Importance de la structure mathématique des modèles statistiques paramétriques pour leur qualité inférentielle et prédictive

Frédéric Gosselin

Présentation donnée par visio lors des journées d'animation du CESCO (café scienti) le 22/03/2021

Importance de la structure mathématique des modèles statistiques paramétriques pour leur qualité inférentielle et prédictive

Frédéric GOSSELIN

INRAE – Nogent sur Vernisson, France

frederic.gosselin@inrae.fr

* Ici, nous ne parlerons que de techniques statistiques basées sur **des modèles**

= *Statistiques paramétriques*

* On suppose que la (densité de) probabilité de la variable à expliquer *y* a une forme précise $g(y|\theta, x)$, qui dépend de θ, le ou les paramètres du modèle statistique, et des variables explicatives *x*.

* A partir d'observations y connues, $g(y|\theta, x)$ définit un **modèle statistique paramétrique.**

on veut estimer θ à partir de $g(y|.,x)$

↳ on va estimer la distribution de probabilité de θ compte tenu de y et x

ou ses caractéristiques
(Théorème Central Limite)

↳ Modèle statistique paramétrique = « inverse » d'un modèle probabiliste

$y \rightarrow \theta$ $\theta \rightarrow y$

Intérêt des modèles statistiques paramétriques

* Prise en compte des multiples « nuisances » dans la relation entre observations et hypothèse (niveau de variation aléatoire, dépendance entre observations – pseudo-réplication –, non-normalité, non-linéarité, dépendance d'une co-variable …)

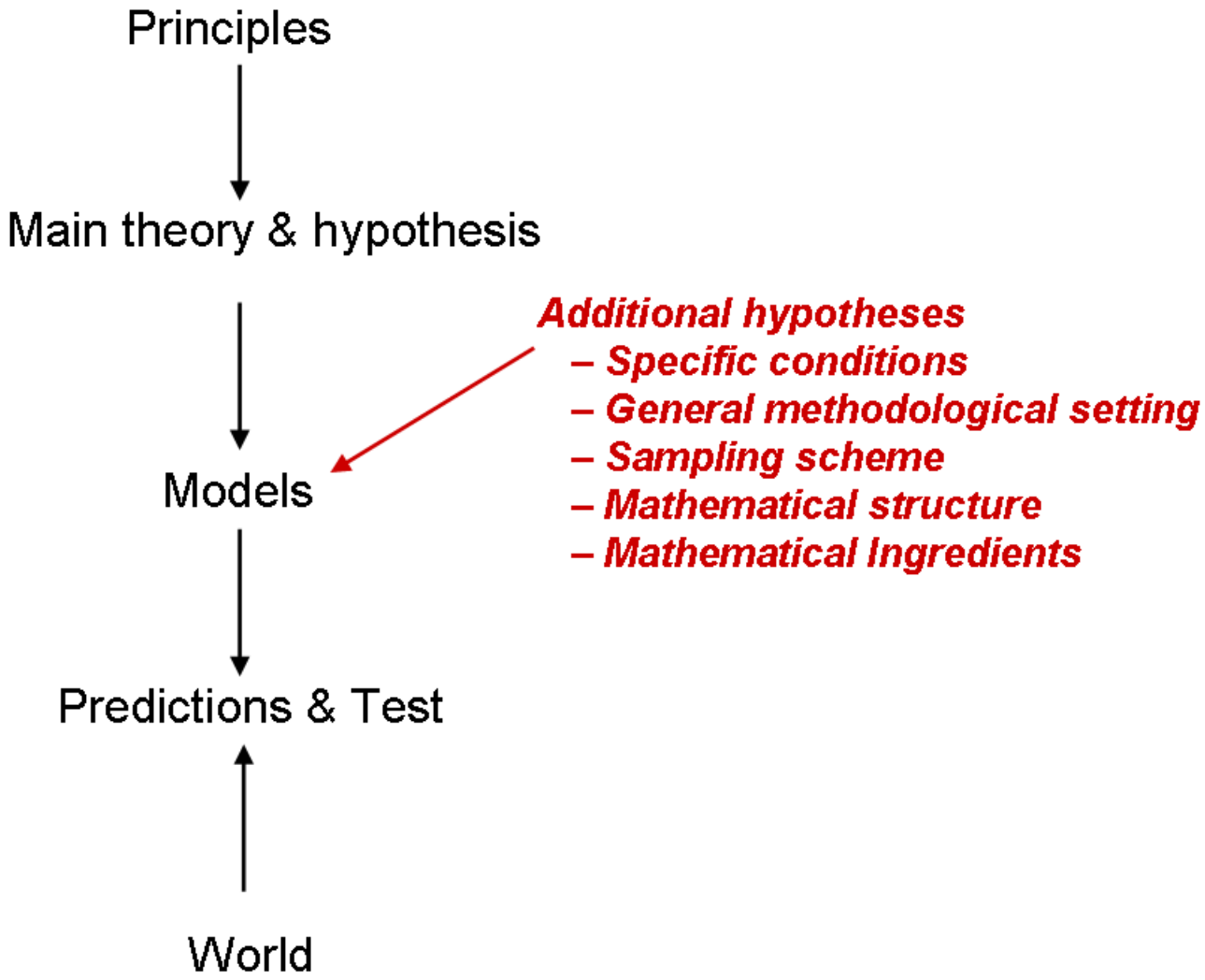
* Richesse de modes d'analyse (tests d'hypothèse principale et auxiliaire, estimation des « effets », comparaison de modèles, prédire de nouvelles observations…)

Limites des modèles statistiques paramétriques

* Hypothèses – parfois – supplémentaires par rapport aux tests non-paramétriques (« normalité », linéarité, indépendance, …)

↳ On parle souvent d'hypothèses auxiliaires pour les distinguer de l'hypothèse principale (écologique…).

* Limites de tout modèle : représente-t-il bien la réalité?


Principles
Main theory & hypothesis
Models
Predictions & Test
World
Additional hypotheses
– Specific conditions
– General methodological setting
– Sampling scheme
– Mathematical structure
– Mathematical Ingredients

Hypothèses mathématiques classiques

* Principales hypothèses étudiées

  – la « Sainte Trinité » (statistique) :

    * **normalité** des résidus
    * **indépendance** des résidus
    * **homogénéité de la variance** des résidus

  – **linéarité** : les variables explicatives x sont reliées linéairement à la variable à expliquer y

  – **x déterministes** : les variables explicatives x n'ont pas d'erreur de mesure

Buts de la présentation

↳ Illustrer l'importance d'hypothèses auxiliaires statistiques pour la qualité de l'inférence ou de la prédiction

- **Distribution de probabilité** des observations
- **Dépendance** des observations
- **Non-linéarité** de la relation entre variables explicatives et observations

Distributions de probabilité

- Un problème maintenant relativement bien maitrisé en écologie: **données de comptage plus dispersées** que la distribution de Poisson (ou distribution binomiale)

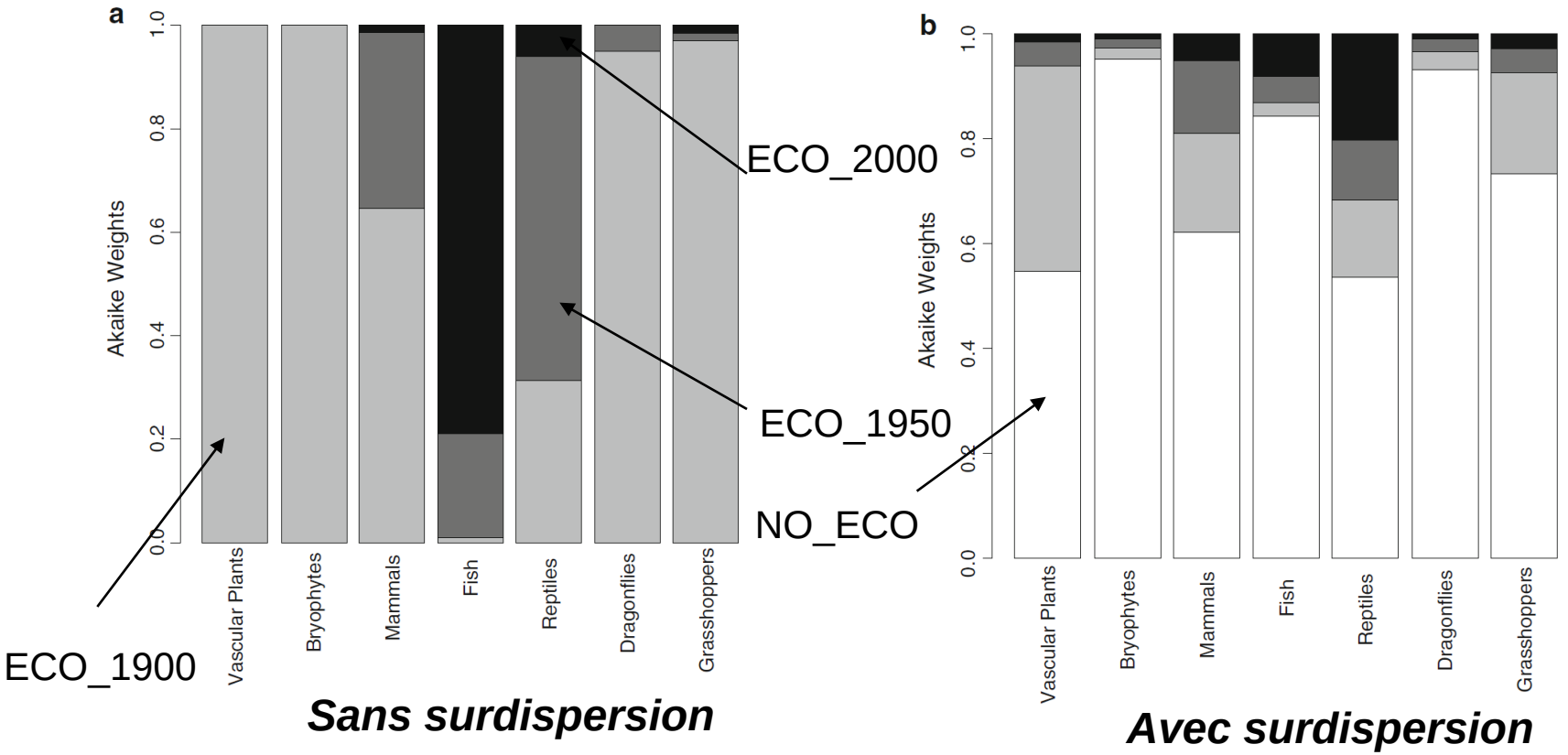
Test	% rejected of 1000
ANOVA	4.9
Poisson regression	12.1
PR overdispersion	36.2
$\{k, m_v\}$ vs. $\{k, m\}$	5.5
$\{k_v, m_v\}$ vs. $\{k_v, m\}$	4.6
$\{k_v, m_v\}$ vs. $\{k, m\}$	4.2

Erreur de type 1 pour données sur-dispersées générées par une négative binomiale

White & Bennetts (1996) *Ecology*

Distributions de probabilité

- Autre exemple avec **distribution binomiale**



Lien proportion d'espèces menacées et paramètres « économiques » de pays

Gosselin (2015) *Biodiv. & Conserv.*

Distributions de probabilité

- Problème lié à hypothèse forte sur le lien variance-moyenne avec Poisson:

$$Var(\mu) = \mu$$

- Résolu pour le sur-dispersé par ajout d'un effet aléatoire « observation » (Poisson-lognormal), l'utilisation d'une quasi-vraisemblance ou de la distribution négative binomiale:

$$Var(\mu) = \frac{1}{p}\ \mu$$

avec $0<p<1$.

Distributions de probabilité

- Beaucoup moins d'outils probabilistes pour **sous-dispersion**

  - Problème très peu pris en compte en écologie. Certains auteurs réticents à envisager de la sous-dispersion (Burnham & Anderson 2002, p.69)

  - Pourtant, des processus écologiques peuvent rendre les processus ponctuels plus réguliers que le processus de Poisson (ex: territorialité)

Distributions de probabilité

Definition	Mean	Variance
Species Richness of Forest Bryophytes	10.1	3.4
Species Richness of Forest Herbs	4.0	4.7
Species Richness of Peri-Forest Herbs	3.0	8.1
Species Richness of Non-Forest Herbs	2.1	7.2
Species Richness of Forest Woody species	4.7	3.1
Species Richness of Peri-Forest Woody species	4.6	3.9

Barbier et al. (2009) Forest Ecol. Manage.

Distributions de probabilité

Groupe écologique	Moyenne (± écart-type) coefficient G Distribution Sur/Sous-disp	Moyenne (± écart-type) coefficient G Négative Binomiale
Bryophytes forestières	0.0060 (± 0.0037)	0.0060 (± 0.0070)
Herbacées forestières	-0.0236 (± 0.0099)	-0.0241 (± 0.0119)
Herbacées péri-forestières	-0.0799 (± 0.0250)	-0.0799 (± 0.0246)
Herbacées non-forestières	-0.1332 (± 0.0329)	-0.1331 (± 0.0333)
Ligneuses forestières	-0.0120 (± 0.0084)	-0.0116 (± 0.0103)
Ligneuses péri-forestières	-0.0346 (± 0.0069)	-0.0341 (± 0.0103)

Gosselin et al. (In Prep.)

Distributions de probabilité

- Besoin d'élargir la gamme de distributions de probabilités disponibles:
  - Ici pour données de comptage sous-dispersées
  - Pour bien d'autres domaines, par ex: données de classes d'abondance-dominance pour la flore :

Ecological Informatics 26 (2015) 18–26

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Analyzing plant cover class data quantitatively: Customized zero-inflated cumulative beta distributions show promising results

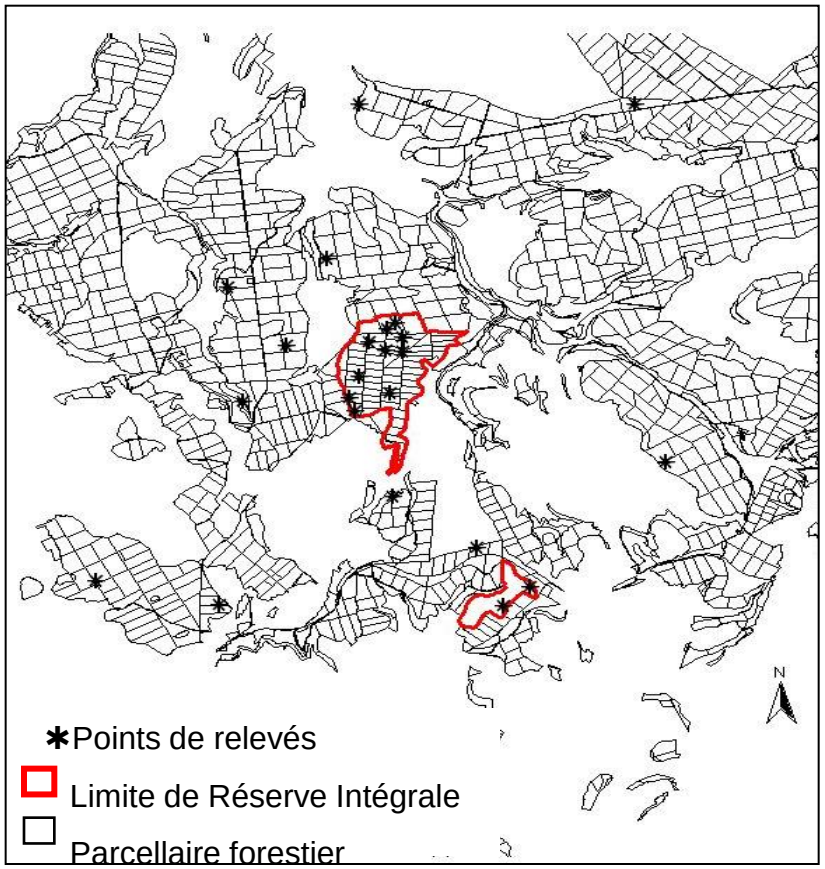
Basile Herpigny, Frédéric Gosselin *

IRSTEA, UR EFNO, Centre de Nogent-sur-Vernisson, Domaine des Barres, F-45290 Nogent-sur-Vernisson, France

Dépendance (entre observations)

Motivation initiale

Projet GNB (Gestion, Naturalité, Biodiversité)



Dépendance (entre observations)

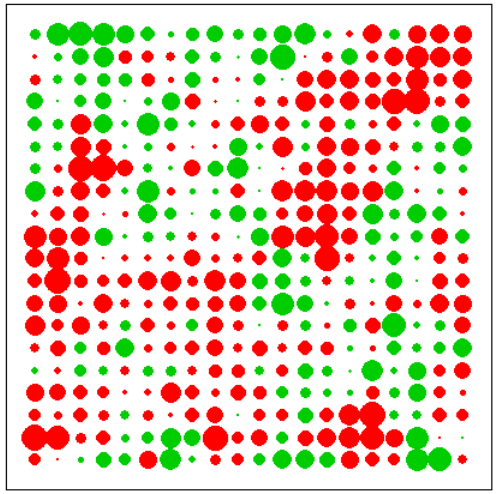
Motivation initiale

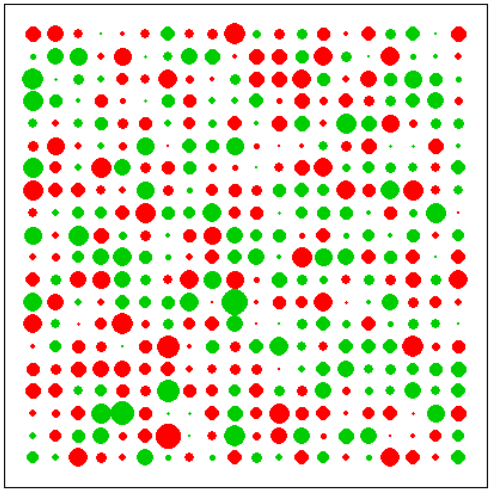
(1) Contexte GNB (appliqué) : données GNB guettées par pseudo-réplication (comme la plupart des données publiées sur le sujet)

(2) Contexte académique: débat en écologie sur l'intérêt d'utiliser des modèles spatialement explicites (Bini et al. 2009 Ecography, Hawkins 2012 J. Biogeog. Vs Beale et al. Ecol. Let. 2010)

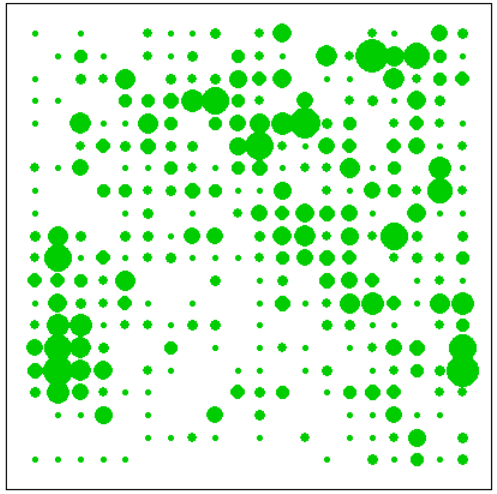
L’autocorrélation spatiale

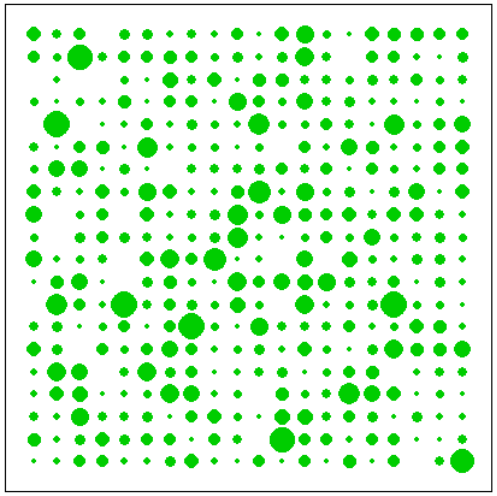
$\mathcal{N}$





$\mathcal{P}$





L'approche par simulation

$$C_{i,j} = \exp\left(-\frac{dist_{i,j}}{\phi}\right)$$

$$\begin{cases} \mathbb{X} = \left(1, X^{(1)}, \ldots, X^{(P)}\right) \\ X^{(p)} \sim \mathcal{N}\left(0, \sigma^2 C\right) \\ \boldsymbol{\beta} = (\beta_0, \ldots, \beta_p) \text{ fixé} \end{cases}$$

$$\begin{cases} \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \\ \boldsymbol{\epsilon} \sim \mathcal{N}\left(0, \sigma^2 C\right) \end{cases}$$

$\mathbb{X}\boldsymbol{\beta}$

Champ spatial des effets fixes

$\boldsymbol{\epsilon}$

Champ spatial des erreurs

$$\boldsymbol{\eta} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

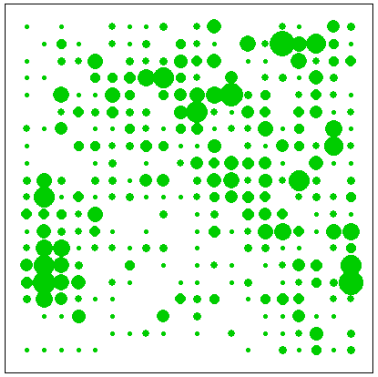
Prédicteur linéaire

$$\forall i \in \{1, \ldots, n\} : Y_i \sim \mathcal{P}(\exp \eta_i)$$

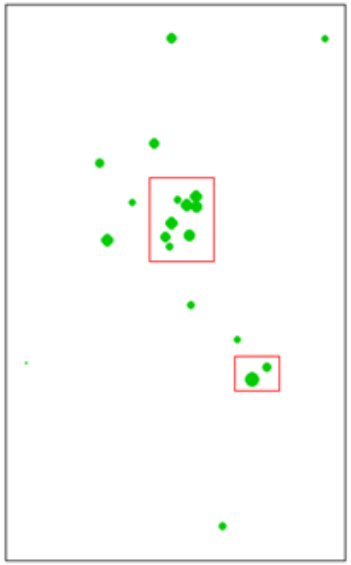
Les différents scénarios

- 16 scénarios de 2 types :

  1) 10 scénarios sur grille : observations **régulièrement** espacées



  2) 6 scénarios sur placettes GNB : observations **irrégulièrement** espacées



zone non-gérée

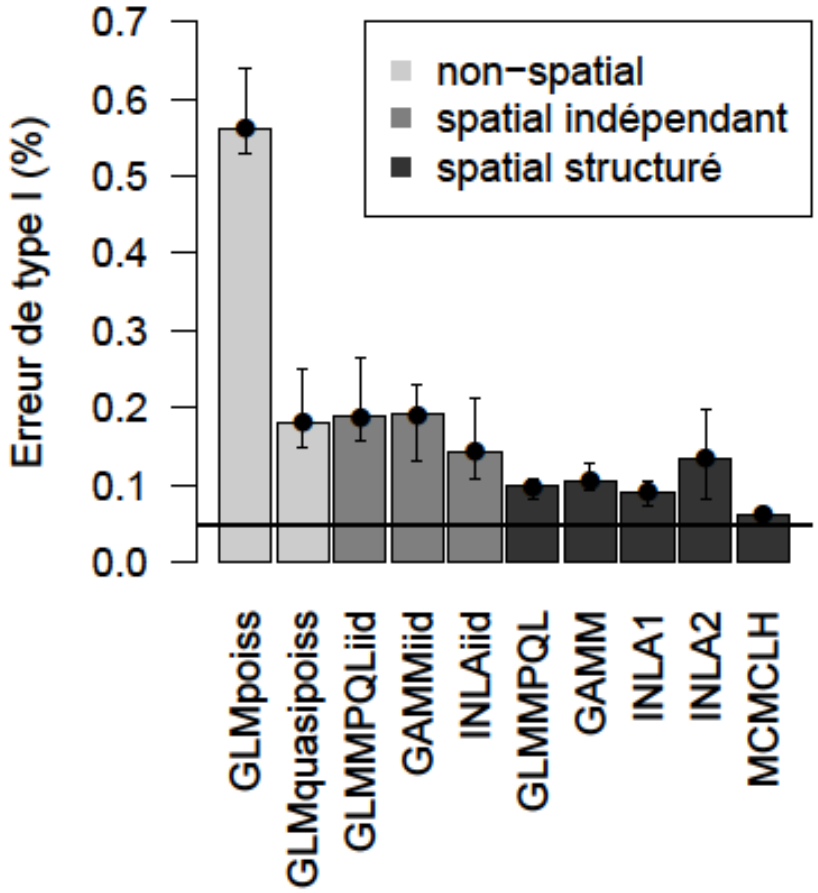
- Scénarios non-stationnaires :

  - portée spatiale non-stationnaire
  - tendances spatiales linéaires

Synthèse des méthodes comparées

Abréviation	Description	Inférence	Classification
GLMpoiss	GLM	fréquentiste (Newton-Raphson)	non-spatial
GLMquasipoiss	GLM Quasi-Poisson	fréquentiste (QML)	non-spatial
GLMMPQLiid	GLM mixte	fréquentiste (PQL)	spatial indépendant
GAMMiid	GAM mixte	fréquentiste (PQL)	spatial indépendant
INLAiid	GLM mixte	bayésienne (INLA)	spatial indépendant
GLMMPQL	GLM mixte	fréquentiste (PQL)	spatial structuré (distances)
GAMM	GAM mixte	fréquentiste (PQL)	spatial structuré (distances)
INLA	GLM mixte	bayésienne (INLA)	spatial structuré (voisinages)
MCMCLH	GLM mixte	bayésienne (MCMCLH)	spatial structuré (distances)

Résultat scénario GNB.4



Synthèse des résultats (I)

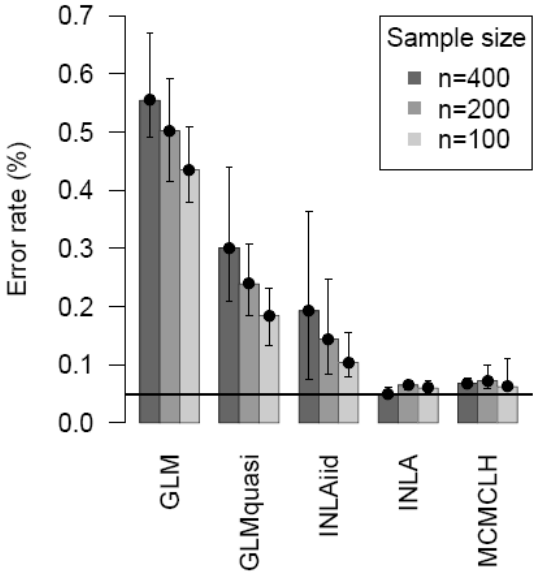
- Spatial structuré > spatial indépendant > non-spatial

- Différences entre méthodes spatiales structurées :
  - Modélisation spatiale : distance > voisinage
  - Inférence : bayésienne MCMC > fréquentiste PQL

- Meilleure méthode : méthode spatiale basée sur les distances avec une inférence bayésienne adaptative MCMCLH

Synthèse des résultats (II)

- **Stabilité des méthodes spatiales (e.g. taille échantillon N) vs problèmes des méthodes non spatiales ne disparaissant pas à N grand**

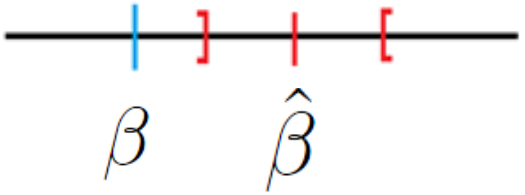


Scenario GRID.4
Saas & Gosselin (2014) *Ecography*

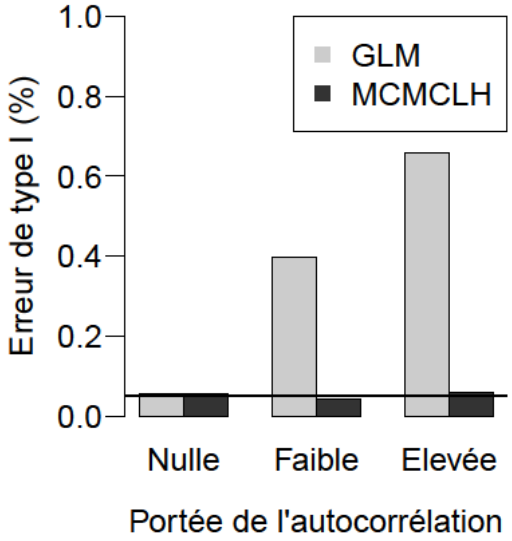
Pourquoi modéliser le spatial ? (1)

1) Sous-estimation de la variance des estimateurs

Intervalles de confiance trop étroits



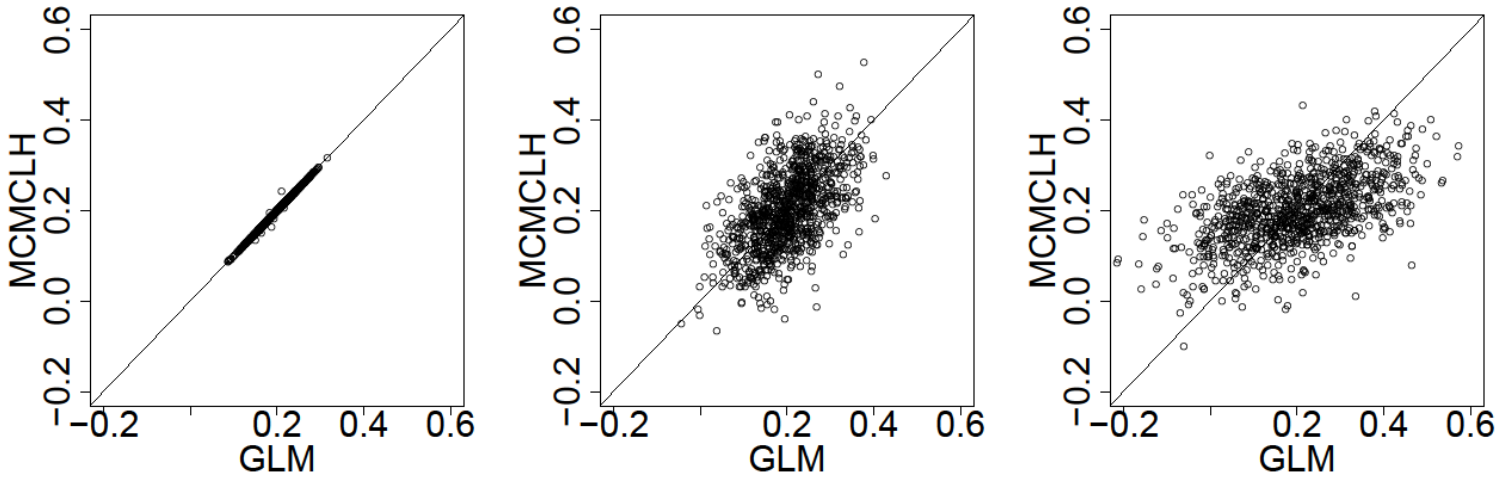
↳ **Erreurs de type I trop élevées**



↳ **Choix de modèles trop peu parcimonieux**

Pourquoi modéliser le spatial ? (2)

2) Estimation imprécise des effets fixes



Augmentation de la portée spatiale

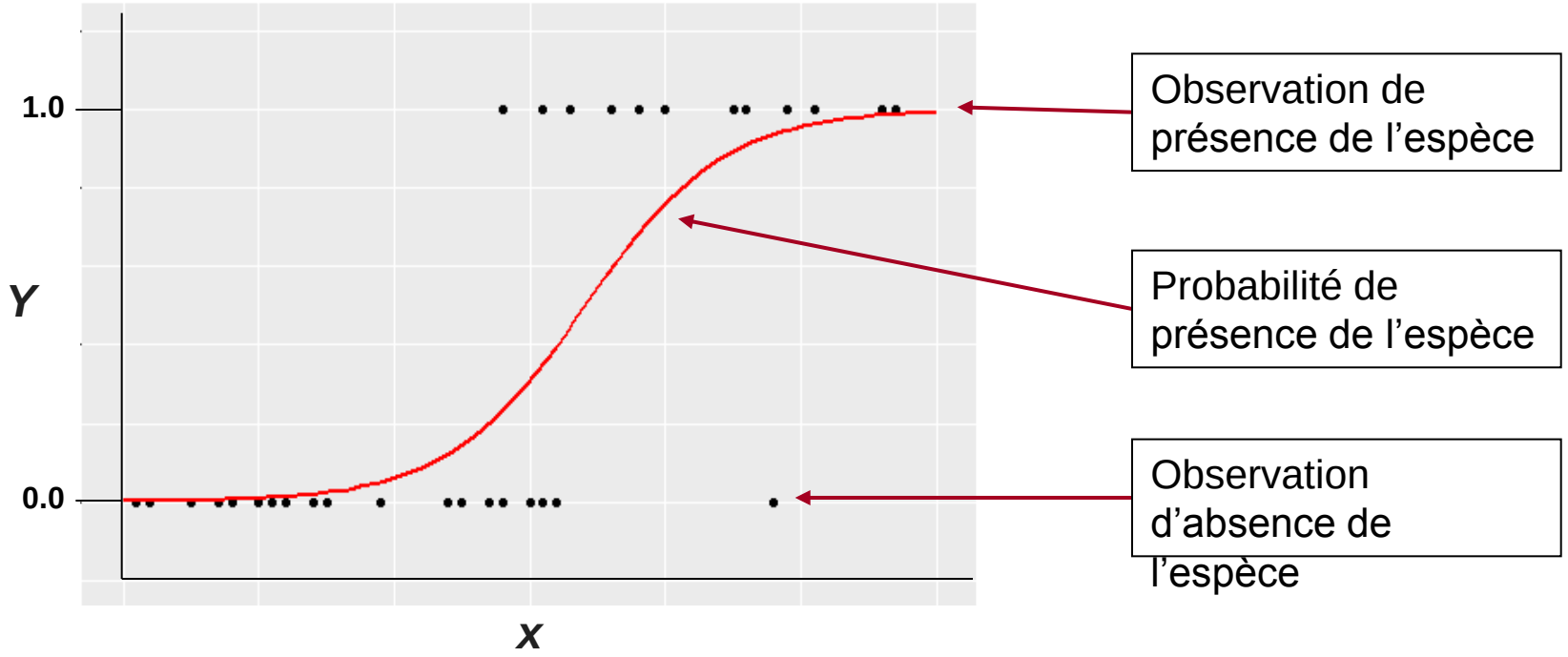
↳ Désaccord avec Bini et al. (2009) et Hawkins (2012): confondent biais et précision et biais et "random shift"

Saas & Gosselin (2014) *Ecography*

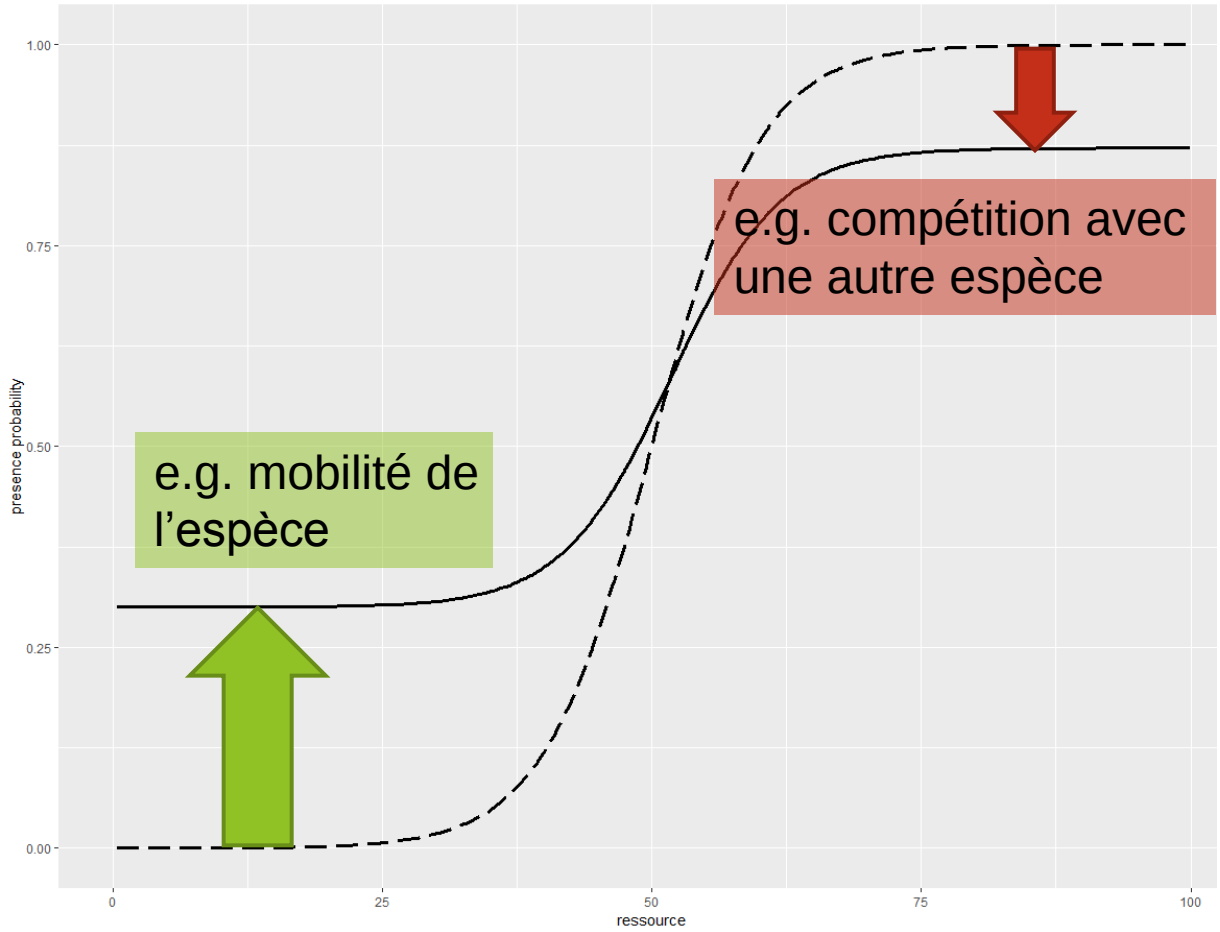
Non-linéarité: fonctions de liens sigmoïdes

Contexte: SDMs sur données binaires

- Modélise la distribution d'une espèce dans l'espace (géographique ou écologique)
- Repose sur des données d'abondance ou **présence/absence**



SDM: exemple de mécanismes



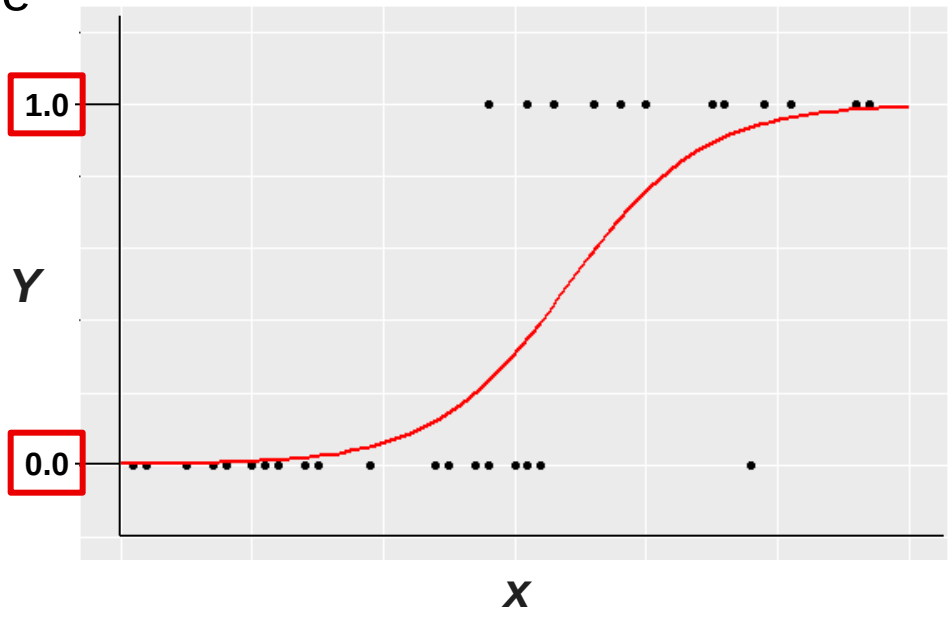
Logistique classique

Logistique avec asymptotes estimées

Données binaires: modèle canonique: GLM

- Approche de référence: GLM avec distribution de Bernoulli
- Logit: Fonction de lien canonique → issue de l'extension du modèle linéaire (LM)
- Inverse logit → logistique classique

  - Asymptote basse = 0
  - Asymptote haute = 1
  - → **Forte hypothèse auxiliaire**



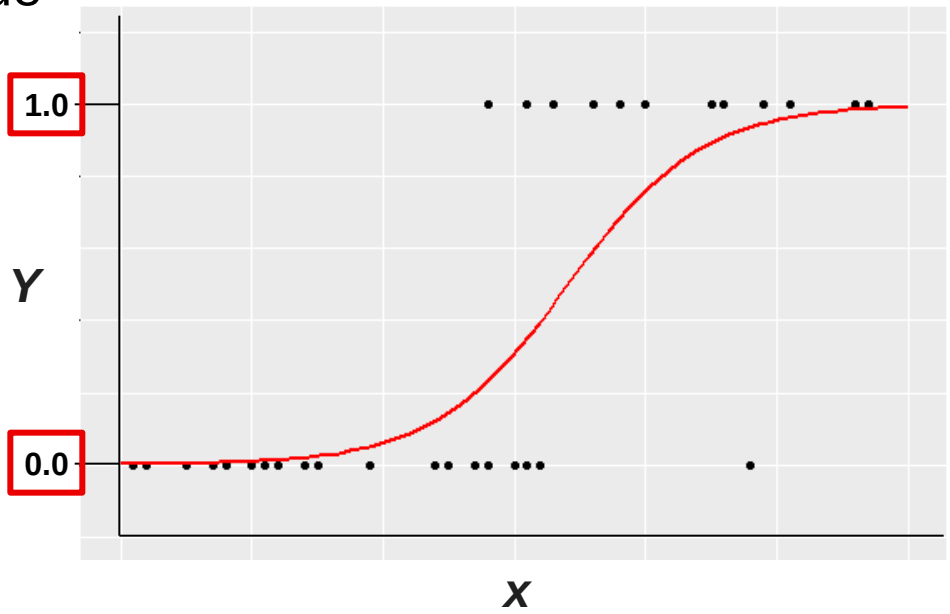
Données binaires: modèle canonique: GLM

- Approche de référence: GLM avec distribution de Bernoulli
- Logit: Fonction de lien canonique → issue de l'extension du modèle linéaire (LM)
- Inverse logit → logistique classique

  ↳ Asymptote basse = 0

  ↳ Asymptote haute = 1

  → **Forte hypothèse auxiliaire**



↳ Test par simulations des impacts de cette hypothèse auxiliaire

Données simulées

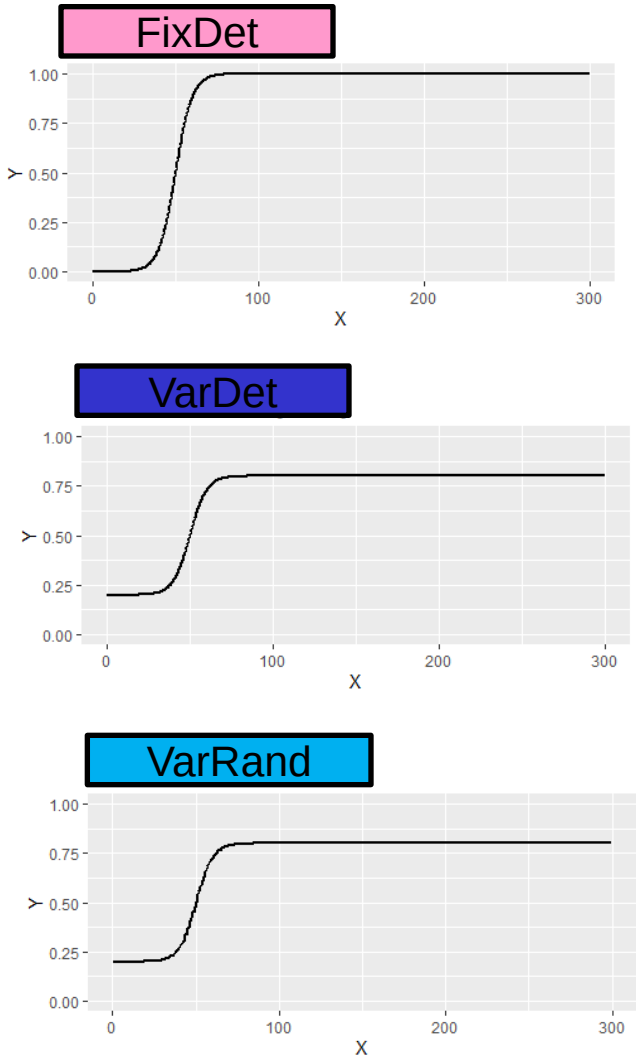
3 Scenarios univariés :

- 2 Scénarios : 10,000 jeux de données
- 1 Scénario (VarRand) : 1,000 jeux de données

Paramètres fixes :
- **ip** (point d'inflexion) = 50
- **sl** (pente) = 0.2

Paramètres variables entre Scénarios :
- FixDet : asymptotes fixées à 0 et 1
- VarDet : asymptotes variables
- VarAléa : asymptotes variables et données aléatoirement distribuées sur le gradient

Paramètres variables entre datasets au sein d'un Scénario :
- **Nobs** variable (entre ≈403 et ≈2981)
- **L et K** aléatoires dans les gammes définies



NSL (pente normalisée) = sl * (K – L)

Analyse des simulations

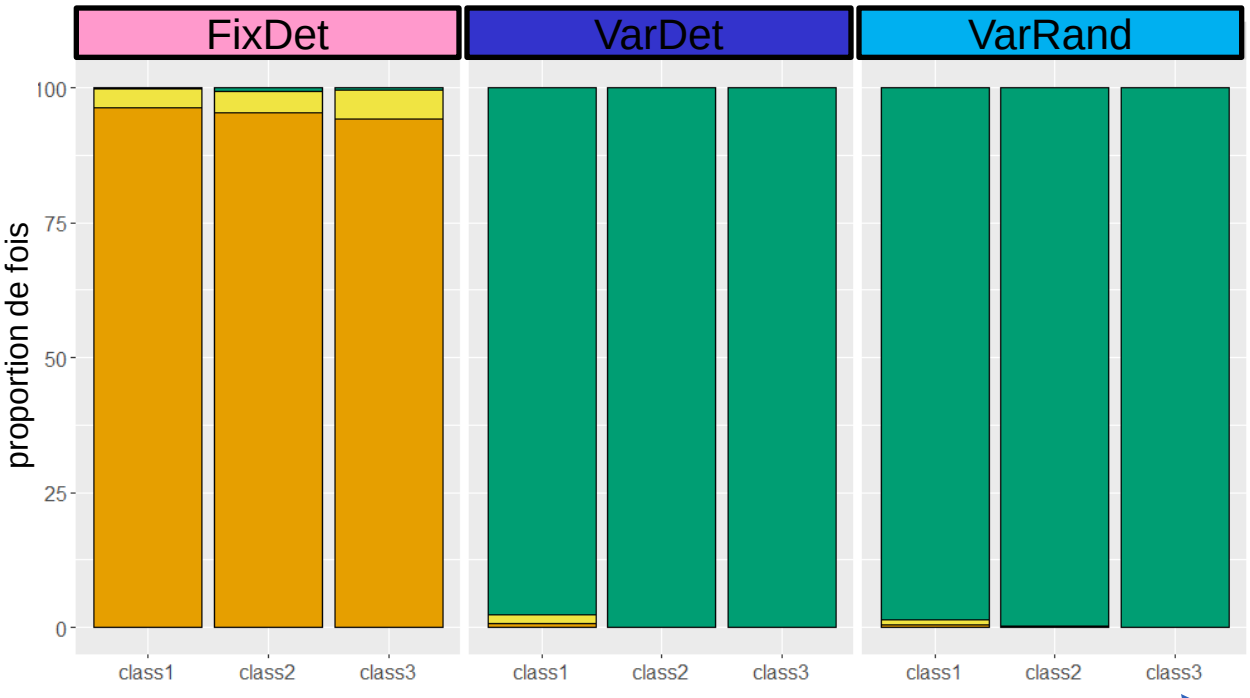
- **AICc (critère d'information d'Akaike corrigé) :**

plus faible AICc → meilleure capacité prédictive du modèle
différence de moins de 2,0 points → les modèles sont équivalents

Paramètre d'intérêt majeur : **pente** (donne la magnitude de la relation)

- Estimation du paramètre :
- comparaison de la précision de l'estimation entre les modèles (précision)
- comparaison avec le paramètre réel (biais)

Comparaison AICc : LestKest / L0K1



Modèles univariés

AICc LestKest < AICc GAM
AICc LestKest = AICc GAM
AICc LestKest > AICc GAM

Classe 1 : nobs Є [147;1092]
Classe 2 : nobs Є]1092;2037]
Classe 3 : nobs Є]2037;2982]

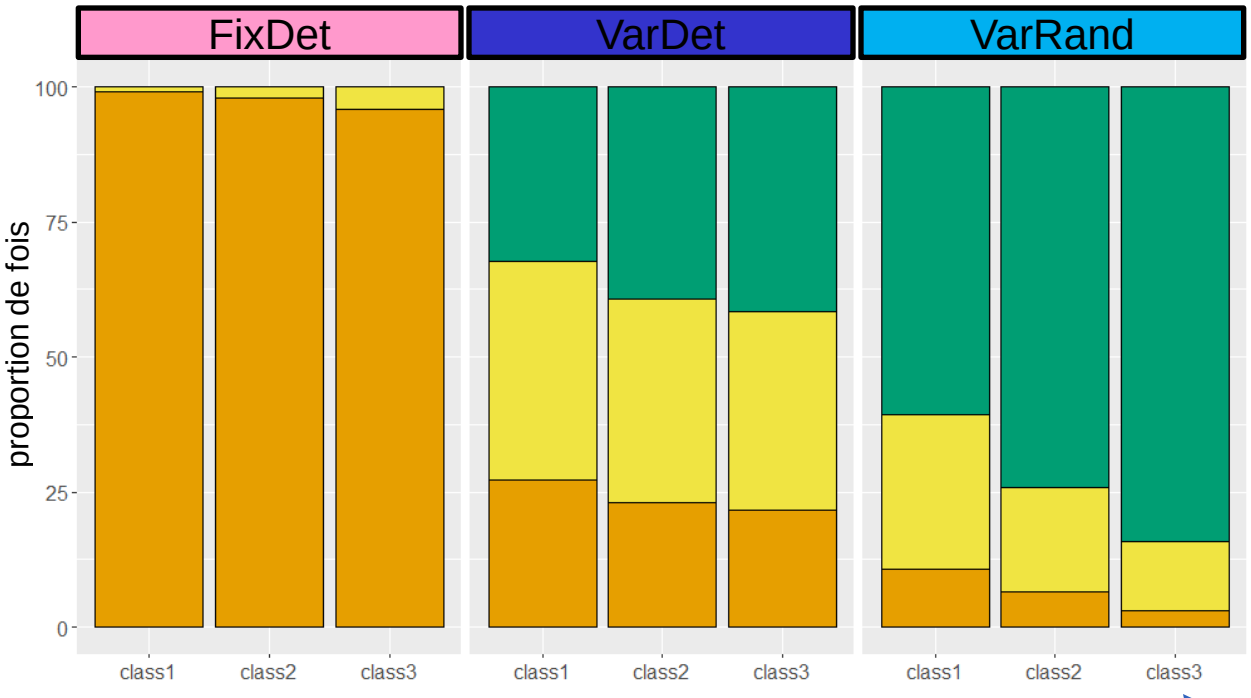
LestKest mieux que L0K1

La proportion de cas où LestKest est mieux que L0K1 augmente avec le nombre d'observations (nobs)

Sauf lorsque :

- les asymptotes sont égales à 0 et 1 dans les données (Scénario FixDet)

Comparaison AICc : LestKest / GAM



Modèles univariés

AICc LestKest < AICc GAM
AICc LestKest = AICc GAM
AICc LestKest > AICc GAM

Classe 1 : nobs Є [147;1092]
Classe 2 : nobs Є]1092;2037]
Classe 3 : nobs Є]2037;2982]

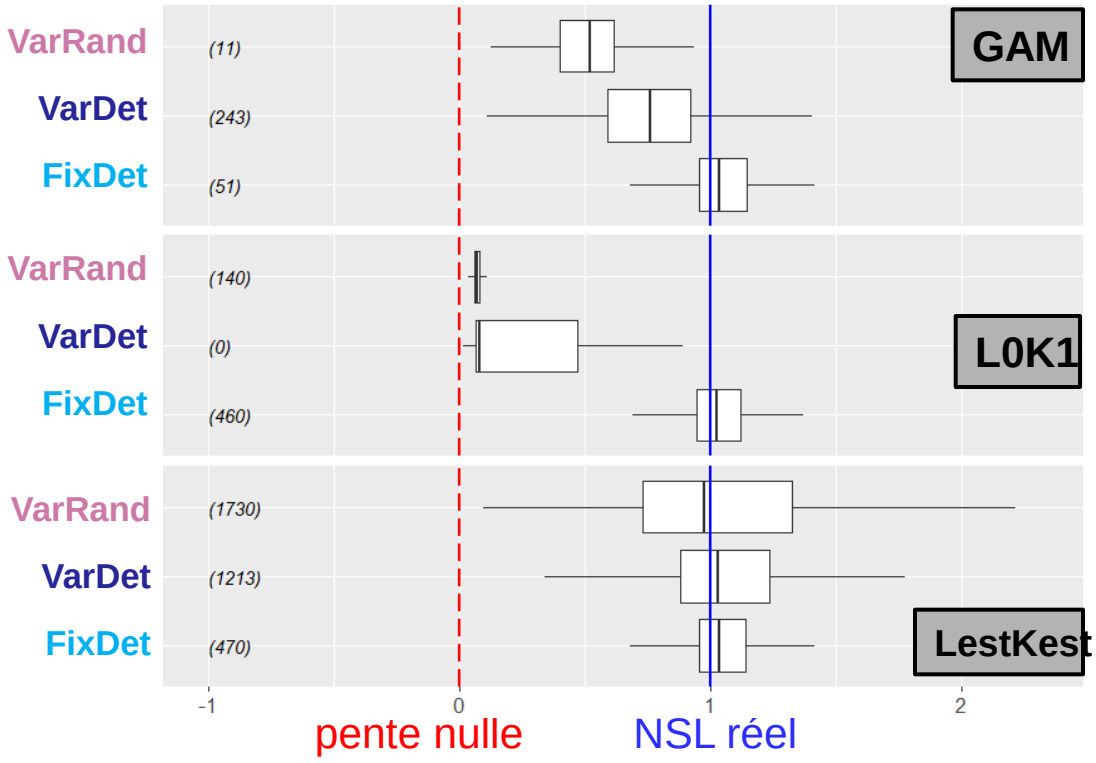
GAM mieux ou équivalent à LestKest

La proportion de cas où LestKest est mieux que L0K1 augmente avec le nombre d'observations (nobs)

Sauf lorsque :

- les données sont aléatoirement réparties sur le gradient (Scénario VarRand)

Estimation: Pente normalisée NSL



Estimation de la pente : LestKest **moins biaisé** que **L0K1** et que **GAM** → PENTE SOUS-ESTIMEE

Surtout lorsque :

- les données sont aléatoirement réparties sur le gradient (**VarRand**)

Sauf lorsque :

- les asymptotes sont égales à 0 et 1 dans les données (**FixDet**)

Godeau & Gosselin (In Prep.)

Conclusions (non-linéaire)

Si au moins une des deux asymptotes est atteinte **ET** différente de **0** ou **1**, alors :

- GLM canonique (L0K1) → Mauvaises capacité predictive et estimation de la pente
- GAM canonique → Mauvaises capacité prédictive et estimation des paramètres sur un gradient aléatoire (et en bivarié)
- Modèle ELUA (LestKest) → Corrige ces problèmes
- Modèle LestKest à stabiliser (stabilité numérique notamment)

Proposition que la fonction ELUA soit intégrée à la trousse à outils de l'analyse des données binaires

Godeau & Gosselin (In Prep.)

Conclusions de l'exposé

- Importance de **faire attention aux hypothèses auxiliaires** des modèles statistiques paramétriques pour une bonne inference et une bonne selection de modèles :
  - **Distribution de probabilité** des observations (voire des variables latentes)
  - **Dépendances** entre données
  - **Fonctions de lien non-linéaires** non-canoniques
- Les modèles semi-paramétriques (GAM) ou l'augmentation du nombre de données ne **résolvent pas tous ces problèmes**
- Développement **d'outils de diagnostic** de la qualité d'ajustement pour essayer de détecter les problèmes (Gosselin 2011 Plos One, Thèse de Thierno Diallo ANR GAMBAS)

Merci pour votre attention

Des questions ?