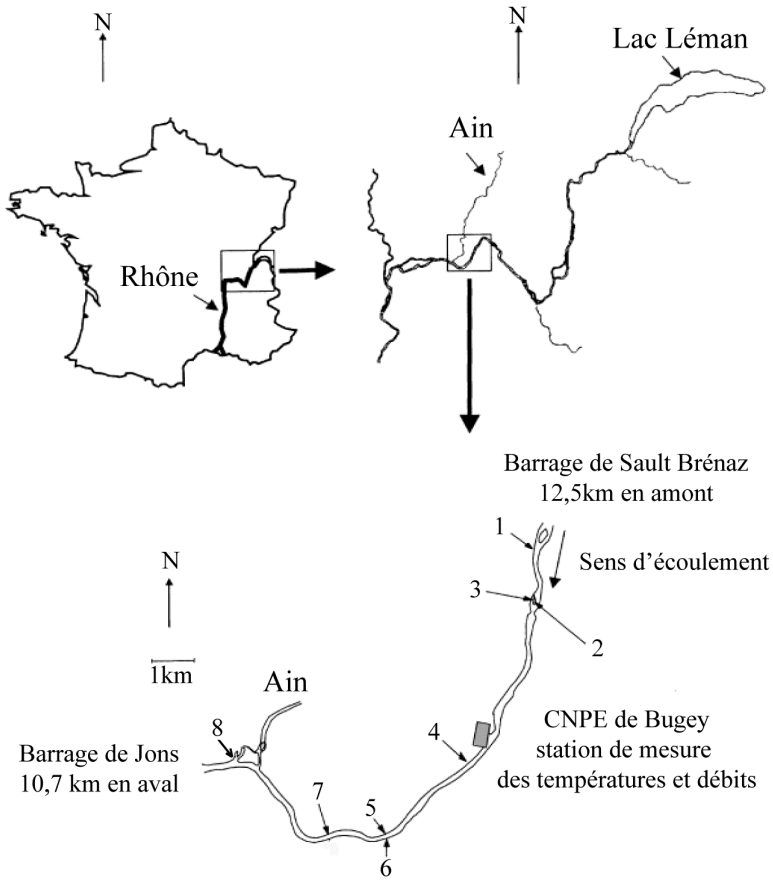


Le site d'étude: le suivi piscicole de Bugey



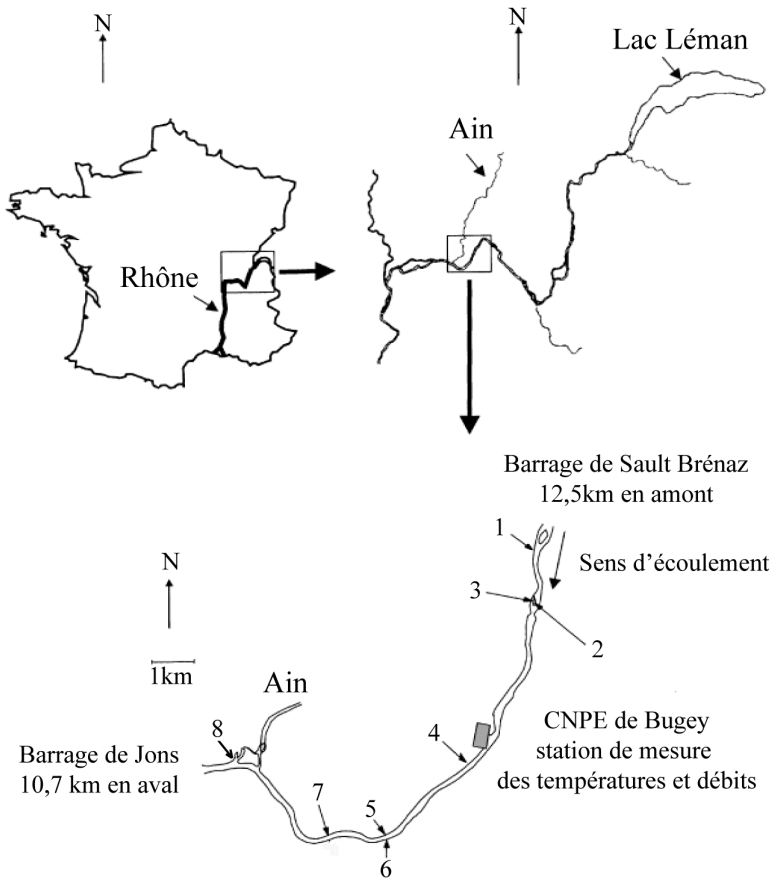
- 1980-2009
- CPEN Bugey (1977-1978)

Le site d'étude: le suivi piscicole de Bugey



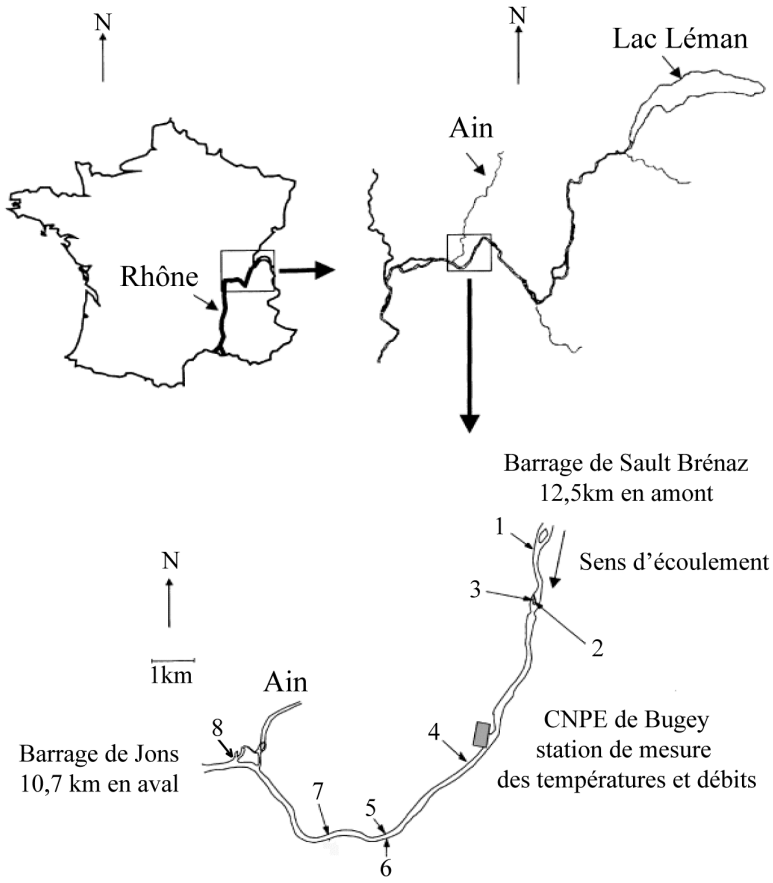
- 1980-2009
- CPEN Bugey
- 8 stations d'échantillonnage
  - Pêches de berges

Le site d'étude: le suivi piscicole de Bugey



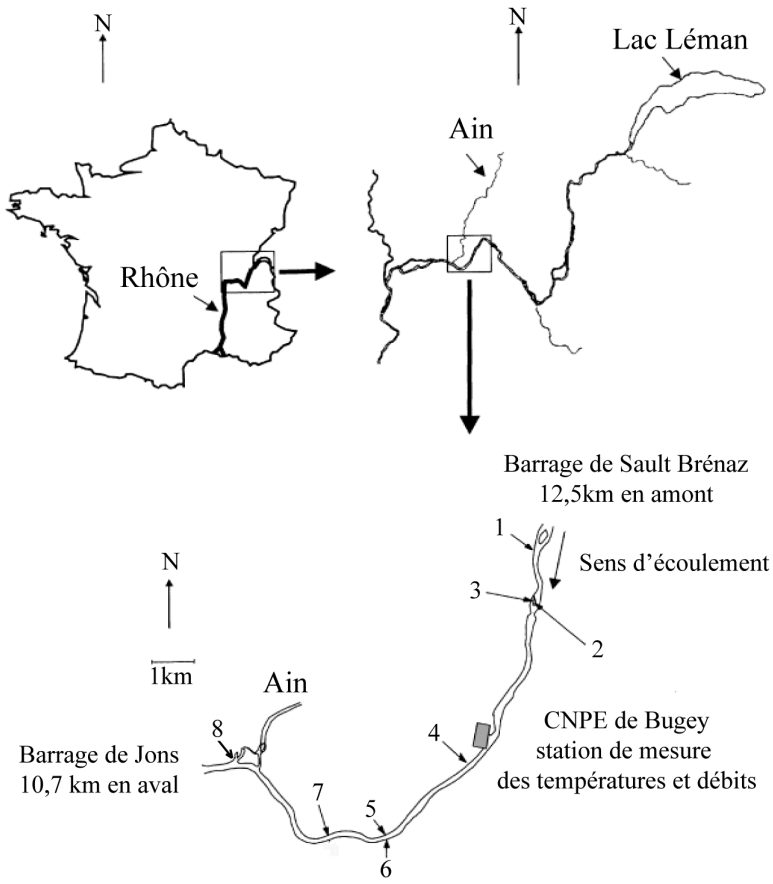
- 1980-2009
- CPEN Bugey
- 8 stations d'échantillonnage
  - Pêches de berges
  - Essentiellement juvéniles

Le site d'étude: le suivi piscicole de Bugey



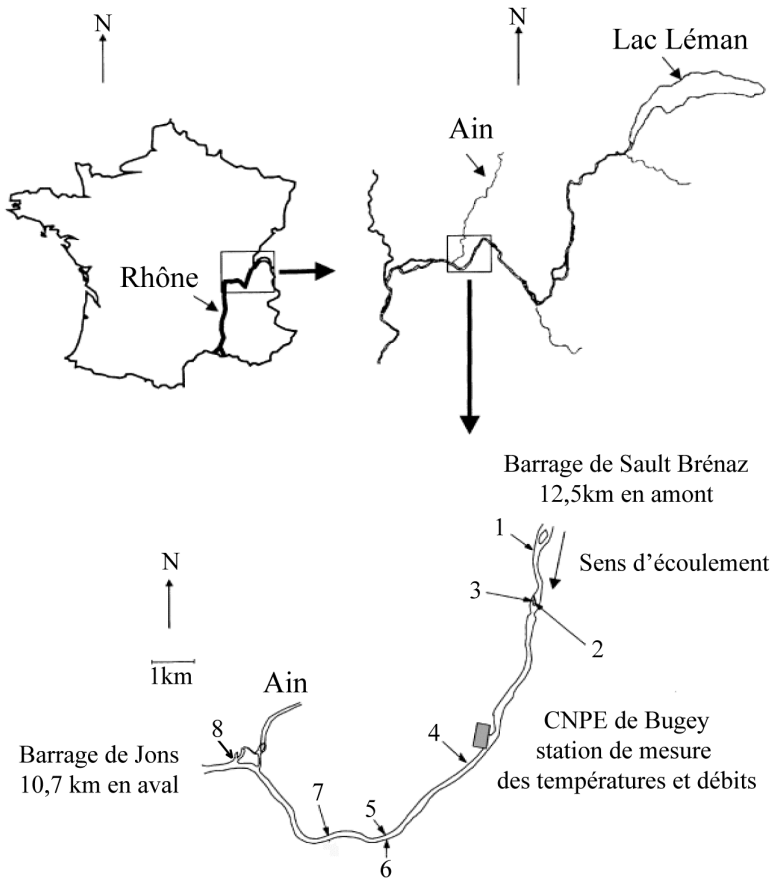
- 1980-2009
- CPEN Bugey
- 8 stations d'échantillonnage
  - Pêches de berges
  - Essentiellement juvéniles

Le site d'étude: le suivi piscicole de Bugey



- 1980-2009
- CPEN Bugey
- 8 stations d'échantillonnage
  - Pêches de berges
  - Essentiellement juvéniles
- 2 saisons
  - Été / automne

Le site d'étude: le suivi piscicole de Bugey



- 1980-2009
- CPEN Bugey
- 8 stations d'échantillonnage
  - Pêches de berges
  - Essentiellement juvéniles
- 2 saisons
  - Été / automne
- 1 station de relevé de température de l'eau et de débit

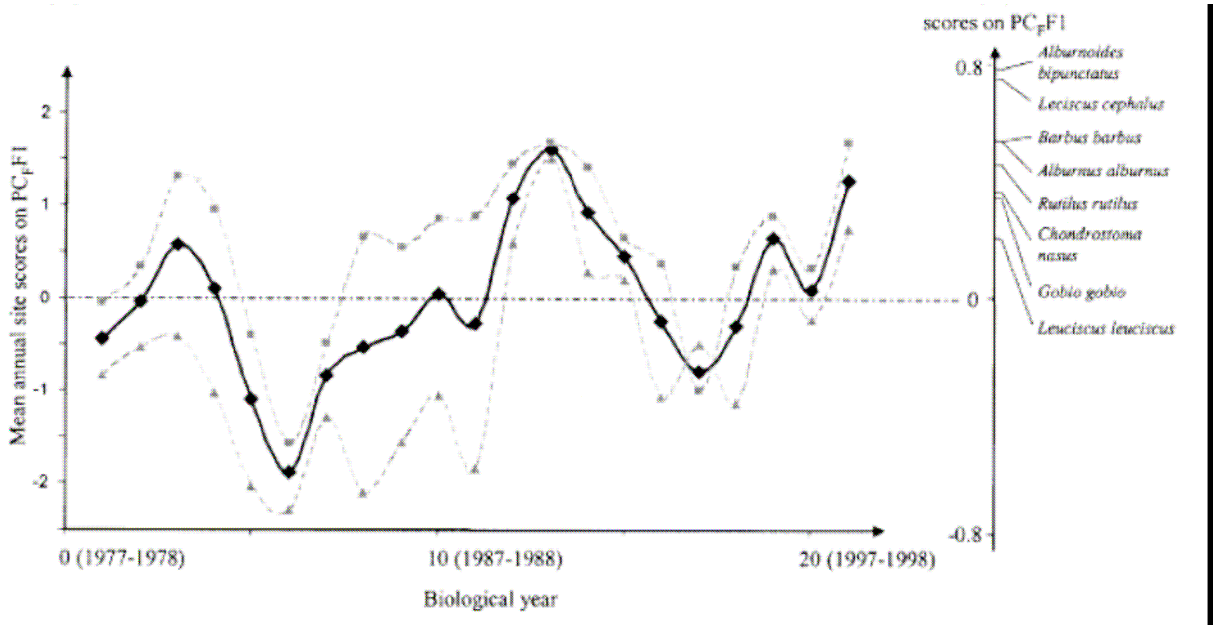
Approches historiques sur les grands cours d'eau

- Approches descriptives (en p. Persat H., Carrel G., Olivier J.M.)
  - Echantillonnage
  - Identification
  - Effets des barrages du Haut Rhône
  - Descriptions de différents patterns temporels ou spatiaux
- Lien patterns – déterminants physiques
  - Débits: mise en évidence d'évènements exceptionnels (Cattanéo, 2001)
  - Thermie: évolutions tendancielles et shift d'espèces

    (Daufresne et al., 2003)

Approches historiques sur les grands cours d'eau

- Approches descriptives (en p. Persat H., Carrel G., Olivier J.M.)
- Lien patterns – déterminants physiques
  - Débits: mise en évidence d'évènements exceptionnels (Cattanéo, 2003)
  - Thermie: évolutions tendancielles et shift d'espèces



dans Daufresne, et al., 2003

Approches historiques sur les grands cours d'eau

- Approches descriptives (en p. Persat H., Carrel G., Olivier J.M.)
  - Echantillonnage
  - Identification
  - Effets des barrages du Haut Rhône
  - Descriptions de différents patterns temporels ou spatiaux
- Lien patterns – déterminants physiques
  - Débits: mise en évidence d'évènements exceptionnels (Cattanéo et al., 2001)
  - Thermie: évolutions tendancielles et shift d'espèces

(Daufresne et al., 2003)

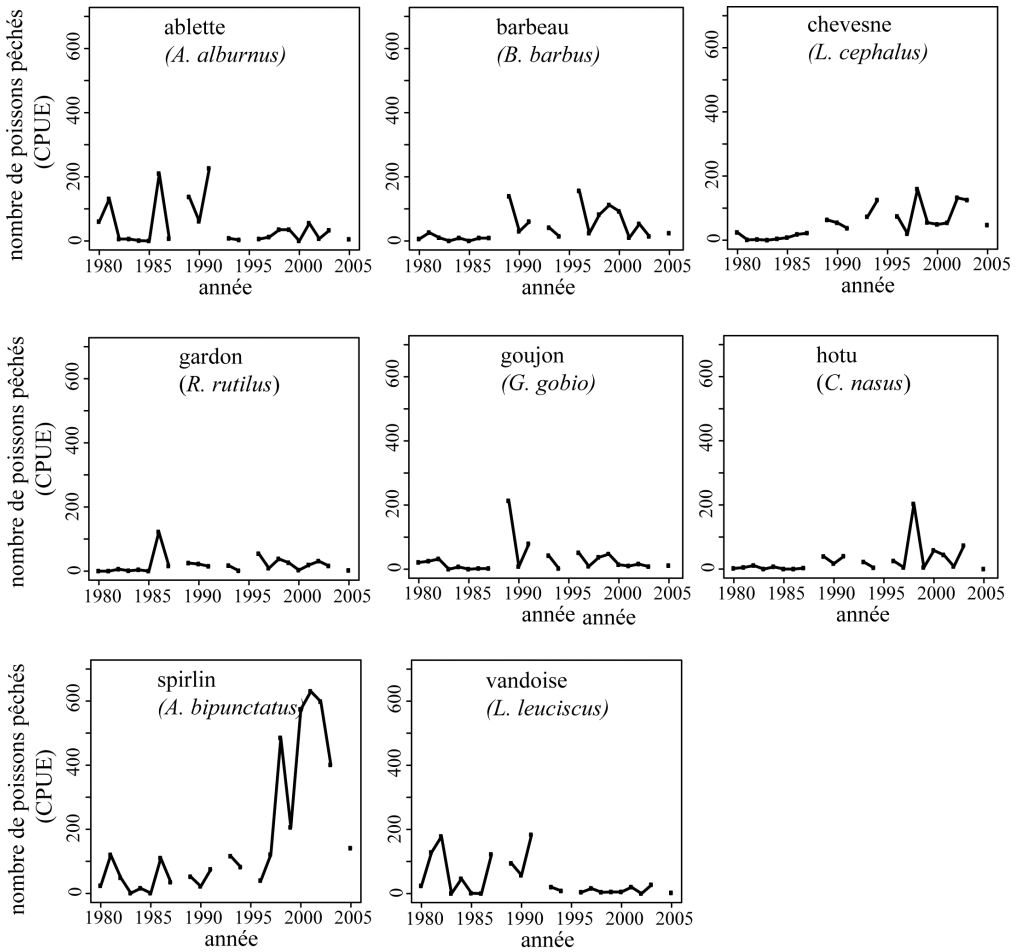
Reste à préciser les causes et conséquences

des variations fines interannuelles

Peut-on quantifier les effets des variations naturelles d'environnement sur l'assemblage de juvéniles?

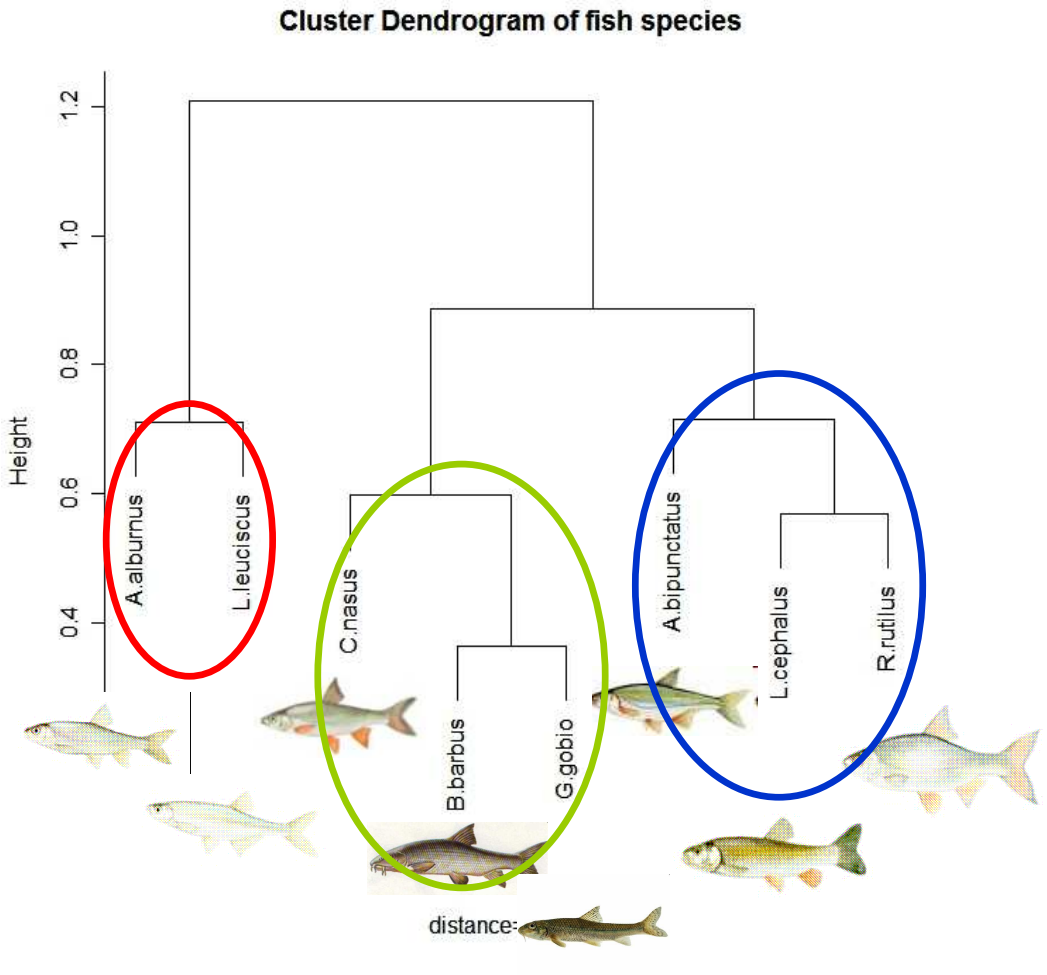
Données biologiques

- Fortes variations interannuelles de recrutement



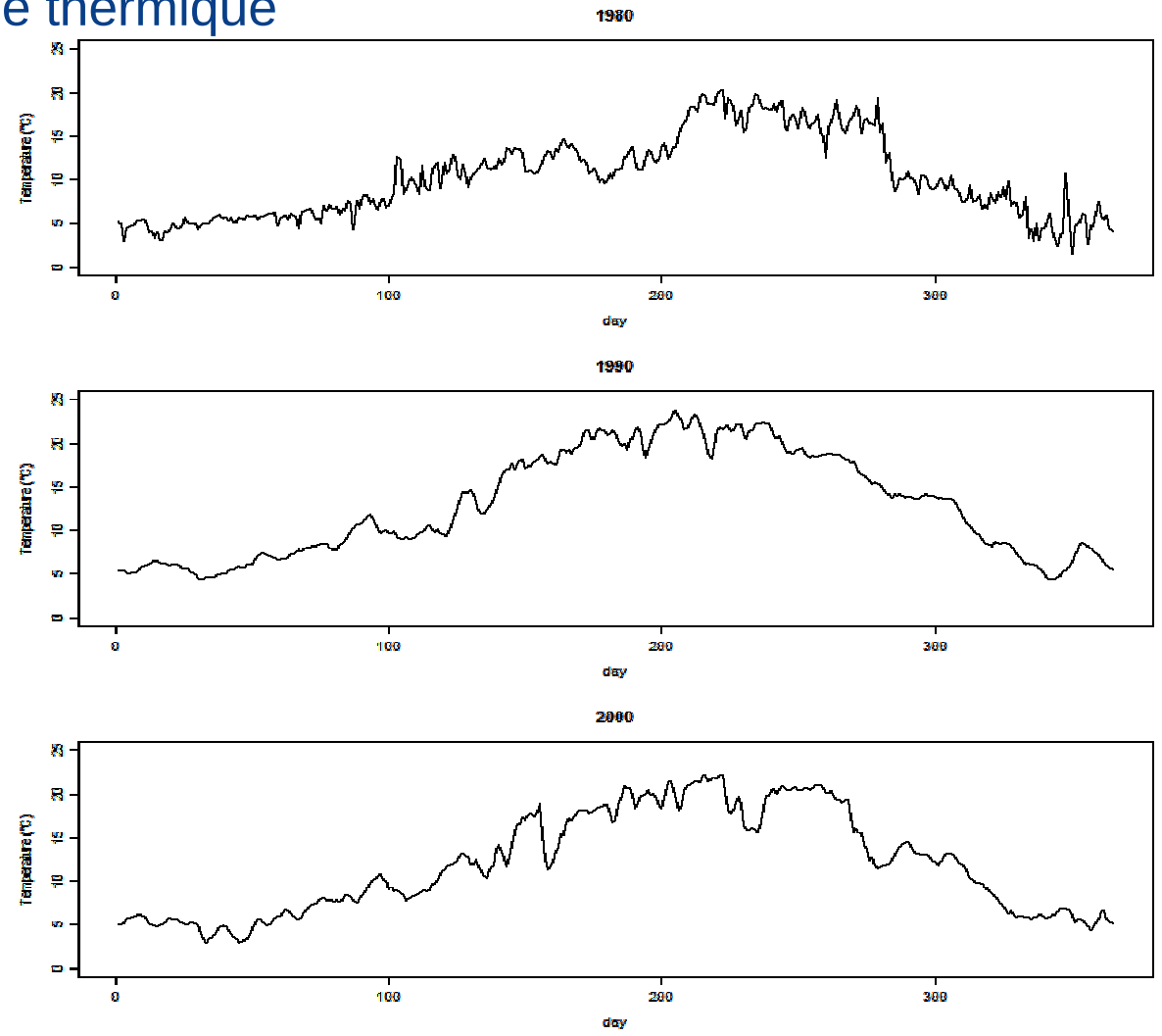
Données biologiques

- 8 espèces
- Synchronisme
- Tau de Kendall
  - Taux de corrélation non-paramétrique
- Cluster
  - Méthode de Ward (minimisation de la variance intra-classe)



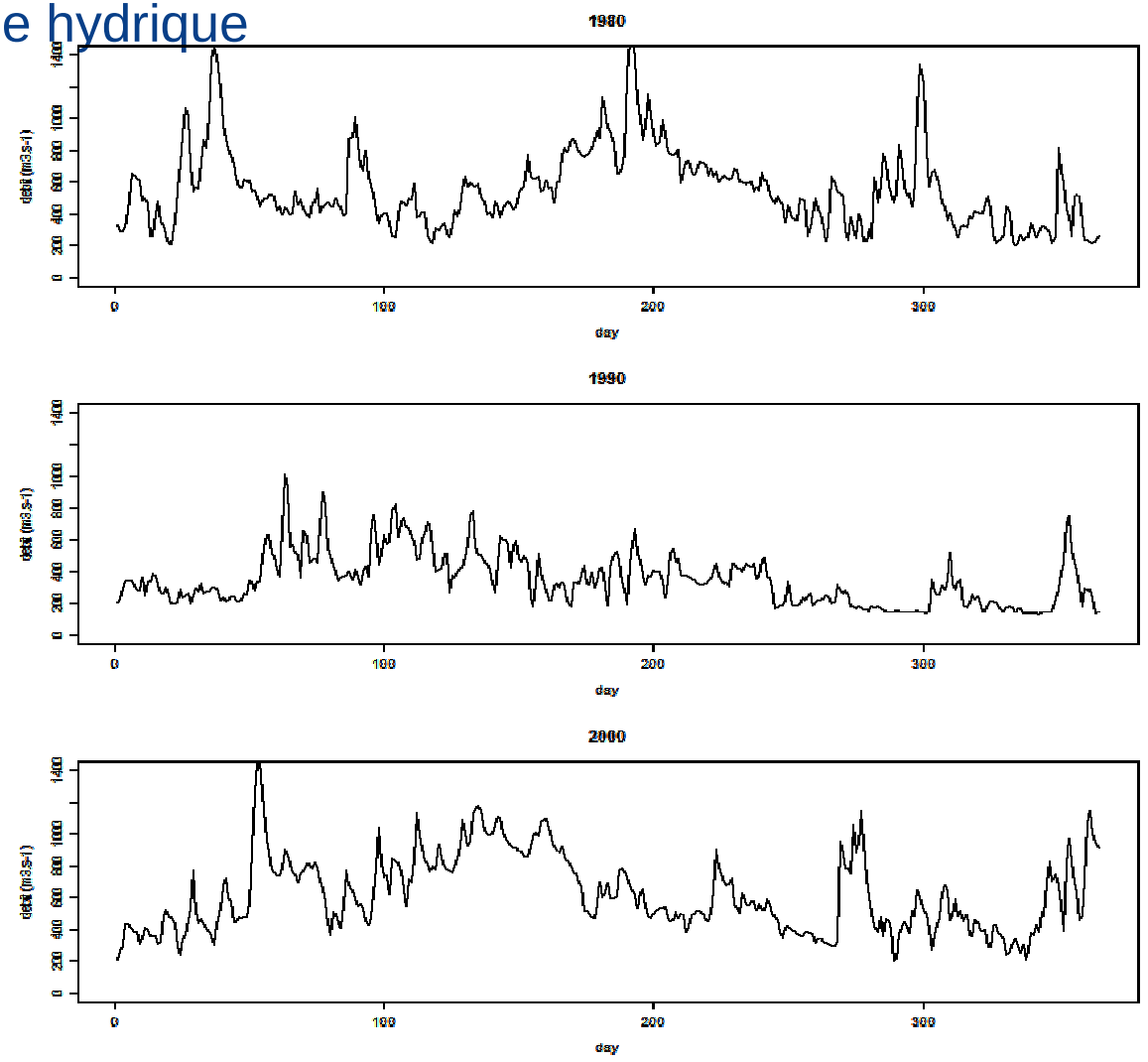
Données environnementales

- Régime thermique



Données environnementales

- Régime hydrique



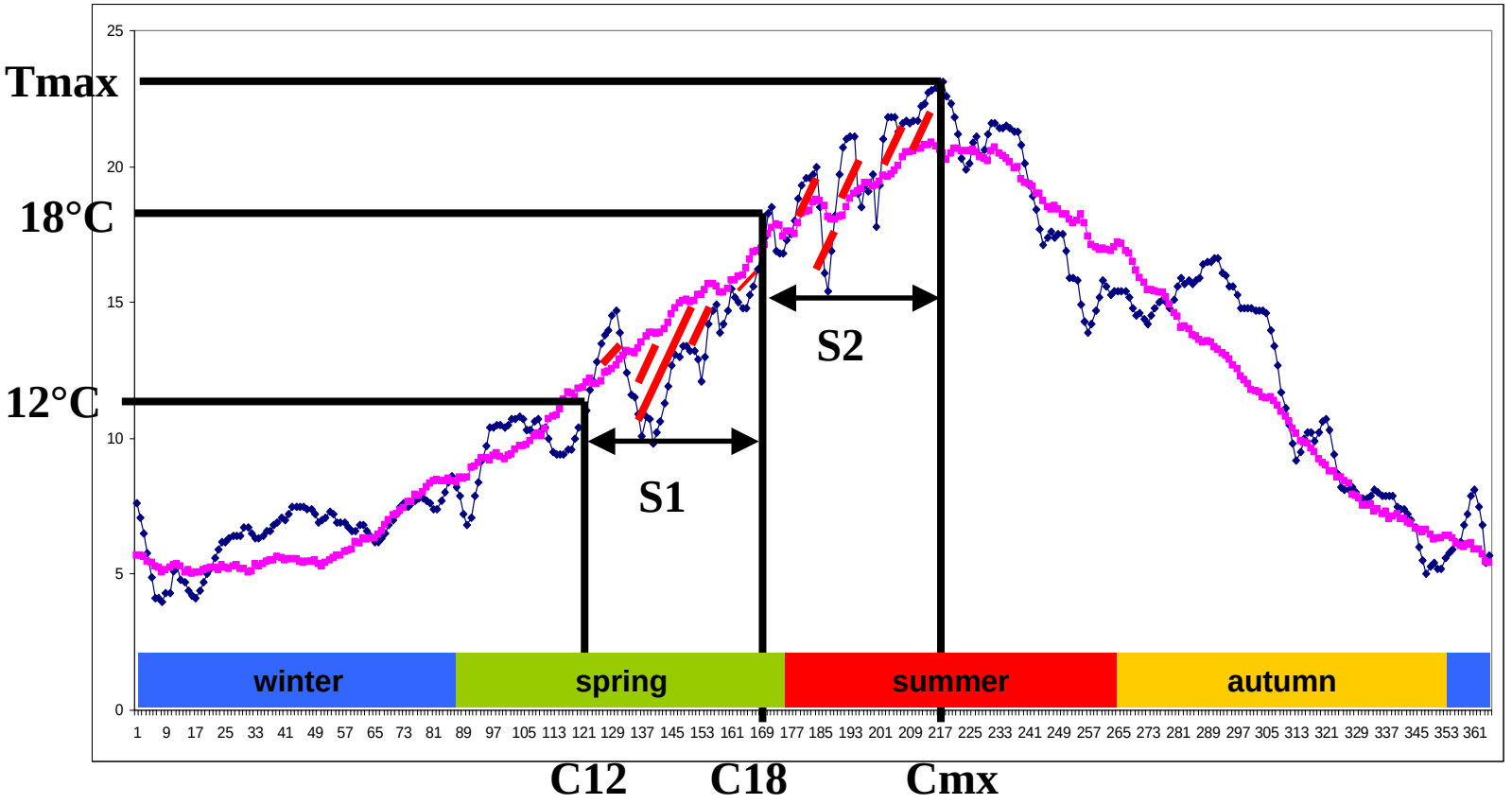
Données environnementales

Variables descriptives synthétiques: thermiques

- Seuils de températures…
  - **12℃** : seuil printanier
  - **18℃** : activité du milieu
  - **Tmax**: température maximale annuelle
- … déterminent 2 fenêtres temporelles

Données environnementales

Variables descriptives synthétiques: thermiques

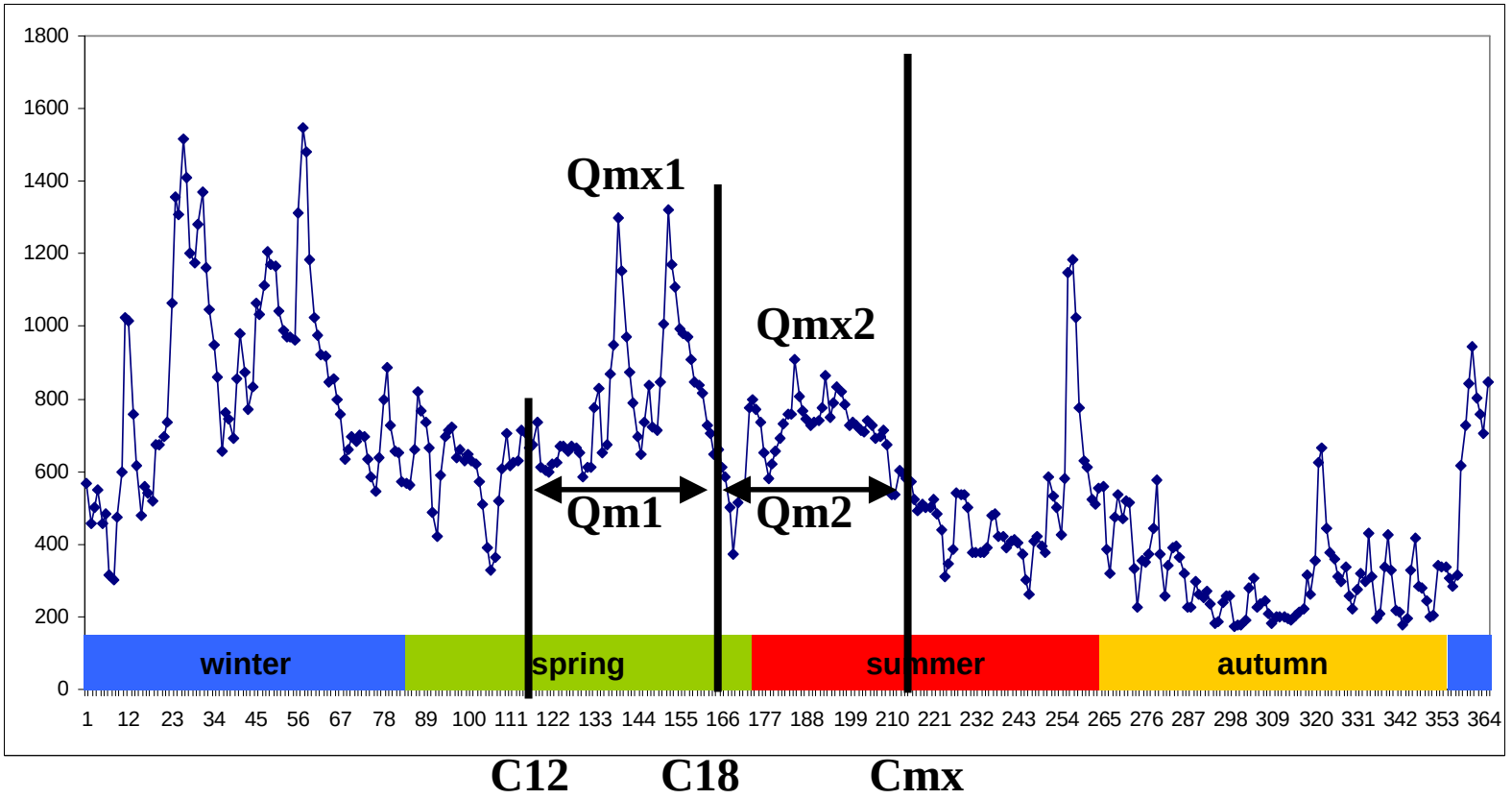


Régime annuel (e.g. 1995)

Régime moyen interannuel

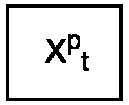
Données environnementales

Variables descriptives synthétiques: débit



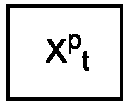
— Régime hydrique (e.g. 1995)

Présentation du modèle hiérarchique : structure

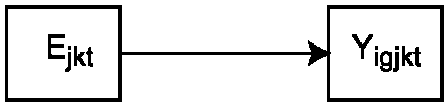


- X^p_t : matrice des régresseurs
  - C12, Cmx, S1, S2, Qm1, Qm2, Qmx2
- Y_{igjkt} : vecteur des observations

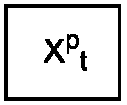
Présentation du modèle hiérarchique : structure



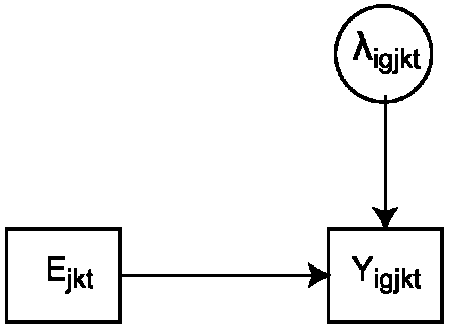
- X^p_t : matrice des régresseurs
- Y_{igjkt} : vecteur des observations
- E_{jkt} : vecteur des temps de pêche



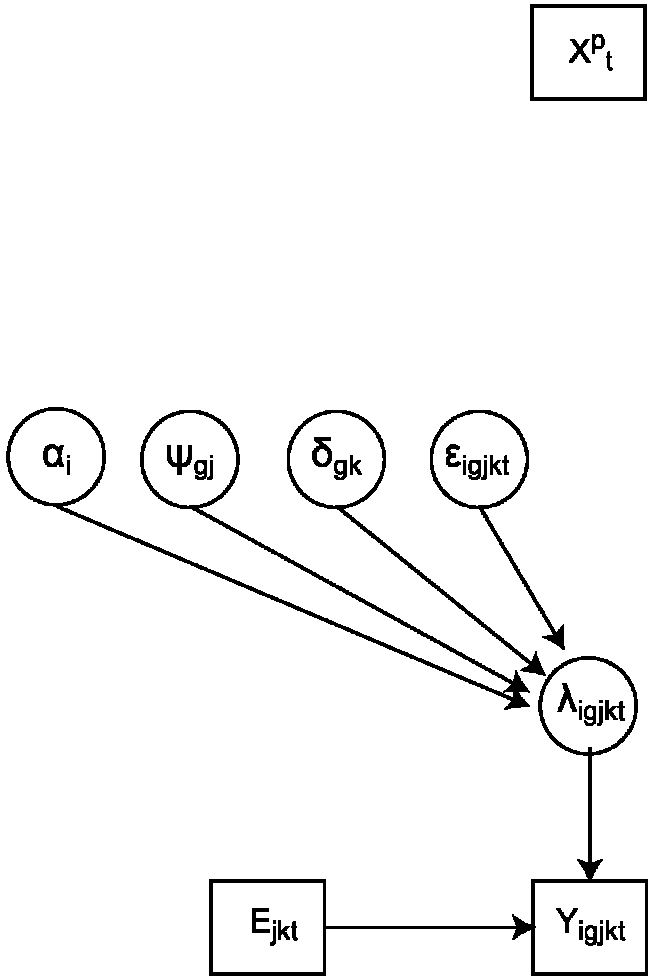
Présentation du modèle hiérarchique : structure



- X^p_t : matrice des régresseurs
- Y_{igjkt} : vecteur des observations
- E_{jkt} : vecteur des temps de pêche
- λ_{igjkt} : vecteur des densités attendues

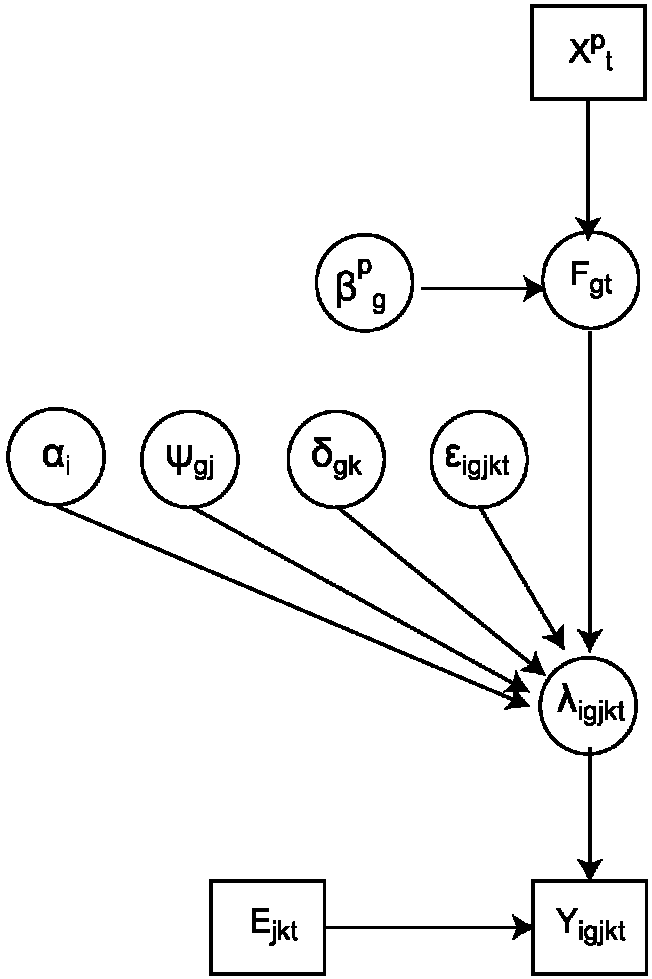


Présentation du modèle hiérarchique : structure



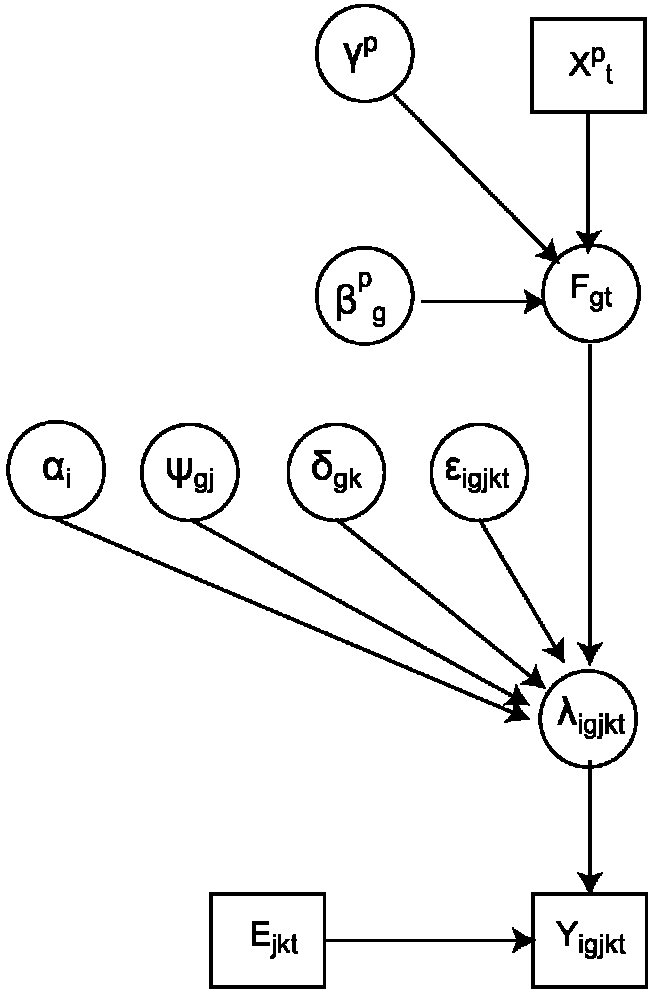
- X^p_t : matrice des régresseurs
- Y_{igjkt} : vecteur des observations
- E_{jkt} : vecteur des temps de pêche
- λ_{igjkt} : vecteur des densités attendues
- Effets fixes : espèce (α), saison (δ), site (ψ)
- ε_{igjkt} : terme d'erreur

Présentation du modèle hiérarchique : structure



- X^p_t : matrice des régresseurs
- Y_{igjkt} : vecteur des observations
- E_{jkt} : vecteur des temps de pêche
- λ_{igjkt} : vecteur des densités attendues
- Effets fixes : espèce (α), saison (δ), site (ψ)
- ε_{igjkt} : terme d'erreur
- β^p_g : vecteur des coefficients de régression
- F_{gt} : Effet de l'environnement de l'année t

Présentation du modèle hiérarchique : structure



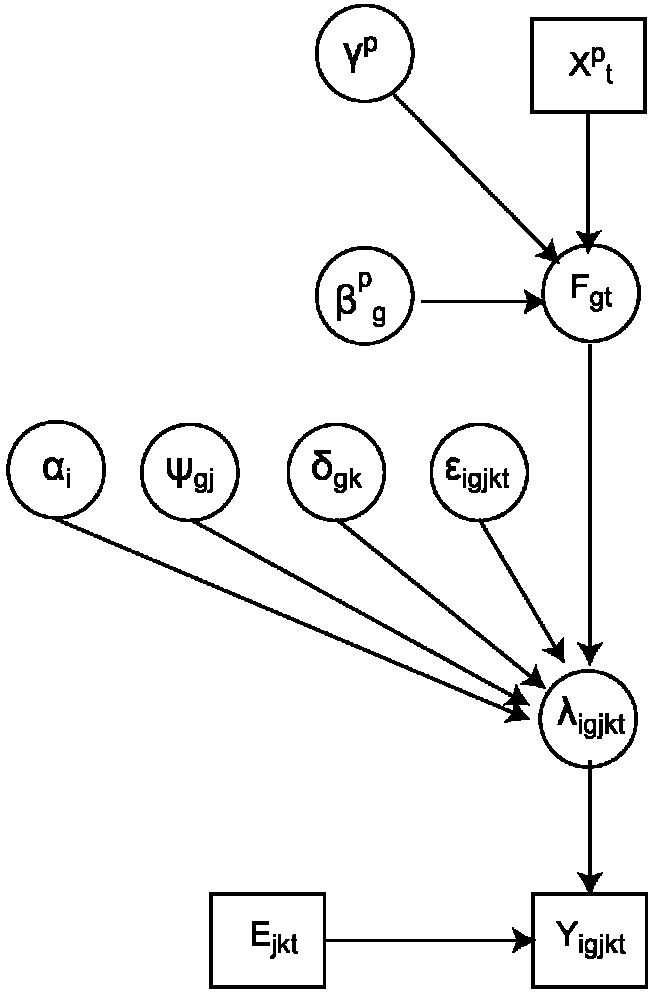
- X^p_t : matrice des régresseurs
- Y_{igjkt} : vecteur des observations
- E_{jkt} : vecteur des temps de pêche
- λ_{igjkt} : vecteur des densités attendues
- Effets fixes : espèce (α), saison (δ), site (ψ)
- ε_{igjkt} : terme d'erreur
- β^p_g : vecteur des coefficients de régression
- F_{gt} : Effet de l'environnement de l'année t
- γ_g^p : indicateur auxiliaire
  - $\gamma^p = 0$, la variable p est absente du modèle
  - $\gamma^p = 1$, la variable p est intégrée au modèle

$$\log(\lambda_{igjkt}) = \alpha_i + \psi_{gj} + \delta_{gk} + F_{gt} + \varepsilon_{igjkt}$$

$$Y_{igjkt} \sim dPois(E_{jkt}.\lambda_{igjkt})$$

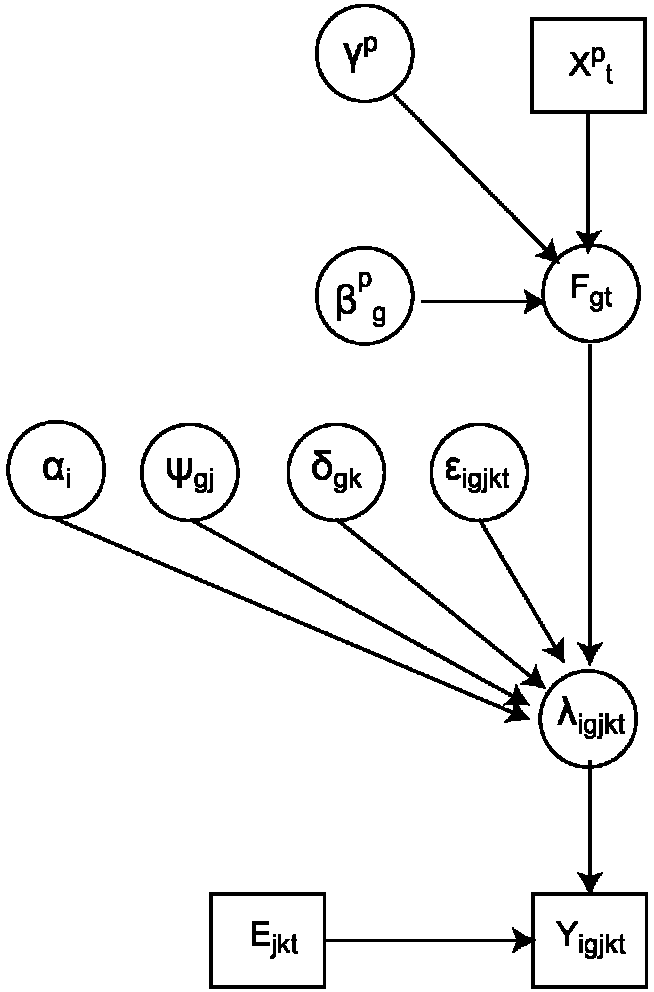
$$F_{gt} = \sum_p \beta_g^p . X_t^p$$

Présentation du modèle hiérarchique : priors



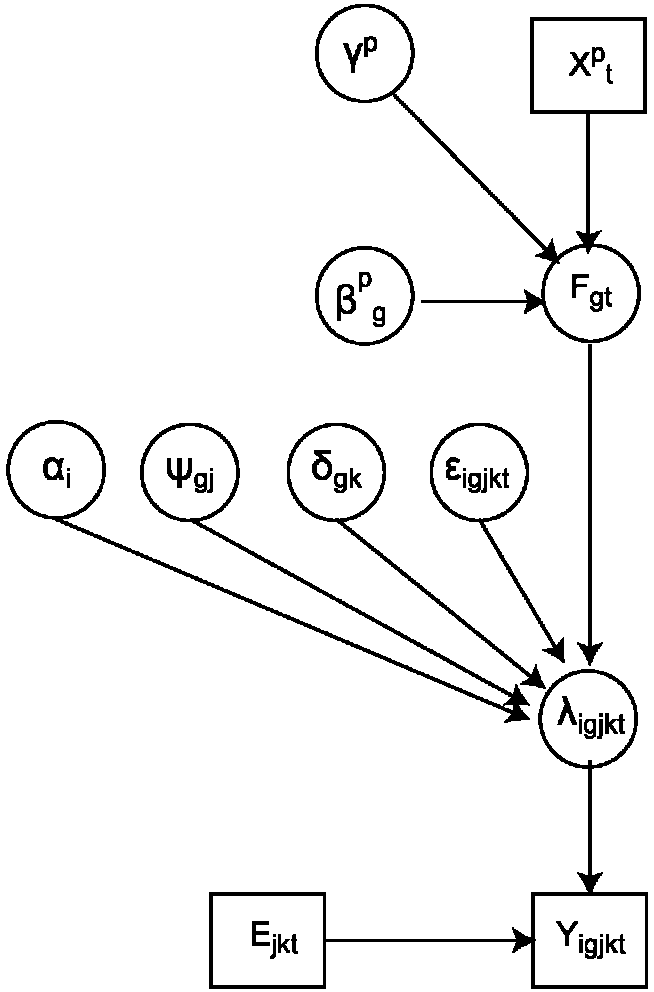
- X^p_t : matrice des régresseurs
- Y_{igjkt} : vecteur des observations
- E_{jkt} : vecteur des temps de pêche
- λ_{igjkt} : vecteur des densités attendues
- Effets fixes : espèce (α), saison (δ), site (ψ)
- ε_{igjkt} : terme d'erreur
- β^p_g : vecteur des coefficients de régression
- F_{gt} : Effet de l'environnement de l'année t
- γ^p : indicateur auxiliaire
  - $\gamma^p = 0$, la variable p est absente du modèle
  - $\gamma^p = 1$, la variable p est intégrée au modèle

Présentation du modèle hiérarchique : priors



- Effets fixes : espèce (α), saison (δ), site (ψ)
  - $\alpha \sim dnorm(0, 10^4)$
  - $\delta \sim dnorm(0, 10^4)$
  - $\psi \sim dnorm(0, 10^4)$
- ε_{igjkt} : terme d'erreur (structure diagonale par cluster)
  - $\varepsilon_{igjkt} \sim dmnorm(0, \Sigma^g)$
  - $(\Sigma^g)^{-1} \sim dwish(Rho, p)$
- γ^p : indicateur auxiliaire
  - $\gamma^p \sim dbern(0.5)$

Présentation du modèle hiérarchique : priors



- Effets fixes : espèce (α), saison (δ), site (ψ)
- ε_{igjkt} : terme d'erreur
- γ^p : indicateur auxiliaire
- β^p_g : vecteur des coefficients de régression
  - Prior « spike and slab »

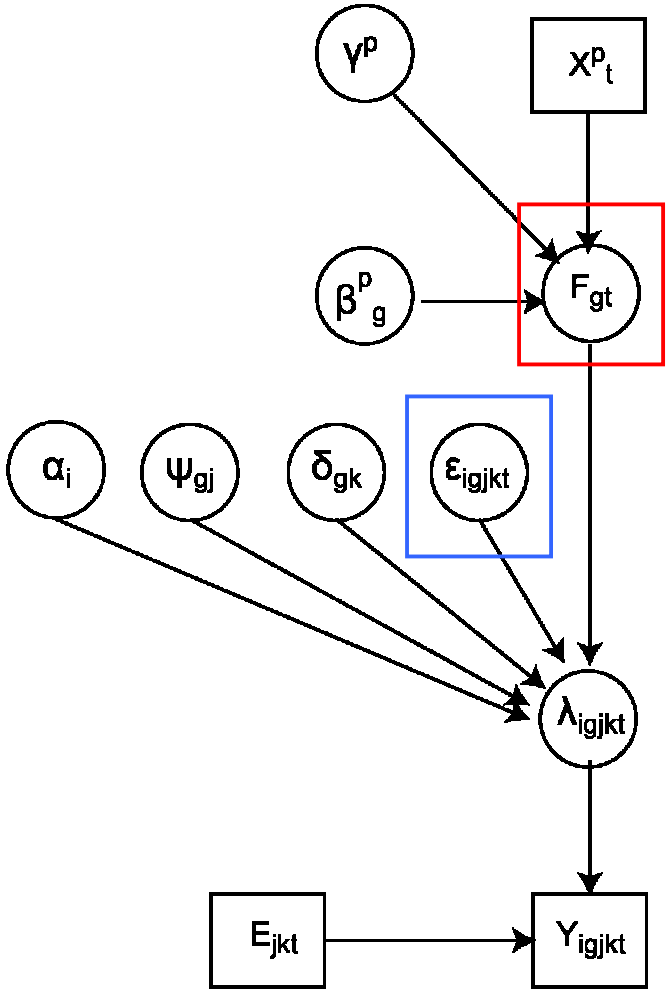
$$\beta_g^p \sim \left(1-\gamma^p\right)'.\pi_{spike}(.)+\left(\gamma^p\right)'.\beta^p.\pi_{slab}(.)$$

  - $\beta^p_g \sim dmnorm(0, B)$

$$B = 10^{-8}.(1-\gamma)+\sigma^{-2}.\left(X'X\right)^{-1}\gamma$$

  - Prior slab de type Zellner
  - Laisse invariant la distribution de X. β
  - $\sigma^{-2} \sim dgamma(10^{-3}, 10^{-3})$
    - Prior Zellner Student

Présentation du modèle hiérarchique : variance



- Variance totale

$$V_{F^g} + \Sigma^g$$

$$V_{F^g} = \sum_p \left(\beta_p^g\right)^2 + 2. \sum_{1 \le p \le p'} \beta_p^g \beta_{p'}^g . corr\left(X^p, X^{p'}\right)$$

- Part de variance expliquée par l'environnement pour l'espèce i

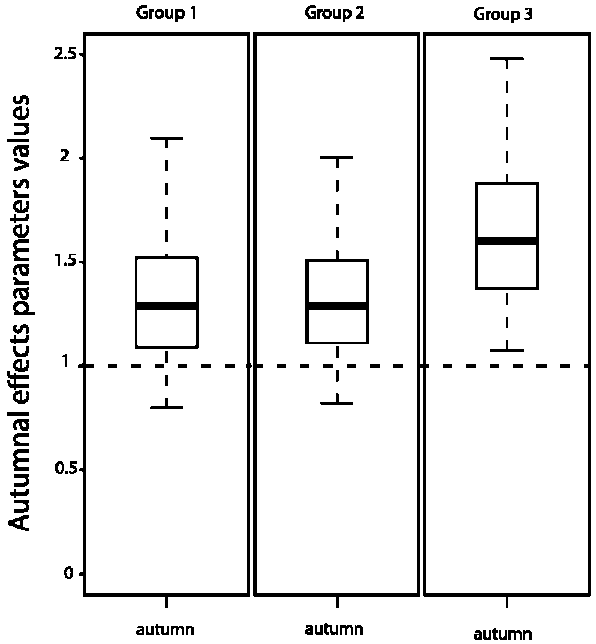
$$r_i = \frac{V_{F^g}}{V_{F^g} + \Sigma_{ii}^g}$$

Résultats

- WinBUGS
  - 10.000 itérations burn-in
  - 20.000 itérations

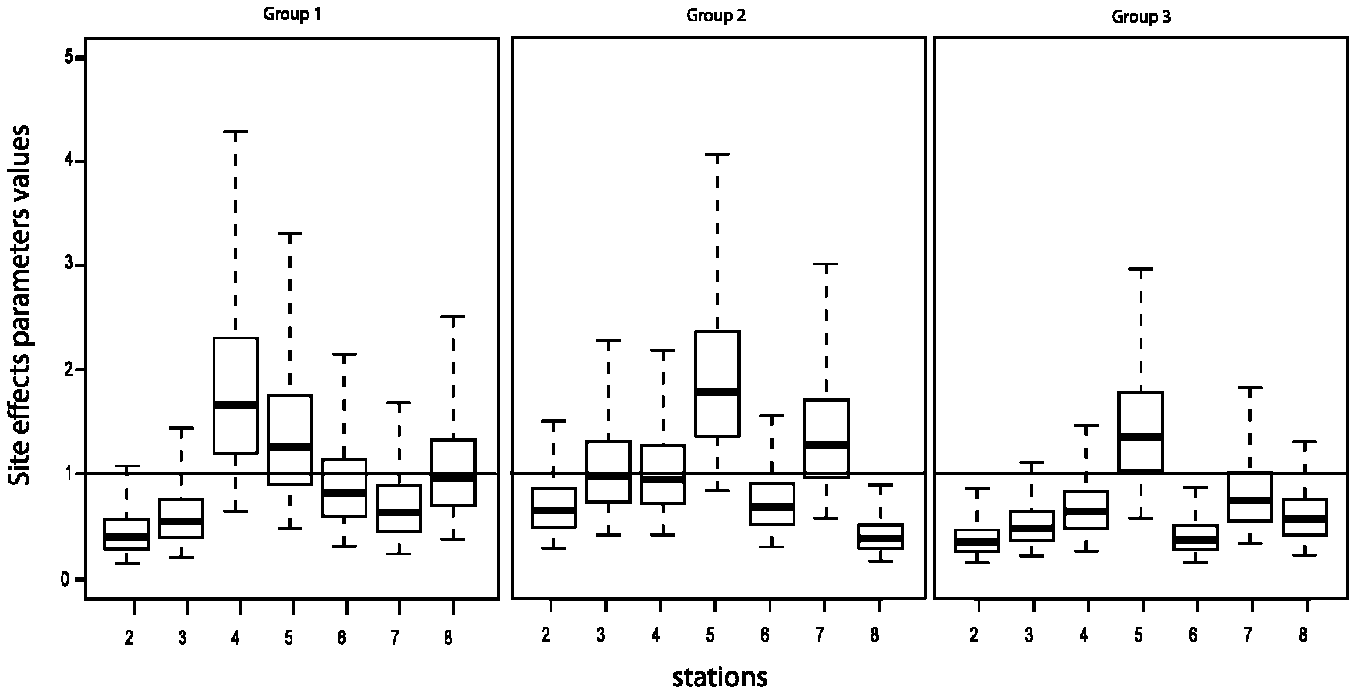
Résultats

- WinBUGS
- Effets saison :
  - Globalement positif



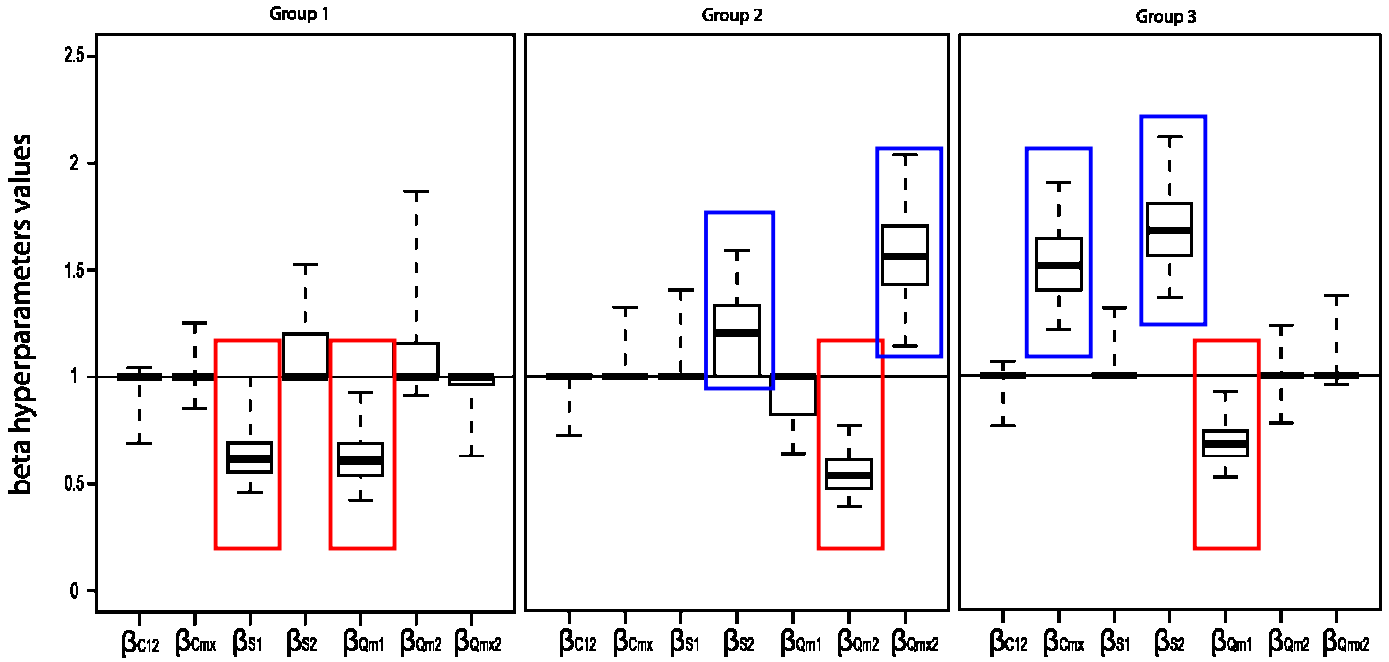
Résultats

- WinBUGS
- Effets saison
- Effets sites
  - Pas de structure particulière



Résultats

- WinBUGS
- Effets saison
- Effets sites
- Effets de l'environnement



Posterior γ

21	21	96	41	99	41	30	
21	15	16	72	39	100	99	
21	100	23	100	98	31	28	

Résultats

- WinBUGS
- Effets saison
- Effets sites
- Effets de l'environnement
- Variance expliquée par l'environnement:

Espèce	R_i (%)
Ablette	4.6
Vandoise	5.0
Barbeau	11.8
Goujon	9.5
Hotu	6.4
Chevesne	14.1
Gardon	7.4
Spirlin	9.2

Conclusions

- Procédure de choix de modèle intégrée adaptée au GLM
  - Recherche stochastique des régresseurs
  - Priors spike / slab
  - Adaptée aux modèles de grande dimension
  - Fonctionne bien même lorsque les variables expliquent une faible part de la variance
- Temps de calcul élevés
  - Surtout la matrice de corrélation X'X
- Améliorations possibles:
  - Effet site aléatoire
  - Interaction environnement * espèce
  - Modèles non-linéaires : autre niveau de complexité
    - GAM, splines, …