Améliorer la pertinence et l'efficacité des modèles statistiques en écologie : extension des fonctions sigmoïdes dans le cadre de l'étude de la distribution de la biodiversité

Ugoline Godeau

UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE Santé, Sciences Biologiques et Chimie du Vivant

Unité de Recherche Ecosystèmes Forestiers - INRAE

THÈSE présentée par :

Ugoline GODEAU

soutenue le : **15 juin 2020**

pour obtenir le grade de : **Docteur de l'Université d'Orléans**

Discipline/ Spécialité : Écologie

Améliorer la pertinence et l'efficacité des modèles statistiques en écologie : extension des fonctions sigmoïdes dans le cadre de l'étude de la distribution de la biodiversité

THÈSE dirigée par :

M. GOSSELIN Frédéric, Ingénieur en Chef des Ponts, des Eaux et des Forêts, INRAE Nogent-sur-Vernisson

RAPPORTEURS :

Mme DELIGNETTE-MULLER Marie-Laure, Professeur des Universités, VetAgro Sup

M. GIMENEZ Olivier, Directeur de recherches, CNRS

JURY :

Mme DELIGNETTE-MULLER Marie-Laure, Professeur des Universités, VetAgro Sup

M. GIMENEZ Olivier, Directeur de recherches, CNRS

Mme PEYRARD Nathalie, Directeur de recherches, INRAE Toulouse

M. REINEKING Björn, Directeur de recherches, INRAE Grenoble

M. GOSSELIN Frédéric, Ingénieur en Chef des Ponts, des Eaux et des Forêts, INRAE Nogent-sur-Vernisson

Je dédie ce manuscrit à la mémoire de mon papa, qui m'a toujours poussée à faire de longues études, en espérant que là c'est assez long pour lui.

Remerciements

Je remercie avant tout **Frédéric Gosselin** qui m'a encadrée et guidée pendant ces quelques années. Tu m'as permis de dédramatiser les mathématiques, et de continuer à consolider mes connaissances en écologie et modélisation mais pas que, tu m'as aussi permis de m'ouvrir à l'épistémologie et à la lecture critique des articles (et du reste), et toujours cela dans des conversations passionnantes et bienveillantes et parfois même philosophiques. Et, au milieu de tout cela, tu m'as surtout appris à persévérer malgré les difficultés.

"No ! Try not ! Do or do not, there is no try." **– Star Wars: The Empire Strikes Back**

Je remercie également chaleureusement mes deux rapporteurs **Olivier Gimenez**, et **Marie-Laure Delignette-Muller** ainsi que les autres membres du jury de thèse, **Nathalie Peyrard**, et **Björn Reineking**, qui ont accepté d'évaluer mon travail.

Merci également aux membres de mon comité de thèse pour leur disponibilité et leur conseils constructifs: **Camille Coron**, **Jérémy Piffady**, **Frédéric Mortier**, **Wilfried Thuiller** et **Frédéric Jiguet**.

Je tiens à adresser mes plus sincères pensées et remerciements à mes co-auteurs pour leurs nombreuses relectures et leur infinie patience : **Christophe Bouget**, **Jérémy Piffady** et **Tiffani Pozzi**.

Un grand merci adressé aussi à **Fabien Laroche**, pour ses conseils, ses discussions et ses explications qui n'ont pas leurs pareils.

"Trust quality of what you know, not quantity" **– Karate Kid**

Merci à **Pierre Onfroy** pour son travail pas toujours facile sur des modèles particulièrement complexes à aborder.

Merci à **Philippe Guillemard** pour son assistance en toute circonstance.

"Help me, Obi-Wan Kenobi. You're my only hope." **– Star Wars**

Merci à **Vicki Moore** pour ses nombreuses corrections des articles.

"I have always depended on the kindness of strangers." **– A Streetcar Named Desire**

De manière générale je remercie toute l'équipe Biodiv de m'avoir accueillie et épaulée pendant ces trois années : **Gwendoline Percel**, **Frédéric Archaux**, **Manon Balbi**, **Marion Gosselin**, **Hilaire Martin**, **Yann Dumas**, **Marie Baltzinger**, **Isabelle Bilger**, **Yoan Paillet** et **Richard Chevalier**.

De manière encore plus générale je remercie tout **EFNO**, **Irstea** (maintenant **INRAE**) du domaine des Barres et de Antony pour leur accueil.

Et parce qu'une thèse ce n'est pas que du travail acharné, mais aussi des moments plus sociaux et de détente …

Je remercie de manière très générale les membres du « **Foodies** » de m'avoir permis de m'évader et me régaler un soir par semaine.

"How was it ? Bangarang" **— Hook**

Merci aussi aux non-permanents qui ont aidé à leur manière à m'acclimater ou tout du moins à supporter la vie dans le Loiret : **Jordan Bello**, **Romain Dous**, **Arnaud Miton**, **Jeanne Menanteau, Barth Dessanges, Manon Chappard** et **Manon Balbi**.

"Life moves pretty fast. If you don't stop and look around once in a while, you could miss it." **– Ferris, from Ferris Bueller's Day Off**

Merci à **Jérémy Cours**, et surtout bon courage…

"Fly you fools." - **The Fellowship of the Ring**

Merci à **Laura Chevaux** et **Nadège Bonnot** pour les nombreux débats (entre autres féministes), bavardages en tout genre et surtout râlages ! Une bonne catharsis ça fait toujours du bien ! Une petite pensée pour les covoiturages avec Laura, moments de détente et de rouspétage supplémentaires.

"Quand le monde te persécute, tu te dois de persécuter le monde." **— Le Roi Lion**

Merci à **Guillem Parmain** qui m'a fait découvrir toutes ses passions, et il y en a beaucoup : les insectes, le boomerang, la forge et les toupies, pour n'en citer que quelques-unes. Et merci pour les soirées jeux et films qui m'ont aéré la tête.

"Get busy living, or get busy dying." **— The Shawshank Redemption**

Merci à **Ushma Shukla** pour ses sages paroles qui m'ont permis de retrouver la voie de la raison quand ma santé mentale commençait à s'effriter. Aujourd'hui tu es maman, mais pendant toute ma thèse tu as aussi joué le rôle de grande sœur.

"Ohana means family. Family means nobody gets left behind." **— Lilo & Stitch**

Merci également à celle qui a joué le rôle de ma petite sœur, **Tiffani Pozzi**, avec qui j'ai passé mon temps à me chamailler (Tic et Tac comme dirait Guillem), à me confier, et à partager mille choses. Tu as commencé par m'accueillir dans ce bureau que j'ai ensuite considéré comme étant le mien pendant plus de trois ans. Puis tu es repartie pour mieux revenir (dans l'autre bâtiment certes, mais le mur invisible et tacite entre ces deux bâtiments ne nous a jamais arrêtées).

"When you're the best of friends having so much fun together. You're not even aware, you're such a funny pair." **— Rox & Roucky**

J'ai une pensée toute particulière pour mes proches qui ont continué à me supporter dans les hauts et les bas, parmi mes amis et ma famille…

Pour commencer merci à **Paul Geoffroy** pour ses conseils et ses relectures. Je remercie donc ici ta

contribution à ma thèse, de m'avoir changé les idées sur demande avec des anecdotes plus qu'incongrues mais surtout d'avoir cru en moi.

"JACK BEAUREGARD: You're sure trying hard to make a hero out of me. - NOBODY: You're that already. You just need a special act, something that'll make your name a legend. - JACK BEAUREGARD: What I don't understand is what difference it makes to you. - NOBODY: If a man is a man, he needs someone to believe in. - JACK BEAUREGARD: I've met all kinds in my life. Thieves and killers. Pimps and prostitutes. Con men and preachers. Even a few fellas that told the truth. The kind of man you're talking about, never. - NOBODY: Maybe you've never met them. Or hardly ever. But they're the only ones who count." **— My name is Nobody**

Merci aussi à mes petits chats et leur amour inconditionnel : Abalone, Rio et Sookie.

"Which pets get to sleep, On velvet mats? Naturalment! The aristocats!" **—The aristocats**

Je remercie ensuite ma sœur, **Zélia**, tu as toujours su me montrer le chemin que je devais emprunter même quand ce n'était pas forcément une évidence. Le petit pec insupportable a bien changé sous ton regard bienveillant et formateur.

"HIGH ALDWIN: Go in the direction the bird is flying! - BURGLEKUTT : He's going back to village! - HIGH ALDWIN: Ignore the bird. Follow the river. " **— Willow**

Je t'aime plus que je ne pourrais le compter…

"How do I love thee? Let me count the ways. One one-thousand. Two one-thousand. Three one-thousand. Four one-thousand. Five..." **— Who framed Roger Rabbit**

Merci à ma **Maman**, une mère dévouée, toujours là quand j'en ai besoin, que ce soit pour vider mon sac ou, à l'inverse, me changer les idées. Merci pour ta tendresse, ta compréhension, ta compassion et ta douceur, mais aussi pour tes blagues/jeux de mots tellement répétés qu'ils sont devenus pour moi comme des rengaines.

"L'humour juif c'est quand ce n'est pas drôle et que ça ne parle pas de saucisses » **— OSS 117 : Rio Ne Répond Plus**

Tu m'as également tout appris et permis de devenir la femme que je suis aujourd'hui

"You drink. Champagne if you're happy. Champagne, if you're sad. You drive a car. Gamble if you want. Own diamonds. Learn how to fire a gun. You travel to Morocco. Take up lovers. Make them suffer. You look a tiger in the eye. And trust without fear. That's what it is to be a woman." **– Jojo Rabbit**

Je remercie enfin l'homme qui partage ma vie, **Sébastien**, dont la présence a toujours réussi à m'apaiser. Tu es toujours là pour me faire voir le bon côté des choses et pour refréner, tant bien que mal, mon côté névrosé.

"I kept asking Clarence why our world seemed to be collapsing and things seemed to be getting so shitty. And he'd say, "that's the way it goes, but don't forget, it goes the other way too". That's the way romance is... Usually, that's the way it goes, but every once in a while, it goes the other way too." **— True Roman**

Avant-propos

Les indices i et j perdent leur point lorsqu'ils sont mis dans des équations avec une accentuation de type barre au-dessus pour notifier un vecteur (ou double barre pour une matrice). Ceci n'est pas intentionnel et est uniquement lié aux limitations de l'éditeur d'équation de word. Ainsi, dans les équations, les indices des formes i et $\bar{\imath}$ (et les indices des formes j et $\bar{\jmath}$), doivent être traités de la même manière.

VOLUME I : Manuscrit

Sommaire

"*The history of science, like the history of all human ideas, is a history of irresponsible dreams, of obstinacy, and of error. ... This is why we can say that, in science, we often learn from our mistakes, and why we can speak clearly and sensibly about making progress there*."

— **Karl R. Popper**, Conjectures and Refutations: The Growth of Scientific Knowledge (1963).

INTRODUCTION GENERALE

I. APPROCHE DE LA MODELISATION

I. 1. Qu'est-ce qu'un modèle ?

L'une des approches pour explorer les questions d'écologie (comme domaine d'étude scientifique), est l'utilisation de modèles. Un modèle est un instrument qui a pour fonction de donner une représentation simplifiée et formelle d'un concept[1] ou d'une réalité, en vue de le ou la décrire (ou visualiser), de l'estimer, de l'expliquer ou de prévoir ses effets (cf. Legay 1997 pour l'aspect instrumental). En écologie, un modèle peut donc être vu comme un instrument qui permet de représenter et comprendre les processus écologiques complexes. La modélisation, comprise comme la représentation d'un phénomène complexe, peut prendre diverses formes telles qu'un schéma, une équation ou un dessin. Elle peut aussi s'exprimer à travers des jeux de rôles conçus pour, par exemple, concilier conservation et agriculture (Moreau et al. 2019).

Cependant, dans la suite de ce manuscrit nous allons explorer uniquement le cas des modèles dits mathématiques (qui seront appelés plus simplement par la suite « modèles ») qui font des descriptions mathématiques (via des équations) du ou des mécanismes ou relations aboutissant aux observations faites du réel (Cherruault 1998). La compréhension écologique repose aujourd'hui pour une part importante sur cette modélisation et, de manière générale,

[1] " Concepts are defined as regularities in events of objects designated by a label […].They rare usually broader and more abstract than the particular instances of events of objects that they encompass. Concepts are constructed from many observations, so they represent an abstraction of the regularity from these observations." (Pickett et al. 2007a)

l'importance des modèles s'est accrue dans l'utilisation des statistiques pour être actuellement omniprésents (Yoccoz 1999).

Les modèles permettent de tester des hypothèses écologiques. C'est la tentative de falsification de l'hypothèse écologique testée, à travers le modèle, qui va permettre de la rejeter ou de l'accepter temporairement (jusqu'à une éventuelle falsification future d'après l'approche Poppérienne de la science - Popper, 1963). Cependant, il est important de souligner que le rejet d'un modèle n'entraine pas nécessairement une falsification de la théorie, puisque le modèle repose sur d'autres hypothèses non reliées à l'hypothèse écologique principale, qu'on appellera hypothèses auxiliaires (induites par exemple par la collecte de données). Un résultat négatif informe que l'ensemble des hypothèses (hypothèse principale testée + hypothèses auxiliaires) est faux, mais sans pour autant indiquer où réside l'erreur (Duhem 1981)[2]. Quine (1951) généralise cette constatation et affirme que chaque hypothèse ne peut être réfutée indépendamment, il n'y a pas d'expérience cruciale qui établisse une fois pour toutes qu'une hypothèse individuelle est fausse, c'est l'ensemble de la connaissance qui est remis en cause[3]. En écologie, il est particulièrement compliqué d'identifier les hypothèses auxiliaires (notamment

[2] « En résumé le physicien ne peut jamais soumettre au contrôle de l'expérience une hypothèse isolée, mais seulement tout un ensemble d'hypothèses ; lorsque l'expérience est en désaccord avec ses prévisions, elle lui apprend que l'une au moins des hypothèses qui constitue cet ensemble est inacceptable et doit être modifiée ; mais elle ne lui désigne pas celle qui doit être changée. » (Duhem 1981 p. 284)

[3] « The dogma of reductionism survives in the supposition that each statement, taken in isolation from its fellows, can admit of confirmation or infirmation at all. My countersuggestion, issuing essentially from Carnap's doctrine of the physical world in the Aufbau, is that our statements about the external world face the tribunal of sense experience not individually but only as a corporate body. » (Quine 1951).

pour les données observationnelles en termes de covariables) qui seraient utiles pour que le modèle puisse bien porter l'hypothèse écologique principale.

Un modèle ne peut être défini qu'en délimitant au préalable son domaine d'applicabilité (ou domaine de validité), c'est-à-dire l'ensemble des conditions prescrites pour lequel le modèle est censé correspondre à la réalité (Loehle 1983). Un modèle sera considéré général si son domaine d'applicabilité est large (Levins 1966, 1993). Le domaine d'exactitude (ou niveau d'exactitude) du modèle représente quant à lui une indication de l'accord attendu entre le modèle conceptuel et la réalité, qui soit consistante avec son domaine d'applicabilité (Loehle 1983). Un modèle sera considéré comme précis si il a un haut niveau d'exactitude (Levins 1966, 1993).

I. 2. La notion d'aléatoire dans les modèles

Un modèle est dit déterministe lorsqu'il part d'un état initial pour arriver à un seul état final. Le système est parfaitement connu, il n'implique aucun phénomène aléatoire, la sortie du modèle est entièrement déterminée par les conditions initiales et les valeurs des paramètres du modèle. L'équation modèle relie i^{ème} observation de la variable expliquée (Y_i) à l'observation correspondante de la variable explicative (X_i), via la fonction f de la manière suivante :

Equation 1 : $Y_i = f(X_i)$

Un modèle est en revanche dit stochastique lorsqu'il intègre une part d'aléatoire, et donc, partant d'un état initial, plusieurs états finaux sont possibles, dont les probabilités sont estimées par le modèle. Les modèles utilisés sont des modèles mathématiques soit probabilistes, soit statistiques. Concernant les modèles probabilistes, la densité de probabilité de la variable à

expliquer Y dépend des paramètres du modèle (θ) et des variables explicatives X avec une structure probabiliste $g(Y|\theta, X)$ et où X et θ sont connus et on cherche Y. Les modèles statistiques possèdent la même structure probabiliste que les modèles probabilistes $g(Y|\theta, X)$, mais où cette fois Y et X sont connus (les Y sont des observations) et on cherche à estimer θ sur la base de X et Y et du modèle. Dans ce type de modèles statistiques, les observations sont supposées être les réalisations de variables aléatoires (Cox 1990). Le modèle statistique constitue donc une représentation d'une population théorique dont est tiré l'échantillon, via des paramètres et des hypothèses sur le type de distribution de ces variables aléatoires (Yoccoz 1999). Il repose sur une famille de distributions de probabilités et d'hypothèses (auxiliaires) sur Y et éventuellement sur les variables explicatives (X_1 à X_n).

La stochasticité peut avoir différentes sources telles que l'imprécision des mesures, une connaissance imparfaite du système (phénomènes inconnus non modélisés) ou le caractère intrinsèquement aléatoire du système (on parle d'aléas intrinsèques). Un aléa intrinsèque constitue une propriété qui est par nature aléatoire (e.g. en biologie, les erreurs de réplication de l'ADN, ou la vision probabiliste de l'expression des gènes - Heams 2009). Ce qui est inclus dans la notion de stochasticité va donc dépendre de la question posée (Bolker 2008). En écologie, les systèmes étudiés sont complexes et vivants, soumis à une évolution constante, on privilégie donc généralement des modèles stochastiques qui rendent compte de cette part d'aléatoire. Ainsi, en écologie la modélisation a une importance majeure comme dans la plupart des domaines scientifiques, mais la prise en compte de l'aléatoire Y est capitale, contrairement à d'autres domaines comme la physique par exemple où le déterminisme est plutôt majoritaire. Les modèles stochastiques ont par exemple une grande importance dans la modélisation de dynamique de

population où l'on peut identifier au moins deux sources de stochasticité, en plus des erreurs de mesures (Lande et al. 2003):

1) la stochasticité démographique (ou variabilité intra-individuelle) qui rend compte des fluctuations aléatoires de la taille de la population qui se produisent parce que la naissance et le décès de chaque individu est un événement discret et probabiliste. Si ces événements sont indépendants, ce type de stochasticité a tendance à se moyenner quand la population est nombreuse et, à l'inverse, à avoir un impact fort sur des petites populations.

2) la stochasticité environnementale, qui rend compte des fluctuations temporelles de la probabilité de mortalité et du taux de reproduction, qui s'applique uniformément à tous les individus de la population. Ce type de stochasticité a globalement le même type d'impacts sur les petites et larges populations.

Ces deux sources de variabilités en dynamique des populations seraient donc des aléas intrinsèques, le résultat d'une mauvaise connaissance du système ou d'une modélisation du système à un grain trop large pour pouvoir en modéliser les détails. En effet, les modèles n'intègrent généralement pas tous les événements pouvant intervenir à l'échelle d'un individu car soit nous n'avons pas les connaissances pour le faire, soit un tel modèle demanderait trop de paramètres.

Cette stochasticité est intégrée dans le modèle en introduisant une loi de probabilité sur la variable étudiée *Y* (Equation 2) et/ou en ajoutant des lois de probabilités sur les paramètres de la fonction ou les variables explicatives (Equation 3).

Equation 2 : Introduction d'une loi de probabilité sur *Y* (L désignant une loi de probabilité) :

$$Y_i \sim L(f(X_i, \overline{\theta_1}), \overline{\theta_2})$$

Avec ~ signifiant que la variable aléatoire suit la probabilité de distribution précisée (ici L).

$\overline{\theta_1}$ et $\overline{\theta_2}$ respectivement les vecteurs de paramètres de la fonction f et de la loi de probabilité L.

Equation 3 : bruitage des paramètres de la fonction f (L_1 et L_2 désignant des lois de probabilité):

$$Y_i = f(X_i, \overline{\theta_1})$$

$$\theta_{1,1} \sim L_1(\overline{\alpha_1}) \; ; \theta_{1,2} \sim L_2(\overline{\alpha_2}) \; \ldots$$

Avec $\overline{\theta_1}$ le vecteur de paramètres de la fonction f, $\overline{\alpha_1}$ et $\overline{\alpha_2}$ les vecteurs de paramètres des lois L_1 et L_2.

L'aléatoire ne concerne pas uniquement les hypothèses auxiliaires et pourrait, dans certains cas, être l'objet d'hypothèses principales et d'inférence (la stochasticité ne pourrait dans ces cas pas se définir comme un simple ajout de « bruit » au modèle) : la forme qu'elle prend dans le modèle peut nous donner des renseignements sur les processus écologiques à l'œuvre (Bolker 2008). Par exemple, des données de comptage suivant une distribution négative binomiale nous informent sur la présence d'une forme de variation environnementale ou une réponse agrégée des individus qui n'aurait pas été prise en compte dans un modèle poissonien (Bolker 2008 et l'exemple de Shaw and Dobson 1995).

I. 3. Les formes des modèles

Nous nous intéressons en particulier aux modèles statistiques paramétriques ou semi-paramétriques. Les modèles statistiques sont définis à travers leurs paramètres θ (de la loi et/ou du modèle) qui peuvent se situer sur un espace de dimension finie, ou sur un espace de dimension infinie.

Les modèles statistiques peuvent prendre diverses formes en fonction de la dimension de l'espace d'état des paramètres. Les deux plus répandus (et que nous aborderons au cours de cette thèse) étant :

- Les **MODELES PARAMETRIQUES** (e.g. modèles de régression linéaire et non-linéaire, modèles de séries chronologiques …) : modèles qui ont une forme probabiliste dépendant d'un nombre fini de paramètres qui se situent dans des espaces de dimension finie. Leur forme est donc prédéterminée (imposée en amont de l'analyse des données).
- Les **MODELES SEMI-PARAMETRIQUES** (e.g. modèles additif généralisé, modèle de régression de Cox…) : modèles qui possèdent des paramètres qui se situent dans des espaces de dimension finie, et des paramètres qui se situent quant à eux sur des espaces de dimension infinie (Bickel et al. 2014, Oakes 2014, Sasieni 2014).

Les modèles semi-paramétriques ne sont pas toujours facilement exprimables en termes quantitatifs. Or, les modèles doivent être construits de manière à permettre une comparaison entre études (Yoccoz 1999). Ainsi, des modèles semi-paramétriques sont difficilement comparables d'une étude à l'autre. Contrairement aux modèles paramétriques avec lesquels on compare des paramètres ayant les mêmes rôles, la comparaison des modèles semi-

paramétriques repose plutôt sur la comparaison des fonctions (par exemple des splines) et est donc bien plus compliquée. Les modèles paramétriques sont quant à eux beaucoup plus comparables d'une étude à l'autre en permettant d'établir des relations à travers des valeurs quantitatives comparables les unes aux autres.

I. 4. Le processus de formation des modèles

Une très grande variété de modèles existe au sein même des modèles statistiques paramétriques et semi-paramétriques (e.g. les modèles linéaires gaussiens, les modèles linéaires généralisés, les modèles non linéaires, les modèles mixtes, les modèles additifs généralisés, les modèles d'équation structurelles ou les modèles autorégressifs, pour ne citer que les plus courants). Le processus de choix et de formation d'un modèle statistique doit reposer sur une réflexion en plusieurs étapes essentielles pour que le modèle réponde réellement à la question posée (Figure 0.1 ; Austin 2007) :

1) Le modèle doit être pensé dans un cadre conceptuel, c'est-à-dire qu'il doit, dans la mesure du possible, être en lien avec des connaissances ou des théories existantes (Yoccoz 1999, Guisan and Zimmermann 2000, Austin 2007) ; mais la détermination du cadre conceptuel logique peut s'accompagner de difficultés (cf. Driscoll and Lindenmayer 2012). Au sein du paradigme, il est possible de dégager des hypothèses qui vont être testées par le modèle (Austin 1999).
2) Le modèle doit être choisi en fonction des objectifs de l'étude (eux-mêmes envisagés dans le cadre conceptuel), c'est-à-dire de la question qui est posée et des attentes sur les

résultats. Il est donc nécessaire de bien définir en amont les attentes vis-à-vis des sorties du modèle (Levins 1966, Starfield 1997, Shmueli 2010, Dickey-Collas et al. 2014). Quatre intentions majeures peuvent se dégager :

- Modéliser pour explorer : on peut vouloir se servir de modèles comme d'outils pour rechercher de nouvelles hypothèses. Afin de remplir cette fonction, on se tournera vers des modèles descriptifs, c'est-à-dire qu'ils ont vocation à synthétiser et/ou représenter les données de manière structurée et organisée. Les modèles descriptifs ne font pas intervenir de théories, et les relations révélées sont de nature corrélatives et non causales (Shmueli, 2010). Les modèles faisant appel à du *data-mining* en sont de bons exemples. Il faut toutefois faire attention à l'utilisation abusive de ce type de pratiques. Ainsi, avec une surabondance de données, il est possible via le *data-mining*, de toujours trouver des relations significatives, perdant ainsi le sens de la significativité (Hand 1998, Harrell 2001). Pour s'en prémunir il est donc préférable de bien réfléchir le modèle dans son cadre conceptuel (cf. premier point). L'idée n'est pas de faire des hypothèses *ad-hoc* mais bien d'explorer les données, pour ensuite poser de nouvelles hypothèses et refaire des analyses plus poussées avec cette fois un modèle construit pour répondre à cette nouvelle hypothèse (et sur de nouvelles données).
- Modéliser pour acquérir du savoir : on peut vouloir se servir de modèles pour essayer de comprendre comment s'articulent les relations entre les éléments du système. Ce type de questions nécessite des modèles explicatifs (ou théoriques d'après Loehle 1983), c'est-à-dire ayant vocation à tester des hypothèses causales

dans des cadres théoriques. Les concepts de domaine d'applicabilité et d'exactitude ne sont donc ici pas nécessairement pertinents puisque les modèles explicatifs se basent sur des théories qui, par définition, devraient avoir une portée universelle et doivent toujours être « vraies » (Loehle 1983). Il existe tout de même des théories en écologie dont le domaine de validité est restreint (e.g. Bersier and Meyer 1994).

- Modéliser pour estimer l'état d'un système : on peut vouloir se servir de modèles afin de tenter de déterminer l'état du système basé sur les observations. Les modèles descriptifs ou explicatifs sont une fois de plus les plus adaptés pour ce type de questionnement. Ils seront notamment beaucoup utilisés dans des questions de recherche appliquée.
- Modéliser pour extrapoler : on peut vouloir se servir de modèles pour faire des prédictions de nouvelles ou futures observations, c'est-à-dire d'extrapoler spatialement et/ou temporellement. Ces modèles prédictifs sont souvent utilisés dans des contextes appliqués et sont peu adaptés à de la recherche scientifique puisque les relations utilisées ne doivent pas nécessairement être réalistes. De plus, ces modèles doivent nécessairement présenter une forme de généralité (dans un domaine de validité large mais bien défini), puisqu'ils nécessitent d'être applicables à des situations en dehors de leur domaine de développement présentant toutefois des conditions similaires. Le modèle doit donc être à l'équilibre entre l'utilisation de relations corrélatives et leur capacité de généralisation à des fins de prédictions en dehors de leur domaine

d'établissement. Les écologues utilisent, par exemple, des modèles de régression multiples comme « outils de calcul prédictif » (d'après Loehle 1983) sans pour autant que les multiples équations aient un sens biologique (Mac Nally 2002). Les modèles basés sur des techniques d'apprentissage automatique (« machine learning ») sont aussi bien répandus dans le but de prédire, car ils possèdent une bonne capacité de prédiction bien que l'on ne connaisse pas les relations sous-jacentes. Ces relations sont purement corrélatives et peuvent être biologiquement fausses. Les modèles prédictifs ont un rôle majeur en écologie car ils permettent de projeter des résultats sur un nouvel espace géographique ou temporel et de faire des recommandations de gestion.

La définition de la question explorée est un prérequis majeur dans le choix du modèle, car dans le cas contraire (lorsque la question est mal définie), le modèle peut aboutir à des conclusions erronées qui peuvent avoir des conséquences non-négligeables notamment lorsque les résultats du modèle servent ensuite à des recommandations de gestion (cf. les exemples de mauvaises pratiques dans Dickey-Collas et al. 2014). Symétriquement, si un modèle doit être formé en fonction de la question initiale posée, il ne peut pas en être séparé, et ne peut donc pas répondre à d'autres questions *a posteriori* (ou avec plus de

difficulté). Le modèle doit être aussi adapté à ses utilisateurs, et à l'utilisation qui en sera faite, notamment si d'autres utilisateurs que le modélisateur initial sont impliqués [4].

D'autre part, la catégorie du modèle se définit en lien avec ses ambitions, et ils peuvent être classés en trois catégories : analytique (lorsque qu'ils se concentrent sur les processus théoriques exprimés de façon mathématique), mécaniste (lorsque qu'ils se concentrent sur les processus réels, les relations de cause à effet connues) ou empirique (aussi appelé phénoménologique). Cette dernière catégorie, que nous allons davantage explorer au cours de cette thèse décrit des modèles qui se concentrent sur la description mathématique des relations dans un contexte précis et de manière réaliste. Ils permettent au final d'obtenir des prédictions sur les situations particulières précises et testables. Ces modèles s'écartent des notions de causalité ou les infèrent à travers des relations de corrélations. Ainsi les modèles empiriques peuvent s'intéresser à des prédicteurs directs (e.g ressource en nourriture, température…) mais aussi des prédicteurs indirects (e.g. indices climatiques à large échelle tels que l'oscillation nord-atlantique – NAO – qui est un phénomène météorologique dans le nord de l'océan Atlantique et qui permet de décrire les variations du régime océan-atmosphère sur cette région). Par ailleurs, ils peuvent permettre d'obtenir des paramètres statistiques non interprétables d'un point de vue théorique (la pente d'une relation linéaire entre deux variables) mais qui peuvent tout de

[4] "Avant de s'engager dans une modélisation l'ingénieur doit impérativement se demander à qui va servir le modèle, s'il s'agit d'un organisme, quels sont ses buts, ses moyens d'actions et ses capacités à faire des mesures. Ce dernier point est primordial. ... Les modèles les plus grossiers sont simples à valider alors que les modèles fins nécessitent tant de mesures pour préciser les fonctions inconnues qu'en pratique leur validation est très mauvaise. Or leur logique déductive et les phénomènes qu'ils font apparaître dépendent de ces choix" (Bouleau 1999).

même être exploitables d'un point de vue pratique (pour donner des recommandations de gestion par exemple). Enfin, ces modèles ne constituent souvent qu'une première étape dans la démarche scientifique, permettant de démarrer une réflexion sur les aspects causaux sous-jacents. D'ailleurs, certains de ces modèles peuvent être composés de parties causales et d'autres parties plus phénoménologiques.

3) Les différentes composantes du modèle doivent également être convenablement considérées. Ces composantes comprennent notamment le domaine d'applicabilité et d'exactitude du domaine (dont nous avons déjà discuté auparavant –cf. Introduction générale I.1). L'échelle d'étude (géographique ou temporelle) constitue également une composante déterminante dans le choix du modèle (Levin 1992, Guisan and Thuiller 2005, Austin 2007). La question posée par l'étude peut se situer à plusieurs échelles qui découlent des objectifs d'utilisation du modèle. Ainsi, les données doivent être adaptées à ces échelles en termes d'étendue (durée temporelle ou aire prospectée) et de résolution (nombre de points par unité de temps ou d'espace). Le modèle couplé aux données utilisées se doivent d'être cohérent avec la question posée initialement (e .g. un modèle basé sur des données à l'échelle d'une communauté ne peuvent pas rendre compte de comportements individuels).

4) Le modèle doit être élaboré de manière à prendre en compte les contraintes pré-identifiées sur les autres composantes du modèle. La disponibilité des données et leur nature vont également affecter la formation du modèle. Par ailleurs, la réflexion en amont sur les objectifs de l'étude (et la finalité du modèle) doit également influencer l'origine des données. En effet, si l'on s'intéresse à des relations causales (e.g. pour expliquer le

système à l'aide d'un modèle mécaniste), des données expérimentales seront indispensables. A l'inverse si l'on s'intéresse à des relations corrélatives (e.g. pour décrire le système à l'aide d'un modèle empirique), des données issues d'observations seront nécessaires. Les ressources disponibles pour l'étude (ressources en compétence, en calcul, en argent, en temps disponible) ont également une incidence sur le choix et la formation du modèle (Loehle 1983). Un modèle de type bayésien sera par exemple souvent plus exigeant en ressource de calcul qu'un modèle fréquentiste. Enfin, les outils d'évaluation de modèles disponibles peuvent avoir un impact sur le choix du modèle.

Bien qu'idéalement basé sur ces principes, le choix du modèle sera aussi habituellement biaisé, influencé par la familiarité du modélisateur pour un type de modèle.

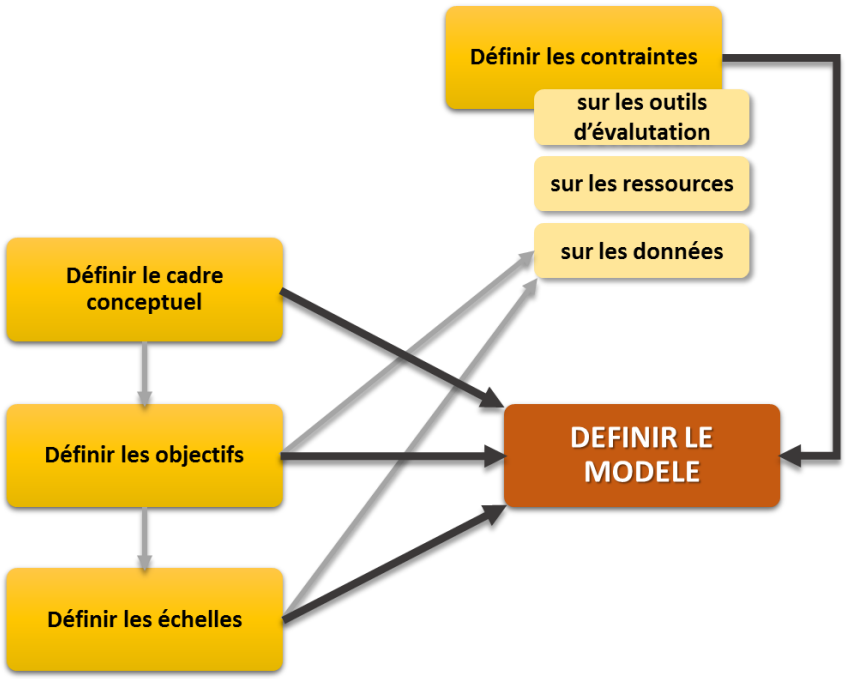


Figure 0.1 : Schéma récapitulatif des étapes de définition d'un modèle et des liens entre ces étapes. Les flèches noires représentent la nécessité du modèle d'être délimité par chacune des étapes définies. Les flèches grises représentent les liens d'ordonnancement entre les étapes.

Les modèles, bien qu'exprimés en langage mathématique, ne sont pas pour autant neutres. Chaque étape de leur formation nécessite des partis pris qui ne sont pas dénués de conséquences (Bouleau 1999). Les modèles se doivent, cependant, de respecter certains critères pour être corrects : être utile (doit répondre à un objectif) et simple (le plus simple possible sans pour autant être simpliste). De manière plus relative, un modèle pourra remplacer un autre (être considéré comme meilleur), s'il est plus général (s'applique à un univers plus large), s'il est plus utile (si l'on se place dans un contexte similaire, alors il rend le premier inutile), s'il permet d'étendre l'utilisation de techniques communes au modèle précédent, ou s'il permet des comparaisons nouvelles (Legay 1997). Dans le but d'établir si le modèle respecte les critères d'utilité et de simplicité, ou s'il pourra avoir la prétention de remplacer un autre modèle, une étape primordiale d'évaluation du modèle est requise.

I. 5. Le processus d'évaluation des modèles

L'évaluation d'un modèle passe par plusieurs phases (Rykiel 1985). La première phase étant celle de la **vérification** du modèle qui est une démonstration que le formalisme de modélisation est correct. Cette phase comporte une part de débogage du programme informatique et contrôle du formalisme mathématique (vérification mécanique), et une part de vérification que ces mêmes programme informatique et formalisme mathématique traduisent bien les idées que l'on veut représenter (vérification logique). En raison de la difficulté à vérifier le modèle en toute circonstance, le modèle est généralement vérifié dans les conditions dans lesquelles il est appliqué et la vérification est donc inapplicable en dehors de ces conditions.

La phase suivante de **calibration** (ou d'**estimation**) est faite par les algorithmes d'optimisation et de convergence. Elle permet l'estimation et l'ajustement des paramètres et constantes du modèle pour améliorer l'adéquation entre les sorties du modèle et le jeu de données.

La phase qui nous intéresse particulièrement est celle de **validation** du modèle, qui décrit le processus qui génère la preuve que le modèle possède une gamme d'exactitude satisfaisante consistante avec les applications prévues du modèle et dans un ensemble de limites prescrites, compatible avec un domaine d'applicabilité donné (Loehle 1983, Rykiel 1985). La validation d'un modèle ne permet pas de démontrer la vérité absolue du modèle, mais que son usage est satisfaisant sous des conditions spécifiques. Le modèle est donc validé dans le contexte spécifique pour lequel il est testé. Globalement, la validation d'un modèle statistique se fait à travers une grande variété de procédures, mais en écologie elle est généralement réalisée via :

- La comparaison de modèles : elle consiste à développer plusieurs modèles, comparer leurs capacités à expliquer les données, et conserver le meilleur de ces modèles (même si celui-ci ne représente que très mal la réalité). Cette approche est souvent favorisée en écologie par rapport à l'approche par falsification d'une hypothèse unique ou d'un modèle unique (Lakatos et al. 1980). Elle est souvent réalisée à travers des critères d'information (AIC, BIC, WAIC…) et permet de choisir un modèle parmi plusieurs.
- La validation opérationnelle : elle consiste à comparer des données simulées par le modèle avec les données réelles de calibration. Elle sert notamment à évaluer les propriétés statistiques du modèle via l'utilisation de Goodness-of-fit par exemple (Gosselin 2011). Cette étape prend généralement place après l'étape de sélection de modèle.

- Robustesse du modèle : elle consiste à valider la cohérence interne du modèle, en s'assurant que le modèle est capable de reproduire l'occurrence, le timing, et la magnitude d'évènements simulés. D'ailleurs, dans le but d'évaluer les modèles, et pour être sûr d'en comprendre les limites d'application (domaine d'applicabilité), Austin (2007) préconise d'utiliser dans un premier temps des données artificielles (simulées en prenant en compte les théories écologiques).

En sciences, le principe de parcimonie préconise de privilégier des théories, des hypothèses ou des modèles les plus simples possibles. Ce principe permet notamment d'éviter un biais fortement redouté : le sur-ajustement des modèles aux données. A cet égard, la validation à travers la comparaison de modèles pénalise les modèles en fonction du nombre de paramètres à estimer (e.g. le critère d'information d'Akaike). Bien que l'utilisation de ce principe paraisse raisonnable, notamment pour la formation de modèles prédictifs qui se doivent de rester généraux, il paraît plus difficile à justifier pleinement dans le cas de modèles descriptifs. Ces modèles décrivent le système étudié, très souvent complexe, et pour lequel il est souvent vain et inutile de vouloir simplifier. Le principe de parcimonie se doit donc d'être appliqué de manière parcimonieuse, et adapté à l'emploi que l'on veut faire du modèle (Coelho et al. 2019).

La phase de validation repose généralement sur l'adéquation du modèle aux données avec lesquelles il a été généré. L'écart entre les données et le modèle théorique peut s'expliquer par les hypothèses auxiliaires introduites (cf. I.2.), et certains auteurs n'hésitent donc pas à, dans certains cas, prioriser la théorie par rapport aux données : « A theory may fail miserably to fit the data, yet be complelling in its explanatony power, elegence, completencess, logic, etc. We may decide that the data are wrong » (Loehle 1983). La priorisation de la théorie par rapport aux

données peut être justifiée dans certains cas mais elle semble aussi minimiser l'approche Poppérienne de falsification des théories (selon Karl Popper une hypothèse est dite réfutable si sa forme logique est telle qu'il est possible de tester son éventuelle fausseté par une expérimentation - Popper 1963).

Enfin, la phase de **crédibilité** rend compte d'un degré suffisant de croyance en la validité d'un modèle pour justifier son utilisation pour la recherche et la décision. Cette appréciation est relative au contexte du modèle et subjective, elle dépend donc des connaissances et d'un jugement individuel non quantifiable. Ainsi, une fois le modèle validé et considéré comme crédible, il est possible de s'intéresser aux sorties réelles du modèle sélectionné et validé, et notamment la magnitude de la relation (dans le cadre de l'aide à la décision par exemple).

II. L'UTILISATION DU NON-LINEAIRE EN MODELISATION STATISTIQUE PARAMETRIQUE

II. 1. Les modèles linéaires et leurs limites

II. 1. a. Modèles de régression linéaire

Le modèle de régression linéaire est un modèle paramétrique qui permet de mettre en relation une variable expliquée avec une (modèle linéaire « simple ») ou plusieurs (modèle linéaire « multiple ») variable(s) explicative(s) :

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ji} + \varepsilon_i$$

$$\varepsilon_i \underset{iid}{\sim} Normal(0,\sigma^2)$$

Avec *i* l'index de l'observation $i \in (1,2, \dots n)$

j l'index de la variable explicative $j \in (1,2, \dots p)$

Y_i la variable de réponse à l'observation i

β_0 l'intercepte

β_j le coefficient de régression de la variable j

x_{ji} la valeur de la variable explicative j à l'observation i

ε_i l'erreur aléatoire (ou terme de perturbation) à l'observation i. Ce terme permet d'ajouter de la stochasticité dans le modèle (cf. I. 2. La notion d'aléatoire dans les modèles). Les erreurs ε sont supposées indépendantes et identiquement distribuées de loi Normal d'espérance 0 et de variance σ^2.

Ce modèle peut aussi s'écrire sous forme matricielle comme ci-dessous :

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad \textbf{avec} \quad E(\boldsymbol{Y}) = \boldsymbol{X\beta}$$

Avec : $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ de dimension (*n*,1), $\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}$ de dimension (*n*,*p*+1)

$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ de dimension (*p*+1,1), $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ et de dimension (*n*,1)

Et $E(\boldsymbol{Y}) =$ l'espérance du vecteur $\boldsymbol{Y}$.

La régression linéaire repose sur quatre hypothèses qui constituent également les limites du modèle :

- hypothèse de la linéarité : la relation entre les deux variables numériques continues est de forme linéaire, β0 et β1 sont constants, il n'y a pas de rupture du modèle.
- hypothèse d'indépendance : les résidus $\boldsymbol{\varepsilon}$ sont indépendants de $\boldsymbol{X}$ et entre eux.
- hypothèse de la normalité : les résidus sont distribués selon une loi Normale de moyenne zéro → $\mathrm{E}(\varepsilon_i) = 0$.
- hypothèse de l'homoscédasticité: les résidus sont distribués de façon homogène (on parle aussi de variance constante ou de résidus identiquement distribués) → $\mathrm{Var}(\varepsilon_i) = \sigma^2$

Le modèle linéaire général est une extension de ce modèle pour lequel **Y** n'est plus un vecteur de la variable à expliquer de dimension (*n*,1), mais une matrice de multiples variables à expliquer de dimension (*n*,*v*) où v est le nombre de ces variables à expliquer. La distribution des résidus est supposée être une gaussienne multivariée.

$$\boldsymbol{Y} = \begin{pmatrix} Y_{11} & \cdots & Y_{1v} \\ \vdots & \ddots & \vdots \\ Y_{n1} & \cdots & Y_{nv} \end{pmatrix}$$

II. 1. b. Modèles linéaires généralisés

Le modèle linéaire généralisé (abrégé GLZM mais plus généralement également abrégé GLM) est une généralisation du modèle de régression linéaire (on se replace ici dans le cas d'un ***Y*** univarié). Le GLM s'est développé à la faveur de résultats mathématiques qui généralisaient certains des résultats connus pour le modèle linéaire à un contexte plus large (mais toujours restreint) en termes de densité de probabilité et non-linéarité. Il permet alors de s'affranchir des hypothèses de linéarité et de normalité. Le modèle linéaire généralisé introduit une fonction dite de lien sur l'espérance de ***Y*** et s'écrit sous forme matricielle de la manière suivante :

$$g(E(\boldsymbol{Y})) = \boldsymbol{X\beta} \qquad \text{ou} \qquad E(Y) = g^{-1}(\boldsymbol{X\beta})$$

Avec : g la fonction de lien et g^{-1} la réciproque de la fonction de lien

*E(**Y**)* l'espérance de ***Y***

Le choix de la fonction de lien (g) et de la distribution de probabilités de ***Y*** dépend de la nature des données (Table 0.1). Dans le cas de données binaires (codé en 0/1, e.g. données de présence/absence d'espèces) ou de données discrètes strictement positives avec un maximum connu (données de comptage e.g. données de richesse spécifique dont le nombre maximal d'espèces possibles est connu), la distribution de probabilité proposée pourra être une Binomiale et les fonctions de liens possibles seront les fonctions logit, probit, et cloglog. Dans le cas de données discrètes strictement positives sans maximum connu (données de comptage également e.g. abondance d'une espèce), la distribution de probabilité proposée sera une Poisson et la

fonction de lien une fonction log. Enfin, dans le cas de données continues positives (e.g. Données de croissance en diamètre des arbres), la distribution de probabilité adaptée sera une Gamma et la fonction de lien une fonction inverse.

Table 0.1 : Distribution de probabilité et fonction de lien disponibles pour étudier, dans le cadre d'un modèle linéaire généralisé, les différents types de données en écologie.

Format des données	**Exemple de données en écologie**	**Distribution de probabilité**	**Fonction de lien *g***
Binaire (codée 0/1)	Données de présence/absence d'une espèce	Binomiale ou quasibinomiale (si sous dispersé)	logit, probit, cloglog
Discrètes strictement positives avec un maximum connu (données issues de comptage)	Données de richesse spécifique dont le nombre maximal d'espèces possible est connu	Binomiale ou quasibinomiale (si sous dispersé)	logit, probit, cloglog
Discrètes strictement positives sans maximum connu (Є **N**) (données issues de comptage)	Données d'abondance d'une espèce ou richesse spécifique	Poisson ou quasipoisson (si sous dispersé)	log
Données continues (Є $\mathbf{R}^{+}$)	Données de circonférence des troncs d'arbres	Gamma ou inverse gaussian	inverse

Le modèle linéaire généralisé à effets mixtes (GZLMM ou GLMM) est une extension du modèle linéaire généralisé qui inclut des effets aléatoires sur le prédicteur linéaire en plus des effets fixes. Cet ajout permet l'analyse d'observations non-indépendantes telles que des données répétées ou imbriquées (e.g. prise en compte d'un effet site). Le modèle s'écrit de la manière suivante :

$$g(E(\boldsymbol{Y})) = \boldsymbol{X\beta} + \boldsymbol{ZU} \qquad \text{ou} \qquad E(Y) = g^{-1}(\boldsymbol{X\beta}) + \boldsymbol{ZU}$$

$\boldsymbol{Z} = \begin{pmatrix} Z_{11} & \cdots & Z_{n1} \\ \vdots & \ddots & \vdots \\ Z_{1q} & \cdots & Z_{nq} \end{pmatrix}$ de dimension (n,q), $\boldsymbol{U} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_q \end{pmatrix}$ de dimension (q,1)

$$\boldsymbol{U} \sim Normal\ (0, \boldsymbol{G})$$

Avec q, le nombre d'effets aléatoires.

$\boldsymbol{G}$**,** est la matrice de variance-covariance des effets-aléatoires, de dimension (q,q).

Les modèles linéaires généralisés et généralisés mixtes ont pour limite d'être restrictifs sur la forme de la relation, à travers les fonctions de lien (et leurs réciproques), fonctions non-linéaires certes mais avec un choix restreint des formes possibles. Les écologues ont donc la possibilité de se tourner vers des modèles non-linéaires.

II. 2. Les modèles non-linéaires

II. 2. a. Modèles non-linéaires semi-paramétriques

L'étude de relations non-linéaires se fait couramment à l'aide de modèles semi-paramétriques. Le modèle additif généralisé (GAM - Hastie and Tibshirani 1986, Hastie 2017) en est un exemple populaire. Il comprend également une distribution de probabilité et une fonction de lien (g) sur l'espérance des $\boldsymbol{Y}$ mais en ajoutant des fonctions (f) sur chaque variable explicative, tel que :

$$g\big(E(Y_i)\big) = \beta_0 + \sum_{j=1}^{p} f_j\big(x_{ij}\big)$$

Les fonctions f sont des fonctions de lissage ou de faible complexité à estimer.

Ces modèles ont le désavantage d'être compliqués à interpréter en termes de paramètres ou de forme de la relation. Ils permettent uniquement de décrire les données (modèles purement à vocation descriptive ou prédictive) et ne permettent pas de faire d'hypothèses sur la forme de la relation, basées sur la théorie.

II. 2. b. Modèles non-linéaires paramétriques

Les modèles non-linéaires paramétriques font intervenir une fonction (F), non-linéaire donnant la valeur de l'espérance de $\boldsymbol{Y}$, choisie par le modélisateur. L'espérance des données d'observations dépend d'une ou plusieurs variables ($\boldsymbol{X}$), et est donc modélisée par une fonction à plusieurs paramètres ($\boldsymbol{\beta}$), chacun des paramètres de la fonction étant la plupart du temps obtenu à partir d'une combinaison linéaire de certaines variables explicatives $\boldsymbol{X}$**.** L'équation peut s'écrire de la manière suivante :

$$E(\boldsymbol{Y}) = F(\boldsymbol{X}, \boldsymbol{\beta})$$

La fonction non-linéaire sera choisie en fonction de la relation attendue entre les variables qui pourra émerger de l'observation des données, d'un modèle précédant défaillant sur le plan de la fonction de lien ou de connaissances biologiques permettant d'émettre des hypothèses (Austin 2007). Ces relations peuvent donc être incluses dans des modèles aussi bien analytiques, qu'empiriques ou mécanistes et dans le but de décrire, expliquer ou prédire.

La plupart des modèles utilisés actuellement sont basés sur des régressions linéaires (e.g. le modèle courant de distribution d'espèce est la régression logistique qui est un cas particulier du modèle linéaire généralisé). Cependant, il existe certains domaines d'études écologiques précis

pour lesquels les modèles non-linéaires sont bien démocratisés, tels que l'écophysiologie, avec les courbes de réponse à un stress (e.g. Verwijst and Fircks 1994, Radanielson et al. 2018), ou le suivi de croissance (Yin et al. 2003, Liu et al. 2018), dont la courbe de la relation est généralement de forme sigmoïdale. La biogéographie utilise également de nombreuses courbes non-linéaires pour étudier la relation entre la taille d'une aire et la diversité en espèce (Connor and McCoy 2001, Dengler 2009a). Cette relation peut également avoir une forme sigmoïdale, mais d'autres formes sont étudiées tels que des courbes exponentielles, puissances….

III. STRATEGIES ET QUESTIONS DE THESE

III. 1. Les modèles étudiés

Dans le cadre de cette thèse, les hypothèses étudiées seront majoritairement basées sur des théories. Cependant, elles seront testées avec des données issues d'observations et les relations mises en évidences seront de nature phénoménologiques. Par conséquent, les modèles abordés seront plutôt empiriques et stochastiques (inclusion de l'aléatoire avec des distributions de probabilité). La plupart des modèles abordés seront de forme paramétrique, afin de pouvoir interpréter les modèles et de pouvoir les comparer entre études (via leurs paramètres), mais pourront être comparés à des modèles semi-paramétriques (GAM) classiquement utilisés dans le type de situations abordées. Enfin, les modèles mis en œuvre auront pour certains davantage vocation à être explicatifs (cf. Chapitre 2, où l'on s'interroge sur l'existence et la forme de la relation entre le bois mort et la richesse spécifique des coléoptères saproxyliques), et pour

d'autres davantage vocation à être descriptifs (estimation fréquente de variances sans hypothèse préalable).

III. 2. Questions et objectifs de la thèse

Le recours aux modèles statistiques dans les études en écologie s'est fortement démocratisé. En effet, l'harmonisation des méthodes dans le milieu scientifique a permis de simplifier les procédures de calculs nécessaires à ces modèles, qui sont désormais accessibles à l'ensemble de la communauté scientifique (Yoccoz 1999).

L'utilisation de « toy-models » (modèles simplistes applicables et appliqués dans des cas d'étude variés) est certes plus simple mais elle est souvent associée à un manque de réflexion autour des propriétés statistiques de ces modèles et des hypothèses écologiques d'intérêts négligées ou inhérentes aux modèles. Yoccoz s'interroge d'ailleurs : « ne présentons-nous pas les modèles sous une forme par trop statistique, en négligeant leur justification biologique (ou médicale, ou sociologique…) et en quoi ils permettent (ou ne permettent pas d'ailleurs) de répondre à nos questions ? », et répond un peu plus loin que « c'est la biologie qui doit davantage nous guider dans le choix de la structure des modèles » (Yoccoz 1999). De même, dans le contexte des modèles de distribution d'espèces, Austin (2007) insiste sur la nécessité d'inclure la théorie écologique lors de la formation d'un modèle, et souligne par exemple que la forme de la courbe attendue des modèles de régression logistique devrait être davantage justifiée par une idée de processus sous-jacent. S'il n'y a pas de réflexion au préalable sur la forme de la relation et les raisons de cette forme, le modèle peut bien s'adapter aux données mais échouer à faire de

bonnes prédictions. Un modèle, même descriptif, doit donc inclure une réflexion sur chacune de ces composantes, en relation avec la théorie écologique ou des arguments logiques. Lors de la formation des modèles, nous avons d'abord une approche plutôt « logique » que nous avons mise en lien le plus possible avec des aspects biologiques. Le but de cette thèse était à la fois d'alimenter la réflexion concernant certains des choix faits lors de la construction des modèles, et sur les implications de ces choix lors de l'utilisation des modèles, et de proposer de nouvelles formes de modèles utiles dans certaines circonstances, visant ainsi à complémenter la trousse à outils du modélisateur écologue.

Pour cela, je me suis plus particulièrement intéressée au cas de l'utilisation de modèles avec fonctions sigmoïdales dans divers contextes. J'ai principalement exploré l'incidence de la position des asymptotes, de la symétrie de la courbe, ainsi que des variations spatiales ou intergroupes de la forme de la courbe de réponse, sur l'estimation des paramètres, l'adéquation du modèle aux données et la capacité prédictive des modèles.

Dans l'optique de mener à bien cet approfondissement autour des modèles avec fonction sigmoïdale, j'ai commencé par mener une réflexion sur le concept de sigmoïde en écologie, son usage et mésusage (Chapitre 1).

Je me suis ensuite placée dans un contexte simple de modèle de biodiversité (relation entre la richesse spécifique et une ressource), pour étudier les possibilités d'amélioration de ces modèles via l'utilisation de fonctions sigmoïdales complexes et notamment dans un contexte où l'on suppose, sur la base d'arguments a priori, que la relation est très vraisemblablement variable

dans l'espace. Je me suis également concentrée sur le processus d'évaluation de tels modèles afin d'en dégager les atouts (Chapitre 2).

Je suis ensuite sortie du cadre de l'écologie pour me tourner vers des considérations statistiques plus méthodologiques et applicables à un nombre de domaines très variés, en me concentrant sur des modèles de type régression logistique afin d'étudier des données binaires. Cette fois, je me suis intéressée aux modèles actuels communément répandus (avec une fonction logistique standard) et certains de leurs problèmes inhérents (notamment l'impact sur l'estimation et la qualité prédictive du modèle), imputables à une hypothèse auxiliaire (imposée par la fonction logistique standard). Nous avons donc simulé les données avec une hypothèse auxiliaire plus souple et générale en proposant une fois de plus l'utilisation de modèles faisant intervenir une fonction sigmoïdale plus complexe pouvant améliorer les résultats et résoudre une part de ces problèmes (Chapitre 3).

Enfin, sur cette base, je me suis de nouveau intéressée à un problème écologique en appliquant le même type de réflexion et de modèles dans un contexte de modèle de présence/absence multi-espèces (Chapitre 4).

Au final, bien que les modèles proposés, ainsi que la démarche d'évaluation les entourant, n'aient pas la prétention d'être parfaits, ni des solutions à tous les problèmes, ils auront je l'espère la qualité d'avoir ouvert une réflexion sur les problématiques les entourant et proposer de nouvelles pistes pour l'écologie et d'autres domaines d'application des statistiques.

CHAPITRE 1

Exploration du concept de sigmoïde en écologie

I. INTRODUCTION DU CHAPITRE 1

Une science se définit comme l'ensemble structuré des connaissances sur un domaine particulier. Cependant, la notion de domaine particulier ne semble pas être appropriée car elle sous-entend une distinction claire entre chaque discipline. Or, il est évident, que les études scientifiques sont en réalité régulièrement à l'interface entre plusieurs disciplines communément pensées séparément mais complémentaires (les sciences cognitives par exemple sont définies comme étant constituées de six disciplines : psychologie, informatique, neurosciences, anthropologie, linguistique, et philosophie ; Miller, 2003). De la même manière, elles empruntent régulièrement des concepts ou des méthodes les unes aux autres.

Le développement de la science est avant tout permis par l'élaboration de concepts bien définis. Une description opérationnelle d'un concept spécifie la gamme des phénomènes que le concept représente, et permet donc de définir le cadre permettant de tester le concept. Une capacité de communication efficace entre les acteurs scientifiques est également l'un des leviers majeurs rendant possible le développement des sciences (Peters 1991). Pour ce faire, une nomenclature harmonisée des termes utilisés est essentielle. Pourtant il n'est pas rare de rencontrer dans la littérature des termes qui posent questions, car mal définis (Slisko and Dykstra 1997), ce qui a pour effet de rendre le discours confus et de freiner le développement des connaissances. C'est d'autant plus vrai en écologie qui est une science relativement nouvelle avec des protagonistes variés. Les termes utilisés peuvent être vagues (définis de manière non précise et applicables à plusieurs concepts sémantiquement proches) ; ambigus (pouvant avoir plusieurs sens non clairement différenciés) ; dépendants du contexte (changeant de sens en fonction de la situation

dans lequel il est utilisé) ; peu spécifiques (généraux, non adaptés à des questions spécifiques) ; indéterminés (sujets à controverse et souvent en compétition avec des termes en apparence synonymes) ou démultipliés (plusieurs termes servent à définir le même concept) (Peters 1991, Hodges 2008, Herrando-Pérez et al. 2014). Cette négligence quant à une utilisation homogène et précise des termes peut potentiellement avoir des conséquences notables sur les avancées scientifiques (en ralentissant le processus de partage du savoir), sur le transfert du savoir (Slisko and Dykstra 1997) et sur les prises de décisions politiques (e.g. le cas des services écosystémiques, Fisher et al. 2007). Les termes et concepts mal définis ou arbitraires utilisés dans la sphère non académique peuvent parfois venir de la sphère académique et avoir des conséquences non négligeables ; c'est d'ailleurs un des rôles importants des scientifiques dans le monde non-académique que de questionner et préciser ces définitions (e.g. Underwood 1995 et leçon 4 de Gosselin 2009). Les tentatives sont donc nombreuses pour mettre en avant les problèmes d'imprécision, et tenter de clarifier et de rendre cohérent les concepts et les termes (quelques exemples généraux et spécifiques : Rykiel (1985), Fauth et al. (1996), Adams et al. (1997), Jax (2005), Dauvin et al. (2008), Hodges (2008), Madin et al. (2008), Middleton and Prigoda (2008), Jax and Hodges (2008), Blackburn et al. (2011), MacGregor-Fors (2011), Ruxton and Schaefer (2011), Herrando-Pérez et al. (2012, 2014). Les ontologies[5] constituent d'ailleurs des moyens efficaces pour harmoniser la littérature et la rendre moins ambigüe (Madin et al. 2008).

[5] "Ontology: a formal model that uses mathematical logic to clarify and define concepts and relationships within a domain of interest (e.g. behavioral ecology)." (Madin et al. 2008)

Par ailleurs, s'il est vrai que parfois, au sein même du domaine d'étude concerné, certaines notions peuvent être mal définies, c'est encore plus vrai lorsqu'il s'agit de termes, ou outils (e.g. les méta-analyses ; cf. Vetter et al., 2013) empruntés à d'autres disciplines comme les statistiques. Yoccoz (1999) fait remarquer à ce propos qu' « une lecture attentive d'articles publiés en écologie, épidémiologie ou médecine conduit à se demander si ne règne pas une méconnaissance des concepts de base des outils statistiques ». La bonne compréhension des outils statistiques, régulièrement utilisés en écologie, devrait être un prérequis. La littérature se doit donc de continuer de standardiser les termes employés afin d'être plus compréhensible.

C'est dans ce sens que nous avons entrepris de définir le terme sigmoïde. Les fonctions ayant une forme de courbe sigmoïdale sont variées (Commun logistic function, Gompertz function, Weibull cumulative distribution, Extrem-value function, Chapman-Richards function, Morgan-Mercer-Flodin function, Cumulative beta-P distribution, Beta growth function…) et elles semblent être relativement bien adaptées à de très nombreux cas d'étude, puisqu'elles permettent de modéliser des relations dont l'effet s'amplifie (croissance ou décroissance qui s'accélère) puis sature (seconde asymptote). Or, au cours de recherches bibliographiques préliminaires sur le sujet, il est apparu clair que le terme sigmoïde n'est pas nécessairement clair et bien délimité, parfois uniquement verbalement définie, et peu rattachée à ses racines mathématiques, entraînant des difficultés de communication et des confusions vis-à-vis du modèle utilisé par les auteurs et des conclusions que l'on peut en extraire. Dans le but de comprendre et de mettre en évidence l'origine de ces problèmes, nous avons effectué une exploration bibliographique non exhaustive afin d'identifier les lacunes ou les inexactitudes dans la définition de sigmoïde, en prenant comme exemple le domaine de la biogéographie. L'identification du problème n'étant

qu’une première étape, nous avons aussi tenté d'établir une définition à la fois accessible et pratique pour tous les écologues. Enfin, pour aider les écologues à appliquer les fonctions sigmoïdes à leurs données, nous avons répertorié et détaillé des fonctions issues de la littérature qui entrent ou pas dans cette définition ainsi que les conditions sur les paramètres pour être de forme sigmoïdale. Le but de cet article est donc d’harmoniser la littérature autour de la définition de sigmoïde afin de permettre aux écologues d’avoir des outils mieux définis et maîtrisés dans leur approche de modélisation.

II. MANUSCRIT 1

Lack of definition of mathematical terms in ecology: the case of the sigmoid class of functions in macro-ecology

Ugoline Godeau[1], Christophe Bouget[1], Jérémy Piffady[2], Tiffani Pozzi[1], Frédéric Gosselin[1]

[1] INRAE, UR EFNO, Domaine des Barres, F-45290, Nogent-sur-Vernisson, France.

[2] INRAE, UR MALY, Centre de Lyon-Villeurbanne, F-69616 Villeurbanne, France.

Statut : soumis dans la revue “Ecology and Evolution”.

Abstract: Defining mathematical terms and objects is a constant issue in ecology; often definitions are absent, erroneous or imprecise. Through a bibliographic prospection, we show that this problem appears in macro-ecology (biogeography and community ecology) where the lack of definition for the sigmoid class of functions results in difficulties of interpretation and communication. In order to solve this problem and to help harmonize papers that use sigmoid functions in ecology, herein we propose a comprehensive definition of these mathematical objects. In addition, to facilitate their use, we classified the functions often used in the ecological literature, specifying the constraints on the parameters for the function to be defined and the curve shape to be sigmoidal. Finally, we interpreted the different properties of the functions induced by the definition through ecological hypotheses in order to support and explain the interest of such functions in ecology and more precisely in biogeography.

Keywords: Sigmoid curve shape, species-area relationship, biogeography, species-resource relationship, curve fitting.

II. 1. Introduction

Using well-defined and uniform terms is a key point in science. Yet, one of the main criticisms that can be made in the science of ecology is the poor definition of terms and concepts or inconstant use within its community (Pickett et al. 2007b, Herrando-Pérez et al. 2017, Kirk et al. 2018). Many concepts do not yet have a consensual definition, and communication is therefore difficult. Furthermore, loosely defined concepts can cause not only an unstable expression of a scientific concept, but can also result in inconsistencies within the concept itself (e.g. Gosselin, 2001). This is why many articles have tried to highlight this problem and to establish precise definitions - i.e. “ecological niche” (Araújo and Guisan 2006a) or “ecological function” (Jax 2005b). However, the problem is not restricted to ecological concepts; it also concerns ecological domains (i.e. “ecological engineering”, cf. Gosselin, 2008) or certain terms and concepts used in ecology and borrowed from other sciences. This is the case for mathematical terms as, for example, the notions of extinction or demographic stochasticity (clarified in Gosselin, 1997; Lebreton, Gosselin, and Niel 2007). Reflections on mathematical definitions make it possible to conceptualize possibilities not yet foreseen (e.g. the importance of dependence between individuals within demographic stochasticity or uncertainty in (McCarthy et al. 1994). In the present paper, we deal with the term "sigmoid" and propose a definition to overcome imprecision problems. Hereafter, we will call “sigmoid” the curve shape that can be represented by different functions, and the “sigmoid class of functions”, the class that contains these functions.

Ecologists often study relationships between two ecological variables (e.g. a biodiversity metric as a function of an environmental variable/predictor). Although, the most often considered form

of these relationships is linear, nonlinear forms have also been used (power, exponential etc.), including sigmoidal forms. In ecology, sigmoidal relationships are generally implicitly used in binomial regressions. However, in the field of macro-ecology and, in particular, in the study of species-area relationships (SARs), explicit sigmoidal forms occur fairly often. Indeed, a sigmoidal shape is very likely to emerge when species richness is related to the area in which the species were sampled (Preston 1962a). Many sigmoidal functions have been developed and used in a SAR context; however, they can also be applied to the study of relationships between biodiversity and a resource gradient other than available habitat area (species-resource relationships, or SReRs). Furthermore, the sigmoidal form of a relationship may prove useful for decision-making in forest or conservation management. Indeed, certain characteristics of the curve can provide management targets like the inflection point or the upper asymptote (Ranius and Jonsson 2007a).

In recent years, numerous articles have been published which review the use of nonlinear functions, including sigmoids, in the field of biogeography and especially for SAR-type relationships (Tjørve 2003a, Dengler 2009b, Tjørve 2009a, Williams et al. 2009a). Unfortunately, no clear definition of the term sigmoid was provided in these publications.

Despite the frequent use of sigmoidal functions, in most cases, there is no proper, accessible definition of what exactly is meant by a “sigmoidal” shape. Classically defined as an S-shape, the sigmoid may seem clear and that is the reason why it is so rarely defined. Yet, the precise characteristics of these curves are not formalized or made explicit. This absence of a clear definition results in a lack of harmonization between papers in ecology, and inconsistencies between articles, or even within one and the same article can ensue. For example, although most definitions include the presence of an upper asymptote (Veech 2000, Tjørve 2003a), Mashayekhi

et al. (2014) define one of their functions (Persistance2) as sigmoidal though it does not have an upper asymptote; this contradicts the general idea of a sigmoid. There is therefore a need to more explicitly define the sigmoidal class of shapes.

Our first goal is to assess the use of the term sigmoid in biogeography studies and highlight the lack of a clear definition. Then, we propose a definition of the term so that its use in the literature is harmonized and no longer confusing. Finally, we justify the definition in relation with ecological theory and we highlight the implications and advantages of this new definition. The two underlying questions are: what characteristics should sigmoid curves exhibit? What functions can be included in the sigmoid class?

II. 2. An obvious lack of a clear definition

The word “sigmoid”, composed of “sigma” and “eidos” (*sigmoeidés* in ancient Greek), means something that has the form of the capital letter sigma (Σ). The term sigmoid is more generally defined as an S-shaped curve. Yet these descriptions, in addition to being vague, are not accurate since the form of an S (or a Σ) is impossible in mathematical curves described by functions. In fact, if we apply an S form to mathematical curves, we notice that we obtain two or three values of f(x) for one x, which is impossible according to the very definition of a function. Moreover, the representation of an S-shaped curve excludes forms that should logically be part of sigmoid curves such as decreasing sigmoid curves.

Given this intrinsic difficulty with the notion of sigmoid, we investigated how authors in ecology have used and define this term. We selected an ecological domain where sigmoid functions are

often explicitly used to describe relationships: biogeography with species-area relationships (conventionally abbreviated as SARs) and species response to ecological gradients within species-resource relationships (abbreviated here as SReRs).

In June 2017, we searched articles accessible via Scopus for a combination of keywords related to sigmoid curves and to the above-mentioned domains of ecology. In some papers, the term sigmoid is not mentioned even if sigmoidal functions are used. Our sigmoid keywords therefore covered a wide range of meanings: we searched for “sigmoid” OR “nonlinear” OR “logistic”. We combined these keywords with other keywords related to the targeted ecological aspect: “SAR” OR “species-area” OR “species-resource” or “biogeography”.

Among the search results, we selected the papers where, according to the title and the abstract, the authors either used sigmoid functions or were interested in a sigmoidal form of relationship.

The 36 selected papers (cf. Annexe I. Table S1.1) were sorted according to the three possibilities: i) papers that did not use a sigmoid family term (“NO” in the second column), ii) papers that used a sigmoid family term but did not define it (“YES” in the second column and “NO” in the third), and iii) papers that either entirely or partly defined the sigmoid (“YES” in the second column and “YES” or “PARTLY” in the third).

As Annexe I. Table S1.1 shows, sometimes authors use sigmoidal function without ever specifically referring to the sigmoid family (13.9 %), but this number may be underestimated due to the difficulty of finding such papers. Most of the time, the authors use a word from the sigmoid family to define their functions (“sigmoid” or “sigmoidal”), but they do not define what they mean by these terms (72.2 %). What is quite surprising is that some authors create new sigmoid functions

and state that their functions have a sigmoidal form, but they never evoke the characteristics implied by this form and included in their function (e.g. Kobayashi 1976, Huisman et al. 1993a).

Finally, only a few authors take the time to define a sigmoid (13.9%), but typically the definition is fragmented or the functions imprecisely characterized, thus giving the impression of an incomplete definition. Sometimes definitions can even be confusing or contradictory.

Preston (Preston 1962a) proposed a descriptive definition of the shape of the sigmoid curve, which gives us an idea of the form but without specifying its properties: “it began at a low slope, steepened considerably, and then became less steep”.

Tjørve (2003, 2009) does not give a complete definition of the sigmoid curve, but does mention some of its characteristics when describing the functions he considers in his study. In Tjørve’s papers (Tjørve 2003a, 2009a), the characteristics common to all sigmoid functions include: i) the presence of an upper asymptote, ii) a lower j-shape (probably implying a lower asymptote), and iii) the presence of an inflection point. Tjørve (2003, 2009) also mentions two characteristics which vary among different sigmoid functions: symmetry around the inflection point, which may or may not exist; and the positions of the inflection point and of the asymptote.

Furthermore, in addition to being incomplete, these "definitions" may present other problems that impede understanding. This is the case when mathematical terms characterizing a mathematical object, here the sigmoid curve, are incorrectly used. For example, some authors erroneously define their sigmoid functions as “convex” (Tjørve 2003a, Gentile and Argano 2005, Tjørve 2009a). Indeed, in mathematics, a curve/function is “convex” if, for any two points A and B of the curve, the segment [AB] is entirely situated above the curve. Conversely, a concave

function is the opposite of a convex function (f is concave if and only if –f is convex). A concave curve is therefore a curve for which, for any two points A and B of the curve, the segment [AB] lies entirely below the curve. Yet, some studies make no distinction between the two curves and use “convex” for both convex and concave forms (Tjørve 2012), then distinguish them with the mentions “downward” or “upward”. Usually, given the properties attributed to the curves defined as convex, the term concave, rather than convex, is clearly the correct term. For example, what Tjørve (2009) described as a "constantly decelerating" convex curve is actually concave, and what he defined as a "J-shape" would correspond to the convex part of the sigmoid curve. This error is common since convex and concave shapes are often respectively described as a hump and a hollow (from the definition of a convex set), which can lead to confusion. Therefore, though the study is very interesting, the discourse is blurred by terms that are confusing (as also pointed out by Dengler 2009b). Consequently, we suggest using mathematical definitions and terms, so that all researchers will refer to the same definition of sigmoid curves.

If one moves away from the literature in ecology, we find that few definitions are easily accessible even in statistical literature. Hill and Lewicki (2006) propose one such definition in their glossary: a sigmoid function is “an S-shape curve, with a near-linear central response and saturating limits” (p724). This definition, which includes the notion of an S-shape discussed above, make it possible to understand the general shape and to accept different forms, but they are not necessarily very clear on which forms are included or excluded when we speak of a sigmoid, and the properties of the functions are not precise. Menon et al. (1996) also start by defining the sigmoid curves as S-shaped; then the authors define two sub-classes of sigmoids: i) simple sigmoids are “odd, asymptotically bounded, completely monotone functions in one variable”, and ii) hyperbolic

sigmoids are “a proper subset of simple sigmoids and a natural generalization of the hyperbolic tangent”. Although detailed, notably when characterizing certain functions, they seem to have forgotten to mention the monotonic character that such a function should have. Moreover, the two defined classes do not integrate all the possible sigmoidal forms; for example, “odd” excludes asymmetric curves and curves that do not intersect the origin. Finally, concerning definitions easily accessible to the general public, dictionaries are not of much better help since, for example, the French dictionary *Le Petit Robert* defines a sigmoid as a "sinuous curve with two waves of growth separated by a point of inflection" (translated from French), a very confusing definition (“Le Petit Robert : Sigmoïde” 2017).

To sum up, very few definitions of sigmoid functions are available in the ecological literature, and they are usually vague, or based on only certain characteristics, or can even contain errors. Therefore, it seems clear that the lack of a time-honored definition, or the use of unstable definitions, can lead to difficulties in producing studies and articles. This is particularly true for bibliographic research and for young researchers and students (PhD or Masters students) who are still forging their knowledge (Herrando-Pérez et al. 2017). It can also sometimes distort communication among collaborators. For example, within our own research group, differences of wording regarding the properties of different curves have surfaced, with misunderstandings of what is meant by “convex” and “concave”.

II. 3. Proposal of a clear definition

Although the definition on Wikipedia is globally correct (Wikipedia n.d.), this website cannot be used as a reference since the page can be modified at any time, making the definition unstable. We have therefore decided to propose a definition, which is stable, understandable for ecologists, and as complete as possible (including as many cases as possible) in this paper. For this purpose, we first looked at the characteristics of the functions used in the literature.

Ultimately, a sigmoid curve is a curve described by a real-valued, univariate function (a function f of a unique real-valued variable x that takes real values y=f(x)), defined over the whole set of real numbers, and which is continuous, infinitely differentiable, monotonic (always either increases or decreases), has at least one inflection point and is bounded on the Y-axis. The term "inflection point" refers to the point where the curve shifts in convexity: from convex to concave or vice versa. The change in slope is continuous and should therefore be distinguished from the term "breakpoint" used by ecologists, which, although we did not find a precise mathematical definition, seems to refer to a non-continuous function (e.g. in change point models, (Quandt 1958, Muggeo 2003).

Its inherent features imply that the sigmoid curve: i) has an upper and a lower asymptote if (x) varies over the set of real numbers; ii) can increase (starting with the lower asymptote and finishing with the upper asymptote, with a positive slope between them, Figure 1.1-a) or decrease (starting with the upper asymptote and finishing with the lower asymptote, with a negative slope between them, Figure 1.1-b); and iii) can be symmetrical or not around the inflection point or points (Figure 1.1-c).

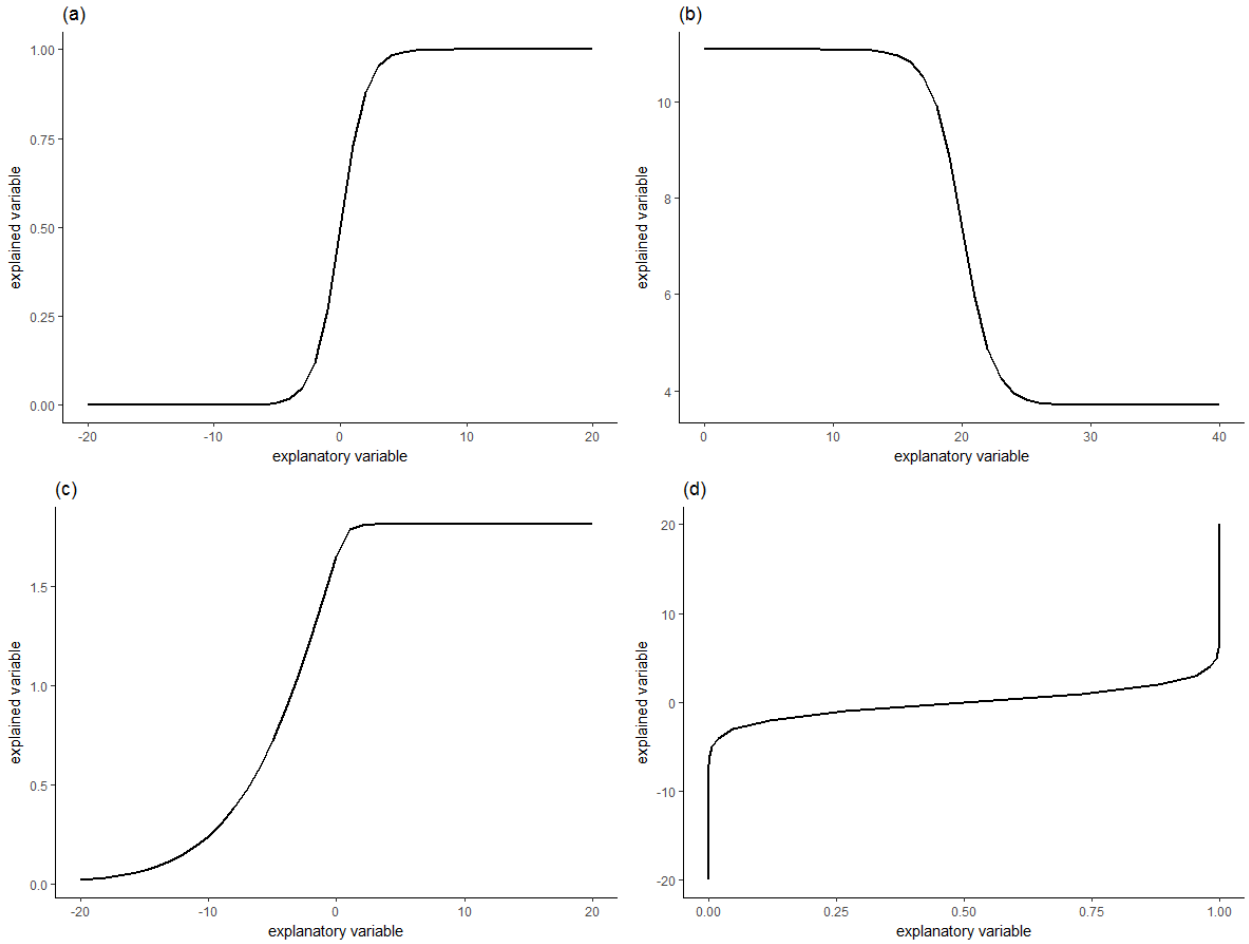


Figure 1.1 : Some possible forms of sigmoids and inverse sigmoids. (a) Simple logistic function, (b) decreasing sigmoid, (c) asymmetric increasing sigmoid, and (d) increasing inverse sigmoid.

We extend the definition given above to two other cases where the explanatory variable (x) is defined on the set of real positive numbers (x≥0) and: i) f(x) is a function of (x) over the entire set of real numbers and has a sigmoid curve, or ii) the above definition for the sigmoid curve applies to f(x) as a function of (x≥0) except for the requirement that f(x) defined over the entire set of real numbers. Indeed, in island biogeography, the function never occurs with negative x-values (since area cannot be negative). In this case, the sigmoid curve has only one of the two asymptotes. Even after extension, however, our definition does not include the case where (x) is bounded on both sides and therefore possesses neither of the two asymptotes (He and Legendre

2002). Note that f(x) as a function of (x) can have a sigmoidal form without f(x) as a function of log(x) or f(exp(x)) as a function of (x) being true, and vice versa.

The class of sigmoid functions includes the functions which, for the given parameters, meet the above definition. The same function may or may not belong to the sigmoid class depending on the value of its parameters. To return to a previous example, the Chapman-Richards function belongs to the sigmoid class if c>1. For other values of c, the function does not belong to the sigmoid class.

The sigmoid class can be divided into two sub-classes: (i) simple sigmoids, containing the functions that give curve shapes with a single inflection point, and (ii) multiple sigmoids containing functions that give curve shapes with several inflection points (i.e. a double sigmoid could fit the phenomenon described in Figure 6 in Lomolino, 2000). There must always be an odd number of inflection points in order to keep the two asymptotes on the Y-axis.

Based on the definition of the sigmoid class that we propose above, we inventoried the classical SAR or SReR functions selected from the prospect we conducted that belong to the simple sigmoid class, at least for some parameter values (see Table 1.1). We also described their characteristics, placing special emphasis on the constraints imposed on the parameter values or explanatory variable to ensure that the function is mathematically defined, is suitable in macro-ecology and does indeed have a sigmoidal form. We also provide the coordinates of the inflexion point, so that readers can distinguish between functions that are sigmoidal only when the whole set of real values for the explanatory variable is considered (i.e. functions with a negative abscissa value of the inflexion point) and those that are sigmoidal even when the abscissa values are positive.

Having a well-established definition of the sigmoid curve and understanding the constraints imposed on the parameter values of the functions which produce sigmoid curves allow us to better apprehend under which conditions a sigmoid function is adapted when one wishes to apply it to a dataset. For example, the Chapman-Richards function is defined only for ($x \geq 0$) and the curve obtained will only be of sigmoid shape when ($c > 1$) (see Table 1.1). Another, more extreme, example combines these two limitations: the persistence2 function. In fact, this function is sigmoid only if ($x > 0$), ($b = 0$) and ($c > 0$).

Another class of functions that is close to the sigmoid class is the class of inverse sigmoid functions. These are bounded on the X-axis and do not have an asymptote over the Y-axis (Figure 1.1-d). These functions have no biological reality in SReR and SAR and are not members of the sigmoid class as we define it. Other curves defined as sigmoid by some authors do not meet the requirements of our definition either, for example, “sigmoid curves […] free of upper asymptotes” (Tjørve 2012).

Table 1.1 : Some characteristics of sigmoidal functions present in the SAR and SReR literature.

	Formula	Constraints on parameters to be defined and relevant to macro-ecology	Further constraints required to be in the sigmoid class	Inflection point	Symmetry around the inflexion point	Lower asymptote	Intersects origin	Direction of the relationship
Common logistic	f(x) = a/(1 + exp(-b*x + c))	a>0	/	x=c/b y=a/2 In other terms y=50% of the upper asymptote	Point symmetry	Zero	No	Increasing (if b>0) or decreasing (if b<0)
Gompertz	f(x) = a*exp(-exp(-b*x+c))	a>0	/	x=c/b y=exp(-1)*a In other terms y=36.8% of the upper asymptote	Asymmetric	Zero	No	Increasing (if b>0) or decreasing (if b <0)
Extreme value	f(x) = a*(1-exp(-exp(b*x+c)))	a>0	/	x=-c/b y=[1-exp(-1)]*a In other terms y=63.2% of the upper asymptote	Asymmetric	Zero	No	Increasing (if b >0) or decreasing (if b<0)
Champan-Richards	f(x) = a*(1-exp(-b*x))^c	a>0, x≥0, c>0, b>0	c>1	x= log(c)/b y= a*(1-1/c)^c)	Asymmetric	/ (irrelevant since x is non-negative)	Yes	Only increasing

Cumulative Weibull distribution	f(x) = a*(1-exp(-b*(x^c)))	a>0, b>0, x≥0	c<0 or c>1	x=((c-1)/(b*c))^(1/c) y=a*(1-exp(-1+1/c))	Asymmetric	/ (irrelevant since x is non-negative)	Yes (if c>0)	Increasing (if c>0) or decreasing (if c<0)
Morgan-Mercer-Flodin (MMF)	f(x) = a*(x^c)/(b+(x^c))	a>0, b>0, x≥0 (with f(0)=a if c<0 to be continuous)	c>1 or c<(-1)	x=((c-1)*b/(c+1))^(1/c) y=a*(1/2-1/(2*c))	Asymmetric	/ (irrelevant since x is non-negative)	Yes	Increasing (if c>0) or decreasing (if c<0)
Cumulative beta-P distribution	f(x) = a*(1-(1+(x/c)^d)^(-b))	a>0, x≥0, c>0, b>0	d>1 or d<(-1/b)	x=c*((-d+1)/(-b*d-1))^(1/d) y=a*(1-(1+(-d+1)/(-b*d-1))^(-b))	Asymmetric	/ (irrelevant since x is non-negative)	Yes (if increasing, d>0)	Increasing (if d>0) or decreasing (if d<0)

Note that models II and III in Huisman, Olff & Fresco (1993), denoted as f(x)=M(1/(1+exp(a+b*x))) and f(x)=M*(1/(1+exp(a+b*x)))*(1/(1+exp(c))), are particular cases of the Common Logistic Function with, respectively, parameter (a) not estimated, and with parameter (a) estimated but with a given maximum value. The Archibald Logistic Function, denoted as f(x)=a/(b+c^x), is equivalent to the Common Logistic Function with (b), (c) and (a) in the Common Logistic Function, respectively equal to (-log(c)), (-log(b)), (a/b) in the Archibald Logistic Function. The He-Legendre Function, denoted as f(x)=a/(b+(x^(-c))), is equivalent to the Morgan-Mercer-Flodin Function with (a) and (b) of the MMF respectively equal to (a/b) and(1/b) in the He-Legendre Function. The type III Holling function, denoted as $f(x)=ax^2/(b^2 + x^2)$, is equivalent to the MMF, with (c) and (b) in the MMF respectively equal to (2) and (b^2) in the Holling III Function.*

II. 4. Ecological justifications and implications of sigmoid curve characteristics

Although some characteristics of the sigmoid definition are justified mainly by mathematical considerations, many can be related to ecological hypotheses or considerations. First, the presence of an inflection point represents the tipping point between the beginning of the gradient where the more X increases, the more the advantage conferred by X is important, and the end of the gradient where the advantage conferred by X allows less and less to overcome other limitations. In SARs, the use of sigmoid curves with this inflection point is justified by the following statement by Lomolino (2000a): "with richness remaining relatively low and apparently independent of area for the smaller islands, increasing rapidly to rise through an inflection point for islands of intermediate size, and then asymptotically approaching, or leveling off at the richness of the species pool for the largest islands". Many other fields of ecology are interested in models that can depict such a pattern (e.g. ecophysiology: Paine et al. 2012). Continuity and differentiability would allow us to formulate hypotheses not only on the mean value of the response variable, but also on the speed (first derivative) or acceleration (second derivative) of the relationship between the response variable and the gradient being studied, which however has not been done so far. The pattern depicted by Lomolino for SARs might have led us to define sigmoid curves only as increasing curves. Yet, we expect that in some areas of ecology the reversed situation might occur and that such patterns would indeed fall into the domain of the sigmoid curve. For example, still in biogeography, a decreasing sigmoid was considered in species-isolation relationships (Hachich et al. 2015). More generally in ecology, the decreasing sigmoidal

curve can be used in the case where the gradient studied has a negative effect on the response variable (e.g. Morante-Filho et al. 2015).

Second, the existence of asymptotes is also very much related to considerations from ecology. The upper asymptote, implying a threshold above which the mean of the response variable (y) cannot go, theoretically reflects the Liebig law of the minimum in ecophysiology and ecology (Paris 1992a, Austin 2007). In this case, the studied predictor would be the first limiting factor , and an increase in this limiting factor would lead to an increase in the explained variable. Then, upon reaching the asymptote, the predictor would no longer be limiting; instead, another unmeasured environmental factor would take over, though its influence would be insufficient to make the explained variable increase any further. Inversely, the presence of a lower asymptote implies that the mean of the response variable cannot be lower than this asymptote. The existence and value of such an asymptote can often be related to the conjunction of the monotonic relationship, the nature of the variable considered and the nature of the system under study. In studies focusing on the response of a single species, the lower asymptote is therefore usually zero (e.g. Huisman et al. 1993a). However, when studying community response, often a lack of a resources does not necessarily imply a total loss of species richness (for example, when studying a system where species are mobile). In such cases, a logistic function where f(x) is a function of log(x), whose lower asymptote is necessarily located at zero (y = 0), is not actually adapted (Godeau et al. n.d.).

The third component of our definition is asymmetry of the curve. Symmetric sigmoid curves, like the common logistic function, are widely used, but more for their ease of modelling than for their underlying ecological theory. Indeed, for bell-shaped curves, (Austin 1976) stated: “there is no a

priori reason to assume that organisms' responses should follow such a symmetrical curve". Diverse phenomena can explain asymmetrical curves (Austin 1990 and Austin & Gaywood 1994 for phyto-ecology) and theoretically supported asymmetry can also appear with sigmoidal curves (e.g. Lim et al. 1998).

II. 5. Conclusion and perspectives

Our literature prospection points out the lack of a clear, stable, universally accepted definition of the sigmoid class of functions in ecology. Some aspects of sigmoid curves are typically ignored (symmetry, direction of the relation, etc.). We also found cases of misuse of convexity to define a curve or a function.

As Jeremy Fox stated "words are imprecise, and so purely verbal models and verbal arguments often are ambiguous or even invalid, even if apparently supported by empirical data (like Elton's verbal arguments about why diversity and complexity beget stability). Mathematics has the virtue of forcing precise definitions of terms, precise and complete specification of assumptions, and rigorous derivation of conclusions" (Fox 2011a). It is therefore unfortunate to accept vague verbal definitions (such as "S-shape" or "J-shape") when one is using a term derived from mathematics.

That is why we have proposed a definition that we hope will allow for better harmonization of what is meant by the term "sigmoid" when describing a curve or a function. In addition to clearly formulating the concept, our definition allows various functions to be united under the same banner (sigmoid class, presented in Table 1.1). This definition also excludes some functions that

were previously considered to belong to the sigmoid family and which, in our opinion, should not be defined as such (sigmoid without an upper asymptote or inverse-sigmoid).

Having clear definitions makes it possible to more clearly reflect on the underlying concepts and theories implied by the functions available, and to visualize the most appropriate form of curve to adopt according to the ecological context. After defining and reflecting on the lower asymptote and asymmetry, the researcher naturally questions the choice of link function in the context of binomial logistic regressions. Classically, users of such tools choose canonical link functions such as the logit or the probit function. These two functions belong to the sigmoid class but they are symmetric around the inflection point and they have pre-specified minimum and maximum asymptotes (respectively 0.0 and 1.0). However, the inherent properties of such link functions could have strong ecological limitations, which would restrict their use in some cases. For example, having a maximum of 1.0 (meaning almost sure presence) along the gradient does not reflect biological situations where, even if local habitat conditions are optimal for the organism, the organism could be absent (e.g. due to dispersal limitation inside a metapopulation; Hanski and Gilpin 1997). Along the same lines, sigmoid and logistic functions are sometimes confused with each other, whereas the latter is nothing more than a particular type of sigmoid (e.g. Hunsicker et al. 2015). Such confusion may prevent researchers from considering other families of functions that fall into the sigmoid class without being logistic.

In future papers, we aim to develop a sigmoid function that incorporates the characteristics retained in this paper: first in an SReR context and second, in binomial logistic regressions. Such development of the sigmoid class might be of more general use in ecology, e.g. by broadening the scope of possibilities in binomial logistic regressions.

Finally, we hope that in future papers, authors who define a new sigmoid function, or use an already existing one, will take the time to specify the properties of the function and to clearly mention their implications and/or justifications in ecological terms.

III. DISCUSSION DU CHAPITRE 1

La biogéographie est un domaine de recherche pour lequel diverses fonctions mathématiques monotones sont utilisées dans les modèles pour expliquer la relation, notamment grâce à des fonctions de formes sigmoïdales, entre la biodiversité et la surface du territoire (ou une autre ressource). Une exploration de la littérature dans ce domaine a permis de révéler une utilisation approximative des termes en rapport avec la classe de fonctions sigmoïdes, imputable à une absence de définition claire et harmonisée au sein de la communauté d'écologues. Ainsi, si les fonctions sigmoïdales sont mal assimilées, les écologues ont tendance à méconnaitre certaines propriétés inhérentes telles que la symétrie, la pente, la direction de la relation ou encore les positions des asymptotes. Par exemple, Medellín & Soberón (1999) ont commencé par appliquer un modèle sigmoïdal à leurs données. Puis, préférant appliquer un modèle plus attendu, ils ont choisi d'exclure certaines données pour que leur jeu de données s'adapte au modèle logarithmique. Fattorini, Maurizi, & Giulio (2012) soulignent qu'il aurait été préférable de ne pas manipuler les données et de conserver un modèle correspondant à l'ensemble du jeu de données. Les données exclues, correspondant à la première partie de la courbe sigmoïdale (où la pente est plus faible), peuvent potentiellement être tout aussi importantes d'un point de vue écologique que les données représentées par le reste de la courbe. En effet, la première partie de la courbe sigmoïdale pourrait refléter divers mécanismes écologiques qui méritent d'être étudiés (problèmes d'échantillonnage, facteurs limitants, exclusions, etc.). Au travers de cet exemple, il devient évident que, si la forme de la courbe sigmoïdale et ses implications ne sont pas

suffisamment reconnues ou définies dans l'esprit de l'écologue, celui-ci risque de faire de fausses déductions et se tromper sur l'interprétation des résultats.

Pourtant les mathématiques ont pour avantage de définir de manière explicite les concepts abordés, et donc d'en obtenir des conclusions logiques et supposément peu variables (Fox 2011b). Une définition claire (à la fois verbale et mathématique) d'un outil, comme proposée dans le manuscrit, permet donc de mieux appréhender les propriétés mathématiques inhérentes à l'outil et de concevoir les implications écologiques de celles-ci. Cette meilleure compréhension des propriétés mathématiques permettra d'implémenter dans les modèles des fonctions sigmoïdales avec des formes plus complexes tout en ayant des hypothèses a priori sur les raisons des formes proposées.

CHAPITRE 2

Les effets aléatoires dans les modèles bayésiens hiérarchiques non-linéaires de forme sigmoïdale complexe

I. INTRODUCTION DU CHAPITRE 2

Lors d'études impliquant de la modélisation, l'un des critères majeurs examinés est le test de significativité statistique de la relation. Lors de ce test, une hypothèse dite « nulle » est posée. Elle représente la position par défaut selon laquelle il n'y a pas de différence notable entre deux groupes. Ce test de significativité statistique est donc souvent utilisé pour tirer des conclusions sur l'existence ou non d'une différence entre les deux groupes étudiés. Or, il est important de noter que les conclusions ne doivent pas être basées uniquement sur ce test qui possède plusieurs limites (Wasserstein and Lazar 2016, Amrhein et al. 2019, Hurlbert et al. 2019, McShane et al. 2019, Wasserstein et al. 2019). Tout d'abord, un test de significativité non significatif ne « prouve » pas l'hypothèse nulle, mais elle échoue simplement à la réfuter. C'est-à-dire que les conditions dans lesquelles il est appliqué ne permettent pas de réfuter l'hypothèse nulle pour différentes raisons potentielles: i) l'hypothèse nulle est en effet « vraie » ; ii) l'intervalle de confiance est trop élevé ; iii) malgré un modèle bien adapté, la pente est trop faible compte tenu des données disponibles pour être déclarée significativement différente de zéro ; iv) il y a une relation entre les deux variables mais le modèle utilisé dans l'analyse statistique n'est pas adapté ; v) pour des raisons dépendantes des données (utilisation d'un échantillon trop faible – cf. iii – ou biaisé, effet observateur…). L'inverse peut également être vrai et un test qui révèlerait des différences significatives peut également provenir d'un faux positif. Ainsi, il ne faudrait pas se hâter de conclure que des variables sont corrélées, parce que le test de significativité est concluant. De plus, dans le langage courant, les résultats de test de significativité ont tendance à être séparés en uniquement deux catégories « non-significatifs » et « significatifs » sans

réellement prendre en compte la continuité qui existe entre ces deux affirmations. De fait, le test de significativité repose sur un seuil de confiance, souvent de 5%. Ce seuil étant choisi arbitrairement, il peut changer d'une étude à l'autre. Enfin, et non des moindres, le test de significativité ne nous renseigne aucunement sur l'amplitude de la relation entre les grandeurs étudiées.

L'étude de la magnitude des effets constitue un véritable ajout à la significativité notamment pour rendre compte de la force des effets. Ce complément est d'autant plus important dans des contextes de gestion. Yoccoz exprime d'ailleurs à ce sujet : « Ici encore, le point n'est pas l'aspect 'statistiquement significatif', qui le devient nécessairement si les données sont en nombre suffisant, mais 'biologiquement significatif', une question qui fait appel au contexte biologique et à laquelle il est souvent délicat, mais essentiel, de répondre » (Yoccoz 1999). Ainsi, la significativité statistique se distingue de la significativité pratique dans le sens où cette dernière nous renseigne sur l'efficacité des pratiques de gestion envisagées. Cette étude de la magnitude permet donc de ne pas prendre de décision de gestion dont les effets seraient trop faibles au regard des objectifs, mais de cibler plutôt des prescriptions dont les effets seraient forts avec le moindre effort. Dans ce but, il est nécessaire de prendre en compte les modifications réalisables par les gestionnaires (pour un exemple de réflexion autour des bryophytes, cf. Bouget and Gosselin, 2017 p48). On distingue donc quatre cas : 1) l'effort consenti pour la gestion est faible et l'effet sur la variable d'intérêt est fort, des pratiques de gestions peuvent être appliquées sans hésitation ; 2) l'effort consenti pour la gestion est faible et l'effet sur la variable d'intérêt est faible, des pratiques de gestions peuvent être envisagées sans grande conviction (en fonction de

l'importance que l'on accorde au résultat) ; 3) l'effort consenti pour la gestion est fort et l'effet sur la variable d'intérêt est fort, des pratiques de gestions peuvent être également considérées (en fonction des ressources et du potentiel d'action des gestionnaires) ; 4) l'effort consenti pour la gestion est fort et l'effet sur la variable d'intérêt est faible, des pratiques de gestions semblent vaines.

Parmi leurs diverses propriétés, les fonctions sigmoïdes présentent une pente qui est variable en fonction d'où l'on se situe sur le gradient (faible au début, puis forte, et faible de nouveau). Ceci a pour conséquence d'obtenir une magnitude des effets différente selon la position sur le gradient. Ainsi, l'étude de la magnitude doit être réalisée à partir de plusieurs points du gradient afin d'en apprécier les variations. Dans un contexte de gestion de ressource par exemple, il est primordial d'étudier la magnitude des effets à partir de points qui soient réalistes, telle que la quantité de ressource actuelle. Il est peu efficace, par exemple, d'étudier la magnitude des effets au niveau du point d'inflexion de la courbe sigmoïde s'il est impossible d'atteindre ces niveaux lors la gestion effective.

Les fonctions non-linéaires, y compris sigmoïdes, permettent de prendre en compte un plus grand nombre de types de variabilités spatiales que le modèle linéaire. En effet, les modèles dits hiérarchiques, incluant des effets aléatoires, permettent de prendre en compte une variation de la relation qui dépend du facteur de hiérarchisation (variabilité spatiale, écologique…). Dans un modèle de régression linéaire seule l'ordonnée à l'origine (intercept) et parfois la pente (beaucoup plus rarement) peuvent varier pour prendre en compte cette variation dans les données. A l'inverse, un modèle non-linéaire peut varier dans son intercepte et sa pente également mais aussi dans sa forme en faisant varier les autres paramètres de la courbe. Le

modèle sigmoïde pourra, par exemple, varier en plus au niveau de la position des asymptotes, ou la forme globale de la courbe.

Nous avons exploré ces deux caractéristiques (magnitude des effets et variabilité hiérarchique) dans un contexte de gestion forestière, en étudiant la communauté d'espèces de coléoptères saproxyliques en fonction du volume de bois mort dans les forêts françaises. Pour cela, nous avons utilisé un indice de biodiversité qui est une métrique résumant la biodiversité avant de l'analyser (e.g. : richesse spécifique, indices de diversité et d'équitabilité, abondance, indices de dissimilarité…). Nous avons choisi d'utiliser la richesse spécifique (c.-à-d. le nombre d'espèces dans une communauté, dans un paysage ou un paysage marin, ou dans une région[6]), pour sa simplicité. Cependant, cette métrique ne résume la communauté correctement que lorsque les espèces qui la composent répondent toutes dans le même sens, en présence-absence, et avec la même magnitude. Nous avons comparé des modèles incluant diverses fonctions (linéaire, exponentielle, et plusieurs fonctions sigmoïdes) tout en faisant varier la forme de la courbe entre massifs forestiers, de toutes les manières possibles pour chaque courbe afin de prendre en compte la variation spatiale. Nous avons en effet supposé que plusieurs aspects pouvaient varier d'un massif forestier à l'autre (tels que la fertilité du sol, la température, l'historique de gestion sylvicole, les espèces d'arbres dominantes…). Ces derniers peuvent non seulement avoir un effet sur la diversité moyenne de la communauté de coléoptères saproxyliques mais aussi sur la forme de relation entre cette communauté et la disponibilité du bois mort (comme observé par Zilliox

[6] Traduit de l'anglais depuis (Colwell 2009) : "The number of species in a community, in a landscape or marinescape, or in a region".

and Gosselin 2014 pour la flore terrestre). Nous avons ensuite étudié la magnitude des effets, afin de regarder si des recommandations de gestion concernant la quantité de bois mort à laisser dans les forêts étaient envisageables.

II. MANUSCRIT 2

The importance of being random! Taking full account of random effects in nonlinear sigmoid hierarchical Bayesian models reveals the relationship between deadwood and the species richness of saproxylic beetles

Ugoline Godeau[1], Christophe Bouget[1], Jérémy Piffady[2], Tiffani Pozzi[1], Frédéric Gosselin[1]

[1] INRAE, UR EFNO, Domaine des Barres, F-45290, Nogent-sur-Vernisson, France.

[2] INRAE, UR MALY, Centre de Lyon-Villeurbanne, F-69616 Villeurbanne, France.

Statut : accepté dans la revue “Forest Ecology and Management”

Abstract: Hierarchical models are used to study the relationship between a response variable and a predictor in structured data. Random effects are meant to capture the structured part of variability among groups of observations. In ecology, random effects are usually incorporated into the intercept. Their application to the other parameters of the curve, especially in nonlinear curves, has been understudied. However, applying random effects to different parameters of the function is of interest, as it allows us to account for variations in the shape of the relationship over groups of observations.

Our study was based on Bayesian models linking the local quantity of deadwood to the local species richness of saproxylic beetles in French forests. Our hypothesis was that it was important to account for inter-forest variations of the relationship to better fit the data. Since a sigmoidal curve seemed adapted to studying this relationship from an ecological point of view, we paid special attention to commonly used sigmoidal functions, but also included two new ones for biogeography originating from ecophysiology (one sigmoid with estimated asymptotes and one with estimated asymptotes allowing asymmetry). We applied various settings of random effects to these different mean functions. We compared, evaluated and interpreted the models and results according to several criteria (WAIC, comparison of significance of the difference in terms of LOOic, goodness-of-fit p-values and magnitude of the effect).

We first found that models without random effects were systematically the worst and that the best model was not necessarily the one with random effect incorporated

into the intercept, as is usually done in ecology. Secondly, we found that, in most cases, for a given mean function, the best model had several random effects, and the model with the most random effects performed nearly as well as the best models. Furthermore, the inclusion of random effects revealed statistically significant relationships between deadwood volume and species richness. Thirdly, we revealed a complementarity between the different assessment criteria, each one giving important information for the selection and interpretation of the models. In conclusion, future forest biodiversity management studies should incorporate random effects into the modeling framework so that more robust conclusions can be made about the relationships, based on complementary post-fitting analysis criteria.

Keywords: Biodiversity relationships; Evaluation criteria; Forest Management; Goodness-of-fit; Model comparison; Spatial variation; Magnitude analysis; Ecological indicator

II. 1. Introduction

Ecology is a science of relationships. The link between a variable of interest (abundance, species richness, survival, recruitment...) and environmental conditions (climate, resources, stress factor…) is therefore very often investigated in ecology. To study these links, statistical models are one primary tool (Clark and Gelfand 2006, Gimenez et al. 2014), and most of the time, we need a response curve that fits the data. Linear models and generalized linear models (GLM) have been very frequently used to identify these links between variables because they are easy to model (Bolker et al. 2009), even if limited.

One way to go further than (generalized) linear models is to use semi-parametric statistical tools such as Generalized Additive Models (GAMs) that allow the model to flexibly transform explanatory variables and thus to handle certain kinds of non-linearity. A second way is to explicitly use nonlinear functions other than the link functions provided by GLMs. In explicit

nonlinear models, the mean function between the linear combination of explanatory variables and the mean of the target variable is nonlinear and can depend on more than one parameter, which can incorporate the explanatory variables – contrary to GLMS & GLMMS. Nonlinear mean functions should be considered more often in ecology for several reasons . First, hypotheses, theories or empirical relationships may require nonlinearity and may suggest a shape or a function for the expected response (Austin 2002, 2007). Second, the purpose and use of the model may require a specific non-linear function, for example, if parameters are easy to interpret graphically, ecologically or in decision-making terms (e.g. in terms of threshold values). Finally, logical or mathematical arguments within an academic framework, can also foster the use of nonlinear mean functions and help determine the shape of these functions (Harrell 2001, Bolker 2008).

Despite some authors' (e.g. Bolker et al., 2013) efforts, explicit nonlinear techniques are still underused in some areas of ecology, for example in the field of species distribution modelling (SDMs) (Oksanen and Minchin 2002, Austin 2007). There are, however, areas of ecology where nonlinear relationships have long been studied: plant growth, landscape ecology (Grêt-Regamey et al. 2014) and island biogeography, including the particular cases of species-area relationships (SARs) (He and Legendre 1996) and species-resource relationships (SReRs) (e.g. Hunsicker et al., 2015; Oksanen & Minchin, 2002; Pausas & Austin, 2001). SReRs are mostly phenomenological, in that they target ecological patterns at macro scales rather than ecological processes that explain these patterns. Huisman et al. (1993), Tjørve (2003, 2009), Dengler (2009), and Williams et al. (2009) collected and confronted different functions available in a context of SReRs. In both SReRs and plant growth, we note the importance of a particular curve shape, the sigmoid. Indeed,

Preston (1962) long ago argued that this form of relationship is likely to occur in the context of SAR. Lomolino (2000) reinforced this idea by describing the possible underlying phenomenon: "with richness remaining relatively low and apparently independent of area for the smaller islands, increasing rapidly to rise through an inflection point for islands of intermediate size, and then asymptotically approaching, or levelling off at the richness of the species pool for the largest islands". In the context of plant growth, the same curve shape can occur because as plants initially grow from small sizes, their resource-gathering capacity increases proportionally (so that their relative growth rate stays approximately constant, leading to exponential growth), while resource or architectural limitations eventually overcome this advantage of greater size. One difference is that sigmoidal functions used in SReRs may be less adaptable to various situations than in plant growth studies, where there has been a real effort to find new functions of sigmoid shapes that are flexible (e.g. Birch, 1999; Paine et al., 2012; Yin et al., 2003). The flexibility of the function makes, for example, asymmetry, or a lower asymptote different from zero possible, which can be a major asset (Godeau et al. n.d.).

In addition to the insufficient use of nonlinear models, another limit in ecology is that the relationship between biodiversity and resource has often been estimated to be constant in time and space. Yet, time or/and spatial variation is pertinent from an ecological perspective. Indeed, ecological systems are complex systems that are related to a variety of gradients; as such, relationships within these systems are expected to vary over space, time (Biggs et al. 2009) or with ecological conditions (Drakare et al. 2006). As an illustration, recent results in ecology have acknowledged that the relationships between surrogate measurements and biodiversity do vary in geographical or ecological space and that these variations should be accounted for in order to

use surrogates efficiently (Zilliox and Gosselin 2014, Pierson et al. 2015). Hierarchical models have been developed to study data when information is available on different levels of observation (Gelman 2004, Biggs et al. 2009, Bolker et al. 2009, Cressie et al. 2009) and parameters vary with grouping structure, which is often associated with space or time. In a nonlinear context, hierarchical models also allow us to take into account the relationship's variations in geographical or ecological space by varying the shape of the curve in space. Yet, in ecology, we seem to often be satisfied with adding variation to the general mean level of the curve by applying random effects only to the intercept (Schielzeth and Forstmeier 2009), and, much more rarely, the slope (called random-slope models, but mostly linear models). The application of random effects to other parameters of the curve, which yields real variations in the shape of the curve with space (be it geographical or ecological), is underused in ecology (Schielzeth and Forstmeier 2009). Much as for spatially auto-correlated random effects (Saas and Gosselin 2014), it is likely that introducing random effects into nonlinear models will have non-negligible effects on the results (Fortin 2013). It may indeed impact both the mean estimate and the standard error, in contrast with linear models in which including random effects only on intercepts simply increases standard errors without any effect on mean estimates. In other scientific fields (e.g. Codd & Cudek, 2014; Cripps & Pecht, 2017), random effects are more systematically included on all model parameters. For example, this is the leading procedure in drug development science, where between-individual variability is fully accounted for in all the parameters of nonlinear functions (e.g. Drikvandi, 2017; Ishibashi et al., 2003; Pillai et al., 2005). In brief, studying ecological relations in a hierarchical nonlinear context makes it possible to account for the flexibility of the relationship through the use of nonlinear functions while taking into account spatial variations.

SReRs may be used to search for critical resources for animal forest communities, e.g. specialist deadwood-associated beetles (hereafter 'saproxylic' beetles), based on an a priori species-energy hypothesis concerning species-habitat associations (Seibold et al. 2016). The relationship between deadwood quantity and the diversity of saproxylic organisms has long been studied in forest ecology. Most results have been obtained using linear statistical tools. Only very few studies involved nonlinear relationships or nonlinear transformations for deadwood (e.g. Martikainen et al., 2000, with field data; Ranius & Jonsson, 2007, with simulated data). Nevertheless, a relationship as described by Lomolino, (2000) for SARs can also be expected in this particular case of SReR (with the saproxylic beetle species richness relationship to deadwood following a sigmoidal curve shape). Indeed, an increasing monotonic relationship, with a zone where the relationship is stronger than at the end of the gradient, has been observed (e.g. Grove, 2002; Martikainen et al., 2000), the increase in species richness with deadwood being limited by the size of the species pool. The slope could also be low at the beginning of the gradient when the resource is too rare (Godeau et al. n.d.). Recent results in deadwood research ecology have acknowledged that the relationships between deadwood and saproxylic biodiversity do vary according to ecological conditions (Bouget et al. 2014). This agrees with Lassauce et al.'s (2011) meta-analysis, which found a weaker ecological correlation between the local quantity of deadwood substrates and species richness in temperate forests than in boreal forests. It is therefore urgent to also fit this relationship with nonlinear tools, and to study whether or not both linear and nonlinear relationships vary in space in order to better apprehend the notion of deadwood threshold and whether it is constant or not throughout space.

In this paper, we tested the possibilities offered by explicit nonlinear, hierarchical Bayesian models phenomenologically linking saproxylic beetle species richness with deadwood volume. We had three main goals. First, we assessed the impact of introducing random effects into an explicit nonlinear context by addressing three questions: (i) Does the introduction of random effects improve the predictive capacity and goodness-of-fit of the model? (ii) Is the intercept the best choice to introduce random effects? (iii) Are the different possible choices in terms of random effects equivalent for a given mean function? If not, what methodology should we use to choose random effects? Secondly, we introduced two new sigmoidal functions for SReRs, based on the function developed by Paine et al. (2012); these functions seemed particularly flexible and we deemed they would be good candidates for application in SReRs. We compared them with different functions, sigmoid and other, and assessed their performance. Thirdly, we tested a new approach to select and interpret statistical models based on model selection, analyses of goodness-of-fit and relationship magnitude. Our general approach was to select the models with the best predictive capacity from a set of models that varied according to mean function and random effects settings, and to complement this selection with two other viewpoints: goodness-of-fit of the model with the data according to specific discrepancy functions, and magnitude of the relationship.

II. 2. Methods

II. 2. a. Datasets

The biodiversity target variable was the species richness (SR) of saproxylic beetles, while the explanatory variable was deadwood volume (gradient X). The deadwood volume reflects the resource's availability to saproxylic beetles, which depend on deadwood for habitat (either during a part or all of their life cycle) or on other organisms that are themselves dependent on deadwood. Our study is based on extensive data compiled from seven different ecological projects (cf. Annexe II. Table S1.1 and Annexe II. Figure S1.1). Inside each site, we followed standardized protocols to collect environmental and entomological data from several sampling plots. The plots were about 0.5 ha in size and were located several hundred meters from each other. The explanatory variable (deadwood volume), was measured by means of a dendrometric survey during the winter period, by measuring downed deadwood of more than 2.5 cm in diameter and standing deadwood of more than 7.5 cm (cf. Annexe II. Section S1.1), as is done for the French indicator 4.5 related to deadwood within the biodiversity criterion for sustainable forest management (Maaf and IGN 2016). To assess the biodiversity target variable, we sampled beetle diversity using a standardized trapping protocol during a single sampling year (the same year as the dendrometric survey for the same project). Flying saproxylic beetles were caught in standardized cross-vane flight interception traps (PolytrapTM) suspended roughly 1.5m above the ground. Each plot had between one and three traps. Each trap was active for six periods during the study, and results were obtained for one to six sampling periods per trap (some traps were destroyed during sampling periods). To shorten the time it took to fit the Bayesian model,

we used data from only one, randomly selected, trap per plot in our analyses. We totaled the results for all sampling periods for that trap, which meant that we had to take the number of sampling periods per trap into account in the model cf. Methods – Main model settings). All saproxylic beetles were identified to the highest possible taxonomic level. With the same taxonomic resolution in all sets, 51 beetle families were considered in the analyses. The dataset was hierarchized by the site factor, corresponding to the forest districts. As this is a geographical factor, each site may differ in several aspects (such as soil fertility, temperature, harvesting history, dominant tree species…), each of which could not only have an effect on diversity but also may have an effect on the relationship between the saproxylic beetle community and deadwood availability itself (as observed by Zilliox & Gosselin, 2014 for ground flora). Overall, our dataset encompassed 589 plots distributed over 33 sites (Annexe II. Table S1.1 and Annexe II. Figure S1.1).

II. 2. b. Functions

We first selected three classical functions used in ecology and biogeography: the linear function, the exponential function and the constant function (our null model). The first is the most classical form of parametric model in statistics while the second is more often used for non-negative discrete variables, especially in the framework of GLMMs (Bolker et al., 2009) – as in our study. Secondly, we selected three sigmoidal functions from the different functions available in the literature for statistical modelling in the context of SARs (Dengler, 2009; Godeau et al., n.d.-a; Huisman et al., 1993; Tjørve, 2003, 2009; Williams et al., 2009): the 3-parameter "logistic" function, the extreme value function (EVF) and the Gompertz function. These three functions

were selected among others because (i) they gave good results in a preliminary frequentist analysis and (ii) they accepted contrasted types of asymmetry around the inflection point. However, the three selected sigmoidal functions had two main limitations. First, they all required a lower asymptote of zero. This entails some constraints on the shape of the function (Figure 2.1.A, B and C). Yet, species dispersal capacity may explain why some individuals can be caught in sample stands they do not originate from (Brin et al., 2009) and can be found where local conditions would not allow it, in our case, in patches without any deadwood resources. Indeed, in Species resource Relationships (SReRs), contrary to a null area in the context of Species Area Relationships (SARs), the absence of resources does not necessarily imply an absence of the studied species. A second limitation of existing sigmoidal functions is that these functions are either symmetrical around the inflection point (the widespread 3-parameter “logistic” function) or, have a fixed asymmetry (as for EVF or Gompertz). These functions therefore lack flexibility in this respect because there is no obvious reason to assume that organisms' responses should follow such a symmetrical curve (for symmetric functions, cf. Austin, 1976 for Gaussian curves; Huisman et al., 1993) and inflexible (for both symmetric avec fixed asymmetric functions).

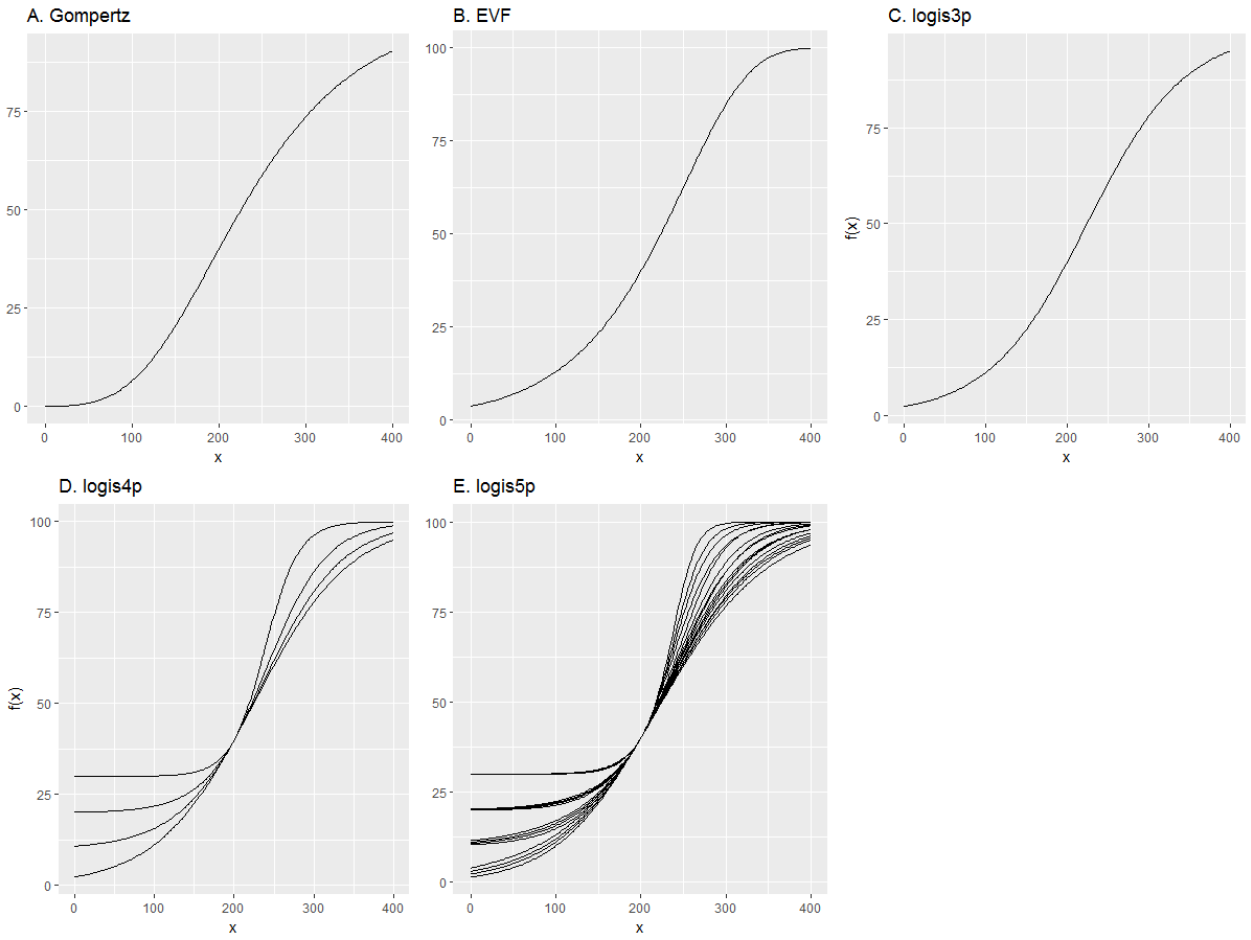


Figure 2.1: Possible shapes of the five sigmoidal functions studied (Gompertz, EVF, logis3p, logis4 and logis5p), when the following two parameters varied: the lower asymptote (L= 0, 10, 20 or 30) for logis4p and logis5p and the asymmetry parameter (e= 0.5, 0.75, 1 or 1.5) for logis5p. The following parameters were held fixed: upper asymptote (K=100), the ordinate at the median of x (here median=200) (M0=40) and the slope at the median of x (slope=80).

In order to overcome these two problems, we applied two new functions for SARs and SReRs. These functions have a sigmoid form as defined by Godeau et al. (n.d.-a). Our first function is a logistic function with four parameters, which makes it possible to control the lower asymptote, unlike the three-parameter logistic function with a lower asymptote fixed at zero. Since this type of four-parameter logistic function has proven to be efficient in various fields (e.g. Loken & Rulison, 2010 in psychometrics), we adapted one parametrization of such a function, used in

physiology by Paine et al. (2012) ("4-parameter logistic" function in Table 1 of Paine et al., 2012). Furthermore, we also applied a 5-parameter logistic function that makes it possible not only to control the lower asymptote but to obtain asymmetrical curves. Even with weak asymmetry, such a function has been found to produce good results compared to a symmetric 4-parameter logistic function (Gottschalk & Dunn, 2005; Ricketts & Head, 1999). We selected and reparametrized (see paragraph below) the 5-parameter function from Paine et al. (2012) (the "5-parameter logistic" function cf. Annexe II. Equation S1.1 and Appendix 2 in Paine et al., 2012), that is also very close from the five-parameters Richards function (Richards, 1959). The number of parameters and the encoding specification of both these functions give more freedom to the possible shapes of the relationship (Figure 2.1.D and E). Finally, both functions have parameters that are easy to interpret with a simple mathematical transformation in order to give clear recommendations to managers. In addition, the function in the five-parameter model (logis5p) can be asymmetric around the inflection point and the direction and intensity of the asymmetry is estimated; this is not the case for the EVF or Gompertz functions. The 5-parameter model is reduced to the four-parameter model when the adjusted data have a symmetrical shape.

There are different ways of defining how parameters and explanatory variables are involved in a function (herein called parametrizations, Bolker 2008), and those different parametrizations have their importance in a Bayesian framework especially because they affect the parameter space and the choice of prior distributions (Gelman, 2004). We tested many different parametrizations for each function (logis4p and logis5p) and retained the one for which the Bayesian model was

most successful in converging and estimating the parameters of the functions without too much MCMC autocorrelation.

II. 2. c. Main model settings

We chose a negative binomial distribution for the distribution of species richness as our data were over-dispersed relative to the Poisson distribution. For reasons similar to those of Lindén & Mäntyniemi (2011), and to avoid Bayesian convergence problems with the negative binomial form that was based on a linear relationship between variance σ^2 and mean μ (noted NB1 in Lindén & Mäntyniemi, 2011), we used a flexible version of the negative binomial. Our parametrization depended on two parameters λ and α and differed from that of Lindén & Mäntyniemi (2011) by its variance parameter (σ^2), since we had: $\sigma^2 = \lambda\,\mu^{1+\alpha}$, with $\lambda \geq 1$ and $0 \leq \alpha \leq 1$, instead of $\sigma^2 = \omega\mu + \lambda\,\mu^2$. This alternative version is closer to the Taylor power law (Eisler et al., 2008; Kilpatrick & Ives, 2003; Lepš, 1993), even though we recognize that both versions could fit ecological data well (Routledge & Swartz, 1991).

We denote as i the index of the site and as j the index of the plot within the site throughout this article. The explanatory variable x in the sigmoidal function was centered at its median (Mx) and scaled by its interquartile (IQx) to ensure good control over the meaning of the parameters involved in the sigmoidal function (Equation 1):

$$\tilde{x}_{i,j} = \frac{x_{ij} - Mx}{IQx} \qquad \textbf{\textit{(eqn 1)}}$$

We also added a correction term to take into account variations in the number of periods per plot in order to account for sampling effort. Indeed, the more sampling periods, the greater the sampling effort. For our data, the number of active sampling periods per plot varied due to trap disturbances (e.g. trap destruction) and this number might have an impact on the number of species captured (Martikainen et al., 2000). The correction term was calculated based on a power function of the number of periods (number of trapping months) for the trap – written as an exponential function of $\widetilde{lNp}_{i,j}$, the normalized logarithm number of periods – with an estimated parameter ρ: $\exp(\rho\ \widetilde{lNp}_{i,j})$, where

$\widetilde{lNp}_{i,j} = (\log(\mathrm{N}_{i,j}) - \mathrm{mean}(\log(\mathrm{N}_{i,j}))/\mathrm{sd}(\log(\mathrm{N}_{i,j})))$.

In the case of the most complex mean function (logis5p), the model was as follows:

$$Y_{i,j} \sim NB\ (F(\tilde{x}_{i,j}, \widetilde{lNp}_{i,j}, \theta_i\ , \varphi), \alpha, \lambda) \qquad \textbf{\textit{(eqn 2)}}$$

with:

$$F(\tilde{x}_{i,j}, \widetilde{lNp}_{i,j}, \theta_i\ , \varphi) = \exp(\rho\ \widetilde{lNp}_{i,j} + h_i)\left[\mathrm{L} + \frac{(\mathrm{K}-\mathrm{L})}{\left(1+\left\{\frac{(\mathrm{K}-\mathrm{L})}{(\mathrm{M}_0-\mathrm{L})}\exp\left[-\frac{r_i}{(\mathrm{M}_0-L)\,(q_i)}\tilde{x}_{i,j}\right]\right\}^{\frac{1}{e_i}}\exp(a_i)\ [q_i]\right)^{e_i}}\right]$$

(*eqn 3*)

and:

$$q_i = 1 - \left(\frac{(\mathrm{M}_0-\mathrm{L})}{(\mathrm{K}-\mathrm{L})}\right)^{\frac{1}{e_i}} \qquad \textbf{\textit{(eqn 4)}}$$

where:

Yij, the j^{th} species richness observation, in i^{th} site

F, parametric function of the structural model

$\tilde{x}_{i,j}, \widetilde{lNp}_{i,j}$, design variables for the j^{th} observation in the i^{th} site

θ_i, model parameters for the i^{th} site (h_i, a_i, r_i, e_i)

φ, vector of other hyper-parameters (ρ, L, K, M_0)

α, parameter of the power in the variance-mean relationship involved in the Negative Binomial distribution

λ, scaling parameter in the variance-mean relationship involved in the Negative Binomial distribution

and where the statistical parameters of the mean function (Equation 3) have the following interpretation (see next section for details on random effects):

h_i, the additional parameter for multiplicative (homothetic) site random effects

a_i, the additional parameter for site random effects that shift the x-scale

r_i, the slope at the inflection point (potentially with a site random effect)

e_i, the asymmetry parameter of the logis5p function (potentially with a site random effect)

p, the correction parameter for sampling effort

L, the lower asymptote

K, the upper asymptote

M_0, the ordinate at the median of $\tilde{x}$.

The logis4p mean function is simplified from the logis5p mean function by setting $e_i = 1$, while the logis3p function is simplified from the logis5p function by additionally setting $L = 0$. The other classical mean functions (linear, exponential, extreme value and Gompertz) were reparametrized so that the parameters would have the same meanings as in the Logis5p model above. All functions are described in Annexe II. Table S1.2.

We mostly selected weakly informative priors or priors corresponding to logical constraints on the possible values of each parameter (Annexe II. Table S1.3). The initial values were chosen in order to respect the a priori distribution while containing little information (Annexe II. Table S1.3).

II. 2. d. Random effects

We first tested the models with no random effect (called no.re) with each functions in order to highlight the effect of not taking into account the hierarchization factor. Then, we fitted the model using different ways to integrate random effects:

1) A model (called re.1a) with a random effect changing the position of the inflection point on the X-axis for sigmoidal functions (Figure 2.2.A) or changing the intercept (for the linear and exponential functions). This corresponds to the typical strategy in ecology according to Schielzeth & Forstmeier (2009).

2) A model (called re.1h) with a random effect introducing a multiplicative (homothetic) variation of the curve. This corresponds to a different, multiplicative global level of species richness among sites without changing the curve shape (Figure 2.2.B). In the case of the exponential function, re.1h is the same as re.1a.

3) A model (called re.1r) with a random effect changing the slope at the median. This corresponds to a variation among sites in the direction and the strength of the relationship (Figure 2.2.C).

4) For each function, a complete model (called full-random-effect model or full.re) with all the random effects on all the parameters of the function (except for the two asymptotes of logis4p and logis5p, which are influenced multiplicatively by a unique random effect as in model re.1h). This corresponds to the following random effects settings:
   - Regarding the exponential function: model with random effects on the slope (parameter r) and the multiplicative random effect (also called re.2hr). For the exponential function, there are no re.2ra, re.2ha or re.3hra models because the multiplicative random effect and the random effect changing the intercept are similar.
   - Regarding the linear function, there are two possibilities: model with random effects on parameter r and on the intercept (also called re.2ra); or the slope (parameter r) and the multiplicative random effect (also called re.2hr). For the linear function, there are no re.2ha or re.3hra models because the multiplicative random effect and the random effect changing the intercept are different but redundant when put together.

- Regarding the sigmoid form functions with three or four parameters (EVF, Gompertz, logis3p and logis4p): model with random effects on the slope (parameter r), on the intercept and the multiplicative random effect (also called re.3hra).
- Regarding the 5-parameter logistic function (logis5p): model with random effects on parameter r, parameter e, the intercept and the multiplicative random effect (also called re.4hrae).

5) All the intermediate possibilities. Each of these model possibilities was named according to a combination of the names of models with simple random effects above (e.g. re.2hr).

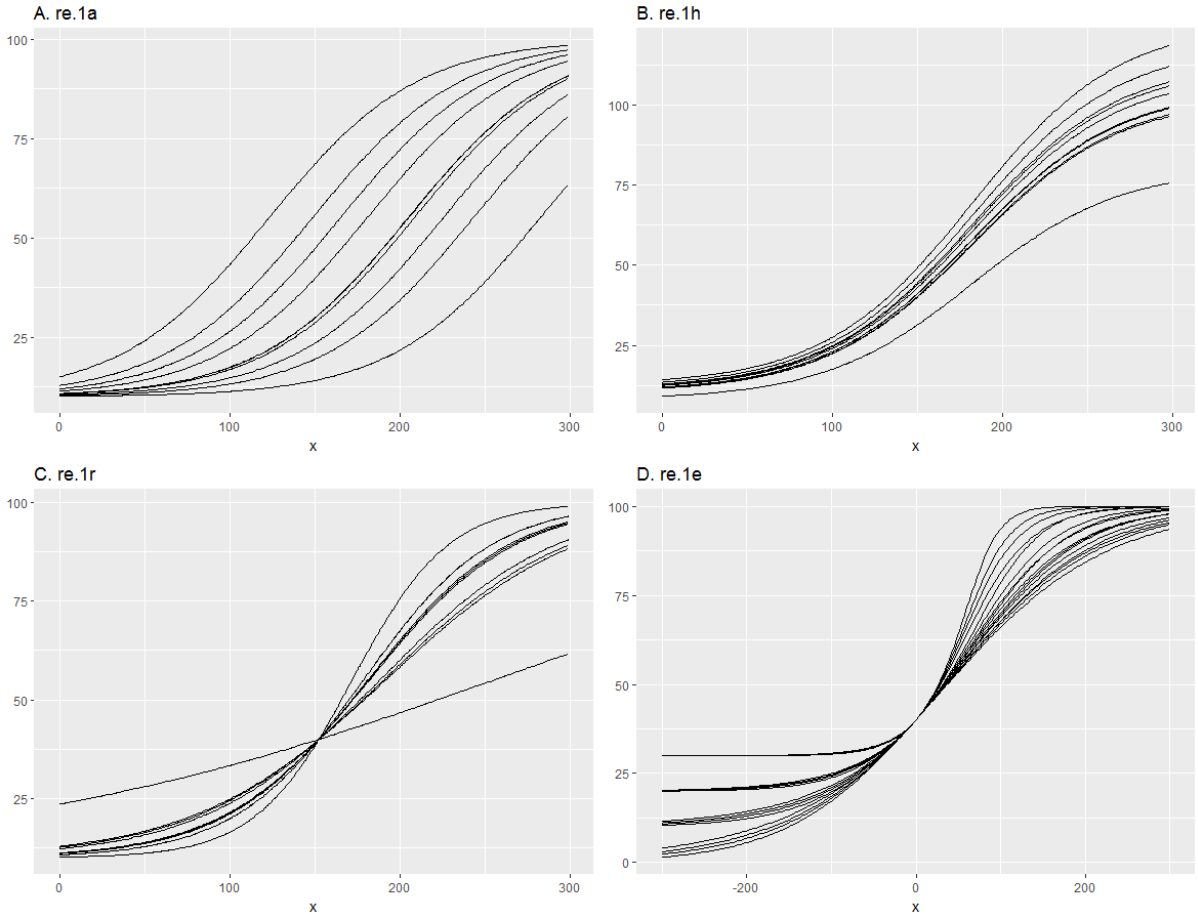


Figure 2.2: Examples of the impact of adding one random effect on the shape of the logis5p curve. The parameters of the function were fixed (r=80, K=100, L=10 M0=40 and e=0). Only the random effect varied.

For logis5p, random effect on parameter e was put only on full.re models. Simple random effect models (re.1e, Figure 2.2.D) and intermediate models containing a random effect on e were not fitted (e.g. re.2he).

We systematically estimated the correlation between the random effects on different parameters (as suggested by Schielzeth & Forstmeier, 2009). WAIC results for models with independent random effects and a comparison between models with and without correlations are presented in Annexe II. Table S2.1 and Annexe II. Table S2.2. We used random effects with a Gaussian distribution – which is the baseline distribution for random effects (Bolker et al., 2009; McCarthy, 2007).

The site random effects in the model stemmed from a multivariate normal distribution with unit variances:

$$\eta_i = \begin{bmatrix} \eta_i^a \\ \eta_i^e \\ \eta_i^h \\ \eta_i^r \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & corr_{12} & corr_{13} & corr_{14} \\ corr_{12} & 1 & corr_{23} & corr_{24} \\ corr_{13} & corr_{23} & 1 & corr_{34} \\ corr_{14} & corr_{24} & corr_{34} & 1 \end{bmatrix} \right) \qquad \textbf{(}\textbf{\textit{eqn 4}}\textbf{)}$$

that were then used in the following equations to incorporate Equation 3 through parameters h_i, a_i, r_i and e_i:

$$a_i = 0 + \sigma_1 * \eta_i^a \qquad \textbf{(}\textbf{\textit{eqn 5}}\textbf{)}$$

$$r_i = r + \sigma_2 * \eta_i^r \qquad \textbf{(}\textbf{\textit{eqn 6}}\textbf{)}$$

$$h_i = 0 + \sigma_3 * \eta_i^h \qquad \textbf{(}\textbf{\textit{eqn 7}}\textbf{)}$$

$$e_i = \exp(\varepsilon + \sigma_4 * \eta_i^e) \qquad \textbf{(}\textbf{\textit{eqn 8}}\textbf{)}$$

Depending on the chosen parametrization, parameters h_i, a_i, r_i and e_i were either fixed among sites (in which case σ and η were removed from equations 5 to 8) or varied between them.

II. 2. e. Model diagnosis and selection criteria

II.2.e.1. Relative model evaluation

The models were compared with Watanabe's (2013) Widely Applicable Information Criterion (WAIC). We applied the classical WAIC presented in Hooten & Hobbs (2015; equations 44 and 45), involving the variances (over the statistical parameters) of the log probabilities. This criterion quantifies the expected predictive accuracy of a Bayesian model. WAIC is analogous to AIC for frequentist models, and similar to the deviance information criterion (DIC) proposed earlier for Bayesian models. It allows for penalization based on the number of parameters in the model (or number of model dimensions); the penalization also accounts for information added by priors. We also looked at the WAIC proposed by Gelman et al., (2014) and results were very similar (see Annexe II. Table S2.1). We considered that two models with a difference of less than two WAIC points were equivalent in terms of predictive capacity (as Kass & Raftery, 1995 did for the Bayesian Information Criterion BIC). We used conditional versions of WAICs throughout (Millar, 2018), i.e. WAICs based on MCMC values of random effects as well as of hyper-parameters. We also calculated marginal WAICs for a subset of models to ensure that they yielded results that agreed with conditional WAICs (Millar 2018; results in Annexe II. Table S2.6, Annexe II. Table S2.7 and Annexe II. Section S2.2). Marginal WAICs are versions of WAICs in which, for each set of MCMC hyper-parameters, random effects are integrated out to yield marginal likelihood. In some settings (especially when there are very few observations per random effect), Millar (2018)

advocated using marginal WAICs rather than conditional WAICs to select models. In order to acknowledge the significance of difference in predictive capacity among the models with no random effects, and the significance of difference in predictive capacity among the models with all random effects, we used Vehtari et al.'s (2017) leave-one-out information criterion (LOOic) to evaluate the statistical significance of the difference in information criterion for selected pairs of models (at a level of 5%).

II.2.e.2. Intrinsic model evaluation

Secondly, criticizing a model is a key step in the modelling process (Box, 1980; Cox, 1997; Hilborn & Mangel, 1997). It is an important procedure, which allows modelers to better comprehend whether a given model fits the observed data in different respects, and which also offers a different perspective than does relative model evaluation, presented above. For this purpose, we used the sampled posterior goodness-of-fit (GOF) p-values (Gosselin, 2011) – based on a sampled value of statistical parameters – on different discrepancy functions applied to normalized quantile residuals (Dunn & Smyth, 1996) of the model and to the random effects in the model (as in Herpigny & Gosselin, 2015). In terms of discrepancy functions, we first diagnosed the distributions of residuals or random effects by using their mean, variance, skewness and kurtosis as discrepancy functions. Then, we used discrepancy functions based on distance correlations (Székely et al., 2007; function dcor in the energy R package - Rizzo & Szekely, 2018) to diagnose: (i) the mean function (with the overall distance correlation dcor between the normalized residuals and the fitted mean values $F\left(\tilde{x}_{i,j}, \widetilde{lNp}_{i,j}, \theta_i, \varphi\right)$ for the sampled parameter or the indicator x_{ij}), (ii) the heteroscedasticity of the model (with the overall distance correlation dcor between the square of the normalized residuals and the fitted mean values) and (iii) the mean function by site

(with the mean over sites of the distance correlation dcor by site between the normalized residuals and the indicator), considering either all sites or the sites with at least 20 observations. Finally, we used two discrepancy functions to diagnose spatial autocorrelation of normalized residuals by site by fitting a generalized least squares model with an exponentially decaying spatial autocorrelation for each site on the normalized residuals (functions corExp and gls in the nlme library with a nugget), then extracting the range and nugget for each site and lastly taking the mean over sites of each of these estimated parameters. For each of these discrepancy functions, we first calculated one sampled GOF p-value per model, based on a single sampled parameter iteration of the MCMC, by (i) calculating the number of times the discrepancy function on observed data was greater than the discrepancy function on replicated data, then (ii) randomizing the associated proportion $\bar{p}$ through a random draw $\tilde{p}$ from a beta distribution based on these counts, and finally (iii) transforming the random proportion $\tilde{p}$ through the function $\tilde{q} = 2\min(\tilde{p}, 1-\tilde{p})$ to concentrate surprising situations on only the region close to 0.0 (compared to around 0.0 and around 1.0 for $\tilde{p}$). Step (iii) implies that these p-values are built to detect departures between the model and the data only when the p-value is close to 0.0 – in contrast with some other GOF p-values that indicate departure when they are close to 0.0 or to 1.0. Then, we calculated 1,000 $\tilde{q}$ values with 1,000 random draws of the statistical parameters from the MCMC and reported the frequencies with which these values were lower than 0.005. We calculated the GOF p-values for the models with no random effects (no.re), with one multiplicative random effect (re.1h), with one random effect on the intercept (re.1a), with all the random effects (full.re) and the best models for each function (best), and compared the p-values between different versions of the models.

II. 2. f. Analysis of the relationships

For the models for which we calculated the GOF p-values, we analyzed the relationship between the indicator and the mean of the modelled richness by analyzing (i) the significance and (ii) the magnitude of the relationship. Magnitude analyses were based on simulations of the variation in mean richness associated to given increases in the indicator, calculated as the ratio of expected mean at the increased value of the indicator over the expected mean at the baseline value of the indicator. This was done on all the iterations of the MCMC. Because our relationship is nonlinear, these sets of ratios were calculated for one level of increase (variation along the gradient x, $\Delta X=10m^3/ha$ corresponding to a realistic level of increase from a management point of view), starting from three different locations along the indicator gradient (Xinit=0, 10 or $70m^3/ha$ corresponding to rounded values stemming from empirical quantiles, cf. Annexe II. Table S1.4 for quantiles correspondence, and Data II.S2 for results), and either within the different sites (i.e. forest districts) or over the entire group of sites (all-sites). For the all-sites analyses, we used two methods: either one that plainly incorporated variability among sites by sampling one site for each iteration of the MCMC, or one that targeted the all-sites mean by considering the geometric mean of the ratios between sites (i.e. forest districts) for each iteration. We also ran the same calculations for the observed gradient values instead of the fixed positions over the gradient (Xobs), by taking the geometric means of the ratios over the observations. We used the same method proposed in Barbier et al. (2009) which defines different intervals of relationships that are considered negligible, positive (non-negligible) or negative (non-negligible). The results of the analyses were conclusive if 95% of the log ratio effects were within the defined intervals. We defined three negligibility intervals ([-0.2;0.2], [-0.1;0.1] and [-0.05;0.05] – the first two the same

as in Barbier et al. 2009 and the last one that was added to detect still smaller relationships). We also defined three intervals to bring out positive relationships ([0.05, Inf], [0.1, Inf] and [0.2, Inf]), and three intervals to reveal negative relationships ([-Inf; -0.05], [-Inf; -0.1], [-Inf; -0.2]). Results were considered inconclusive with respect to magnitude if none of the above cases was met. For significance analyses, we compared the logarithm of the ratios involved in magnitude analyses with 0.0 using Bayesian quantiles based on beta random draws as in (Gosselin, 2011).

Finally, we reported and interpreted the estimates of the hyperparameters in the best model. For each forest district with at least 20 observations, we visually represented fitted values of Y and modified observed values $Y_{i,j}$ as a function of gradient x for the four following cases: i) linear models with no.re, ii) linear models with re.1a, iii) linear model with the best set of random effects, and iv) the best model overall. To do this, we simulated response Y for each observation of x 100 times, based on 100 different sets of model output parameters and with a fixed sampling period number of four. On the graphs, we showed the mean of the 100 values and ten responses generated by ten of the simulations.

II. 2. g. Platform, packages and MCMC settings

We worked in a Bayesian framework because of the adaptability of this approach when working with complex models. Hierarchical statistical models were built with the nimble R-package (de Valpine et al., 2017). We used the MCMC algorithm, more precisely the Metropolis-Hastings random walk algorithm, for all our parameters except for the two dispersion parameters of the Negative Binomial, which were highly correlated and were sampled conjointly with the Nimble Block sampler. We systematically controlled the initial value of the variance of the proposal

distributions (parameters scale and propCov), then used a modified RW sampler (described in Miasojedow et al., 2013 for multivariate sampling, but transposed to univariate sampling). In addition, to start with a model nearly without any random effects that would force the mean structure to fit that of the data, we took very low initial values for the level of over-dispersion and for the standard errors of random effects. The standard deviation of random effects was not updated (i.e. sampled in the MCMC) for the first 2,000 iterations to force the algorithm to first fit the model without random effects before introducing random effects into the model. The results of the MCMC were analyzed with the coda R-package (Plummer et al., 2006).

The default burn-in period was 50,000 iterations. At the beginning of model fitting, the default thinning rate was 120 and we fitted 15 chains. Based on the experience gained during fitting in terms of model convergence and independence of MCMC outputs, we replaced these two figures by 1,000 and 10 respectively. For all the models, the final number of parameter outputs per chain was 1,000. We estimated the convergence of the MCMC chains by applying the diagnosis published by Gelman & Rubin (1992) but with a lower threshold as suggested by Vats & Knudson (2018): we increased the burn-in period if the diagnosis metric was above 1.015 for any of our hyper-parameters. We also calculated the effective sample size to estimate the autocorrelation of the MCMC sample based on the ratio of the Times-series standard error over the Naive standard error: we increased the thinning interval if this was equal or above 2.0 for any of the hyper-parameters.

II. 3. Results

II. 3. a. Random effects and predictive power of the model

For each mean function, the WAIC score of the models with no random effect (no.re) was much higher than the WAIC of the models with one random effect or more (Table 2.1). The models with no random effect (no.re) were equivalent for all functions in terms of WAIC (all within a range of 1.8 WAIC points) except for logis5p, which did not perform as well as the others, with a difference of 4.6 from the best model (constant mean function; Table 2.1). Most functions were not significantly different from each other in terms of LOOic at the level of 5%, except for the constant model, which was significantly better than all the other mean functions, and logis5p, which was significantly worse than all the other functions (Table 2.2).

Table 2.1: WAIC score for all tested models.

WAIC	Const	Linear	Exp	Gompertz	EVF	Logis3p	Logis4p	Logis5p
no.re	4798.1	4799.8	4799.9	4799.6	4799.6	4799.7	4799.3	4802.7
re.1r		4776.4	4772.5	4755.5	4730.6	4747.3	4652.4	4657.3
re.1h	**4333.6**	4333.3	4333.4	**4328.6**	**4324.0**	**4326.8**	4324.3	4324.0
re.1a		4330.2	4333.4	4330.7	4331.4	4330.3	4330.7	4331.0
re.2hr		**4327.5**	**4327.5**	**4326.6**	**4324.6**	**4326.3**	**4318.3**	**4318.8**
re.2ra		**4328.5**		4328.7	4328.5	4328.6	4328.5	4328.7
re.2ha				4329.5	4327.0	4328.9	4325.8	4324.8
re.3hra				**4328.2**	4326.8	**4327.8**	4320.8	**4320.8**
re.4hrae								4321.1

In bold, for each mean function (i.e. column), the best models and models within 2.0 points of WAIC. Underlined, the best overall model. Re.1 is used for a model with only one random effect and re.2 or more correspond to models with several associated random effects, "h" corresponds to the multiplicative random effect, "r" to random effect on the slope, "a" on the intercept and "e" on the parameter controlling the asymmetry.

Table 2.2: Differences in terms of LOOic (and standard error) between the model in column and the model in row, for models with no random effect (no.re) corresponding to the different mean functions.

	Constant	Linear	Exp	Gompertz	EVF	Logis3p	Logis4p
Linear	-0.8 (0.1)*						
Exp	-0.9 (0.1)*	0.0 (0.1)					
Gompertz	-0.7 (0.2)*	0.1 (0.1)	0.1 (0.1)				
EVF	-0.7 (0.2)*	0.1 (0.1)	0.1 (0.1)	0.0 (0.1)			
Logis3p	-0.8 (0.2)*	0.1 (0.1)	0.1 (0.1)	0.0 (0.0)	-0.1 (0.1)		
Logis4p	-0.6 (0.2)*	0.2 (0.1)*	0.3 (0.2)	0.1 (0.1)	0.1 (0.1)	0.2 (0.1)*	
Logis5p	-2.0 (0.5)*	-1.2 (0.5)*	-1.1 (0.5)*	-1.2 (0.4)*	-1.3 (0.4)*	-1.2 (0.4)*	-1.4 (0.4)*

** indicates models significantly different at the 5% significance level. A positive value indicates that model in the row is better than the model in the column, and conversely a negative value indicates that the model in the column is better than the model is the row.*

For each mean function separately, the models with no random effect were different from all other models. Indeed, the difference in terms of WAIC score between the model with no random effect and all the other models was between 23.4 (for the linear function with the re.1r model) and 481.0 (for the logis4p function with re.2hr), and the difference in terms of LOOic between the model with no random effect and all the other models was significant at a level of 5% (Table 2.3).

Table 2.3: Differences in terms of LOOic (and standard error) between the model in column (corresponding to the specified function in the column name and with no.re) and the model in row (corresponding to the function specified in the column name and the random effect specified in the row name).

Random effects	Constant	Linear	Exp	Gompertz	EVF	Logis3p	Logis4p	Logis5p
re.1r		11.1 (4.3)*	13.2 (4.6)*	20.2 (6.0)*	32.4 (7.3)*	24.2 (6.8)*	56.6 (12.1)*	62.2 (11.5)*
re.1h	232.1 (18.0)*	233.0 (18.2)*	233.0 (18.1)*	235.3 (18.4)*	237.6 (18.5)*	236.2 (18.5)*	237.3 (18.5)*	238.8 (18.5)*
re.1a		234.6 (18.2)*	233.0 (18.1)*	234.3 (18.1)*	233.8 (18.2)*	234.4 (18.2)*	234.1 (18.1)*	235.3 (18.2)*
re.2hr		235.7 (18.1)*	235.7 (18.0)*	236.0 (18.1)*	237.1 (18.4)*	236.1 (18.2)*	240.1 (18.0)*	241.3 (18.0)*

re.2ra	235.2 (18.1)*	235.0 (18.0)*	235.0 (18.1)*	235.0 (18.0)*	234.9 (18.1)*	236.2 (18.1)*
re.2ha		234.8 (18.2)*	236.0 (18.3)*	235.1 (18.2)*	236.4 (18.3)*	238.3 (18.3)*
re.3hra		235.2 (18.1)*	235.8 (18.2)*	235.4 (18.1)	238.8 (18.0)*	240.1 (18.0)*
re.4hrae						240.0 (18.0)*

** indicates models significantly different at the 5% significance level. A positive value indicates that the model with random effects (in the row) is better than the model with no random effect (in the column), and conversely a negative value indicates that the model no.re is better than the model is the model with random effects. For random effect abbreviations, see Methods or the legend for Table 2.3.*

In terms of the GOF p-value analysis, the models with no random effect (no.re) had several shortcomings related to heteroscedasticity, the quality of the mean function and spatial autocorrelation (in Table 2.4), respectively heteroscedasticity, link.bysite20 and range.bysite). The mean function in particular was very unsatisfactory since for all functions more than 70% of the model parameter iterations were problematic (more than 90% for six functions over the eight). For all the no.re models and for all Xinit values, the relationship between deadwood and species richness did not appear to be significant for either global metrics or forest-by-forest (Figure 2.3, Data II.S2), except for one forest for logis5p Xinit3 = 70m^3/ha. The magnitude of the effect was very close to 1.00 meaning that adding 10m^3/ha of deadwood would not increase much saproxylic beetle species richness.

Table 2.4: Proportion of time that 1,000 sampled posterior p-values $\tilde{q}$ were below a threshold of 0.005, indicating a significant mismatch between the test statistics on observed data and on replicate data.

Mean function	random effects	kurtosis	heteroscedasticity	link.bysite20	range.bysite	nugget.bysite
constant	no.re	0.000	0.652	0.999	0.345	0.080
linear	no.re	0.001	0.606	0.995	0.313	0.056
exp	no.re	0.001	0.396	1.000	0.294	0.030

gomp	no.re	0.009	0.377	0.936	0.275	0.067
EVF	no.re	0.009	0.494	0.915	0.222	0.041
logis3p	no.re	0.001	0.454	0.937	0.297	0.063
logis4p	no.re	0.009	0.562	0.874	0.322	0.061
logis5p	no.re	0.001	0.599	0.741	0.249	0.032
linear	re.1a	0.306	0.000	0.795	0.398	0.127
exp	re.1a	0.276	0.001	0.995	0.566	0.125
gomp	re.1a	0.247	0.002	0.705	0.345	0.054
EVF	re.1a	0.173	0.002	0.420	0.358	0.088
logis3p	re.1a	0.192	0.002	0.614	0.353	0.057
logis4p	re.1a	0.149	0.002	0.354	0.366	0.126
logis5p	re.1a	0.247	0.002	0.548	0.394	0.107
constant	re.1h (full.re)	0.124	0.000	1.000	0.559	0.093
linear	re.1h	0.258	0.001	0.966	0.529	0.113
exp	re.1h	0.242	0.002	0.992	0.460	0.129
gomp	re.1h	0.288	0.001	0.409	0.289	0.124
EVF	re.1h (best)	0.304	0.004	0.110	0.249	0.080
logis3p	re.1h	0.411	0.001	0.225	0.254	0.091
logis4p	re.1h	0.362	0.000	0.074	0.215	0.069
logis5p	re.1h	0.292	0.002	0.049	0.227	0.053
linear	re.2hr (best/full.re)	0.131	0.003	0.242	0.303	0.079
exp	re.2hr (best/full.re)	0.062	0.003	0.133	0.213	0.065
gomp	re.2hr (best)	0.273	0.000	0.131	0.228	0.093
logis3p	re.2hr (best)	0.229	0.000	0.099	0.241	0.037
logis4p	re.2hr (best)	0.080	0.003	0.024	0.143	0.026
logis5p	re.2hr (best)	0.068	0.003	0.009	0.161	0.025
linear	re.2ra (full.re)	0.174	0.003	0.322	0.336	0.110
logis4p	re.2ra	0.102	0.001	0.206	0.315	0.084
logis4p	re.2ha	0.256	0.001	0.075	0.292	0.096
gomp	re.3hra (full.re)	0.217	0.001	0.178	0.295	0.082
EVF	re.3hra (full.re)	0.125	0.003	0.094	0.242	0.059
logis3p	re.3hra (full.re)	0.213	0.000	0.196	0.239	0.068
logis4p	re.3hra (full.re)	0.072	0.000	0.043	0.206	0.044
logis5p	re.4hrae (full.re)	0.083	0.000	0.011	0.162	0.050

Kurtosis = kurtosis of normalized random quantile residuals; heteroscedasticity = link between the square of normalized random quantile residuals and the mean function; link.bysite20 = mean dependence between the normalized quantile residuals and deadwood volume for each site with more than 20 observations; range.bysite (and respectively nugget.bysite) = mean of the estimate of the range (respectively, the nugget) parameter in spatially autocorrelated GLS models of the normalized quantile residuals for each site for which we had spatial coordinates. For random effect abbreviations, see Methods or the legend for Table 2.3. Models containing the full set of random effects for the mean function concerned are denoted full.re. The best model in terms of WAIC (Table 2.3) for the mean function concerned are denoted best . Values above 0.1 (10%) that we considered problematic are underlined.

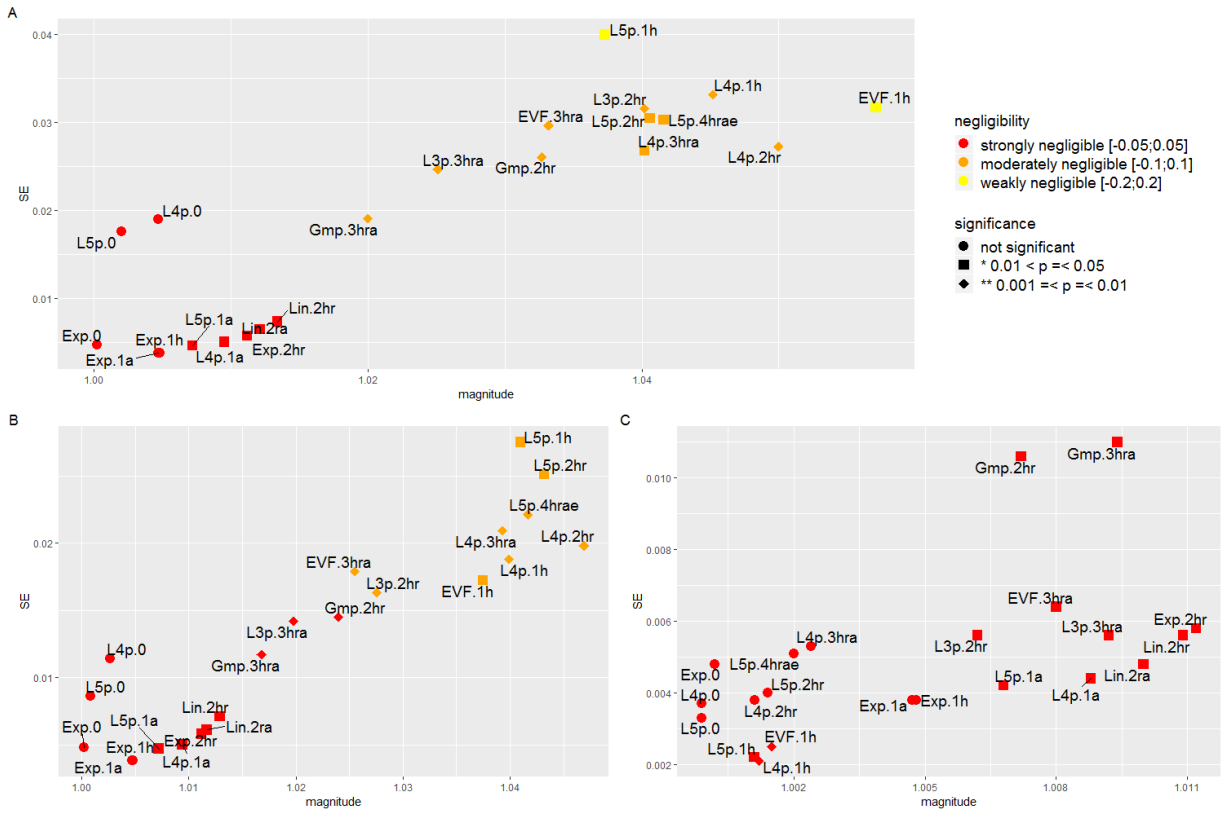


Figure 2.3: Standard deviation of the magnitude of the effect as a function of magnitude when adding 10m3/ha of deadwood to initial deadwood levels of 0m3/ha (A), 10m3/ha (B) and 70m3/ha (C). Negligibility and significance for the effect of this addition are shown. Note that due to the asymmetry of the credibility intervals, there is a degree of mismatch between levels of significance and standard errors. For legibility, model names have been shortened by suppressing "re." and transforming "no.re" into "0". Cst = constant function; Exp = exponential function; EVF = extreme-value function, Gmp = Gompertz function; L3p = Logis3p; L4p = Logis4p and L5p = Logis5p. In order not to overload the figure, we only present results for the exponential, logis4p and logis5p models with no.re, re.1a and re.1h, and all best and full.re models for each mean function. The complete results are available in Data II.S2.

The representations of the fitted and observed values as a function of gradient x for three linear models (no.re, re.1a and re.2hr, Figure 2.4) revealed an improvement in the fitting quality when adding at least one random effect.

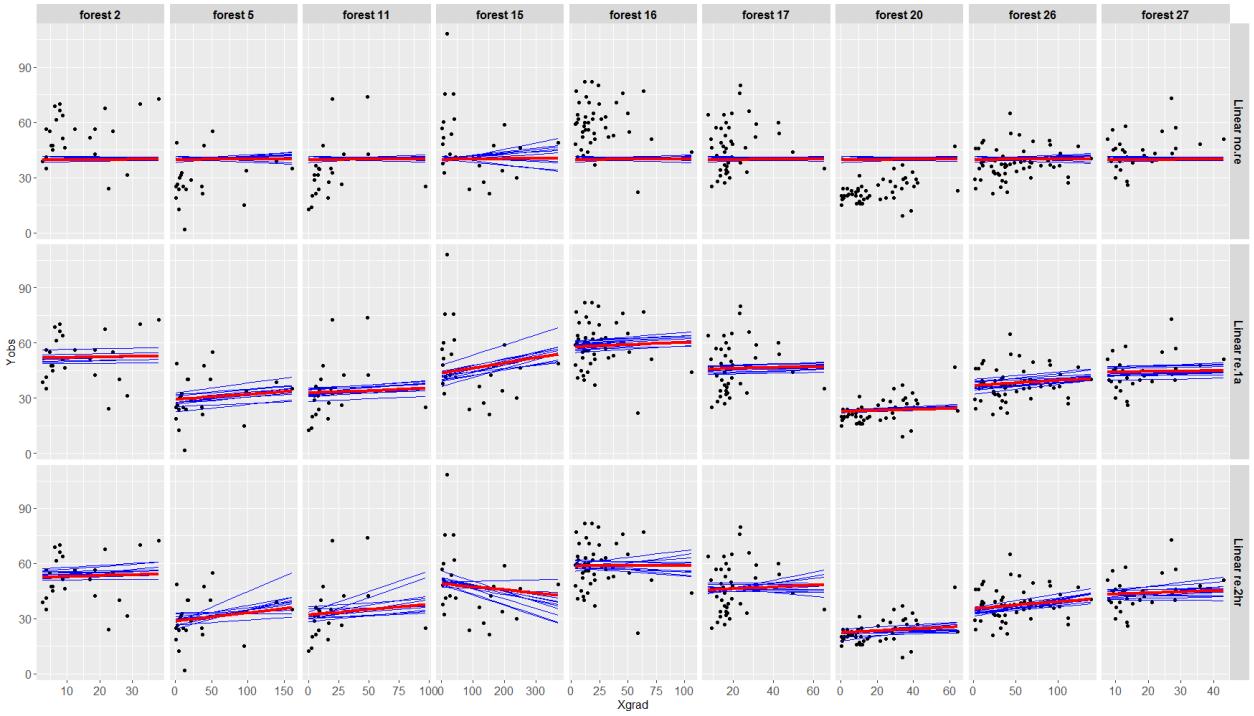


Figure 2.4: Fitted species richness response for ten sets of outputs (blue lines), mean response of 100 sets of outputs (red line, not necessarily of sigmoidal or even monotonic shape) and observed (black dots) response as a function of gradient x (deadwood volume in m3/ha) for forests with more than 20 observations and for three linear-function models: the models with no random effect (no.re), with one random effect on the intercept (re.1a), and with the best set of random effects (one random effect on the slope and the multiplicative random effect (re.2hr that is also one of two linear the full.re models). X-axis scale varies among forests. The theoretical response curves were derived from simulations for which the number of active periods was set at four.

II. 3. b. Comparison between common models with random effect on the intercept only and other single random-effect models

In the case of the constant mean function, there is only one possible model with a random effect so no comparison was made. For all the other mean functions with one random effect, the model with the highest WAIC score (i.e. the worst model) was the model with one random effect on the slope (re.1r) (Table 2.1). Regarding the two other possibilities for the linear mean function, the model with one random effect on the intercept (re.1a) had a lower WAIC (i.e. was better) than

the model with one multiplicative random effect (re.1h) (Table 2.1). For the exponential function, the two models with one random effect (re.1a and re.1h) had the same WAIC (i.e. were equivalent). Finally, for sigmoidal functions (Gompertz, EVF, Logis3p, logis4p and logis5p), the lowest WAIC was found for the model with a multiplicative random effect (re.1h) (Table 2.1). Nevertheless, the differences between the model with one random effect on the intercept (re.1a) and the model with the multiplicative random effect (re.1h) were not statistically significant for any functions (Annexe II. Table S2.3).

We did not interpret models with one random effect on the slope (re.1r) in terms of GOF p-values and magnitude because these models were not satisfying in terms of WAIC and because the only one we tested for magnitude (logis4p) presented aberrant results (Data S2). The other models with one random effect (re.1a and re.1h) had shortcomings concerning kurtosis, the quality of the mean function and spatial autocorrelation (respectively kurtosis, link.bysite20 and range.bysite in Table 2.4). The lack of fit for the mean function (link.bysite20) concerned a much lower proportion of parameter values than the models without random effects, except for the exponential, constant and linear models with multiplicative random effects (re.1h) and the exponential model with random effect on the intercept (re.1a). Half of the models also had mild spatial autocorrelation nugget issues (nugget.bysite) but this did not seem to depend on the type of random effect. For the three Xinit values studied, in most models with one random effect (r.1h or re.1a), the relationship between species richness and deadwood volume was globally significant (Figure 2.3) and for almost every site (except for three cases: re.1a and re.1h models with exponential mean function, and linear re.1h model, Data II.S2). However, the effect had a low magnitude with a maximum of 1.0571 (for the re.1h model with EVF mean function and at

Xinit=0m^3/ha), translating to an increase of 5.7% in the previous species richness for an additional 10m^3/ha of deadwood.

The representations of the fitted and observed values as a function of gradient x (Figure 2.4) for linear models with one random effect on the intercept (re.1a) and a more complete set of random effect (re.2hr) revealed a lack of fit for model re.1a, which was partly resolved with model re.2hr. This was especially clear for forest 15.

II. 3. c. Overall comparison of models differing by their set of random effects

As shown earlier, the models with no random effect (no.re) were all equivalent across all functions in terms of WAIC. When one or more random effects were added, differences among mean functions appeared since the difference in terms of WAIC became greater than 2.0 (for sigmoidal mean functions and particularly logis4p and logis5p). Nevertheless, when looking at the significance of the differences in LOOic values, unlike the models with no random effect (Table 2.2), the models with all the random effects combined (full.re) were not significantly different from one another (Table 2.5).

Table 2.5: Difference in terms of LOOic (and standard error) between the model in column and the model in row, for models with all random effects (full.re) corresponding to the different mean functions.

	Constant	**Linear (re.2ra)**	**Linear (re.2hr)**	**Exp**	**Gompertz**	**EVF**	**Logis3p**	**Logis4p**
Linear (re.2ra)	2.3 (2.3)							
Linear (re.2hr)	2.8 (2.6)	0.4 (1.3)						
Exp	2.8 (2.6)	0.5 (1.4)	0.0 (0.4)					
Gompertz	2.4 (2.6)	0.0 (0.7)	-0.4 (1.0)	-0.4 (1.2)				
EVF	3.0 (2.9)	0.7 (1.3)	0.2 (1.3)	0.2 (1.5)	0.6 (0.8)			
Logis3p	2.5 (2.7)	0.2 (0.9)	-0.3 (1.0)	-0.3 (1.2)	0.1 (0.3)	-0.5 (0.6)		
Logis4p	6.1 (3.6)	3.8 (2.6)	3.3 (2.1)	3.3 (2.2)	3.8 (2.2)	3.1 (1.8)	3.6 (2.1)	
Logis5p	5.9 (3.8)	3.6 (2.8)	3.2 (2.2)	3.2 (2.3)	3.6 (2.4)	2.9 (1.9)	3.4 (2.2)	0.2 (0.2)

** indicates models significantly different at the 5% significance level. A positive value indicates that model in the row is better than the model in the column, and conversely a negative value indicates that the model in the column is better than the model is the row.*

Generally, the best model (with the lowest WAIC value) for each mean functions had more than one random effect, except for EVF (re.1h; Table 2.1). For linear and exponential mean functions, the models with all random effects (full.re) were the best (Table 2.1). For Gompertz and Logis3p mean functions, the models with all random effects (full.re) were as good as the best model in terms of WAIC (difference in WAIC for the best model and the full.re ≤ 2.0; Table 2.1), and the difference was not significant in terms of LOOic (Table 2.6). For EVF and logis5p mean functions, the difference in WAIC score between the full.re and the best model was more than 2.0 points, but still very close (respectively 2.8 and 2.2; Table 2.1), and remained non-significant (Table 2.6). Finally, for the logis4p mean function, the difference in WAIC score between the full.re and the best model was also more than 2.0 points and still very close (2.5, Table 2.1) but the difference was significant (Table 2.6).

Table 2.6: Differences in terms of LOOic (and standard error) between the model with all random effects (full.re) and the best model in terms of WAIC (specified in the first row) for each different mean functions.

Loo.diff (se)	Linear	Exp	Gompertz	EVF	Logis3p	Logis4p	Logis5p
full.re/best	re.2hr/re.2ra 0.4 (1.3)	full.re/re.2hr 0.0 (0.0)	full.re/re.2hr 0.8 (1.0)	full.re/re.1h 1.8 (1.9)	full.re/re.2hr 0.7 (1.0)	full.re/re.2hr 1.3 (0.6)*	full.re/re.2hr 0.6 (0.8)

** indicates models significantly different at the 5% significance level. A positive value indicates that the "best model" is better than the full.re model. For random effect abbreviations, see Methods or the legend for Table 3. Models containing the full set of random effects for the mean function concerned are denoted full.re.*

All the models investigated in terms of GOF p-value had autocorrelation issues (range.bysite in Table 2.4). Though the models with one random effect (here re.1a and re.1h) did not have the heteroscedasticity issue that was found in the model with no random effect, they did have an additional problem with kurtosis. With more than one random effect (full.re or best model if different from full.re), kurtosis and/or quality of the mean function (link.bysite20) issues tended to disappear (Table 2.4). In general, full.re models seemed to have fewer issues than other models. Furthermore, when different from best models, models with full random effects (full.re) had the same shortcomings (in a same range of value of GOF p-values) as the best models (Gompertz, logis3p, logis4p and logis5p mean functions) or even fewer issues when the best model was with one random effect (EVF, with better quality of the mean function and fewer kurtosis problems; Table 2.4).

All the models without random effects (no.re) revealed a statistically non-significant relationship between deadwood volume and species richness. Conversely, all models with one or more random effects revealed a statistically significant relationship between deadwood volume and

species richness at least at 5% level (Figure 2.3 and global metric in Data II.S2; except for the three cases mentioned earlier: re.1a and re.1h exponential models and re.1h linear model).

No model showed non-negligible effects (whether positive or negative). Like the models with no random effects (no.re), all linear and exponential models and sigmoidal models with one random effect on the intercept (logis4p and logis5p re.1a) had a very negligible effect (belonged to the interval [-0.05; 0.05]) at the beginning of the deadwood gradient (Figure 2.3.A). Still at the beginning of the deadwood gradient (Figure 2.3.A), significant sigmoidal models with one multiplicative random effect (re1.h) were weakly (within the interval [-0.2; 0.2]) or moderately (within the interval [-0.1; 0.1]) negligible. Finally, sigmoidal models with two or more random effects alternated between weakly and strongly negligible effects. At the middle of the gradient (Figure 2.3.B), no models had weakly negligible effects, only moderately to strongly negligible. At the end of the gradient (Figure 2.3.C), all the models (including all functions and all sets of random effects) had strongly negligible effects.

II. 3. d. Best mean function and random effect setting

We compared the models not only by functions, but also by set of possible random effects. We noticed that the models with the logis4p mean functions appeared in the best models within a range of 2.0 WAIC points for all combinations of random effects (Table 2.1, read by lines). Logis5p was not as represented because the difference in WAIC with the best model was more than 2.0, but it was still very close (3.2 for re.2hr and 2.9 for re.3hra). In addition, when looking at full.re models, not all mean functions were equivalent; we were able to divide them into three categories: (i) models with shortcomings in autocorrelation (range.bysite), kurtosis and quality of

the mean function (constant model, both linear models, Gompertz and logis3p); (ii) models with shortcomings in autocorrelation and either in kurtosis or quality of the mean function (exponential and EVF); (iii) models with shortcomings only in autocorrelation (logis4p and logis5p). When we compared the models by mean function, we noticed that the ones with a multiplicative random effect and random effects on the slope (re.2hr) appeared in the best models within a range of 2.0 WAIC points for all the mean functions (Table 2.1).

Finally, the best model overall based on WAIC was the re.2hr model with the logis4p mean function (underlined model in Table 2.1), followed very closely by logis5p with the same random effects. Differences in terms of LOOic between the best model overall (logis4p re.2hr) and other logis4p models (with different random effect settings) or other re.2hr models (with different functions), were not all significant at the 5% level, although values were not far from significance for most functions (except for EVF and logis5p, Annexe II. Table S2.4). Logis4p and logis5p (full.re and best model re.2hr) were the least problematic in terms of GOF p-values with only autocorrelation issues.

Excluding the end of the gradient (Xinit=70m^3/ha), the re.2hr logis4p model showed a significant relationship between species richness and deadwood volume at a level of 1% (Figure 2.3). This model also had one of the highest values for magnitude of the effect of adding 10m^3/ha to x (1.05, Figure 2.3 and Data II.S2). Full random effects Logis4p and logis5p models (full.re corresponding to re.3hra and re.4hrae respectively), and the best model logis5p (re.2hr) also showed significant relationships at levels between 1% and 5% and revealed a magnitude of the effect very similar to the re.2hr logis4p model (around 1.0410, Figure 2.3 and Data II.S2). In addition to these magnitude analyses, which were the result of the different parameters of the model, the

estimated parameters for the best model are interpreted in Annexe II. Table S2.5 and Annexe II. Section S2.1.

Finally, from a graphical point of view (Figure 2.5), comparing the re2.hr logis4p model (the best model overall) and its linear counterpart (linear model with re.2hr random effects, the best linear model), the logis4p model simulations seemed to better fit real observations than those generated by the linear function model, at least for several forests (e.g. forest 5, 15 and 20).

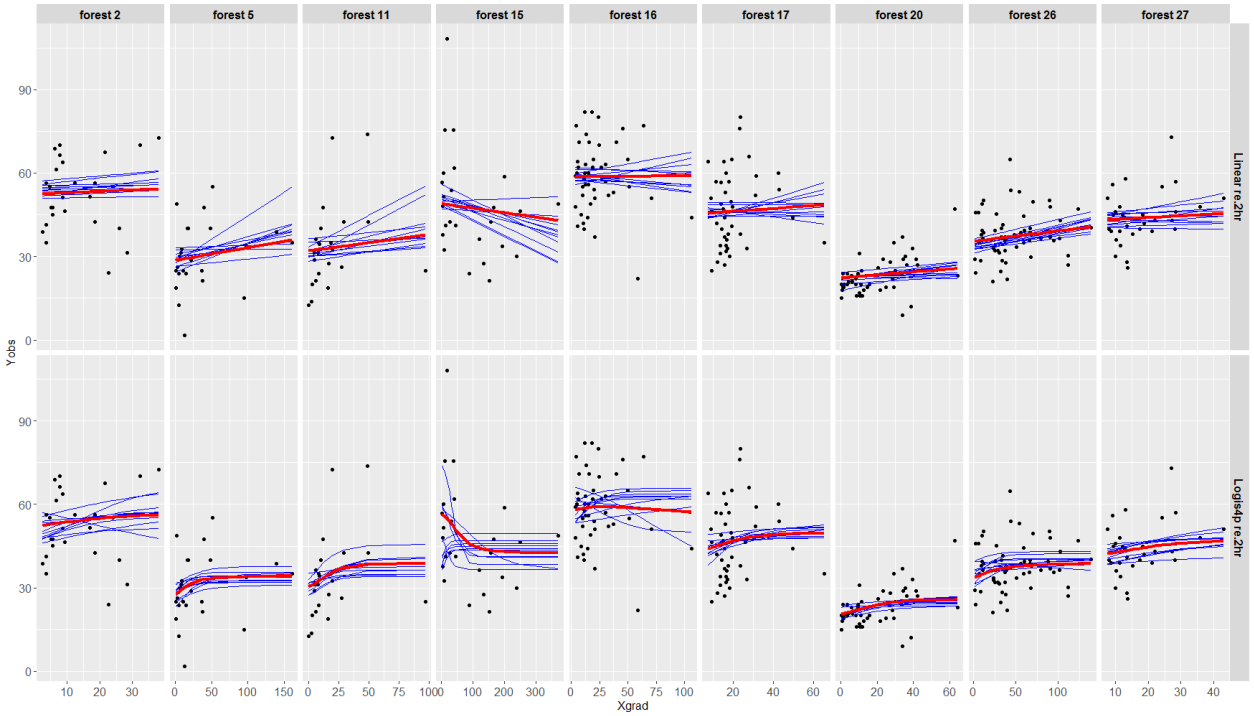


Figure 2.5: Fitted species richness response for ten sets of outputs (blue lines), the mean response of 100 sets of outputs (red line, not necessarily of sigmoidal or even monotonic shape) and observed response (black dots) response as a function of gradient x (deadwood volume in m3/ha) for forests with more than 20 observations and for two models with re.2hr random effects: the model with a linear mean function (best linear model) and the model with a logis4p mean function (best logis4p model and best model overall). X-axis scale varies among forests. The theoretical response curves were derived from simulations for which the number of active periods was set at four.

II. 4. Discussion

Models should be developed to meet the specific use case, with careful consideration of nonlinearity, and should incorporate random effects in an appropriate manner. By comparing various models, including different mean functions and different sets of random effects (from none to as many as possible), we were able to show that 1) the way random effects are incorporated into the model matters (cf. section II.4.a and II.4.b); 2) using and developing relevant nonlinear functions can make a difference (cf. section II.4.c); and 3) the selection and analysis of models should be performed with complementary methods (cf. section II.4.d).

II. 4. a. The introduction of random effects makes it possible to detect significant variations in the dataset

For every mean function tested, the model with no random effect was the worst model in terms of WAIC and the difference with all the other models was statistically significant in terms of LOOic.

All the models with no random effect seemed equivalent, including the constant models, when looking at the WAIC scores (Table 2.1). Moreover, among all the models with no random effects, the constant model was significantly better than all the other mean functions, according to the LOOic test of significance of the differences (Table 2.2). This means that if we do not take into account the effect that sites can have on the relationship between explanatory and response variables, no pattern will be detected and the effects of deadwood will be very small and insignificant (Figure 2.3). In contrast, adding at least one random effect allows us to discriminate functions with WAIC, and especially to discriminate between the constant model with random

effects and the other models involving relationships of different shapes. This echoes other results on biodiversity indicators that stressed the fact that the relationships between biodiversity and indicators, or surrogate variables, do indeed vary in space (Larrieu et al., 2018; Pierson et al., 2015). A model with no spatial structure at all (without random effects) would have given a non-significant relationship, whereas more elaborate models (with at least one random effect) provided stronger and more significant relationships (cf. infra). The apparition of significance thanks to the introduction of random effects in nonlinear models contrasts sharply with what is known by ecologists on the effect of random effects in linear or generalized linear models (references in Fortin, 2013; Saas & Gosselin, 2014): the inclusion of random effects in linear models does not usually change the mean estimators, but merely increases the standard errors of the fixed-effect parameters, which usually decreases the statistical significance of these effects. What we found in our study is exactly the reverse: the inclusion of random effects revealed more significant fixed-parameter effects. In addition to being manifest from a numerical point of view, these findings were also visually clear, since the models with at least one random effect (re.1a) fit the data better, despite varying conditions in the different forest districts (Figure 2.4).

The interpretation of the results of hierarchical non-linear models can follow two lines: a first “mechanistic” one in which elements of the non-linear model are related to the mechanisms implied by the hypotheses and are interpreted as such; and a “phenomenological” line in which non-linearity and random effects are introduced to give a better relationship with the data, in which case only parts of the model are interpreted, those that are of use in practice. It is this last line which we have adopted in this paper and the inclusion of random effects allowed us to reveal stronger and more significant relationships.

II. 4. b. Multiple random effects give better results than an automatic application of random effects to the intercept

This being said, the method generally used in ecology is to add random effects to the intercept of the curve (Schielzeth & Forstmeier, 2009). In the case of our dataset, this implies that the value of saproxylic species richness varies based on a systematic shift on the ecological gradient – here related to deadwood volume. The difference between two given sites could be explained by a limiting deadwood factor variable among sites. This is depicted by a translation of the curves along the X-axis. This idea is not absurd, but unfortunately, it reduces the difference between sites to this single possible effect. Other ways to incorporate random effects are conceivable, by modifying the shape of the curve instead of by simply shifting its relative position on the X-axis. The form of the relationship could also differ according to site with, for example, a different slope (as in Schielzeth & Forstmeier, 2009), which could imply a varying sensitivity to the explanatory variable. In the case of our studied functions, random effects can indeed be added to the parameter controlling the intercept and the slope as explained above, but other possibilities must also be taken into account; for instance, a random effect could transform the shape of the curve in a multiplicative way on the y-axis. Schielzeth & Forstmeier (2009) demonstrated that the use of random effects on the intercept, in addition to being biologically unjustified and too restrictive, when applied to an unsuitable dataset could cause estimation errors, especially on type I errors. Three cases are particularly at risk: (i) when the variation among sites in the slopes is strong; (ii) when within-sites scatter around the individual regression line is low; (iii) when the number of measurements taken from the same site is important.

In general, our study revealed that the automatic application of random effects only on the intercept gave similar or less robust results, in terms of both predictive capacity (WAIC) and analysis of the relationship, with magnitudes of effect that were much lower in the case of sigmoidal models.

This suggests that merely applying random effects to the intercept, one of the simplest random-effect structures and currently the most popular one in use in ecology, is actually a sub-optimal strategy. These results converge with those of Schielzeth & Forstmeier, (2009) who advise ecologists to stop being satisfied with the inclusion of a single random effect on the intercept to take into account spatial variability of biodiversity relationships (as done by Johnson, 2014 and Nussey et al., 2007). Choosing random effects requires working on a case-by-case basis. Ideally, it would be necessary to ground the choice of where to put random effects on a priori reasoning or ecological assumptions (Austin, 2002). Yet, it is not always easy to completely specify in ecological terms how the shape of the curve varies according to the factor of hierarchization. The best approach might be to compare models that incorporate all the different combinations of random effects. It is also possible to apply random effects to several or all of the parameters of the function in the same model. This is what Schielzeth & Forstmeier (2009) finally recommended for simple linear models; in their study, including random effects on both the intercept and the slope in the model, gave satisfactory results. In a context similar to ours, i.e. non-linear Bayesian modelling, Li et al. (2012) also demonstrated that a model with two kinds of random effects (their model 3) gave better results than the two corresponding models with each of the random effects included separately (their models 1 and 2). Our results are consistent with these two previous studies; we generally had better WAIC results for models with multiple random effects than for

models with only one random effect (except for the EVF mean function). In our case, models with multiple random effects also allowed us to eliminate some shortcomings revealed by GOF p-values concerning kurtosis and the poor quality of the mean function.

In our study, we also found that the full-random-effect models (full.re) either performed as well as the best model (sigmoidal mean functions except for EVF) or were themselves the best model (linear and exponential mean functions) in terms of WAIC. They also had either the same number of or fewer issues revealed by GOF p-values. This was also the case for magnitude results where the best model and the full-random-effect model showed very similar significance and magnitude for the relationship. It therefore appears likely that a robust solution would be to automatically apply all the possible random effects in the model, as is done in pharmacodynamics (e.g. Pillai et al., 2005). Another solution would be:

1) to first compare models for different mean functions with all random effects included (full.re) because the full-random-effect model makes it possible to discriminate between the mean functions and select the preferred mean function(s) based on WAIC and GOF p-values (keeping an eye on the significance of the difference based on LOOic);
2) secondly, to try all possible alternatives to incorporate random effects into the preferred mean function(s) (in order to simplify and improve the final model).

This method would have a good chance of revealing the best mean function-random effect combination without testing all the possibilities, which would be very long and tedious.

II. 4. c. New sigmoid mean functions (Logis4p and logis5p) are promising for fitting species-resource relationships

In addition to generally having better predictive power than other mean functions (WAIC), the logis4p and logis5p functions with some combinations of random effects (especially re2.hr and full.re models) allowed us to reduce the autocorrelation issue and improve the coherence between the statistical model and the data relative to the kurtosis and the mean function. This was not the case for models with only one random effect or no random effects. Other mean functions that came close to being correct in terms of GOF p-values were the extreme value function and the exponential function with random effects on all the parameters – with the exponential function yielding relationships of lower magnitude and statistical significance. We therefore propose a new parametrization for logis4p and logis5p (Equation 3 for logis5p and Annexe II. Table S1.2 for logis4p) that clearly controls the parameters of the function so that we understand where to put the random effects (Figure 2.2). This was achieved through an analysis of the derivatives of the function. The parametrization is also written to avoid numerical issues associated with power functions. We therefore propose that this version, or its enhancements, should be preferred to the one provided by Paine et al. (2012).

Markedly, classic linear or generalized linear approaches (respectively, the linear and exponential mean functions) gave relationships that were of lower significance and magnitude than their sigmoidal counterparts, especially when the multiplicative random effect was incorporated (alone or in association with others). These results provide additional alternatives to GLMMs (Bolker et al., 2009) and advocate for considering more often truly nonlinear approaches in this

endeavor. Testing or validating surrogate variables or indicators should therefore incorporate an explicit consideration of potential spatially variable nonlinearities. Fortunately, a comparison of WAIC scores and GOF p-values with suitable metrics targeting the mean function allowed us to better detect potential problems and guided us in this approach (cf. infra). Although less obvious than for the purely numerical results (WAIC, GOF p-values, significance and magnitude), the improvement in fit added by the best model is still perceptible on a visual representation (Figure 2.5).

Not only are the two functions (logis4p and logis5p) interesting in the case studied, they may also be useful for the study of presence / absence data. We have previously mentioned the case where a species can be present on a patch despite an absence of the studied resources thanks to their mobility (cf. Methods - Functions subsection). This case can also be observed during the study of source-sink systems, where the resource is not present or not in sufficient quantity in the patch to maintain a viable population, but where individuals of the species are nevertheless present in the patch due to immigration from another patch where the resource is sufficient (Gosselin, 1996; Pulliam, 1988). The opposite is also possible: the species may be absent from a patch where the resource is present due to stochasticity in demographic processes or the absence of another resource not taken into account and yet necessary. The logis4p and logis5p functions, contrary to the other sigmoidal functions studied, allow such data to be fitted.

II. 4. d. Multiple selection and interpretation criteria are complementary

Although the results in terms of WAIC score alone showed that all the models with no random effects were equivalent – since they were all within 2.0 WAIC units of each other – the observation

of the significance of the difference in terms of LOOic told a different story: here the models were found to be significantly different at the 5% level. Conversely, for models with all random effects, the opposite effect occurred: the models were different in terms of WAIC score – with a group of two models that dominated the other ones (logis4p re.2hr and logis5p re.2hr) – and were globally equivalent when we looked at the LOOic significance of the difference. The new approach based on the statistical significance test of LOOic between models therefore gave different results from the more classical WAIC approach (two models are similar if they are within 2.0 WAIC units of each other). This could be due to the introduction of random effects that would induce uncertainty that in turn would increase the standard error of the difference of the models, thus affecting the statistical significance of their comparison in terms of LOOic predictive power. The two criteria might therefore complement each other. Indeed, we can classify the models according to their WAIC or LOOic score, while using the significance of the LOOic difference between the models to know if they are statistically significantly different in terms of predictive power. To refine model selection, we can turn to GOF p-values, which provide new information. Where WAIC score and LOOic difference significance reflect the predictive ability of the models, the GOF p-values allow us to detect problems internal to the models, on metrics other than the deviance of the model. One important point in our approach with GOF p-values is to devise diagnostic functions (which are technically called discrepancy functions; Gosselin, 2011) that are well suited for the case at hand. Here, it appeared that a bare study of the dependence between normalized residuals and the explanatory variable (as used in Harrell, 2001; Herpigny & Gosselin, 2015) did not provide sufficient power to criticize the model relative to its mean function. Only the study of the mean dependence between normalized residuals and the explanatory variable x

by site, for sites with a minimum number of observations (here: 20), was able to provide us with a sufficiently accurate gauge to diagnose models relative to the mean function. This calls users to pay special attention to the conception of their discrepancy measures in relation to the components of the models they want to diagnose. Unlike the WAIC scores, the GOF p-values allow us to know if the models are good as such, and not in comparison with other models. Thus, by comparing the GOF p-values obtained, it is possible to classify the models according to the quantity and severity of their shortcomings and to better understand the improvements brought by different variations of the model (here: mean functions and random effects). GOF p-values seem at their best when relatively comparing different statistical models on the same dataset rather than in an absolute approach, where only one model is criticized for a given dataset. Indeed, discrepancies between statistical models and data are expected to emerge, especially when data are numerous, since a statistical model is only a simplification of reality. A joint analysis of the three criteria (WAIC, LOOic and GOF p-values) allows us to classify the models according to their predictive capacity, keeping in mind the significance of the differences, and the relevance of the model itself relative to the data. The three criteria are therefore complementary.

Once the best model(s) is (are) selected with these three criteria, its (their) results need to be studied. The form of the relationship between the two variables is already studied through model selection and GOF p-values, but we wanted to better apprehend the relationship between deadwood volume and saproxylic beetle species richness. In ecology, many studies focus on the significance of the relationship in terms of the p-value. However, this approach alone is known to be insufficient in applied ecological problems (Barbier et al., 2009; Yoccoz, 1991). We need a complementary approach to interpret results that is based on the analysis of the magnitude of

the effects. When the relationship is significant, here between deadwood volume and mean saproxylic beetle species richness, the magnitude of the relationship gives additional information. It allows us to discriminate between three situations: 1) cases where a given variation in the explanatory variable would only give negligible variations in the target variable; 2) cases where the variation would be non-negligible; and 3) cases where we cannot conclude on the magnitude of the relationship. In our case, models with no random effect showed an insignificant and very negligible relationship, and logically the constant model turned out to be the best in terms of WAIC (and significantly better than the others in terms of LOOic). Models with no random effect also turned out to have many internal problems, revealed through the GOF p-values, and therefore did not satisfy all the selection criteria. For the models with random effects on the intercept, the relationship was significant but also negligible and therefore failed to show an interesting relationship between deadwood volume and species richness. This was in agreement with the results obtained by Schielzeth & Forstmeier (2009). Finally, the models that seemed the strongest in terms of magnitude, but still remained below our threshold of non-negligibility, were the best models in terms of WAIC and partly of GOF p-values. There was a sort of convergence of all the criteria (GOF p-value, WAIC, significance and magnitude of the relationship) around these models, with logis4p re.2hr, logis4p re.3hra, and logis5p re.2hr standing out as the best models according to these criteria.

II. 4. e. Some limitations on the general use of our models

In spite of their great potential, our hierarchical Bayesian models might be difficult to use in more general conditions. First, our case study was particularly adapted to non-linear modeling, and

ecological datasets do no always fit this profile. Fitting complex models can become a real challenge when: (i) the dataset does not contain enough observations to properly estimate all the parameters of the complex models; (ii) the dataset contains so many observations that fitting this kind of model becomes extremely slow; (iii) the functional form is not well established; and (iv) the purpose is to describe the response to many predictor variables that can potentially be included in the different parameters of the non-linear function. Second, we did not encounter too many difficulties with the priors and formulations we used, probably because the data dominated the priors in our case. We recognize that this might not be the case in other conditions, particularly for the parametrization of the correlation matrix through uniform priors on individual correlations (see Annexe II. Table S1.3). Third, we recognize that LOOic might be preferable to WAIC in general (Vehtari et al., 2017). However, in our case study, both methods gave the same results.

Our objective herein was to provide a methodological line of approach; we did not therefore develop the interpretive aspect of the parameters, which the explicit nonlinear formulation allows compared to GAMs. However, this methodological strategy will be of interest for future research intended to study the values of the output parameters in more detail from an ecological point of view. Moreover, even though we grounded our study in a phenomenological, rather than mechanistic, context, explicit nonlinear modeling with more complex models, can enhance reflection on the form of the relationships under study, unlike GAMs.

Finally, for our dataset, all the models had autocorrelation issues and did not take into account measurement errors. It would have been interesting to build a new model to take into account both measurement errors and spatial autocorrelation among plots (as done in Saas & Gosselin,

2014). Unfortunately, this would have required greater calculation resources and computing support than were available to us. For this particular study, we chose to focus our reflection on random effects.

II. 5. Conclusion

In this paper, the main assumptions we made were that: (i) the relationship would be monotonic (we expect mostly increasing), that there would be a zone where the relationship is stronger than at the end of the gradient; and (ii) the relationship would have a varying shape in space.

We studied the impact these two assumptions had on the results of the model in terms of predictive capacity, inference (magnitude analysis) and the fit of the model to the data (goodness-of-fit p-values). The use of the three selection criteria seems essential because they are complementary and their joint use makes it possible to choose the models which are simultaneously the best in terms of predictive power (WAIC score), significantly better than the other models in terms of predictive power (significance of the difference of LOOic) and that are robust in themselves and not only in comparison with others (GOF p-values). Moreover, the use of GOF p-values makes it possible to identify problems that can potentially be fixed (in our case, autocorrelation). In the same way, results should be analyzed with two criteria: the study of the significance of the relationships (as is generally done), but also the study of the magnitude of the effects. The latter provides major results that can be used in decision-making.

The joint use of the three selection criteria and the interpretation criteria showed that random effects should not only be understood as intervening on the average of the answer (random effect

on the intercept), but also on its shape, or even on several parameters of the curve. Ideally, we should have a theoretical knowledge of the studied system in order to choose its random effects (Austin, 2002), or, if we do not have this knowledge, test them all. This being unrealistic, we recommend choosing the best mean function(s) based on complete models, then testing the different possible random effects on the selected mean function(s). This method makes it possible to go beyond the automatic use of random effects on the intercept, which gave unsatisfactory results. Finally, we insist that, for nonlinear functions, the addition of random effects is crucial. Indeed, depending on the choices made about random effects, flat relationships (here with no random effect) or truly sigmoidal relationships (with one or several random effects) are fitted – with the latter fitting data better. Thus, far from merely adding noise to the estimators, the introduction of random effects in nonlinear relationships can actually reveal the relationship more clearly than in models with no random effects. Thanks to the addition of random effects, the two functions we propose appear to be better than any other function we tested, and will undoubtedly be of great interest in ecological studies. The next logical step in our work would be to introduce additional covariables for each of the parameters containing random effects to explain and reduce the level of variation among sites associated with these random effects.

As models are essential to the appropriate prediction and estimation of biodiversity estimators, appropriate model development is necessary. Models should be developed to meet specific use case, with careful consideration of nonlinearity, and also should incorporate random effects in an appropriate manner.

III. DISCUSSION DU CHAPITRE 2

L'ajout d'effets aléatoires dans les modèles linéaires ne permet que de faire varier le niveau moyen de la variable étudiée (en faisant varier l'ordonnée à l'origine) et la force de la relation (en faisant varier la pente). En revanche, l'ajout d'effets aléatoires dans des modèles utilisant des fonctions non-linéaires complexes permet de faire varier complètement la forme de la courbe pour prendre en compte une variation dans la forme de la réponse en fonction de la position dans l'espace géographique ou écologique. Ces effets aléatoires appliqués à divers paramètres de la courbe doivent être davantage considérés, si possible toujours avec une base théorique (ce qui n'était pas vraiment le cas dans cette étude).

Les modèles développés dans ce chapitre s'adaptent bien aux forêts directement étudiées tandis qu'ils peuvent être plus approximatifs pour prédire ce qu'il se passe dans d'autres forêts, non observées (car nous n'avons pas d'information pour déduire a priori la valeur des effets aléatoires qui leur sont associés). Pourtant, au niveau de la population (ici la population des forêts en France), nos modèles sont plus rigoureux, estimant plus précisément les paramètres de moyenne et de variance avec lesquels sont générés les effets aléatoires. Cela signifie que le modèle permet de gagner en précision et réalisme au niveau de l'ensemble de la population (y compris sur la distribution de probabilité des effets aléatoires) tout en augmentant l'incertitude au niveau d'une nouvelle forêt non observée.

Comme nous l'avions présumé, intégrer des fonctions sigmoïdes plus complexes et surtout plus flexibles dans des modèles de type SAR peut s'avérer très efficace. En effet, cet ajout permet de produire de très bons résultats, en termes de capacité prédictive (AIC), de qualité d'ajustement (Goodness-of-fit) et d'analyse de la relation (significativité et magnitude). En conséquence, les modèles intégrant ce type de courbes, et plus généralement des fonctions non-linéaires peu contraintes, nécessitent d'être davantage considérés lors de l'analyse de relations écologiques, tout en alimentant la réflexion sur les raisons qui peuvent mener à cette forme de relation. Leur efficacité pourrait s'étendre au-delà de la biogéographie. En effet, dans le cadre d'approches linéaires ou linéaires généralisées (e.g. approches de type Structural Equation Modeling - SEM) où les relations sont traitées principalement comme étant linéaires, des liens significatifs pourraient émerger à tort. La relation imposée comme étant linéaire ne serait pas en mesure de couvrir la relation réelle entre les deux variables. Elle serait ainsi compensée dans le lien entre deux autres variables faisant apparaître une relation significative, bien que non-existante en réalité, entre ces deux variables. A titre d'exemple, en étudiant la communauté à l'aide de modèles linéaires généralisés, Paillet et al. (2018) concluent à un rôle plus fort des micro-habitats que du volume de bois mort sur les espèces de coléoptères saproxyliques. Les conclusions seraient-elles similaires en envisageant des relations non-linéaires ?

Malgré des résultats très prometteurs, les modèles développés demeurent limités dans leur utilisation (pour plus de détails cf. Chapitre 2 - II.4.e Some limitations on the general use of our models). Pour autant, notre étude avait pour objectif de mener une réflexion et de proposer une approche méthodologique quant à l'utilisation des effets aléatoires dans les modèles non-

linéaires. L'utilisation de tels modèles s'adaptant bien à nos données, l'une des prochaines étapes serait de les appliquer à des données de communautés saproxyliques comprenant plus d'espèces (incluant en plus des coléoptères saproxyliques, des données de mycologie et de bryologie), afin d'étudier précisément le lien entre les communautés saproxyliques et la quantité de bois mort dans les forêts françaises, avec pour ambition, à termes, d'émettre des recommandations de gestion. Nous envisageons également de prendre en compte dans les modèles le caractère bruité du volume de bois mort (la variable explicative X – e.g. Model II Legendre and Legendre 1998 p. 504) ; nous nous attendons à ce que la relation estimée soit plus forte. Enfin, l'aboutissement de ce travail sur l'étude de la relation entre le volume de bois mort et les communautés saproxyliques en forêts françaises, serait de pouvoir expliquer les variations aléatoires de la forme des relations entre les deux variables par des variables écologiques. Ce travail permettrait d'être plus précis notamment au niveau des prédictions sur des forêts non observées, et de donner des recommandations plus fiables. Il permettrait également de s'écarter un petit peu plus d'une approche phénoménologique pour entamer une réflexion sur les aspects causaux sous-jacents (cf. Introduction – I.4).

Dans un contexte similaire d'analyse de la réponse des espèces à un gradient (que nous abrégeons SReRs), Huisman et al. (1993) ont également mené une réflexion poussée sur de possibles formes de relations (décrites par des fonctions appelées HOF), dont notamment leur modèle III qui applique une fonction sigmoïdale dont l'asymptote haute est estimée. Notre modèle complète cette réflexion par : (i) une extension des formes sigmoïdales possibles (via notamment l'estimation de l'asymptote basse et de l'asymétrie) ; (ii) une considération particulière envers

l'inclusion d'effets aléatoires (peu fait sur des fonctions HOF) ; (iii) et une utilisation conjointe de plusieurs métriques de comparaisons et d'étude de la relation notamment à travers le GOF p-value et l'analyse de la magnitude.

Ce type de questionnements apporte de nouvelles pistes de réflexions et permet de développer de nouvelles solutions pour étudier la réponse des espèces à un gradient. En outre, ces fonctions sigmoïdes élaborées pourraient également être intégrées dans des modèles linéaires généralisés binomiaux qui reposent normalement sur une fonction de lien canonique peu flexible, la fonction logit (réciproque de la fonction logistique commune dont les asymptotes basses et hautes sont fixées respectivement à zéro et un). Remplacer cette fonction logistique, par exemple, par une fonction sigmoïdale dont les deux asymptotes seraient estimées, pourrait permettre une plus grande flexibilité dans l'estimation de la réponse possible et éventuellement une meilleure estimation.

CHAPITRE 3

Nouvelles formes de fonctions logistiques pour les modèles de régression logistiques

I. INTRODUCTION DU CHAPITRE 3

Les modèles linéaires généralisés binomiaux (binomial GLM), avec une loi de Bernoulli, sont utilisés pour décrire des relations pour lesquelles la variable expliquée est de forme binaire, c'est-à-dire qu'elle présente uniquement deux états, généralement codés en 0 et 1 (McCullagh and Nelder 1989, Agresti 1990, Collett 2002, Cox 2018). Les fonctions de lien disponibles pour étudier de telles données, dans le cadre de GLMs, sont les fonctions logit, probit et cloglog (cf. Table 0.1). La fonction logit est la fonction de lien canonique. Elle est la réciproque de la fonction logistique commune, qui est symétrique autour du point d'inflexion et possède des asymptotes fixées à zéro et un. Ainsi, la fonction de lien logit permet de transformer l'ensemble des probabilités sur l'intervalle $]0; 1[$, en l'ensemble des réels, c'est-à-dire l'intervalle $]-\infty; +\infty[$, qui est l'ensemble d'états possibles de la combinaison linéaire. De manière générale, les fonctions de lien canoniques peuvent être utiles car elles sont faciles d'utilisation, et semblent souvent bien adaptées, permettant de transformer l'espace d'état des données en l'ensemble des réels. Cependant, les asymptotes fixées à zéro et un de la logistique commune constituent une hypothèse auxiliaire forte, puisqu'elle contraint la forme de la relation. Or, dans certains cas, il paraît peu probable que les asymptotes fixées à zéro et un (impliquant une tendance vers un échec obligatoire pour une valeur du prédicteur s'approchant de moins l'infini et une tendance vers un succès obligatoire pour une valeur prédicteur s'approchant de plus l'infini) soient adaptées à tous les jeux de données. Nous pouvons prendre pour exemple un cas d'étude souvent exploité par A. Gelman dénommé « Red State-Blue State » qui présente le résultat des votes aux élections américaines en fonction du revenus des électeurs (Gelman et al. 2010, Gelman 2011).

Une relation positive émerge entre le revenu médian des ménages des comtés d'un Etat (le Texas) et la proportion d'électeurs votant Républicain dans ce comté. Cependant, il paraît déraisonnable d'affirmer que pour un revenu moyen tendant vers plus l'infini, la proportion des électeurs du comté votant Républicain tendrait vers 100% (Figure 3.1.A.). D'après une rapide analyse graphique, les asymptotes semblent en effet plutôt situées autour de 18% et 91% (Figure 3.1.B.). Une fonction de lien peu flexible avec des asymptotes fixées à zéro (0% des électeurs votent Républicain) et un (100% des électeurs votent Républicain) paraît donc ici peu adaptée.

Par ailleurs, le modèle basé sur la réciproque logistique commune, échouerait à adapter une telle fonction sur ce type de données. En effet, en tentant de faire concorder les asymptotes fixes aux données, le modèle échouerait à interpréter la partie du gradient pour laquelle la probabilité de succès tendrait vers zéro, mais où des « succès » seraient observés, et la partie du gradient où la probabilité de succès tendrait vers un, mais où des « échecs » seraient toujours observés. Ainsi, le modèle estimerait très mal la pente de la fonction pour s'adapter à ces « succès » et « échecs » résiduels (cf. Figure 3.1.A). La contrainte des asymptotes fixées enlevée, le modèle serait beaucoup plus efficace à estimer correctement la pente (cf. Figure 3.1.B).

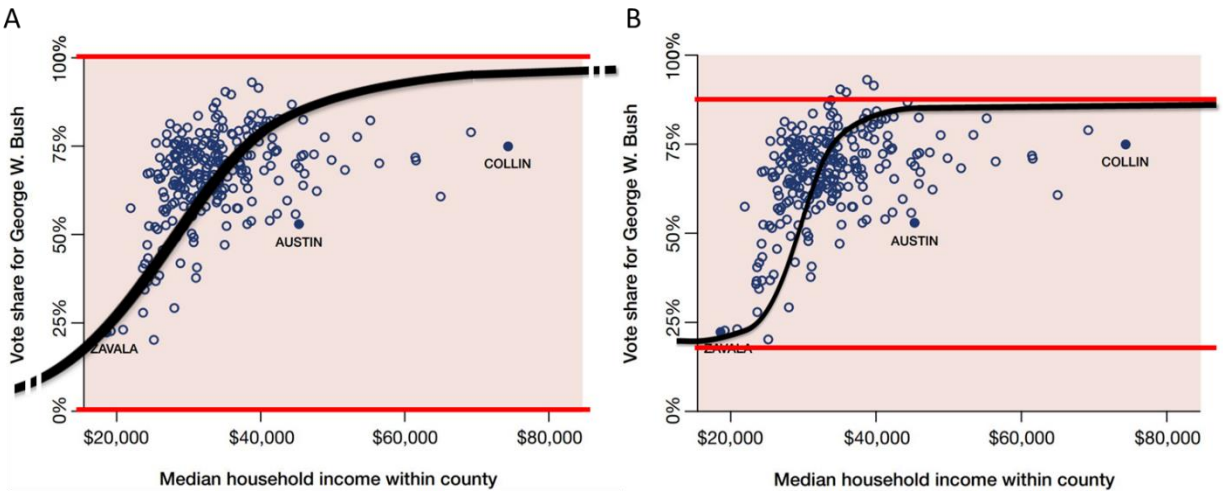


Figure 3.1: Proportion des votes pour George W. Bush (candidat Républicain) en fonction du revenu médian des électeurs dans les différents comtés de l'État du Texas en 2000. Figure modifiée d'après (Gelman 2011). Les lignes rouges représentent sur la figure A les asymptotes imposées par le modèle logit, et sur la figure B les asymptotes théoriques observées graphiquement. Les courbes noires représentent schématiquement les courbes issues de la fonction de lien et des paramètres estimés par le modèle, dans un cas où la fonction logistique est imposée (A) et dans un cas où une fonction logistique avec asymptotes estimées est appliquée (B).

Pour résoudre ce problème nous avons donc voulu développer un modèle dans lequel les asymptotes de la fonction logistique ne seraient pas fixées à zéro et un mais estimées et pouvant également prendre des valeurs très proches de zéro et un lorsque les données l'exigeaient.

II. MANUSCRIT 3

Generalized Linear Misleading: the need for a logistic function with estimated asymptotes to complement the classical logistic function for binomial GLMs

Ugoline GODEAU[1*], Frédéric GOSSELIN[1]

[1] INRAE, UR EFNO, Domaine des Barres, F-45290, Nogent-sur-Vernisson, France.

Statut : en preparation

Abstract: In many domains, researchers use binomial Generalized Linear Models (GLMs) to study the relationships between a binary outcome and a set of predictors. The canonical link function is the logit function, which transforms a value between zero and one into a real variable. However, its inverse function, the classical logistic function, has asymptotes fixed at zero and one, and the impacts of this assumption have never been studied. Herein we propose new models with logistic functions that have estimated, not fixed, asymptotes. We study its impact on model performance compared with GLMs and Generalized Additive Models (GAMs). Through a simulation study, we illustrate the improvement that these new models offer in terms of predictive capacity, parameter estimations and goodness of fit, compared to the classical logistic model. When we simulated data with these new link functions, we observed an important underestimation of the slope parameter for the logit model; this underestimation can have a strong impact if the statistical model is used quantitatively. We also found that the model with estimated asymptotes can equal or even surpass GAMs in terms of predictive capacity. In particular, our simulations highlighted the inability of GAMs to adequately model data i) when there are at least two predictors and the link function used is inadequate; or ii) when the gradient is not uniformly distributed. Finally, we tested and compared these models on real data from various research domains. We found that cases where the asymptotes differ from zero and one can indeed be found in real data. After it has been developed further, the new logistic function we propose, called the “Estimated Lower Upper Asymptote” logistic, or ELUA logistic, may very well be of major interest in the study of binary data in contexts where the asymptotes of the logistic function are not necessarily zero and one.

Keywords: Logit; ELUA logistic; Bias; Predictive capacity; Generalized linear model; Generalized additive model; Regression; Binary data; Simulation study; Real data.

II. 1. Introduction

Binary data describe cases where each observation (response variable y) has two possible outcomes, usually called “success” (y=1) and “failure” (y=0). Binary data arise frequently in many areas of research such as survival analyses (in physiology, treatment testing…) and presence-absence (occurrence of pathogens in epidemiology, species in ecology, diseases in medicine, behavior in sociology or ethology…), among others. This type of data follows a Bernoulli distribution, with one parameter “p”: the probability of success for one trial and for one individual P(y=1) = p (McCullagh and Nelder 1989, Agresti 1990, Collett 2002, Cox 2018).

In order to explain or predict the occurrence of the two states (success v. failure) through a set of predictors, binary data are generally analyzed with Generalized Linear Models (GLM, McCullagh and Nelder 1989, Agresti 2015). These models rely on four elements: a univariate random variable, its probability distribution (response variable with independent observations), a linear predictor (a linear combination of predictors) and a link function (relating the linear predictor to the mean of the random component). In the case of binary data, the link function gives the mean probability of success by transforming the linear predictor into values between zero and one. When analyzing binary data, the canonical GLM is the logistic Bernoulli regression model in which the link function is the standard logit, the most common function to transform a variable of probability between zero and one into a real variable. The inverse of the logit function is the

classical logistic function, that can be written as $f(x) = 1 / (1 + e^{-x})$, with a lower asymptote of zero (the observation has no chance of being a success) and an upper asymptote of one (the observation can only be a success). The logit function is the canonical GLM link function associated with binomial distribution because it has simple theoretical properties, its results are easy to interpret and it is well adapted to retrospectively collected data (McCullagh and Nelder 1989).

Even though alternative link functions (probit, complementary log-log link functions and even the asymmetrical log-log link function) have been proposed and compared to the logit function in the literature (McCullagh and Nelder 1989, Huettmann and Linke 2003), to our knowledge no study has ever evaluated the impact of the fixed asymptotes (0.0 and 1.0) in the classical logistic inverse of the logit link function. Indeed, the classical logistic function makes two strong assumptions related to asymptotes: 1) the probability of having a "success" unavoidably tends to one when the predictor tends to plus infinity; 2) conversely, the probability of having a "failure" unavoidably tends to one ("success" reaches zero) when the predictor tends to minus infinity. These two assumptions may not be always pertinent. Indeed, in some cases there may be logical, or theoretical, reasons to challenge them. For example, in political science, when studying the relationship between votes for Democrats in the US and voter income, it might be inappropriate to impose that all people with very high incomes vote Republican and all people with low incomes vote Democrat. Imposing the classical logistic function in such conditions may hinder good fit, and could even result in underestimating the slope. A more flexible position of the asymptotes in the inverse of the link function may be more appropriate. In this respect, in some fields like biology, logistic functions with an estimated (not fixed at one) upper asymptote have already proven to

be efficient on binary data (Price et al. 2019). Nevertheless, no study to date has seriously investigated the impact of these two fixed asymptotes on model fit and parameter estimations.

In order to overcome the limitations of GLMs, general additive models (GAMs) are often used. They are a generalization of the linear regression model, where the linear combination of explanatory variables is replaced by the sum of the smooth functions (Wood 2006, Hastie 2017). However, this model is semi-parametric, not fully parametric; therefore, it cannot account for theoretical hypotheses related to curve shape. This makes it difficult to access parameters that would otherwise easily indicate relationship magnitude or that are suited for meta-analyses. That is why GLMs containing an explicit nonlinear link function with estimated asymptotes seem to be promising alternatives. They would allow the model to estimate either asymptotes similar to those of the classical logistic function (zero and one), or different from these extrema (see Price et al. 2019), while retaining access to parameter estimations and controlling the shape of the relationship.

This paper investigates how five different models perform on simulated binary data, under different circumstances for the predictor gradient. We compared a classical GLM, with lower and upper asymptotes fixed at zero and one, to three models with estimated lower and/or upper asymptotes; we also compared the four models to GAMs. To this end, simulations provide stable comparative material between the different models because, by knowing their expected outputs (e.g. the true value of statistical parameters; e.g. Saas and Gosselin 2014) they allow studying the relative performance of each model and its intrinsic quality at the same time. To justify extrapolating our results to real-world multidisciplinary situations, we also tested three of the five models on real datasets from various fields. The aims of this paper are to show, in cases where

the lower and/or upper asymptotes for binary data are not zero and one: 1) that using classical GLMs can result in erroneous estimations, especially in a strong underestimation of the slope, and this has a strong impact if the statistical model is used quantitatively; 2) that models with estimated lower and upper asymptotes perform equivalently, or better, both in terms of predictive power and estimations; 3) that GAMs also generally do a good job when there is only one explanatory variable, but have serious drawbacks, resulting in a serious underestimation of slope, as soon as there are two variables or when the explanatory variable is not evenly placed on the gradient; and 4) that such situations can be found in real datasets and that the new models we developed are an interesting new tool for statistical modelers. We hope that in the light of our results, researchers will be more inclined to consider models with estimated asymptotes, instead of classical GLMs, to avoid misestimating other parameters (slope and inflexion point) and misinterpreting the results for datasets where the explained variable is binary.

II. 2. Material and Methods

Binomial GLMs are classically used with canonical link functions like logit, probit and cloglog. Here, we restrict our attention to the logit function, and it inverse the classical logistic function:

$$f(x) = \frac{1}{1 + \exp(-x)}$$

We argue that this function has a strong embedded assumption: that asymptotically the values reached by the function should be 0.0 and 1.0. We hypothesized that this assumption would have strong implications on the estimation of statistical parameters in cases where the data generated

do not correspond to this assumption. In particular, we consider cases where the lower and/or upper asymptotes differ from 0.0 and 1.0 respectively. More specifically, we will consider:

$$f(x) = L + \frac{K - L}{1 + \exp(-x)}$$

II. 2. a. Simulated Datas

To test our models, we generated multiple success-failure (1/0) simulation datasets in ten different scenarios. The first eight scenarios involved simulated univariate datasets based on a common logistic function: $f(x) = L + ((K - L)/(1 + exp(-sl * (x - ip))))$ and with a slope parameter (*sl*) fixed at 0.2, the abscissa at inflexion point (*ip*) fixed at $x = 50$, and a binomial distribution. The number of observations was randomly drawn between exp(6)≈403 and exp(8)≈2981 from a uniform distribution on the log scale. We generated datasets with varying features to test models in contrasting cases, which could reflect different types of real datasets (cf. Annexe III. Figure S1.1).

The first simulation scenario with 1,000 datasets (cf. Table 3.1 Scenario1) fixed the lower asymptote of the logistic function (*L*) at 0.0 and the upper asymptote (*K*) at 1.0 with a predictor gradient equally distributed from 0.0 to 300.0.

The second simulation scenario with 10,000 datasets (cf. Table 3.1 Scenario2) was generated on the same gradient, but with upper and/or lower asymptotes that varied. In one third of the cases, the lower asymptote was equal to 0.0 and the upper asymptote was randomly uniformly drawn between 0.6 and 0.9. In another third of the cases, the lower asymptote was randomly uniformly drawn between 0.1 and 0.4 and the upper asymptote was fixed at 1.0. Finally, in the last third of

the cases, lower and upper asymptote were both randomly uniformly drawn between the same values as described above (respectively between 0.1 and 0.4 and between 0.6 and 0.9).

The same simulation scenario was used on other predictor gradients in order to treat cases where the inflexion point was at the beginning, the middle or the end of the gradient, and to account for cases where the gradient was or was not totally covered. The chosen gradients ranged from 0.0 to 100.0 (cf. Table 3.1 Scenario3), from 0.0 to 60.0 (cf. Table 3.1 Scenario4), from 40.0 to 300.0 (cf. Table 3.1 Scenario5), from 40.0 to 100.0 (cf. Table 3.1 Scenario6) and from 40.0 to 60.0 (cf. Table 3.1 Scenario7).

Another simulation scenario with 1,000 datasets (cf. Table 3.1 Scenario8) was generated on a gradient ranging from 0.0 to 300.0 but randomly distributed between the bounds and with lower and upper asymptotes randomly uniformly drawn between the same value ranges as described above (cf. Scenario2, L between 0.1 and 0.4 and K between 0.6 and 0.9).

Finally, two simulation scenarios with 10,000 datasets were generated with lower and upper asymptotes randomly uniformly drawn as described above for Scenario2, but with the response variable drawn from a bivariate logistic regression. The two predictors X_1 and X_2 were linearly combined with respective slopes equal to 0.2 and 0.1 ($CL = 0.2 * X_1 + 0.1 * (X_2 - mean(X_2))$, and the linear combination *CL* was used as a predictor in the logistic function with same inflexion point *ip* at *cl*=50.0, ($f(x_1, x_2) = L + ((K - L)/(1 + exp(-(cl - 0.2 * ip))))$). In the first multivariate simulation scenario (cf. Table 3.1 Scenario9), data were generated based on two predictors: X_1 with a [0.0;300] gradient and X_2 with a [0.0;100.0] gradient. In the second

multivariate simulation scenario (cf. Table 3.1 Scenario10), data were generated based on two predictors: X_1 with a [40.0;100.0] gradient and X_2 also with a [0.0;100.0] gradient.

Table 3.1: List of simulation Scenarios with their characteristics in terms of number of datasets simulated, gradient of predictor X and values taken by lower (L) and upper (K) asymptotes.

Name of the simulation scenario (type)	Number of datasets simulated	Gradient of predictor *X*	Values taken by *L* and *K*
Scenario1 (univariate)	1,000	0-300	*L*=0.0 and *K*=1.0
Scenario2 (univariate)	10,000	0-300	1/3 of cases *L*=0.0 and *K* ∈ [0.6;0.9] 1/3 of cases *L* ∈ [0.1;0.4] and *K*=1.0 1/3 of cases *L* ∈ [0.1;0.4] and *K* ∈ [0.6;0.9]
Scenario3 (univariate)	10,000	0-100	
Scenario4 (univariate)	10,000	0-60	
Scenario5 (univariate)	10,000	40-300	
Scenario6 (univariate)	10,000	40-100	
Scenario7 (univariate)	10,000	40-60	
Scenario8 (univariate)	1,000	0-300 (random)	*L* ∈ [0.1;0.4] and *K* ∈ [0.6;0.9]
Scenario9 (multivariate)	10,000	0-300 (X_1) and 0-100 (X_2)	1/3 of cases *L*=0.0 and *K* ∈ [0.6;0.9] 1/3 of cases *L* ∈ [0.1;0.4] and *K*=1.0 1/3 of cases *L* ∈ [0.1;0.4] and *K* ∈ [0.6;0.9]
Scenario10 (multivariate)	10,000	40-100 (X_1) and 0-100 (X_2)	

Curve shapes produced by univariate simulation scenarios are presented in Annexe III. Figure S1.1. Link function used for univariate simulations: $P(Y = 1|X = x) = L + ((K - L)/(1 + exp(-sl * (x - ip))))$, *with P(Y=1|X=x) the probability of success given x, L and K respectively the lower and upper asymptotes, sl the slope and ip the inflexion point. Link function used for multivariate simulations:* $P(Y = 1|X_1 = x_1, X_2 = x_2) = L + ((K - L)/(1 + exp(-(cl - 0.2 * ip))))$, *with CL the linear combination of predictors X_1 and X_2, with respective slope coefficients coef1 and coef2:* $cl = coef1 * x_1 + coef2 * (x_2 - mean(x_2))$.

II. 2. b. Real Datasets

Real datasets were chosen with respect to the following criteria:

- Free of cost and easily accessible (i.e. that were found with a simple internet search);
- Taken from various fields of research;
- With a binary response variable and a numeric predictor gradient;
- With a row for each observation;
- With a varying number of observations but with at least 50 observations;
- Adapted to analysis through a GLM with a logit function and a Bernoulli distribution.

With respect to the first two criteria described above (accessibility and variety of fields), R packages are good sources of datasets. To cover numerous fields, we chose datasets from different R packages through a Google search for "R Datasets" in April, 2019. The first results yielded a website [7] with information about several R packages and containing valuable information about datasets like "has_binary" and "has_numeric", two of our other criteria. We selected all the datasets with a "YES" in these two columns, then retained only the ones where the explained variable was clearly binary and at least one explanatory variable was numeric (through clear examples, personal knowledge or minimum research on references). When a dataset was duplicated in two different R packages, we only kept one. When two or more response variables were identified, they were treated as different datasets. One dataset

[7] http://vincentarelbundock.github.io/Rdatasets/datasets.html

identified as simulated was removed. We thus selected 52 datasets for the univariate analysis. For the multivariate analysis, we removed the datasets where only one predictor was available, making a total of 39 datasets. We had no a priori knowledge of the selected datasets regarding which explanatory variables to use, or what form of relationship to expect (or even if they were monotonic).

II. 2. c. Modelization approach

Models on univariate datasets. Five types of models (Table 3.2) were tested on the univariate datasets (Scenario1 to Scenario8), all based on the Bernoulli distribution:

1) A model equivalent to a classical GLM (L0K1) with the lower and upper asymptotes of the classical logistic function respectively fixed at nearly 0.0 and nearly 1.0.
2) A model (L0Kest) where the lower asymptote was fixed at nearly 0.0 and the upper asymptote was estimated: the "Estimated Upper Asymptote" (EUA) logistic.
3) A model (LestK1) where the lower asymptote was estimated and the upper asymptote was fixed at nearly 1.0: the "Estimated Lower Asymptote" (ELA) logistic.
4) A model (LestKest) where both lower and upper asymptotes were estimated: the "Estimated Lower & Upper Asymptotes" (ELUA) logistic.
5) A general additive model (GAM).

When fixed, the lower and upper asymptotes were respectively set to nearly 0.0 and nearly 1.0 because exact values of 0.0 and 1.0 created numerical singularities and prevented the models

from converging. Asymptotes were fixed at $\mathrm{L} = 1 - 1/(1 + \exp(-30))$ and $\mathrm{K} = 1/(1 + \exp(-30))$.

Table 3.2: List of model names and their associated lower (L) and upper (K) asymptote treatments.

Model Name	Link function name	L	K
L0K1	Logit	≈0.0	≈1.0
L0Kest	EUA logit	≈0.0	estimated
LestK1	ELA logit	estimated	≈1.0
LestKest	ELUA logit	estimated	estimated

Concerning the models with estimated *L* and/or *K*, because uniform distribution of priors was not easy with the "TMB" R package (Kristensen et al. 2018), we used intermediate values of the asymptotes, called *Lp* and *Kp*, to recalculate them. We put priors on *Lp* and *Kp* with a normal distribution and mean and standard deviations respectively equal to 0.0 and 100.0. Estimated lower (*L*) and upper (*K*) asymptotes were then recalculated from these parameters as follows:

$$L = \frac{1}{1 + exp(-Lp)}$$

$$K = L + (1 - L) * \frac{1}{1 + exp(-Kp)}$$

Finally, to analyze observations (Y_o) with the parametric link function as the mean, we used binomial distribution parameterized via logit(prob) (binom_robust in R), because it is numerically stable for probabilities close to 0.0 or 1.0, when working on a logit scale. The parametric function of the model was similar for all models (in Table 3.2):

$$F(X_o, L, K, sl, ip) = \log\left(\frac{L + \frac{K-L}{1.0 + \exp\left(\left(-sl * (X_o - ip)\right)\right)}}{1 - L - \frac{K-L}{1.0 + \exp\left(\left(-sl * (X_o - ip)\right)\right)}}\right)$$

$$Y_o \sim Br(F(X_o, L, K, sl, ip), 100)$$

With: *F,* the parametric function of the model

o, the observation index

X_o, predictor at observation *o* (where *X* is the vector of predictors)

K, the upper asymptote

L, the lower asymptote

sl, the relative slope at the inflexion point (for the logistic function normalized to L=0.0 and K=1.0)

ip, the abscissa at inflexion point

$Y_o \sim Br(F(X_o, L, K, sl, ip), 100)$ means that Y_o is considered a random variable that follows a (robust version of the) binomial distribution with a mean parameter $\mathrm{F(X_o, L, K, sl, ip)}$ and a variance parameter of 100.0.

Finally, for the GAM, in each case, we compared restricted cubic splines (rcs), with their acknowledged advantages, and P-splines (ps), for their connections with mixed models and Bayesian analysis. We tested both types with different spline dimensions (from two to ten), and

kept the best model overall in terms of AICc. We calculated the general tendency (the slope) of the best selected model as follows:

$$NSL = 4 * \left(\max_{o}(predicted) - \min_{o}(predicted)\right) * \max_{o}\left(\frac{predicted_{X+\varepsilon} - predicted_{X-\varepsilon}}{2 * \varepsilon}\right)$$

With, $predicted$, the values predicted by the GAM on gradient X

$predicted_{X+\varepsilon}$, the values predicted by the GAM on gradient $X + \varepsilon$

$predicted_{X-\varepsilon}$, the values predicted by the GAM on gradient $X - \varepsilon$

ε, the value added to (or removed from) each *x* on gradient *X*. We used $\varepsilon = 0.00001$

Models on multivariate datasets. For multivariate cases (Scenario9 and Scenario10), we first estimated GAM models. We tested these GAMS with two types of splines (ps and rcs) and applied various spline dimensions (from two to ten) to each of the two predictors in the same way, and kept the best model overall in terms of AICc. We then adapted and tested models L0K1 and LestKest by estimating the coefficients of each *X* variable implicated (resulting from the association of weights of the linear combination and the general slope *sl*), in addition to *L* and *K* for the LestKest model.

$$F(CL_o, L, K, ip, coef_1) = \log\left(\frac{L + \frac{K - L}{1.0 + \exp(CL_o - ip * coef_1)}}{1 - L - \frac{K - L}{1.0 + \exp(CL_o - ip * coef_1)}}\right)$$

$$Y_o \sim B\left(F\left(X_{o,i}, L, K, ip, coef_i\right), 100\right)$$

With: *i*, the predictor index

$X_{0,i}$, predictor *i* at observation *o* (with $\bar{\bar{X}}$ being the matrix of predictors for all observations, and $\overline{X_{1,}}\overline{X_{,i}}$ the vector of predictor *i* for each observation)

CL_o, the linear combination of predictors at observation *o:* $CL_o = \sum_i X_{o,i} * coef_i$

$coef_i$, the slope coefficient of predictor *i*

Here, in the link function *F*, *ip* was multiplied by $coef_1$ to facilitate interpretation, because *ip* represented the inflexion point relative to variable $\overline{X_{1,}}\overline{X_{1,}}$ (leading to the formula: $(coef_1 * (X_{1,} - ip)$). We chose this specific formulation because alternative parametrizations gave equivalent or worse results.

Models for real datasets. For the real datasets, we fitted the L0K1 model, the LestKest model and two different types of GAM models. The first type of GAM was the same as for the simulated datasets (i.e. the best from either ps or rcs with two to ten degrees of freedom), therefore assuming a free form (called GAM-free). The second type was a P-spline GAM forced to be monotonic (called GAM-monotone), for which we kept the best from an increasing (mpi) or decreasing (mpd) form for univariate cases, and the best of four possible combinations (mpi-mpi, mpd-mpd, mpi-mpd and mpd-mpi) for multivariate cases. We fitted GAM-monotone in addition to GAM-free because the latter could have provided better results simply because the relationship was not monotonic. Since the GLM methods we studied were by nature monotonic, we wanted to compare them with monotonic GAMs, while allowing some flexibility in curve shape. We only used the GAM-free results to check for monotonicity.

For univariate analyses, for each dataset and each model, we tested all the different variables as predictors and selected the best one based on its associated AICc. For multivariate analyses, we

kept the variable selected by the univariate analysis, tested all the other predictors when several were available, and kept the model with two predictors with the best results in terms of AICc. We only tested multivariate models with two explanatory variables to avoid multiplying the number of GAM-monotones.

II. 2. d. Numerical methods

We chose the Template Model Builder (TMB) tool in the R software (R Core Team 2018), "TMB" R package (Kristensen et al. 2018), for its rapidity (to process the 82,000 datasets generated), its thorough and diverse documentation, and its compatibility with various optimization methods. To optimize the TMB models, we applied five methods ("nlminb", "L-BFGS-B" and "nlm" from base R; "Rcgmin" and "ucminf" from the "Rcgmin" and "ucminf" packages - Nash 2014, Nielson and Mortensen 2018) to different draws of initial values (cf. Annexe III. Table S1.1) and selected outputs resulting from the one with the maximum value of likelihood. P-splines were from the "mgcv" package (Wood 2011), and restricted cubic splines (rcs-splines) from the "rms" package (Harrell 2019). We ran monotonic GAMs with the "scam" function from the "scam" package (Natalya 2019).

II. 2. e. Model evaluation criteria

II. 2. e. 1. Relative performance evaluation

The Akaike information criterion (AIC) is a metric used to compare the relative predictive capacity of statistical models by measuring deviance from the data, penalized with the number of parameters to prevent over-fitting. It is an approximation of the (relative) expected Kullback–

Leibler distance based on Fisher's maximized log-likelihood (Burnham and Anderson 2002). For each model and each dataset, we calculated the corrected version of the metric (AICc - Hurvich and Tsai 1989) based on: the number of parameters to estimate in the model (*np*), the negative-log-likelihood (*NLL*) and the number of observations (*nobs*).

$$AICc = 2 * np + 2 * (NLL) + 2 * np * \frac{np + 1}{nobs - np - 1}$$

II. 2. e. 2. Intrinsic performance evaluation

In order to evaluate the accuracy of the estimations, we first looked at the ratio between the mean estimates and the true value of the parameters, the parameters being better estimated when this ratio approaches 1.0. We also measured Type I errors to evaluate the quality of inference by measuring error rates for inference. For each parameter, model and dataset, we calculated the type I error at a level of 5% (and a level of 1%), based on the proportion of simulations in which the 95% (respectively 99%) confidence interval included the true value of the parameters (considering the mean estimate and its standard error and assuming that the estimator had a Gaussian distribution). To make these two intrinsic performance evaluations comparable among models, we used the normalized slope ($NSL = sl * (K - L)$) instead of the actual slope.

Moreover, for one univariate dataset (dataset 1 in Scenario3) and four multivariate datasets (datasets 1 and 7 in Scenario9; and datasets 1 and 9 in Scenario10), we drew 10,000 draws of the parameters from a normal multivariate distribution based on the mean estimates and variance-covariance matrix at the optimum (to take into account the correlation between parameters). We plotted the mean curve shape estimated by the LestKest and L0K1 models (with their 0.025 and

0.975 confidence intervals) and compared it to the true shape, and the shape predicted by GAM (obtained with the "predict.gam" function in the "mgcv" package). We chose to plot datasets linked to extreme cases because they seemed more representative of the cases we had encountered during data analysis.

Finally, we computed the plug-in p-value (Robins et al. 2000) to gauge whether the data were consistent with the model at the optimum parameter values. To do this, we measured how often discrepancy functions calculated with observed data were greater than discrepancy functions calculated with replicated data, drawn from the model with the optimum parameters. Discrepancy metrics were applied on the randomized quantile residuals of data (Dunn and Smyth 1996, Gosselin 2011), denoted as Ynorm, because the resulting p-values have a uniform distribution if the model used to analyze the date is the same as the model used to generate the data (Johnson 2007). We estimated the plug-in p-values with 10,000 data replications. Our discrepancy metrics targeted: i) the distribution probability (with the mean, variance, skewness and kurtosis of Ynorm); ii) the link function through Hoeffding's dependence (function "hoeffd" in Hmisc package; Harrell 2001) between Ynorm and the fitted value (corYnormlink) and between Ynorm and the explanatory variable (corYnormx) ; and iii) heteroscedasticity through Hoeffding's dependence between the square of Ynorm and the fitted value (corYnormsqlink). We also calculated meanYnorm and corYnormlink for three different parts of the curve: the central part (between the 10% and 90% quantiles of the fitted values), the lower part (below the 10% quantile of the fitted values) and the upper part (above the 90% quantile of the fitted values). Due to insufficient computing time, we calculated these p-values on a subsample of the first 100 datasets in each simulation scenario. Based on this subsample, we calculated the frequency with which

the plugin p-values detected issues at a level of 5%, 1% and 0.1%, and highlighted cases that departed from these nominal levels with probabilities of 5% and 1%.

II. 3. Results

II. 3 . a. Comparison between models on univariate simulated datasets

II. 3 . a. 1. Comparison of the relative quality among models (AICc)

On datasets where lower and upper asymptotes were set respectively at zero and one, model L0K1 performed much better than all other TMB models in terms of predictive capacity (AICc) (Scenario1 in Figure 3.2 and in Annexe III. Figure S2.1 and Annexe III. Figure S2.2). Strong discrepancies appeared among the models' predictive performances when asymptotes were different from zero and one, and depended on the gradient considered (Scenario2 to 8 in Figure 3.2 and in Annexe III. Figure S2.1 and Annexe III. Figure S2.2):

- When the data covered the whole gradient but with a large proportion of the data on the upper asymptote (Scenario2 and Scenario8), L0Kest performed better and LestKest considerably better in terms of AICc than L0K1. This was always the case, whether the gradient was uniformly (Scenario2) or randomly distributed (Scenario8, especially for L0Kest). LestK0 did not perform any better than L0K1 on average.
- When the data covered the whole gradient and was centred (Scenario3), L0Kest and LestK1 were both equivalent to L0K1. However, LestKest performed much better than L0K1.

- When the data covered only the beginning of the gradient (Scenario4), L0Kest performed poorly compared to L0K1; however, LestK1 performed better than L0K1, and LestKest performed even better.
- When the data covered only the end of the gradient (Scenario5 and Scenario6), LestK1 performed poorly compared to L0K1; however, L0Kest performed better than L0K1 and LestKest still performed better.
- Finally when the data covered only the middle of the gradient and reached no asymptotes (Scenario7), L0K1 performed either better than or equivalently to all other models.

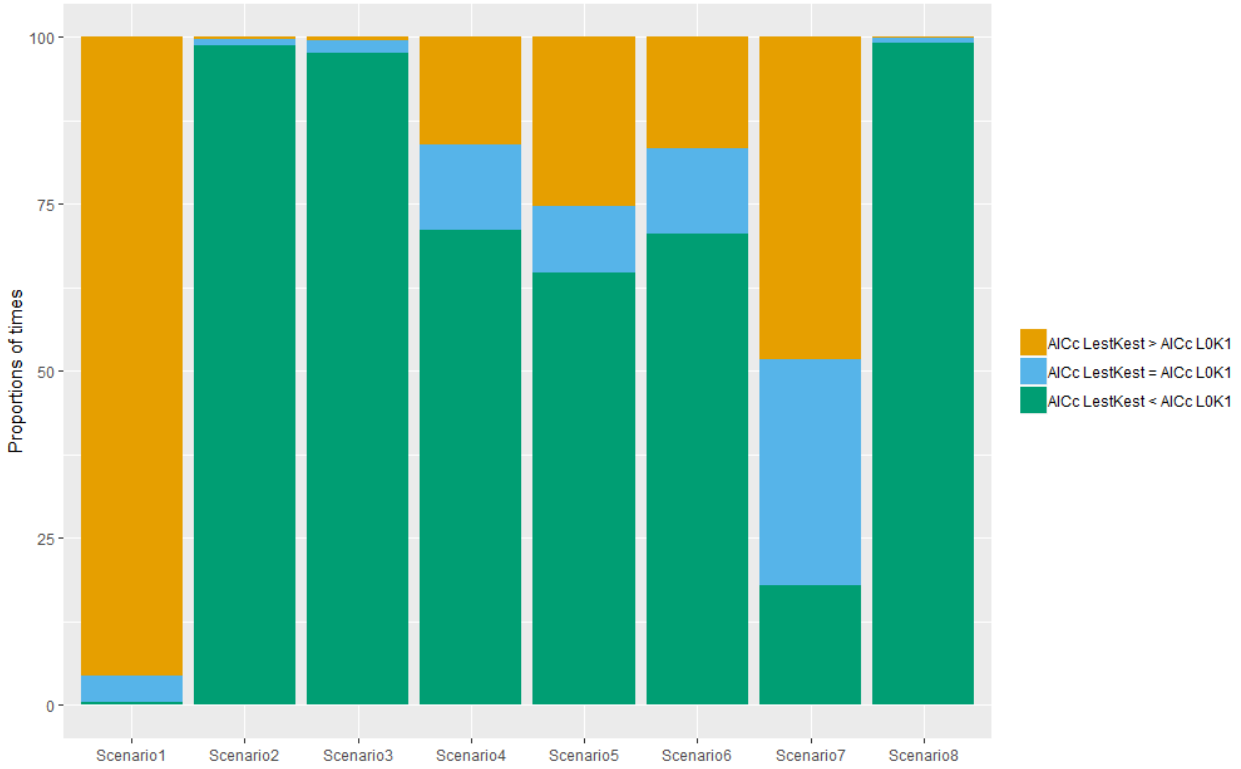


Figure 3.2: Proportion of times, for each simulation scenario, where the model LestKest had less good predictive capacity than L0K1 (i.e. an AICc greater by at least 2.0 points, in orange); where the predictive capacity of models LestKest and L0K1 was equivalent (AICc values were within 2.0 points, in blue); and where LestKest had better predictive capacity than model L0K1 (an AICc lower by at least 2.0 points, in green).

Regarding GAMs, P-spline models usually performed equivalently (78.7% of the time), sometimes better (17.5%) and rarely worse (3.8%) than RCS models. We therefore kept the P-spline models to use as the GAM results for further analyses. When we compared the results of these P-spline models with the LestKest model, we found that the GAMs usually performed equivalently to or better than the LestKest model (Figure 3.3). This was particularly true for data where the asymptotes were set at zero and one (Scenario1), however, it was also the case when the asymptotes were variable but the data did not cover the whole gradient (Scenario4 to 7), and when the data covered the whole gradient but was centered (Scenario3). However, when the data covered the whole gradient but most of the observations were concentrated on the upper asymptote (Scenario2), LestKest performed a bit better. Finally, LestKest performed noticeably better on data randomly distributed along the gradient (Scenario8).

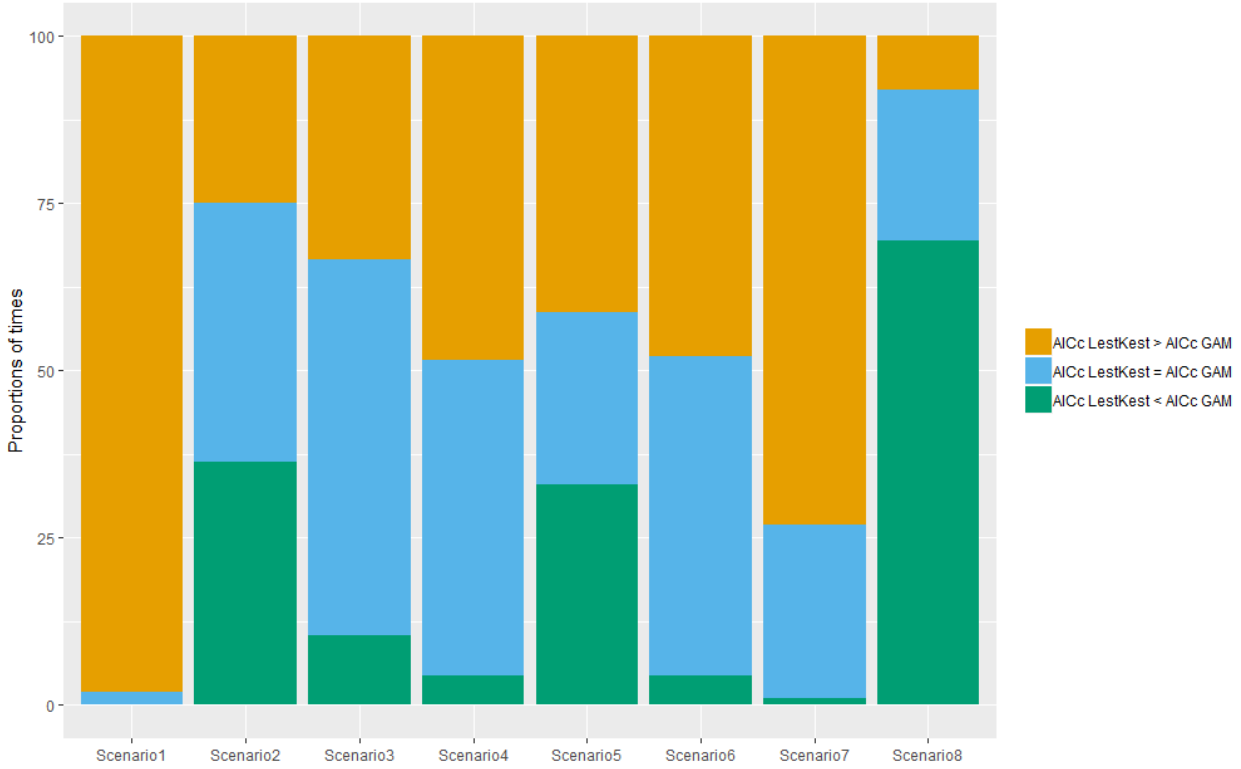


Figure 3.3: Proportion of times, for each simulation scenario, where the LestKest model had less good predictive capacity than GAM (an AICc greater by at least 2.0 points, in orange); where the predictive capacity of the LestKest and GAM models was equivalent (AICc values were within 2.0 points, in blue); and where LestKest had better predictive capacity than the GAM model (an AICc lower by at least 2.0 points, in green).

Logically, differences in AICc between LestKest and L0K1 became more pronounced (and more variable) as the number of observations increased (Figure 3.4 and Annexe III. Figure S2.3). The LestKest model was considerably more efficient compared to the L0K1 model when only one asymptote was variable. When both asymptotes were variable, the LestKest model remained better but approached L0K1, probably because it was difficult to estimate both asymptotes. We also noticed that, in the few cases when the LestKest model was equivalent to, or less good than, the L0K1 model, there were usually few observations. This was not the case when we compared LestKest to GAM: here, the difference in AICc remained stable even as the number of observations increased (Figure 3.5). However, whenever the number of observations were categorized, an increasing number of observations increased the number of cases where LestKest was better than both L0K1 and GAM, whatever the simulation scenario (cf. Annexe III. Figure S2.3 and Annexe III. Figure S2.4).

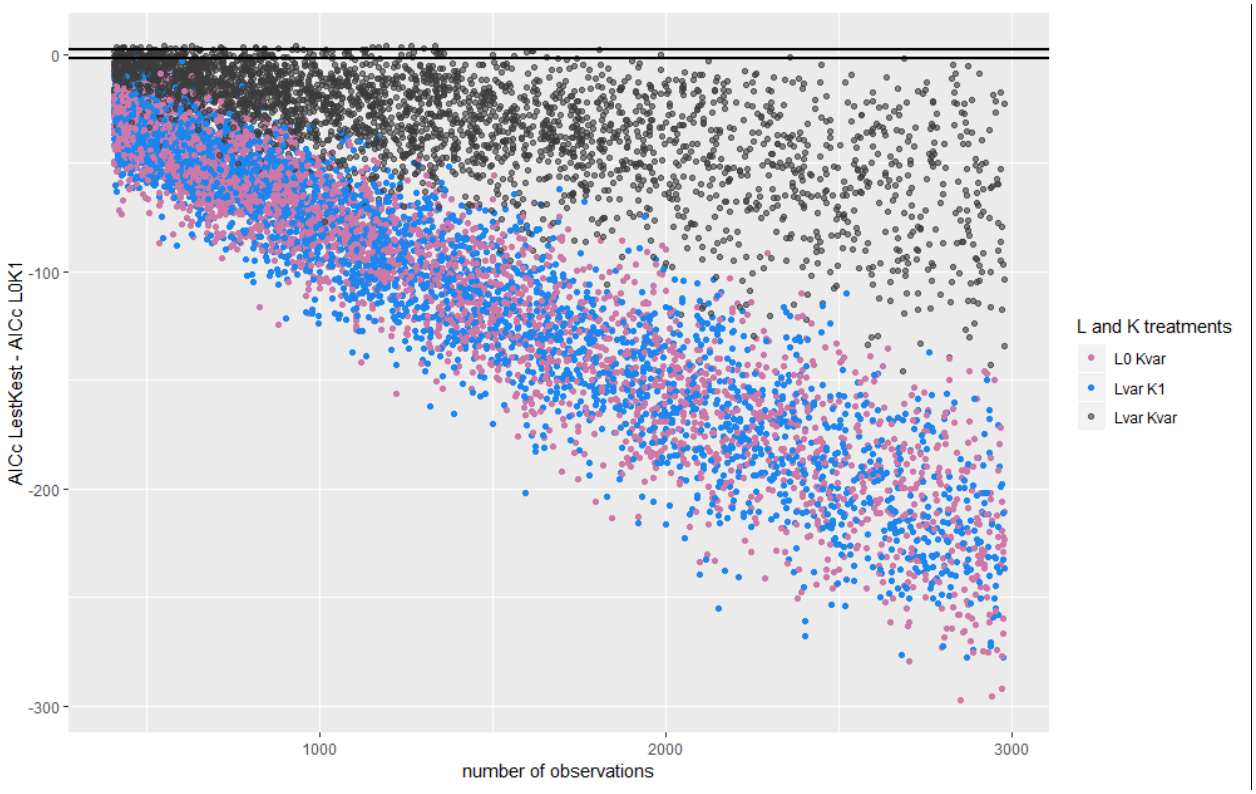


Figure 3.4: AICc difference between the LestKest and L0K1 models as a function of the number of observations in the datasets of Scenario3 (gradient = 0-100) for three cases: L was fixed at 0 and K was variable in data generation (L0 Kvar in pink); L was variable and K was fixed at 1 in data generation (Lvar K1, in blue); and both L and K were variable in data generation (Lvar Kvar, in grey). The points above the lines represent cases where L0K1 was better than LestKest; points between the two lines represent the cases where L0K1 and LestKest were equivalent (a difference in AICc of less than two units); and points below lines represent the cases where LestKest was better than L0K1.

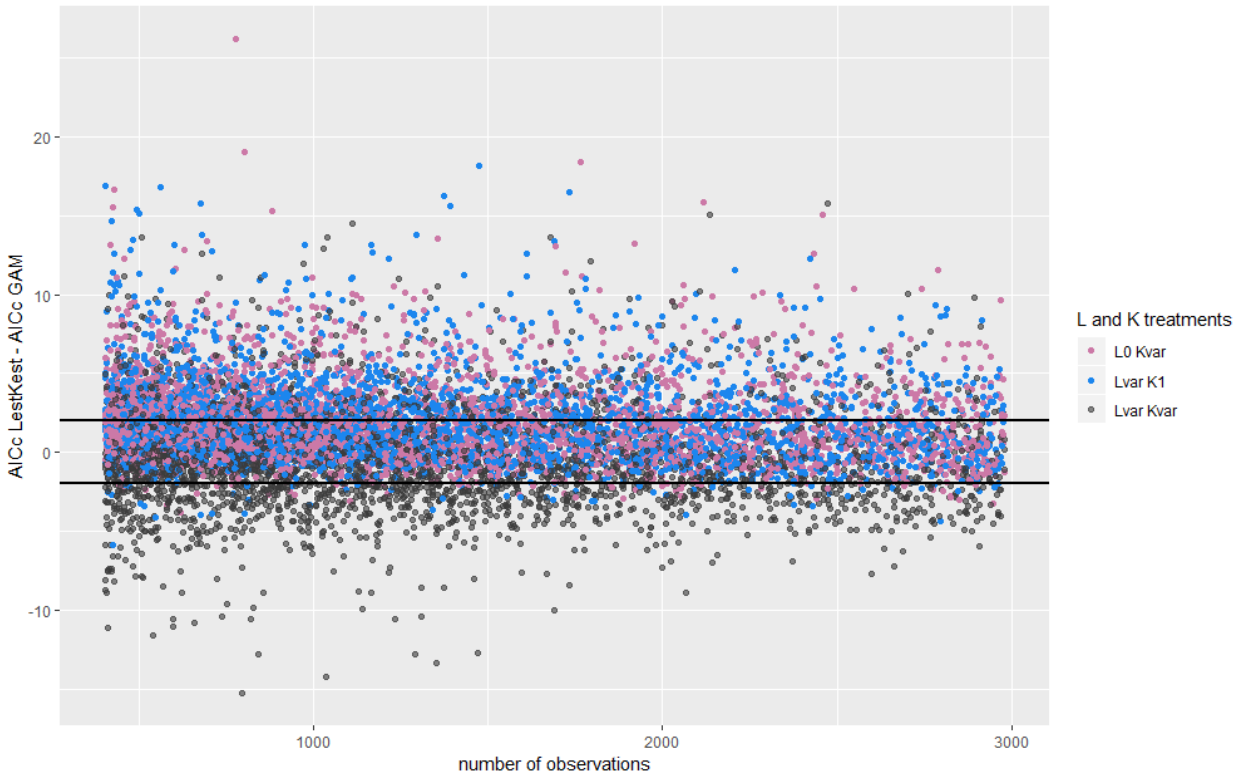


Figure 3.5: AICc difference between LestKest and GAM models as a function of the number of observations in the datasets of Scenario3 (gradient = 0-100) for three cases: L was fixed at 0 and K was variable in data generation (L0 Kvar in pink); L was variable and K was fixed at 1 in data generation (Lvar K1, in blue); and both L and K were variable in data generation (Lvar Kvar, in grey). The points above the lines represent the cases where GAM was better than LestKest; the points between the blue lines represent the cases where GAM and LestKest were equivalent (a difference in AICc of less than two units); and the points below lines represent the cases where LestKest was better than GAM.

II. 3 . a. 2. Comparison of intrinsic quality among models (Type I error and parameter estimations)

When we compared Type I errors with a nominal level of 1% (for tables of results at level 1% and 5%, see Annexe III. Table S2.2 and Annexe III. Table S2.2), significant departures appeared in many cases. Though type I errors at the 1% level for parameters *NSL* and *ip* in models L0Kest and LestK1 were globally equivalent for data covering the whole gradient (Scenario1 to Scenario3 with

exception of Scenario8), but they differed for data with truncated gradients (Figure 3.6). Indeed, L0Kest had fewer type I errors at the 1% level when the data mainly covered the end of the gradient, and LestK1 had fewer type I errors at the 1% level when the data mainly covered the beginning of the gradient. In all cases, the L0K1 model had more Type I errors at the 1% level than all the other models (except for *ip* in Scenario8 and *NSL* in Scenario7 as mentioned above), even when the asymptotes were set at 0 and 1 (Scenario1 – except for *ip* where type I errors were not significantly different from other models at the 5% level: p-Value for L0Kest = 0.3377638, p-Value for LestK1 = 0.8267631, p-Value for LestKest = 0.6852818). In contrast, using the LestKest model drastically reduced Type-I errors at the 1% level for slope and inflexion point in all Scenarios. For example, in Scenario2, for model L0K1, true values for the inflexion point (*ip*) and the slope (*NSL*) were respectively 69% and 99% of the time outside the confidence interval of the estimated values. In contrast, for the LestKest model, true values for *ip* and *NSL* were outside the confidence interval of the estimated values less than 5% of the time. Moreover, LestKest also had fewer Type-I errors at the 1% level for parameters *L* and *K* than did the L0K1 model (and generally, the other models) although this was less true when the asymptotes were not covered by the data (e.g. *K* in Scenario4). The same phenomena were observed for type one errors at the 5% level (Annexe III. Table S2.3). Important inferential problems appeared for parameter *NSL* in all the models where observations were randomly drawn (Scenario8), though these problems were less severe for the LestKest model (Figure 3.6). Finally, though LestKest was, on the whole, much better than the other models in terms of Type-1 errors, it should be noted that Type-1 error levels frequently departed from the nominal rate of 1%, thus indicating persistent inferential problems even with LestKest (see last column of Annexe III. Table S2.2 and Annexe III. Table S2.3).

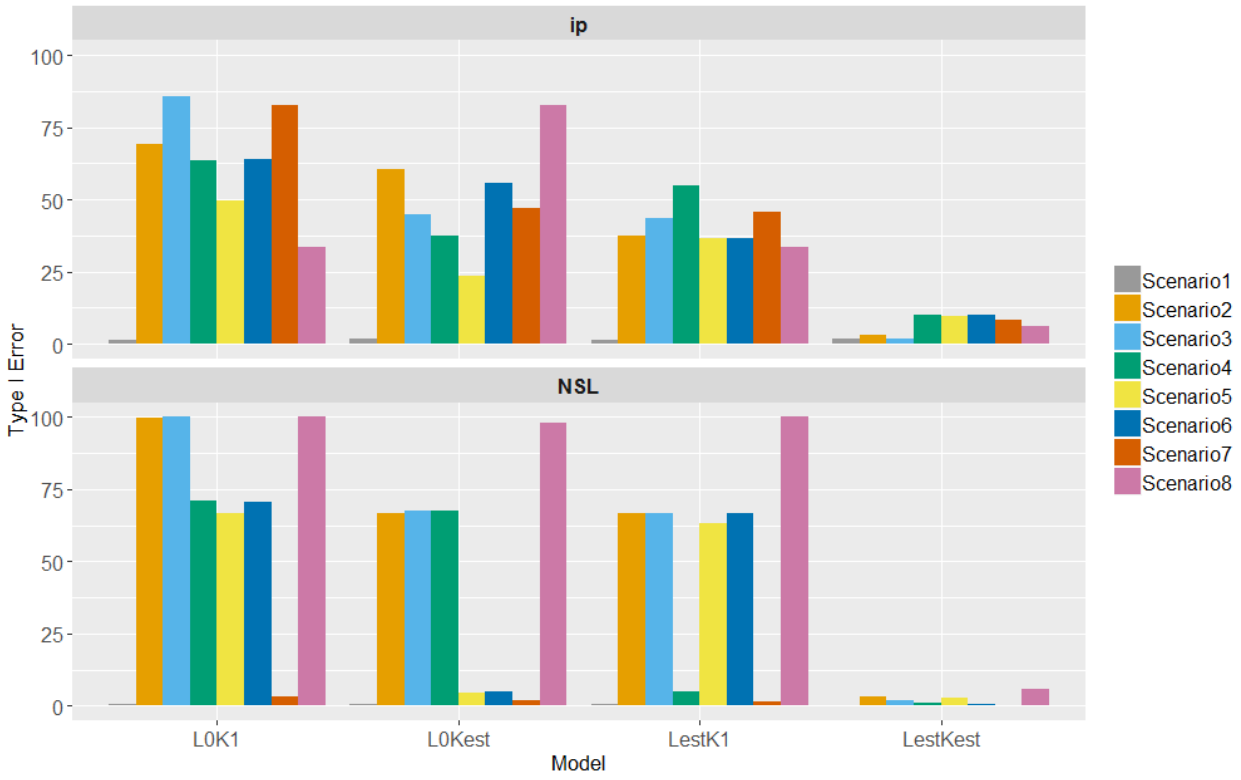


Figure 3.6: Type-I errors (in percentage of times that the true value was outside the confidence interval) at the 1% level for parameters ip and NSL in each univariate simulation scenario and for models L0K1, L0Kest, LestK1 and LestKest. For the significance of the difference at 1% level, see Annexe III. Table S2.3.

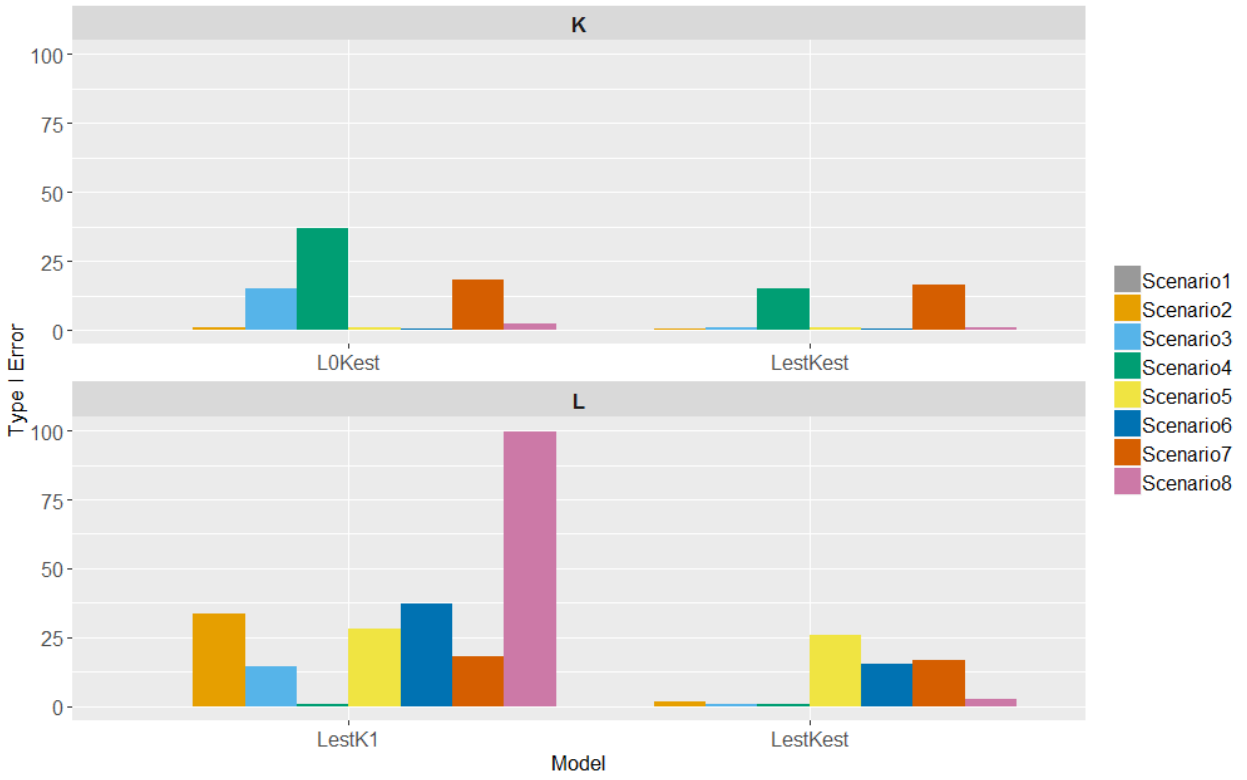


Figure 3.7: Type-I errors (in percentage of times that the true value was outside the confidence interval) at the 1% level for parameters K and L in each univariate simulation scenario and for models LestKest, L0Kest (only for parameter K) and LestK1 (only for parameter L). For the significance of the difference at 1% level, see Annexe III. Table S2.2.

The LestKest (*ip* in Figure 3.8 and *NSL* Figure 3.9) and GAM models (only *NSL*, Figure 3.9) were much better able to estimate parameters than the L0K1 model. Indeed, the optimum values (of *ip* and *NSL*) were closer to the real values of the parameters, and *ip* was estimated more precisely with LestKest than with L0K1 (Figure 3.8). The L0K1 model underestimated slope (represented by normalized slope, *NSL*) in all the simulated datasets except for Scenario1 (with upper and lower asymptotes at 0 and 1 respectively) and Scenario7 (almost linear and with no asymptotes). This underestimation even approached zero in some cases (implying a flat relationship). We also observed much stronger variations in the difference between optimum estimates and true values for the estimated parameters for L0K1 models than for LestKest models. This was especially true

for *ip*: LestKest estimations were very stable, unlike those resulting from L0K1 (Figure 3.8 and Figure 3.9). Although GAM was globally as good as LestKest in estimating the normalized slope, we noticed that for the Scenario8 simulation, GAM severely underestimated *NSL*, while LestKest estimations, were closer to reality, even though variable.

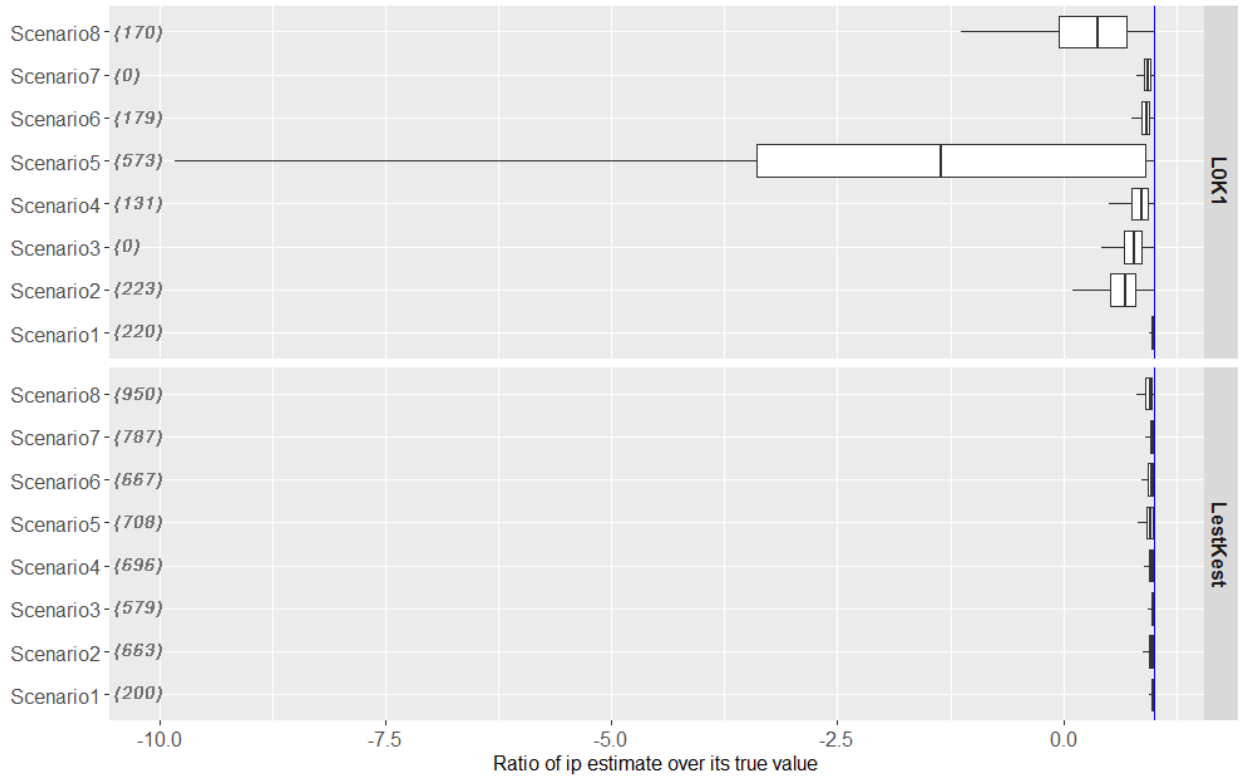


Figure 3.8: Boxplot showing the ratio of the optimum estimates of inflexion point (ip) over its true value for the L0K1 and LestKest models. The blue line represents the target (= 1 when estimates were equal to the true value). Outliers were excluded for ease of reading. The number of excluded outliers is in parenthesis in italics.

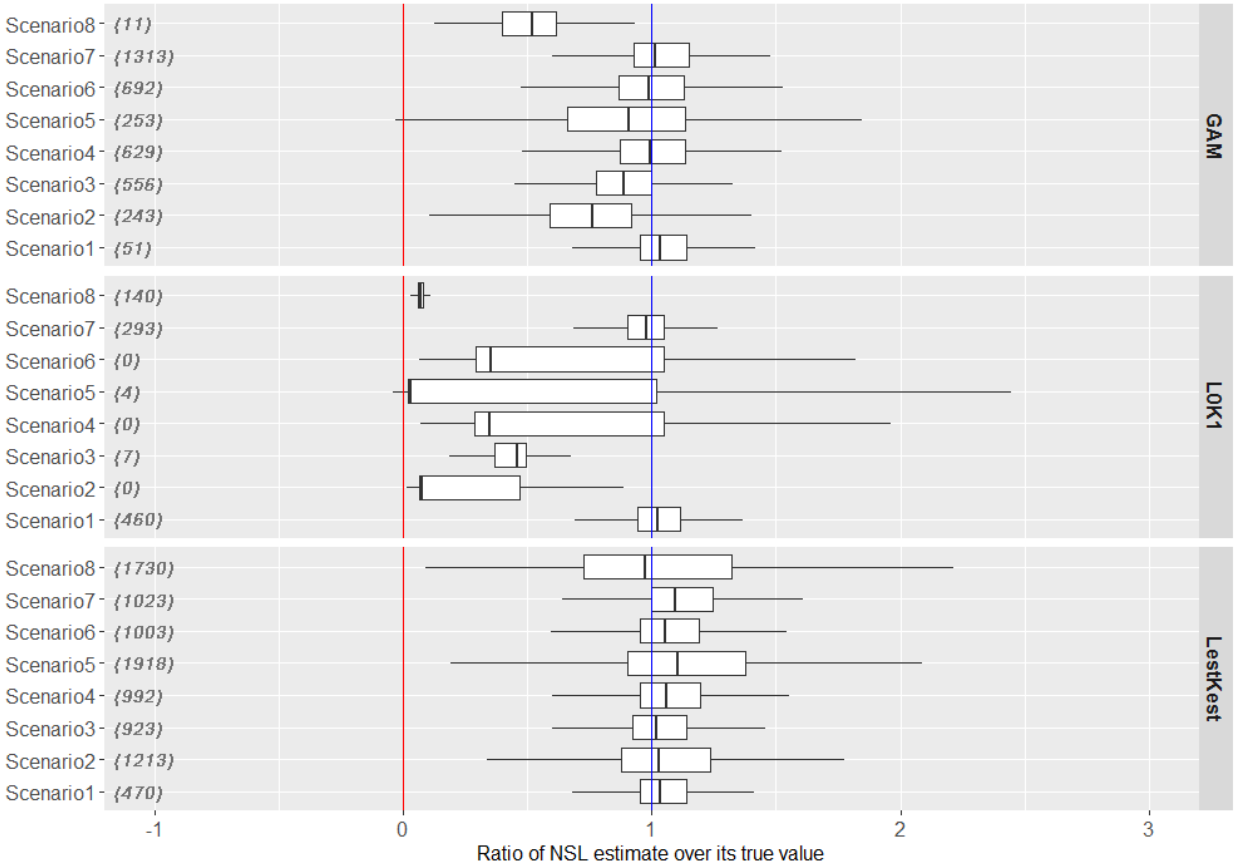


Figure 3.9: Boxplot showing the ratio of optimum estimates of the normalized slope (NSL) over its true value for models L0K1, LestKest and GAM. The blue line represents the target (= 1 when estimates were equal to the true value). The red line represents a flat relationship (normalized slope estimated to be equal to 0). Outliers were excluded for ease of reading. The number of excluded outliers is in parenthesis in italics.

When *K* was set to 1.0 during the simulation and *L* was variable (LvarK1, points in blue inFigure 3.10), the L0K1 model tended to strongly underestimate the normalized slope (*NSL*), and to slightly underestimate the inflection point (*ip*) for all *L* values. Both these biases (on *ip* and *NSL*) decreased slightly with decreasing values of *L* (approaching 0.0). When *L* was set to 0.0 during the simulation and *K* was variable (L0Kvar, points in pink in Figure 3.10), the slope was also underestimated, while the inflection point was slightly overestimated. Both these biases also

decreased slightly with increasing values of *K* (approaching 1.0). When *L* and *K* both varied during the simulation (LvarKvar, in grey in Figure 3.10), the slope was even more strongly underestimated, though the inflection point was nearly accurate with only a slight overestimation (when *L* was close to 0.0) or a slight underestimation (when *K* was close to 1.0). Conversely, with the LestKest model, the inflection point was very accurately (all values of ratio close to the expected value of one) and precisely estimated (not very variable) whatever the asymptote(s) variable during the simulation (grey, pink and blue points mingled), and whatever their value (constant over the *X*-axis). On the other hand, the normalized slope estimation was much more erratic because, although the estimate generally did not seem to be under or overestimated, it was very variable (very variable values of ratio in Figure 3.9 and Figure 3.10), especially when *L* and *K* were both variable (grey points). Again the variability of the estimation of *NSL* did not depend on the value of variable asymptote(s) (also constant on the *X*-axis of Figure 3.10).

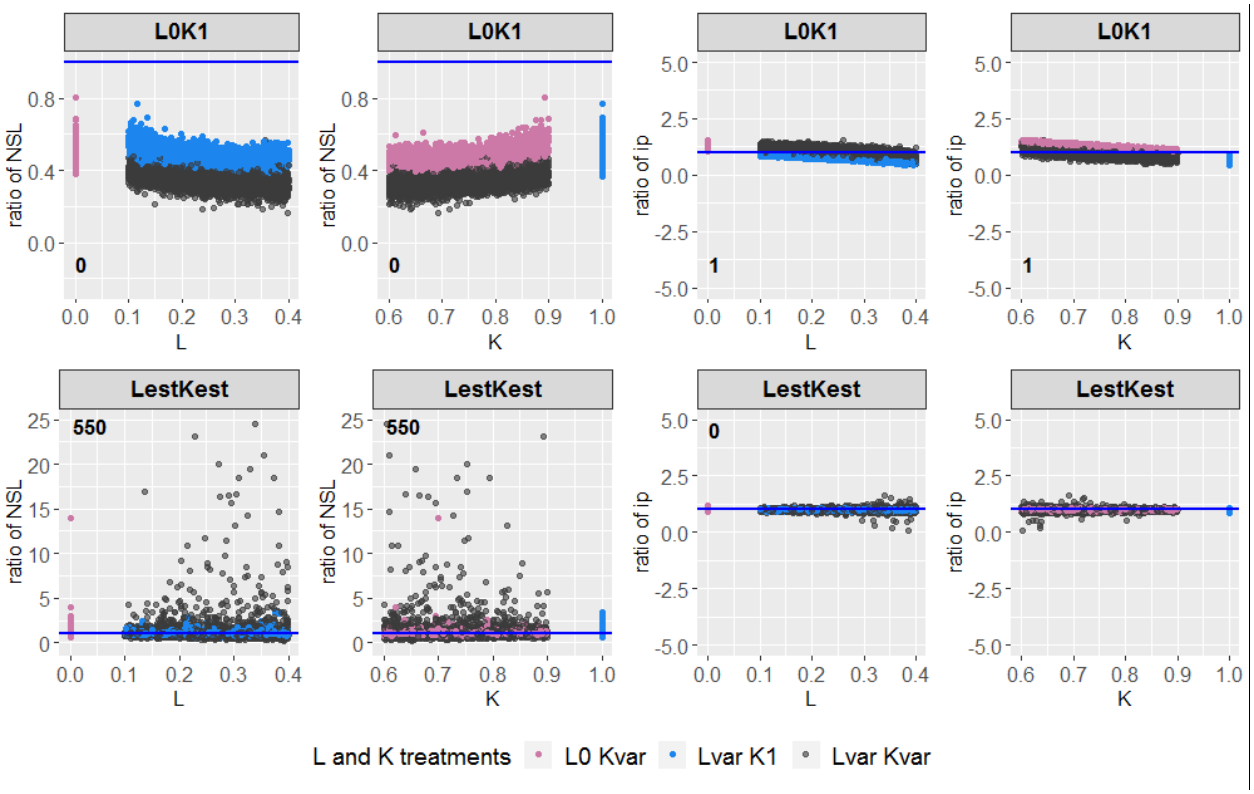


Figure 3.10: Ratio of the optimum estimates over their true values for normalized slope (NSL) and inflexion point (ip) as a function of the true values of L and K for Scenario3 (gradient = 0-100) and for models L0K1 and LestKest. Windows of four plots have been reduced to zoom in and exclude extreme values for ease of reading. The number of excluded values is specified in bold (either at the top left of the window, or the bottom left). The blue line represents the target (= 1 when estimates were equal to the true value).

The misestimation of parameters with model L0K1 induced a poor estimated curve shape compared to LestKest, where predicted curve shape was close to the true shape (Figure 3.11). The curve shape produced by GAM also fitted the true shape well, except for confidence interval of the two asymptotes that were larger . Also, the upper asymptote was not predicted as reached, as it should have been.

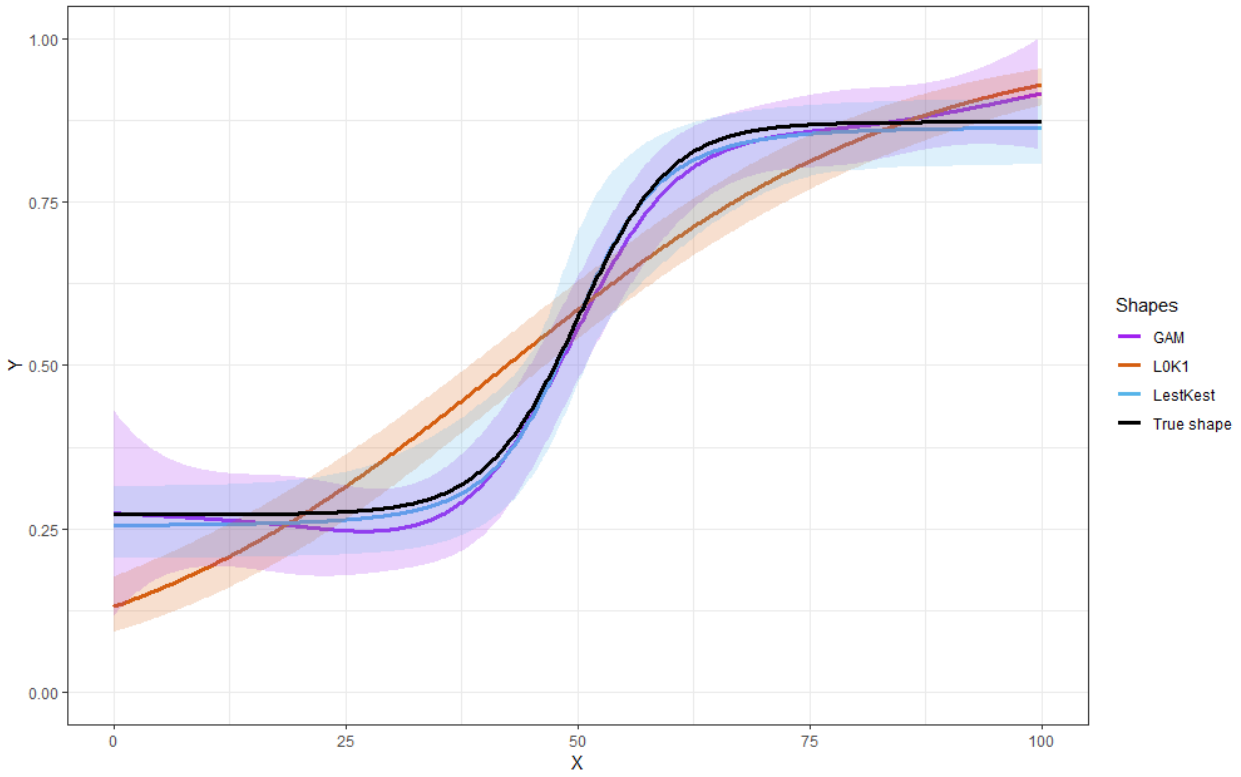


Figure 3.11: Curve shapes predicted by models L0K1, LestKest and GAM for the first dataset in Scenario3 compared to the true shape that generated the data. The confidence interval of the curve shapes are between quantiles 0.025 and 0.975.

II. 3 . a. 3. Comparison of intrinsic quality among models (plug-in p-values)

We detected no problem for any discrepancy measure and any model under Scenario1 (cf. Annexe III. Table S2.4). Scenario7 was also nearly descrepancy-free, except for the lower part of the curve where departures were detected more frequently than expected for all the models. We also detected issues with the link function of GAM in this Scenario7 (and some slight issues for LestKest as well). Model L0K1 had link function problems in all scenarios other than Scenarios 1 and 7 – either on the whole gradient or in specific places on the gradient. Models L0Kest and LestK1 had significant departures from expected values in these scenarios as well, though the departures were less strong, or even insignificant, in the scenarios where the two models had the

best fit (Scenarios2, 5, 6 and 8 for L0Kest; Scenario4 for LestK1; see Annexe III. Figure S2.1 and Annexe III. Figure S2.2). Regarding the LestKest and GAM models, some significant departures from expected values were also detected in the other scenarios relative to the link function but they were of less magnitude than for the other models.

II. 3 . b. Comparison between models on multivariate simulated datasets

II. 3 . b. 1. Comparison of the relative quality among models (AICc)

In multivariate cases, GAM always performed as well as or better than model L0K1 (never performed less well – Figure 3.12, Model1=GAM; Model2=L0K1). In the vast majority of cases, the LestKest model performed better than L0K1 (Figure 3.12, Model1=LestKest; Model2=L0K1). Finally, contrary to univariate datasets where GAM usually performed as well as or better than LestKest, in multivariate cases, the LestKest model performed considerably better than GAM (Figure 3.12, Model1=LestKest; Model2=GAM).

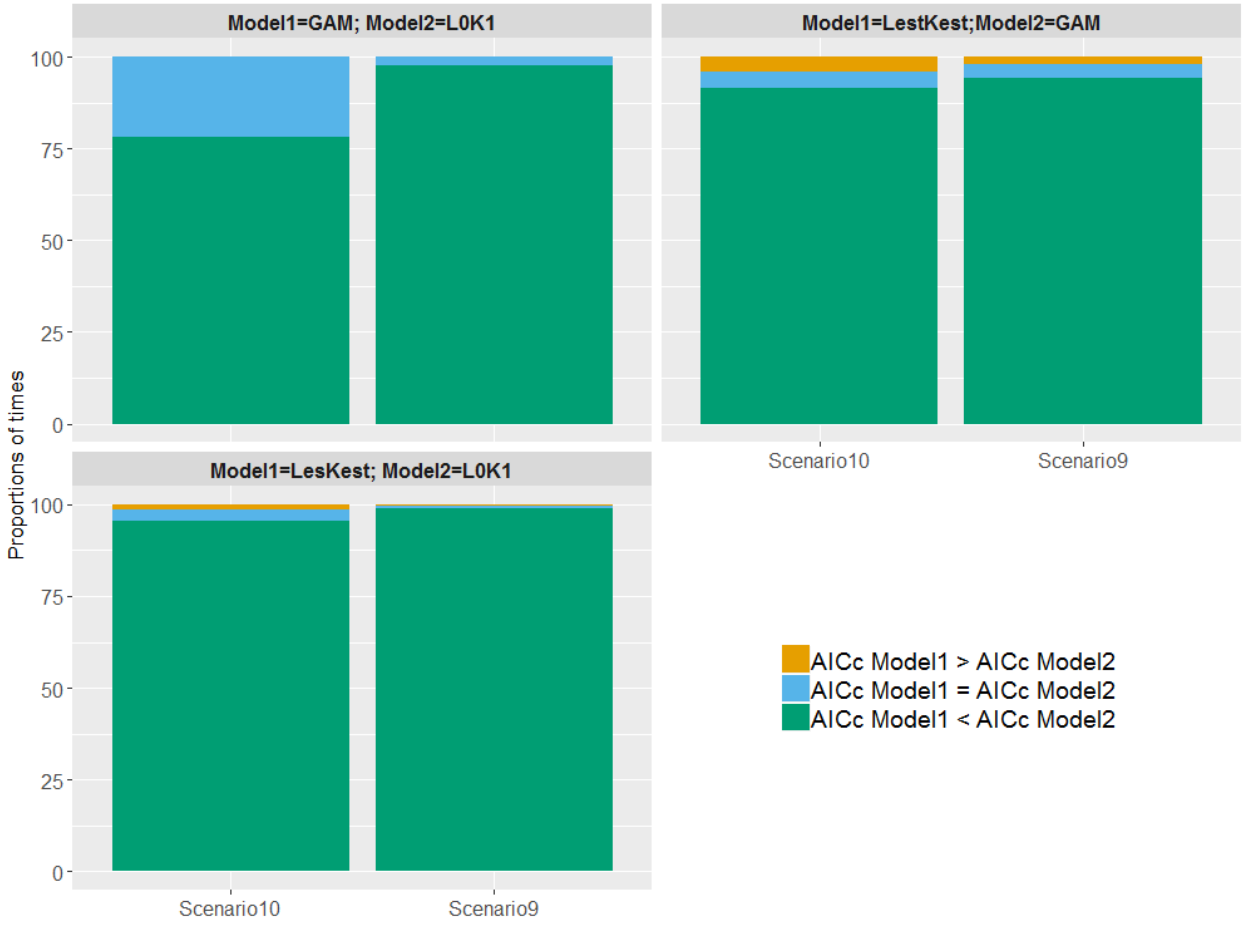


Figure 3.12: Proportion of times, for each simulation scenario where: i) model Model1 had less good predictive capacity than Model2 (an AICc greater by at least 2.0 points, in orange); ii) the predictive capacity of Model1 and Model2 were equivalent (AICc values within 2.0 points, in blue); and iii) Model1 had better predictive capacity than model Model2 (an AICc lower by at least 2.0 points, in green).

As in the univariate cases, for multivariate datasets, the difference in AICc between the LestKest and L0K1 models became more and more important as the number of observations increased (example for Scenario9 in Annexe III. Figure S2.5). On the other hand, unlike the univariate cases, this increase in AICc difference with the number of observations also occurred between the LestKest and GAM models, to the advantage of the LestKest model (example for Scenario9, Figure 3.13). Once again, generally speaking, an increasing number of observations, categorized and

uncategorized, always (for both multivariate Scenario series) induced an increasing proportion of cases where LestKest was better than the L0K1 and GAM models (Figure 3.13, Annexe III. Figure S2.5 and Annexe III. Figure S2.6).

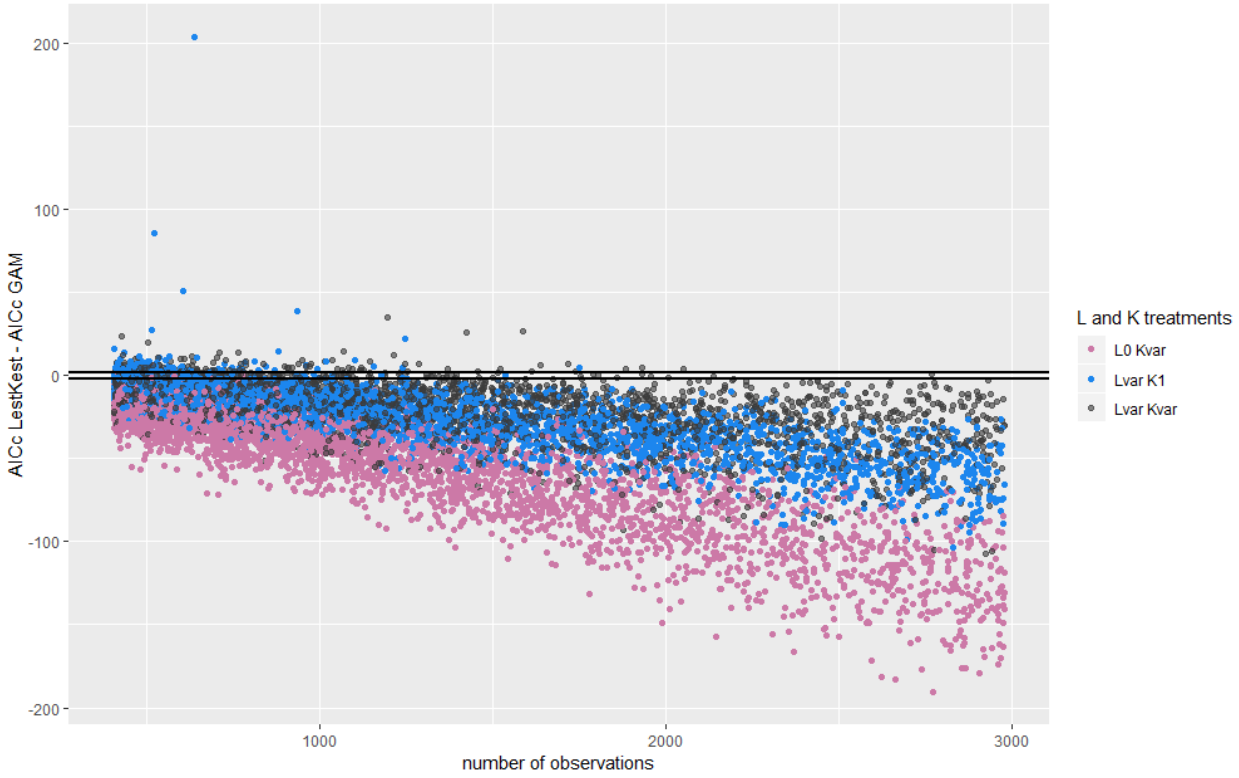


Figure 3.13: AICc difference between models LestKest and GAM as a function of the number of observations in the datasets for Scenario9 in three cases: L was fixed at 0 and K was variable during data generation (L0 Kvar in pink); L was variable and K was fixed at 1 during data generation (Lvar K1, in yellow); and L and K both varied during data generation (Lvar Kvar, in grey). Points above the lines represent the cases where GAM was better than LestKest, points between the two lines represent the cases where GAM and LestKest were equivalent (a difference in AICc of less than two units) and points below lines represent the cases where LestKest was better than GAM.

II. 3 . b. 2. Comparison of intrinsic quality among models (parameter estimations)

As observed for univariate datasets, LestKest produced fewer Type-I errors for inflexion point (*ip*) estimates at the nominal 1% level (Figure 3.14 and Annexe III. Table S2.5) and the 5% level

(Annexe III. Table S2.6) than did model L0K1 for both simulation scenarios. Around 70% of time, true *ip* values fell outside the confidence interval of estimated values with L0K1 (Type-I error at the 1% level Figure 3.14) for both multivariate datasets. Meanwhile, with the LestKest model, true *ip* values were outside the confidence interval of estimated values only around 3% of the time (Type-I error at the 1% level, Figure 3.14). The percentage of times that true values for both normalized slopes ($NSLX_1$ and $NSLX_2$) were outside the confidence interval of their estimated values was also very low for the LestKest model (less than 4%) compared to the L0K1 model (more than 75%). Finally, for the *K* and *L* parameters, the percentage of times that their true values fell outside the confidence interval of the values estimated with the LestKest model was also low (at the 1% level, Type-I errors occurred less than 2.5% of the time).

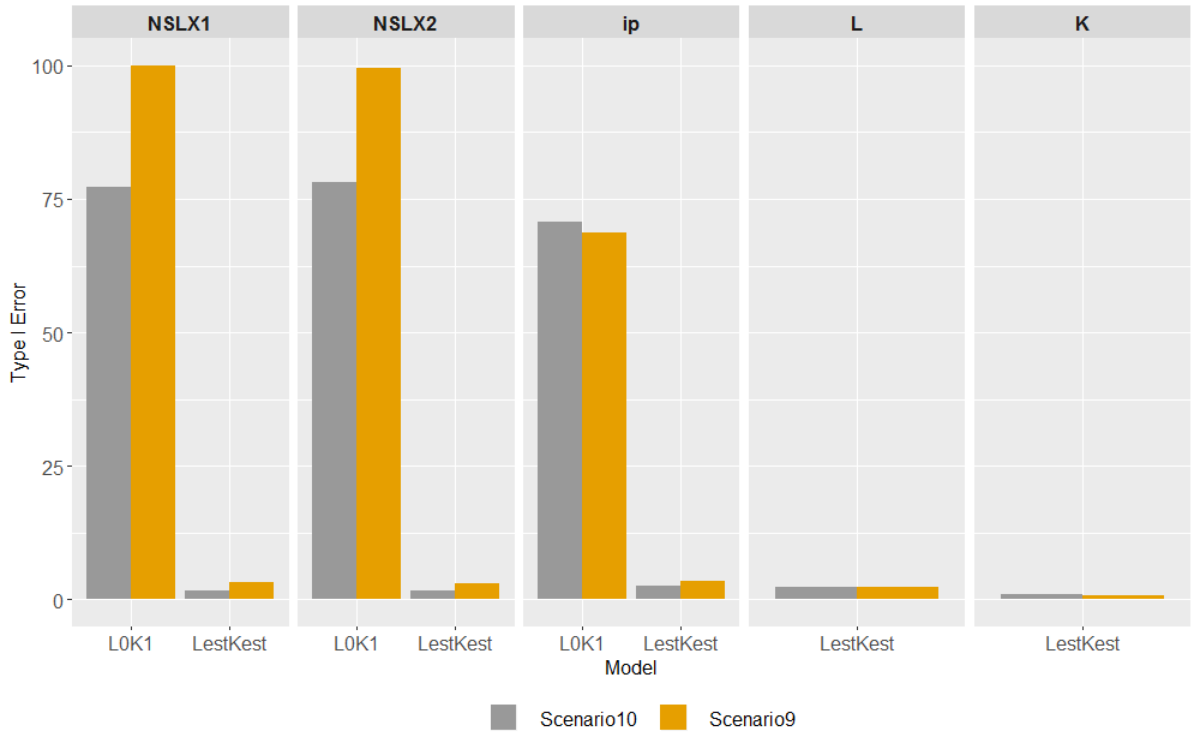


Figure 3.14: Type-I errors at the 1% level for parameters NSLX1, NSLX2, ip, L and K in both multivariate simulation scenarios for LestKest and L0K1(excluding parameters L and K).The significance of the difference at 1% is indicated in Annexe III. Table S2.5.

With LestKest, the parameters of the multivariate datasets ($NSLX_1$, $NSLX_2$ and *ip*) were well estimated, especially the inflexion point, which was both accurate and precise (Annexe III. Figure S2.7), contrary to the estimations with L0K1 where $NSLX_1$ and $NSLX_2$ were underestimated and *ip* varied. GAM also tended to underestimate normalized slope parameters, though less severely than L0K1 (Figure 3.15).

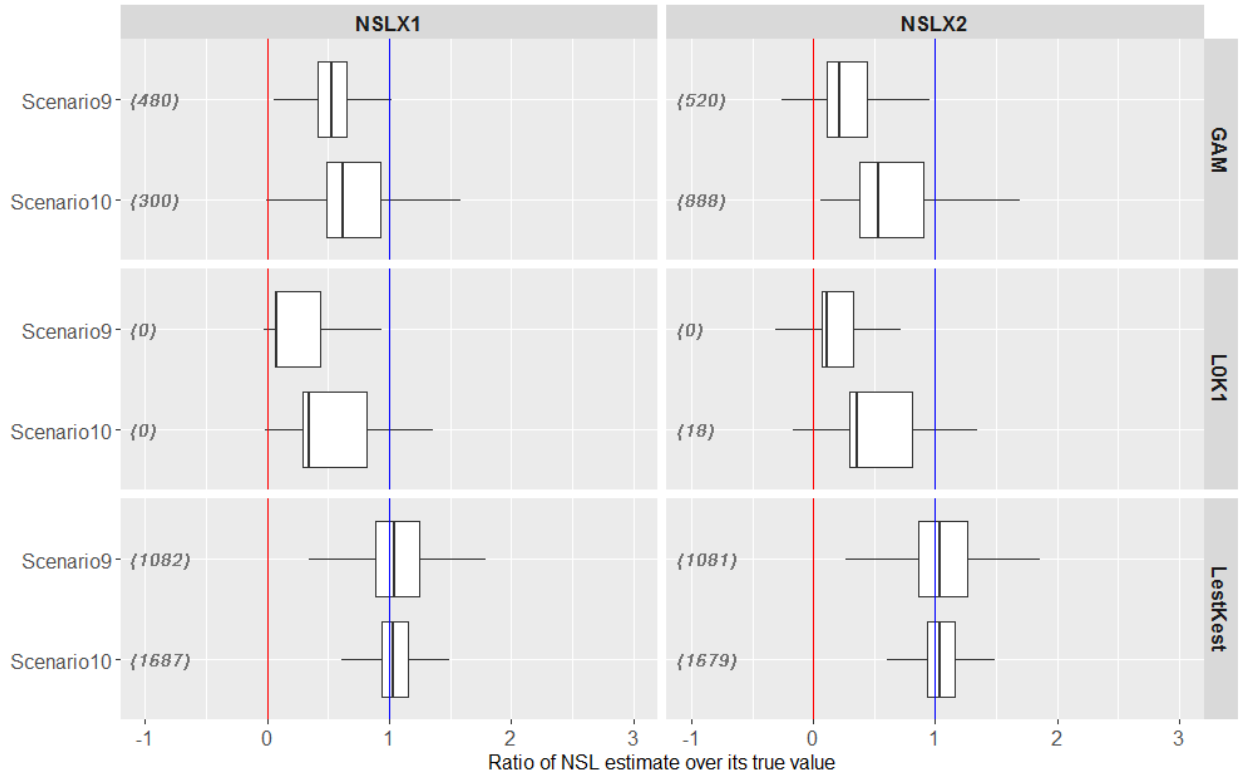


Figure 3.15: Boxplot showing the ratio of the optimum estimates of normalized slopes (NSLX1 and NSLX2) over their true values for models L0K1, LestKest and GAM. The blue line represents the target (= 1 when estimates were equal to the true value). The red line represents the case where the associated X gradient had no effect on the Y response (normalized slope estimated as zero). Outliers were excluded for ease of reading. The number of excluded values is given in parenthesis in italics.

In most cases, curve shapes produced by model LestKest were very close to the true shape of the data (Annexe III. Figure S2.8). Nevertheless, confidence intervals were often wide due to extreme

parameter estimates (for example Annexe III. Figure S2.8, Scenario10.9, see also the outliers in estimate boxplots Annexe III. Figure S2.7). The curve shapes produced by GAM were either very close to the true shape (even if less close than for LestKest - Annexe III. Figure S2.8, Scenario9.1 and Scenario9.7) or, on the contrary, very close to the shape estimated by L0K1), which, was always quite different from the true shape (similar to the results found for univariates) (Annexe III. Figure S2.8, Scenario10.1 and Scenario10.9).

II. 3 . b. 3. Comparison of intrinsic quality among models (plug-in p-values)

For model L0K1, strong departures from expected frequencies were detected in both Scenario9 and Scenario10, with discrepancies mostly targeting the link function (Annexe III. Table S2.7). For GAMs, similarly strong departures were only detected in Scenario 10, while LestKest remained free of discrepancy, both for Scenario 9 and 10.

II. 3 . c. Comparison among models on univariate and multivariate real datasets

When applied to various binary datasets (Annexe III. Table S2.8), models L0K1 and LestKest had balanced AICc results in univariate cases. Indeed, 35% of the time (18/52) L0K1 was better than LestKest, 29% of the time (15/52) they were equivalent (with a difference of less than 2.0 points), and 37% of the time (19/52) LestKest was better than L0K1. In multivariate cases (Annexe III. Table S2.7) AICc for L0K1 was better than LestKest in 36% of cases (14/39), but there were fewer cases where L0K1 and LestKest were equivalent (8/39 – 21%) and more cases where LestKest was better (17/39 – 44%). When comparing GAM-monotone AICc results to LestKest AICc results, in

univariate cases, GAM-monotone was better 65% of the time (34/52), the two models were equivalent 27% of time (14/52) and LestKest was better than GAM-monotone 8% of the time (4/52). The contrast was less severe for multivariate cases: GAM-monotone was better 51% of the time (20/39), LestKest and GAM-monotone were equivalent 33% of the time (13/39) , and LestKest was better than GAM-monotone 15% of the time (6/39). In both multivariate and univariate cases, in the vast majority of cases, at least one of the two asymptotes estimated by the LestKest model (Annexe III. Table S2.8 and Annexe III. Table S2.9) was different from the fixed values of 0.0 and 1.0 (for lower and upper asymptotes respectively). This difference occurred in more than 95% of the univariate cases and more than 80% of the multivariate cases, even when we take the standard deviation into consideration. The best variables selected were not always the same among models (Annexe III. Table S2.8 and Annexe III. Table S2.9), but there did not seem to be any specific pattern: LestKest was sometimes better than the other models (i.e univariate RealData12), and sometimes less good (i.e. univariate RealData19).

II. 4. Discussion

II. 4. a. A promising new model in the toolbox

In simulated univariate cases, as expected, all the models provided identical estimates and performed equally well in Scenario1 where parameters *L* and *K* were fixed at 0.0 and 1.0 respectively. We found that the conventional model with no estimation of the asymptotes (L0K1) did not fit well data with *L* and/or *K* values other than 0.0 and 1.0. Indeed, the normalized slope estimates for L0K1 were strongly biased (underestimated) and inflexion point estimates suffered

from a slight underestimation and a bimodal tendency (Annexe III. Figure S3.1). Slope may have been underestimated because, had its true value been respected, the model would have imposed an upper asymptote fixed at 1.0 on data without consistent success observations even when the upper asymptote is reached. The same reasoning holds for the lower asymptote. Consequently, the model considered that the fixed upper asymptote (1.0) had not been reached and reduced the slope to fit this profile. These are serious problems because slope is a very interesting parameter for investigating magnitudes of the effect of the predictor on the response variable. If misestimated, erroneous conclusions can be drawn with severe consequences (e.g. misevaluation of the effectiveness of a treatment). GAM showed better performance in terms of AICc and estimations of the normalized slope compared to L0K1, except in the case of datasets where the gradient was not regularly distributed (as in Scenario8). Models with estimated *L* and/or *K* (L0Kest, LestK1 and LestKest) provided consistently better intrinsic and relative performances compared to L0K1, with an even greater gap for LestKest, the best performer. The gap in terms of predictive capacity between L0K1 and LestKest widened as the number of observations increased. In general, a small number of observations was detrimental to GLM performance for binary data (Annexe III. S3). A sufficient number of observations is always necessary, especially for the LestKest model, which estimates a larger number of parameters. When the gradient was shifted to the right (with more data on the part of the gradient that reached the upper asymptote) or not complete (with more of the end of the gradient), it was better to estimate the upper asymptote K, and even better both K and L, as opposed to retaining fixed values. Conversely, when the gradient was shifted to the left (with more data on the part of the gradient that reached the lower asymptote), it was better to estimate the lower asymptote L, and even better both K and L, as

opposed to retaining fixed values. Finally, in datasets where the whole gradient was represented and centered, though L0Kest and LestK1 were equivalent to L0K1, and LestKest (that estimates both asymptotes) had a much better fit. Since we usually do not know in advance which part of the gradient our observations will cover, it appears preferable to estimate both asymptotes to ensure better results (only one is not enough). Indeed, LestKest was the only model that was likely to estimate the parameters correctly for all gradient cases. LestKest also handled data with lower and upper asymptotes at zero and one as well as data with different lower- and upper-asymptote values. That is reassuring because, in real-world studies, we do not necessarily have a priori knowledge of the asymptotic values. For all the criteria studied (AICc, accuracy of the estimator, Type-I errors, plug-in p-values), we found that GAM was always at least as good as the LestKest model for univariate scenarios, excepted for datasets where the gradient was not regularly distributed (as in Scenario8, where the LestKest model had better results).

For multivariate cases, we chose to compare only models L0K1 (corresponding to GLM), GAM and LestKest (which actually include L0Kest and LestK1). In simulated multivariate cases, with at least one asymptote different from either 0.0 or 1.0, the L0K1 performed poorly in terms of AICc (compared to the other two models) and in terms of Type-1 error and estimation of parameters. GAM showed better performance than L0K1 in terms of AICc, but remained far behind LestKest, which again revealed a much better predictive performance. Furthermore, LestKest was much more efficient in estimating the parameters of the curve than GAM (slope) and L0K1 (slope and inflection point). As in the univariate cases, LestKest suffered much less from error than did L0K1 when estimating the inflexion point and both normalized slopes ($NSLX_1$ and $NSLX_2$), and was much

more accurate in estimating the parameters of the curve than GAM (slope) and L0K1 (slope and inflection point).

II. 4. b. Limitations and prospects

In spite of its good predictive capacity as demonstrated in these preliminary analyses, the LestKest model revealed two main limitations. First, the optimization procedure was long and tedious, and required many adjustments to obtain one good optimization (i.e. using a variety of optimization methods, various sets of initial values, repeating the process until a finite value of likelihood function was obtained – see codes in Data III.S1). Second, extreme parameter values and abnormal variance-covariance matrices were occasionally obtained, leading to poorly estimated parameters, making it impossible to calculate RMSE, and producing many outliers. Third, though Type-I error departures from nominal rates appeared with the LestKest model, these could be linked to numerical issues, and might be solved with a Bayesian approach (Saas and Gosselin 2014). The LestKest model is therefore an interesting concept, but which still requires further development, especially to improve optimization. Although using the LestKest model in a Bayesian context was not possible in this study (due to the amount of datasets analyzed), we believe that a Bayesian approach could potentially improve the model's estimates and reduce or even eliminate extreme values.

LestKest model did not perform as well on real datasets as on simulated data. Indeed, the GAM-monotone and L0K1 models competed well with LestKest, even in multivariate cases. Nevertheless, it is important to acknowledge that we chose our real datasets randomly, without any upstream reflection on the variables used (both explained and explanatory) or on the

expected form of the relationship. We sought to show that, even though rare, it is possible to find relationships with asymptotes that different from 0.0 and 1.0. The fact that we found such cases, without any a priori selection, proves that the LestKest model can be interesting. It therefore seems very likely that the model would be of major interest when applied to datasets for which we have reason to believe that the relation did not reach one of the two asymptotes, For example, the probability of voting Democratic in the US as a function of one's income, mentioned in the introduction, would be a very interesting case for our LestKest model, and is one that we would like to investigate in the future.

LestKest is also a good candidate for a 0/1 data fit, especially in a monotonic context. It is all the more interesting (especially compared to GAM) when studying multivariate predictors. Indeed, while GAM is quite flexible in a univariate approach, it creates misestimation in a multivariate approach, whereas LestKest is adapted to both. The LestKest model should therefore be added to the toolbox for univariate analyses, along with L0K1 and GAM. LestKest is a very credible alternative to L0K1 for monotonic contexts, and it is not sure that the asymptotes are 0.0 and 1.0.

In our case study, applying GAM to the data was not a goal in itself (since we sought to use a truly nonlinear relationship with accessible parameters) but was rather to provide a reference model for comparing results and improving the features of the model we developed. Using GAM as a comparison tool seems to be an interesting strategy in general when developing nonlinear models. We also learned that GAMs tend to perform inaccurately when the additive part is supported by an inadequate link function: when either there is more than one explanatory variable (multivariate case) or the explanatory variable is distributed randomly over the gradient (Scenario 8). These are new perspectives on GAM to our knowledge.

II. 5. Conclusion

Our analyses suggest that there are promising new ways to study binary data via more extensive non-linear regression models rather than simple GLM-type models. Our approach was novel in that we estimated both the lower and upper asymptotes of the logistic function (called the ELUA logistic function, for “Estimated Lower & Upper Asymptotes”). The results look very promising for simulated monotonic data, though the model still requires some adjustments. In particular, our results show that the classical binomial GLM suffers from a strong underestimation of slope when the true asymptotes are not 0.0 and 1.0, with important implications for application. They also show that binomial GAMs behave better than GLMs, except when the explanatory variable is irregularly spaced or when there is more than one explanatory variable. The new approach we propose does not have these problems. Though our results for uncontrolled real data were more mitigated, the LestKest model remains interesting if there is upstream reasoning on variables and estimator outputs. The LestKest model developed in this paper, with its associated ELUA logistic function, does not solve all potential problems, but it does constitute an additional tool in the modeler’s toolbox, which, when applied in the appropriate cases, enhances modeling accuracy. Our LestKest model provides a better predictive capacity (AICc) and a better inference (estimates and errors), especially for slope. The model also brings up a new entity on which scientists will be able to study: the asymptotes of logistics in logistic regression models on binary data.. Indeed, given the central role of the study of binary data in all fields of research, we are convinced that developing new functions is a major way forward. These new functions should integrate estimated asymptotes, as we did in this paper, or account for asymmetry.

Komori et al. (2016) also improved analysis of binary data with promising results by implementing an asymmetric logistic regression model to account for data complexity. To take into account all the intricacies of complex datasets, it would be interesting to combine both improvements and so, to implement asymmetric estimated-asymptotes logistic regression models (with functions such as the 5-parameter logistic model in Godeau et al. n.d.).

III. DISCUSSION DU CHAPITRE 3

Lors de cette étude, nous avons proposé une nouvelle approche prometteuse qui vient compléter la palette d'outils disponibles pour étudier les données binaires. Notre nouvelle approche consiste à estimer l'asymptote inférieure et supérieure de la fonction logistique, réciproque de la fonction de lien logit. Les résultats semblent très prometteurs sur des données monotones simulées, puisque le modèle permet, contrairement au GLM binomial, de (i) bien estimer (en termes de biais et d'erreur de Type 1) la pente de la relation, l'ordonnée à l'origine et la position du point d'inflexion ; (ii) d'avoir un modèle plus cohérent avec les données (adéquation) et (iii) d'améliorer la capacité prédictive du modèle. Le modèle développé permet également une amélioration par rapport aux modèles de type GAM dans le cas où au moins deux variables explicatives sont utilisées dans le modèle ainsi que dans le cas d'une seule variable explicative non uniformément distribuée (ce qui est souvent le cas en écologie).

Ainsi, avec ce nouveau modèle, l'idée était de mettre en évidence l'existence d'une hypothèse auxiliaire forte, ainsi que son impact sur l'estimation et la qualité prédictive du modèle. De plus, ce nouveau modèle permet de proposer une alternative à travers une famille d'hypothèses auxiliaires plus souples et générales (asymptotes estimées à partir des données), qui puisse par ailleurs porter des hypothèses principales sur les raisons sous-jacentes pouvant expliquer la position des asymptotes estimées. Concernant le cas d'étude « Red state-Blue State » de A. Gelman, l'asymptote haute serait différente de un car on prévoit que malgré des revenus moyens très élevés dans l'État, des disparités entre les électeurs subsistent au niveau du choix du vote. Le nouveau modèle donne ainsi naissance à une nouvelle quantité d'intérêt scientifique,

l'asymptote haute, dont on pourrait par exemple chercher à expliquer les variations d'un État à l'autre des États-Unis et cette variabilité pourrait être expliquée par d'autres variables (e.g. la religion, l'origine ethnique…).

Outre l'hypothèse auxiliaire forte sur la position des asymptotes, la fonction de lien canonique logistique commune fait également une hypothèse auxiliaire quant à la symétrie de la relation. En effet, la relation est supposée symétrique autour du point d'inflexion. Or, il n'y a pas d'argument a priori pour forcer une telle limitation. Certains cas symétriques doivent certes se rencontrer dans les données, mais des cas où la relation est asymétrique semblent tout aussi concevables. En ce sens, Komori et al. (2016) ont développé un modèle GLM autorisant une asymétrie estimée (cf. également Godeau et al., n.d., sur des données de nature différente). Si l'on observe d'ailleurs le cas d'étude Red state-Blue State, une relation asymétrique (dont la pente avant le point d'inflexion est plus importante qu'après le point d'inflexion), mais toujours avec les asymptotes estimées, semble encore plus cohérente avec les données (Figure 3.16). Dans de futures recherches, il serait donc intéressant de se pencher sur le développement de modèles alliant estimation des asymptotes et asymétrie afin d'obtenir un ajustement précis des données et de pouvoir tirer des conclusions plus justes.

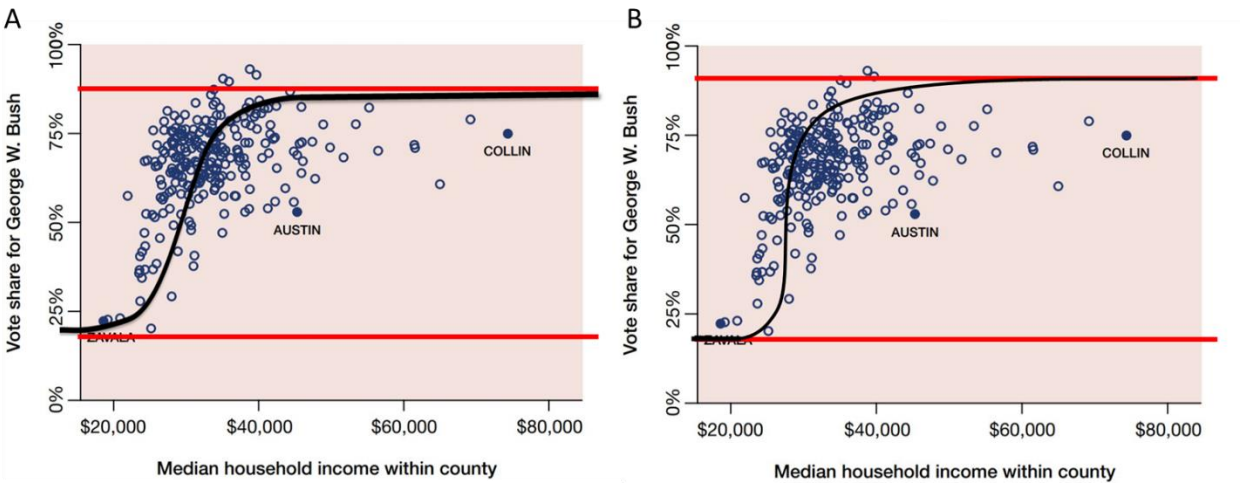


Figure 3.16: Proportion des votes pour George W. Bush (candidat Républicain) en fonction du revenu médian des électeurs dans les différents comtés de l'État du Texas en 2000. Figure modifiée d'après (Gelman 2011). Les lignes rouges représentent les asymptotes potentielles observées graphiquement. Les courbes noires représentent schématiquement les courbes issues de la fonction de lien que pourrait potentiellement estimer le modèle, dans un cas où une fonction logistique avec asymptotes estimées est appliquée (A) et dans un cas où est appliquée une fonction logistique dont les asymptotes et l'asymétrie sont estimées (B).

En écologie, lorsque l'on étudie la distribution des espèces (SDM), les données binaires analysées représentent souvent des données de présence/absence de ces espèces. Afin d'établir la distribution (spatiale ou selon un gradient écologique) de l'espèce étudiée, il s'agit de mettre en lien ces données de présence/absence avec des données environnementales. Dans cette optique, les GLMs binomiaux permettent de faire le lien (linéaire) entre une combinaison linéaire de variables explicatives et le logit de la probabilité de présence/absence de l'espèce. Cependant, l'utilisation de la fonction de lien logit, et de sa réciproque la logistique commune, sous-tend que la probabilité de présence de l'espèce tend vers 1.0 lorsque la combinaison linéaire de prédicteurs environnementaux tend vers l'infini (et vers 0.0 lorsque la combinaison linéaire de prédicteurs environnementaux tend vers moins l'infini). Or, il est écologiquement réaliste de penser que ce n'est pas toujours le cas, et qu'il s'agit une fois de plus d'une hypothèse auxiliaire forte, lorsque

l'objectif principal est d'étudier la distribution de l'espèce et son rapport avec l'environnement. Par exemple, la probabilité de présence d'une espèce peut ne pas tendre vers une asymptote de 1.0 : (i) si une autre espèce entre en compétition avec l'espèce étudiée au cours du gradient inspecté, (ii) si une ressource qui n'a pas été intégrée aux variables explicatives devient limitante (ii) à cause du fonctionnement intrinsèque de la population étudiée (e.g. : fonctionnement en métapopulation).

Le modèle présenté dans ce chapitre propose d'affiner la méthode des GLMs en estimant les asymptotes de la logistique. Il pourrait également être appliqué pour l'étude de la distribution d'espèces en occurrence. En effet, si l'on se réfère à nos résultats sur les données réelles (Annexe III. Table S2.8 et Annexe III. Table S2.9), le modèle avec estimation des asymptotes performe mieux que le GLM classique pour représenter la relation entre la présence/absence de la grenouille *Pseudophryne corroboree* (RealData7) et la distance en mètres par rapport à la population existante la plus proche (variable V5 ; AICc GLM = 244.80 ; AICc LestKest = 235.15). Le modèle développé permet également d'améliorer les résultats obtenus avec deux variables explicatives, en mettant en avant une autre variable explicative d'intérêt par rapport au GLM (le modèle GLM sélectionne en priorité V2, le point de référence de latitude, et obtient un AICc de 226.93 ; tandis que le modèle LestKest sélectionne en priorité V8, la température minimale moyenne au printemps, et obtient un AICc de 211.71). De plus, le modèle estime des asymptotes différentes de 0.0 et 1.0 (L = 0.16 ± 0.036 et K = 0.70 ± 0.085 en univarié ; L = 0.15 ± 0.032 et K = 0.76 ± 0.051 en multivarié). Enfin, notre modèle n'a pas un impact positif en univarié par rapport au GLM sur le second jeu de données étudiant la distribution d'une espèce (RealDataset33 : présence/absence d'anguilles en fonction du temps de la marée basse). Cependant, il performe

un mieux en multivarié (en ajoutant la densité en crustacés) et estime une asymptote haute fortement différente de 1.0 ($K = 0.21 \pm 0.063$). Il est donc probable que, de manière générale, le modèle avec estimation des asymptotes permette une amélioration dans l'estimation des relations dans un contexte d'étude de la distribution d'espèces.

Pour aller encore plus loin, lors de l'étude de la distribution jointes de plusieurs espèces (JSDM), il serait envisageable d'implémenter de tels modèles pour lesquels la (les) asymptote(s) estimée(s) pourrai(en)t être variable(s) entre espèces ou groupes d'espèces.

CHAPITRE 4

Nouvelle forme de fonction sigmoïde pour les modèles de présence/absence multi-espèces

I. INTRODUCTION

I. 1. Les modèles de distributions d'espèces

La modélisation de la distribution des espèces (SDM) consiste à modéliser statistiquement la distribution d'une espèce dans l'espace (géographique ou écologique) à l'aide de données environnementales dans le but de mieux la comprendre et de pouvoir la prédire. Ce type de modélisation repose sur le concept de niche écologique. La niche écologique d'une espèce décrit l'habitat qui remplit les conditions environnementales qui permettent à l'espèce de satisfaire ses besoins minimums de manière à ce que la population locale persiste (Chase et al. 2003, Araújo and Guisan 2006b). On distingue deux niches, la niche fondamentale qui représente l'aire où l'espèce perdure indéfiniment si l'on ne prend pas en compte la compétition ; et la niche réalisée qui est incluse dans la niche fondamentale et représente l'aire où l'espèce n'est pas exclue par la compétition (Hutchinson 1957, Figure 4.1). La niche réalisée peut à l'inverse être plus large que la niche fondamentale lorsque des phénomènes de facilitations sont impliqués (e.g. Padilla and Pugnaire 2006, Figure 4.1).

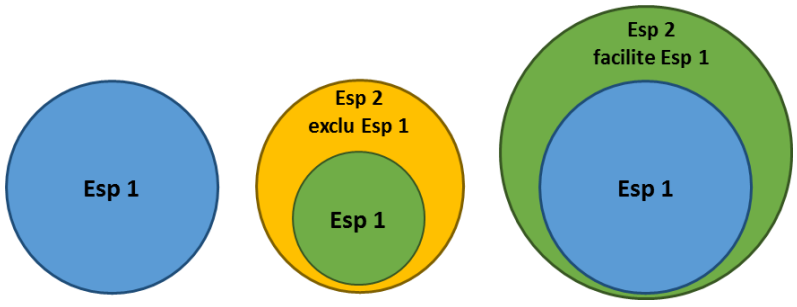


Figure 4.1 : Représentation théorique de la niche fondamentale (en bleu) et la niche réalisée (vert) de l'espèce 1 et la différence entre les deux niches par exclusion par l'espèce 2 (en jaune)

Les modèles de distribution d'espèces décrivent plutôt la niche réalisée, mais quand ils prennent en compte explicitement l'espace, ils permettent d'appréhender la notion de distribution de l'espèce (Araújo and Guisan 2006b).

Les modèles de distributions d'espèces (SDMs) peuvent être superposés (stack SDMs) pour obtenir un aperçu de la distribution des différentes espèces composant une communauté, mais chaque espèce est étudiée indépendamment.

Les modèles de distribution d'espèces peuvent prendre en compte une forme d'« interaction » entre espèces de manière unidirectionnelle (Figure 4.2), quand l'étude de la relation entre environnement et présence/absence (ou abondance) de l'espèce cible inclut la présence ou l'abondance d'autres espèces dans l'environnement (que nous appellerons SDMui pour « SDM with unidirectional interaction », e.g. Araújo and Luoto 2007, le Roux et al. 2014).

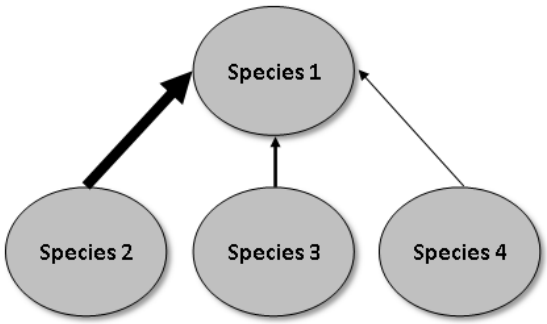


Figure 4.2: schéma théorique et simplifié de la nature des interactions entre les espèces pour les modèles de distribution d'espèces avec interactions unidirectionnels (SDMui). Ici il n'y a pas d'interaction entre les espèces, mais des effets de la présence ou de l'abondance des autres espèces sur une espèce d'intérêt.

I. 2. Les modèles multi-espèces

Les modèles multi-espèces (ou modèles de distribution joints - JSDMs) sont quant à eux des modèles non-espèce-centrés, c'est à dire qu'ils ne se concentrent pas sur une unique espèce (contrairement au SDMs) mais sur un cortège d'espèces. Ils étudient simultanément (comme les modèles SDMs superposés) et conjointement plusieurs espèces de la communauté et peuvent intégrer une forme de dépendance entre les espèces (comme les SDMui). Contrairement aux SDMs superposés, ils permettent avant tout de mutualiser de l'information contenue dans les données, de potentiellement prendre en compte des corrélations entre les espèces, et d'avoir une vision fonctionnelle et/ou hiérarchique des données de biodiversité (communautés, traits, espèces) (Ovaskainen and Soininnen 2011, Pollock et al. 2014). Lorsqu'elles sont intégrées, les dépendances entre espèces permettent de mettre en avant de la cooccurrence entre espèces. Cette cooccurrence peut être due soit à un partage de réponses à un environnement non décrit (une ou plusieurs variables environnementales non prises en compte), soit à d'autres processus écologiques ou évolutifs tels que des interactions biotiques. Une dépendance positive entre les espèces peut signifier : un phénomène de facilitation, de l'atténuation de la concurrence, une réponse commune positive à une variable environnementale non prise en compte dans le modèle. A l'inverse, une dépendance négative entre les espèces peut signifier : une interaction négative (compétition, prédation…), une réponse commune négative à une variable environnementale non prise en compte dans le modèle (Pollock et al. 2014).

Ces modèles étant lourds d'un point de vue numérique, ils sont assez récents, leur développement ayant été permis grâce aux avancées statistiques et technologiques permettant

de plus grandes puissances de calcul. Les modèles multi-espèces peuvent prendre des formes diverses en fonction de la nature des dépendances intégrées entre les différentes espèces. Dans le but d'avoir une vision globale des différentes formes de modèles multi-espèces, nous avons défini les différents types de modèles rencontrés dans la littérature. Toutes les catégories de modèles décrites ci-après ne sont pas exclusives les unes des autres (un modèle peut appartenir à deux catégories à la fois) et la liste des catégories n'est pas exhaustive. Nous y avons accoudé quelques d'exemples de publications utilisant un modèle appartenant à la catégorie abordée. Nous avons également tenté de représenter schématiquement la nature des dépendances entre les espèces dans chaque catégorie de modèles.

Les modèles présentés reposent sur une fonction de lien (f) et suivent donc une logique de modèle linéaire généralisé (GLM).

Paramètres communs à tous les modèles :

Les lettres grecques sont utilisées pour designer des paramètres estimés. $\bar{}$ désigne un vecteur et $\bar{\bar{}}$ désigne une matrice.

- f est la fonction de lien (dont le choix dépendra de la nature des données, cf. Table 0.1).
- i l'indice du niveau d'observation considéré qui correspond, en grande majorité des cas, au site dans les JSDMs (et que nous désignerons donc par « site » dans la suite).
- j l'indice de l'espèce
- μ_{ij} la valeur de la réponse moyenne de l'espèce j au site i

$\overline{x'_i}$ le vecteur ligne correspondant à la ligne de la matrice $\bar{\bar{x}}$ au site i, avec $\bar{\bar{x}}$ la matrice des prédicteurs environnementaux

π l'ordonné à l'origine

Paramètres facultatifs :

Trois paramètres peuvent être ajoutés de manière additive dans la combinaison linéaire, afin de prendre en compte une information partagée entre les espèces sur la variable réponse :

α_i un effet du site partagé par toutes les espèces

β_{0j} un effet espèce

$(\overline{x'_i} * \overline{\Omega_0})$, avec $\overline{\Omega_0}$ un effet environnemental partagé par toutes les espèces (**dimension : nombre de prédicteurs**)

Ces trois paramètres peuvent être fixes ou aléatoires (mais pas tous fixes en même temps pour éviter les singularités). Dans le second cas il est nécessaire de spécifier une distribution de probabilité. Ici par exemple si α_i et β_{0j} sont aléatoires et qu'une distribution gaussienne est choisie :

$$\alpha_i \sim N(0, \sigma_\alpha^2) \qquad \beta_{0j} \sim N(0, \sigma_{\beta 0}^2)$$

Dans les catégories présentées ci-après, les modèles sont présentés en incluant ces trois termes (en bleu), mais les mêmes modèles peuvent être écrits avec aucun, uniquement un, ou deux de ces termes.

I. 1. a. Modèles autécologiques (Autecological models)

L'autécologie désigne l'étude d'organisme individuel ou d'espèce individuelle, ou d'une population, en relation avec son environnement (en dehors de la communauté) (d'après Van der Klauuw, cf. Dubbeldam 2007).

Les modèles autécologiques lient la réponse de diverses espèces à un ou plusieurs prédicteurs environnementaux de manière idiosyncratique. C'est-à-dire qu'elles répondent de manière spécifique – sans lien avec les autres espèces – à l'influence de ces prédicteurs environnementaux. Ces modèles ne prennent pas en compte de forme de dépendance entre les espèces. Ces modèles respectent la forme d'un GLM ou GLMM. L'information partagée se situe dans les paramètres $\overline{\overline{CV}}$ et $\overline{\Omega_0}$.

$$f(\mu_{ij}) = \pi + \alpha_i + \beta_{0j} + (\overline{x'_\iota} * \overline{\Omega_0}) + (\overline{x'_\iota} * \overline{\beta_{1J}})$$

$$\overline{\beta_{1J}} \sim MN(\bar{0}, \overline{\overline{CV}})$$

Ce modèle peut aussi prendre la forme :

$$f(\mu_{ij}) = \pi + \alpha_i + \beta_{0j} + (\overline{x'_\iota} * \overline{\beta_{1J}})$$

$$\overline{\beta_{1J}} \sim MN(\overline{\Omega_0}, \overline{\overline{CV}})$$

Avec :

- $\overline{\beta_{1J}}$ l'effet de chaque prédicteur environnemental sur l'espèce j. Il s'agit du vecteur des coefficients de régression (**dimension : nombre de prédicteurs**).

$\overline{\overline{CV}}$ la matrice de variance-covariance qui quantifie dans quelle mesure les espèces varient dans leur réponse au(x) prédicteur(s) (éléments diagonaux) et la covariance en réponse à des couples de prédicteurs (éléments non diagonaux) (**dimension: nombre de prédicteurs x nombre de prédicteurs**).

Exemples de publications utilisant cette catégorie de modèle :

- Ovaskainen and Soininnen (2011) : équation 1 (sans l'effet site) et l'équation 4 (avec l'effet site en effet aléatoire).
- Brown et al. (2014) : modèle appelé « spp*env ».

I. 1. b. Modèles traits-dépendants (Traits-dependent models)

Les modèles traits-dépendants lient la réponse de diverses espèces à un ou plusieurs prédicteurs environnementaux, selon des traits définis (caractéristiques biologiques / écologiques) et éventuellement aussi idiosyncratiquement. Ces modèles ne prennent en compte de la dépendance entre les espèces que via un effet site commun et via le(s) trait(s) qu'elles partagent Figure 4.3). Le(s) trait(s) étudié(s) peut(euvent) également être un niveau taxonomique auquel l'espèce appartient. Ces modèles peuvent prendre deux formes en fonction de s'ils n'incluent (TDM-A) ou non (TDM-B) une réponse idiosyncratique des espèces.

Ecriture du modèle de type TDM-A :

$$f(\mu_{ij}) = \pi + \alpha_i + \beta_{0j} + (\overline{x'_i} * \overline{\Omega_0}) + \overline{x'_i}(\overline{\gamma_{traits}} * \overline{traits_j})$$

Avec :

$\overline{traits_j}$ le vecteur de la valeur des traits de l'espèce j (**dimension : nombre de traits**)

$\overline{\overline{\gamma_{traits}}}$ la matrice des effets de chaque prédicteur sur chaque trait (**dimension : nombre de prédicteurs x nombre de traits**).

Ecriture du modèle de type TDM-B :

$$f(\mu_{ij}) = \pi + \alpha_i + \beta_{0j} + (\overline{x'_i} * \overline{\Omega_0}) + (\overline{x'_i} * \overline{\beta_{1j}})$$

$$\beta_{0j} \sim N(\overline{\gamma_{0traits}} * \overline{traits_j}, \sigma_0^2) \quad \text{et/ou} \quad \overline{\beta_{1J}} \sim N(\overline{\gamma_{1traits}} * \overline{traits_j}, \overline{\sigma_1^2})$$

Avec :

$\overline{\gamma_{0traits}}$ et $\overline{\gamma_{1traits}}$ les vecteurs des ordonnées à l'origine pour chaque trait (**dimension : nombre de traits**)

Exemples de publications utilisant cette catégorie de modèle :

- Gelfand et al. (2005) : modèle appelé « equation 5 » de la forme TDM-B, avec β_{0j} qui possède une distribution dépendante des traits (appelé Ψ_k).
- Jamil et al. (2013) : modèle (cf. équation 2) de la forme TDM-B avec β_{1j} qui possède une distribution dépendante des traits (appelé β_j).
- Brown et al. (2014): modèle de la forme TDM-A appelé « trait*env » avec une intercepte globale étant le même pour tous les sites et espèces (b0).
- Brown et al. (2014) : modèle des formes TDM-A et autécologique appelé « trait*env + spp*env ».

- Caradima et al. (2019) : modèle mSDM avec le taxon d'appartenance comme trait étudié.

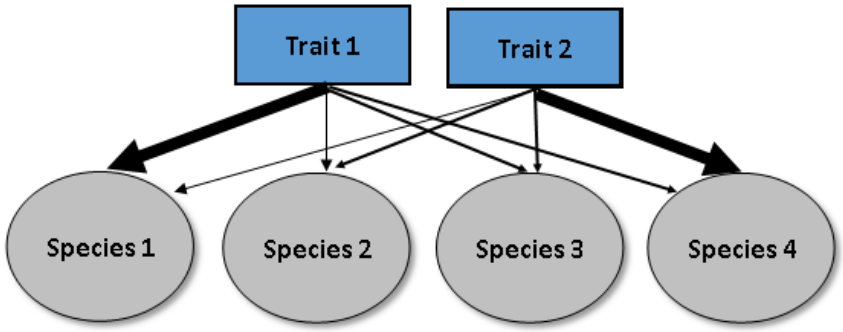


Figure 4.3 : schéma théorique et simplifié de la nature des interactions entre les espèces pour les modèles traits-dépendants. Ici il n'y a pas de dépendance entre les espèces mais un effet des traits sur les espèces.

I. 1. c. Modèles synécologiques déterministes à axe commun (Axed synecological deterministic models)

La synécologie désigne l'étude d'un groupe d'organismes, un groupe de plusieurs espèces, ou communautés, en relation avec leur environnement (d'après Van der Klauuw, cf. Dubbeldam 2007).

Les modèles synécologiques déterministes à axe commun lient la réponse de diverses espèces à un ou plusieurs prédicteurs environnementaux sur un ou plusieurs axes communs qui représente(nt) la ou les combinaison(s) linéaire(s) des prédicteurs, auxquels les espèces répondent de manière idiosyncratique (Figure 4.4). Il existe également une réponse commune à toutes les espèces, pas nécessairement portée par les axes communs (à travers le terme $(\overline{x'_\iota} * \overline{\Omega_0})$).

$$f(\mu_{ij}) = \pi + \alpha_i + \beta_{0j} + (\overline{x'_\iota} * \overline{\Omega_0}) + (\overline{x'_\iota} * \bar{\bar{\delta}}) * \overline{\beta_{1j}}$$

Avec :

$\bar{\bar{\delta}}$ la matrice représentant les axes (combinaisons linéaires des prédicteurs environnementaux) (**dimension : nombre de prédicteurs x nombre d'axes**)

$\overline{\beta_{1J}}$ le vecteur des effets des différents axes sur l'espèce j (**dimension : nombre d'axes**)

$(\overline{x'_i} * \bar{\delta})$ représente donc l'effet des prédicteurs à travers les axes, et

$(\overline{x'_i} * \bar{\delta}) * \overline{\beta_{1J}}$ la réponse de l'espèce j aux axes communs

Exemple de publication utilisant cette catégorie de modèle :

- Harris (2015) : avant réduction et transformation non-linéaire

Figure 4.4: schéma théorique et simplifié de la nature des interactions entre les espèces pour les modèles synécologiques déterministes axés. Ici les interactions entre espèces sont portées par un unique axe.

I. 1. d. Modèles synécologiques à variables latentes (Latent variable synecological models)

Les modèles synécologiques à variables latentes (ou facteurs latents) lient la réponse de diverses espèces à un ou plusieurs prédicteurs environnementaux et une ou plusieurs variable(s) latente(s)

de prédicteurs (non observés) qui induit(sent) la corrélation entre les espèces. Les variables latentes représentent des prédicteurs manquants et l'axe principale (ou les axes principaux) de co-variation de réponse des espèces (Figure 4.5).

$$f(\mu_{ij}) = \pi + \alpha_i + \beta_{0j} + (\overline{x'_\iota} * \overline{\Omega_0}) + (\overline{x'_\iota} * \overline{\beta_{1j}}) + (\overline{z'_\iota} * \lambda_j)$$

$$z'_i \sim MN(\bar{0}, \overline{\overline{Id}})$$

Avec :

$\overline{z'_\iota}$ la transposée de la matrice $\bar{\bar{z}}$ au site i, où $\bar{\bar{z}}$ est la matrice des variables latentes (**dimension : nombre de variables latentes *d***).

$$\overline{z_\iota} = (z_{i1}, \dots, z_{id})$$

d le nombre de variables latentes (à faire varier par l'utilisateur pour tester des modèles avec un nombre différent de variables latentes).

MN la distribution de probabilité normale multivariée

$\overline{\overline{Id}}$ la matrice identité (**dimension : *dxd***).

$\overline{\lambda_j}$ le vecteur des coefficients qui quantifie le lien entre la réponse de l'espèce et la variable latente (**dimension : nombre de variables latentes *d*)**

$$\overline{\lambda_j} = (\lambda_{j1}, \dots, \lambda_{jd})$$

Afin d'éviter l'invariance de rotation et assurer l'identifiabilité des paramètres, tous les éléments triangulaires supérieurs de la matrice $\bar{\bar{\lambda}}$ doivent être fixés à zéro et les éléments diagonaux doivent être positifs (Huber et al. 2004, Niku et al. 2017).

Exemples de publications utilisant cette catégorie de modèle :

- Warton et al. (2015) : LVM + librairie « gllvm »
- Thorson et al. (2015), avec une structure d'autocorrélation des facteurs latents
- Hui (2016) : modèle de la librairie « BORAL »
- Ovaskainen et al. (2016)
- Ovaskainen et al. (2016a)
- Ovaskainen et al. (2017)
- Tikhonov et al. (2017)
- Niku et al. (2017)
- Caradima et al. (2019) : modèle appelé « jSDM CT1 » et « jSDM CT2 ».

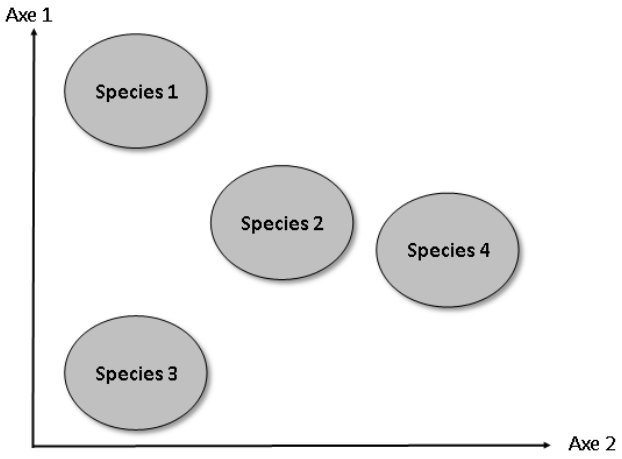


Figure 4.5 : schéma théorique et simplifié de la nature des interactions entre les espèces pour les modèles synécologiques à variables latentes. Ici les interactions entre espèces sont portées par les deux axes latents.

I. 1. e. Modèles synécologiques aléatoires (Random synecological models)

Les modèles synécologiques aléatoires lient la réponse de diverses espèces à un ou plusieurs prédicteurs environnementaux, avec addition d'un effet aléatoire dépendant de l'espèce par site et par espèce.

$$f(\mu_{ij}) = \pi + \alpha_i + \beta_{0j} + (\overline{x'_i} * \overline{\Omega_0}) + (\overline{x'_i} * \overline{\beta_{1j}}) + \upsilon_{ij}$$

$$\overline{\upsilon_i} \sim \mathrm{MN}(\overline{0}, \overline{\overline{\Sigma}})$$

Avec :

- υ_{ij} l'effet aléatoire pour le site i et l'espèce j
- $\overline{\overline{\Sigma}}$ la matrice de covariance (**dimension : nombre d'espèces x nombre d'espèces**)

Ces modèles peuvent avoir une forme dite non-structurée, dans ce cas la matrice $\overline{\overline{\Sigma}}$ est entièrement estimée (Figure 4.6); ou une forme structurée, dans ce cas la matrice $\overline{\overline{\Sigma}}$ possède une forme paramétrique qui dépend de la distance entre les espèces (Figure 4.7; e.g. distance phylogénétique ou fonctionnelle).

Exemples de publications utilisant cette catégorie de modèle :

- Pollock et al. (2014) : modèle de forme non structurée
- Warton et al. (2015) : Joint model for abundance – GLMM forme non structurée

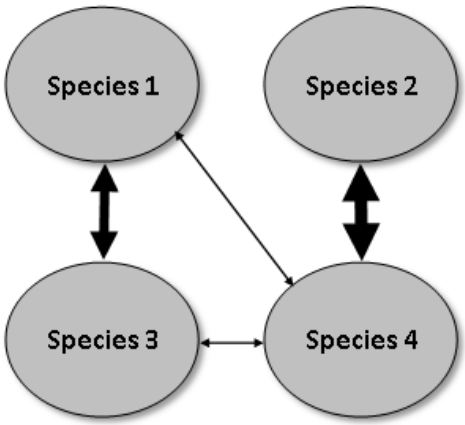


Figure 4.6: schéma théorique et simplifié de la nature des interactions entre les espèces pour les modèles synécologiques aléatoires non-structurés. Ici les interactions entre espèces sont portées par la matrice de covariance et sont symétriques (c'est-à-dire, par exemple, que l'espèce 1 aura un effet de même magnitude sur l'espèce 3 que l'inverse).

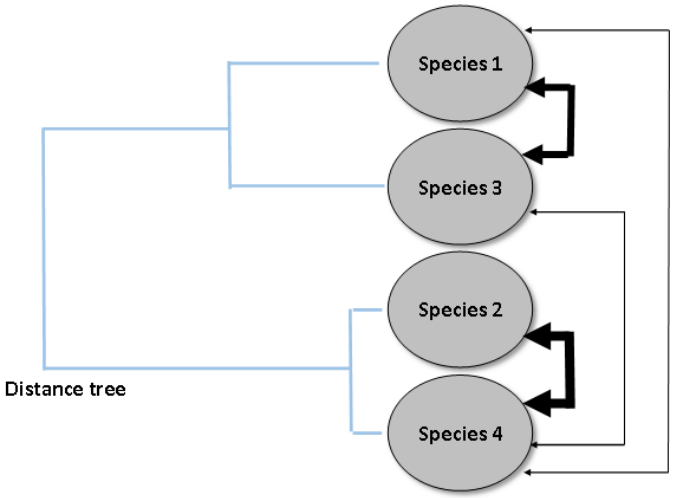


Figure 4.7: schéma théorique et simplifié de la nature des interactions entre les espèces pour les modèles synécologiques aléatoires structurés. Ici les interactions entre espèces sont portées par la matrice de covariance, sont symétriques et dépendent de la distance entre les espèces.

I. 3. Etude de cas : réponse des communautés de carabes à la conversion en futaie régulière de chênes en forêt française

I. 2. a. Question et stratégie de recherche méthodologique

Malgré un intérêt certain de chacun de ces types de modèles, nous nous sommes tournés vers des modèles à variables latentes (cf. Chapitre 4 – I.2.d), pour leur flexibilité d'une part, mais surtout pour leur rapidité de calcul. Nous nous sommes particulièrement focalisés sur des modèles multi-espèces de type présence-absence, c'est-à-dire sur des données binaires de présence (codé 1) ou absence (codé 0) des espèces. Nous avons donc intégré les développements faits précédemment (cf. Chapitre 3) sur les modèles GLM binomiaux, en incorporant l'estimation des asymptotes dans la fonction non-linéaire réciproque de la fonction de lien. L'estimation des asymptotes a été faite de différentes manières : (i) fixée à 1.0, (ii) estimée commune à toutes les espèces, (iii) estimée mais variable entre groupes d'espèces, (iv) estimée mais variable entre espèces et enfin, (v) estimée mais variable entre espèces avec une moyenne par groupe d'espèces. Enfin, nous avons aussi intégré une réflexion sur l'impact du choix dans le nombre et dans les niveaux d'intégration des variables latentes. Pour cela nous avons observé les répercussions de ces choix en comparant la capacité prédictive des différents modèles, et en étudiant la façon dont sont prises en compte les asymptotes variables.

Pour effectuer ces développements, nous nous sommes intéressés à un cas d'étude écologique alliant des questions de conservation et de gestion forestière : la réponse de la communauté de carabes à la conversion d'un taillis sous futaie en futaie régulière de chênes.

I. 2. b. Question de recherche écologique

Le taillis sous futaie (Figure 4.8) est un régime sylvicole qui associe les régimes de futaie et de taillis. On y trouve donc deux types de végétation bien distincts : la cépée issue de la régénération végétative, et la futaie issue de la régénération sexuée. Aux siècles derniers, ce traitement représentait une bonne réponse aux besoins domestiques et industriels puisque le taillis fournissait du bois de feu, et la futaie du bois d'œuvre. Cependant, la forte réduction du besoin en bois de chauffage a dévalorisé le taillis et n'a pas permis au régime de taillis-sous-futaie de rester assez rentable. À partir de la deuxième moitié du 19eme siècle en France, une grande campagne de conversion a été mise en place, visant à convertir les taillis sous futaie en futaie régulière. La futaie régulière est composée à terme de grands arbres adultes issus de semis et dont tous les arbres des essences principales sont d'âges proches. Elle permet la production de grosse quantité de bois d'œuvre et est donc plus rentable dans le contexte actuel. La conversion se fait via une succession de coupes régularisantes, dont la dernière est la coupe définitive (Dauffy-Richard et al. 2010). En futaie régulière, à la fin de chaque cycle, l'ensemble du peuplement est coupé (coupe de régénération), en général par le biais de coupes progressives réparties dans le temps. Ainsi, la futaie régulière passe successivement par plusieurs stades au fur et à mesure de la repousse : le stade semis (stade 1 Figure 4.8), de fourré, de gaulis (stade 2 Figure 4.8), de bas-perchis (stade 3 Figure 4.8), de haut perchis (stade 4 Figure 4.8), de jeune futaie (stade 5 Figure 4.8), et enfin celui de futaie adulte (non représenté).

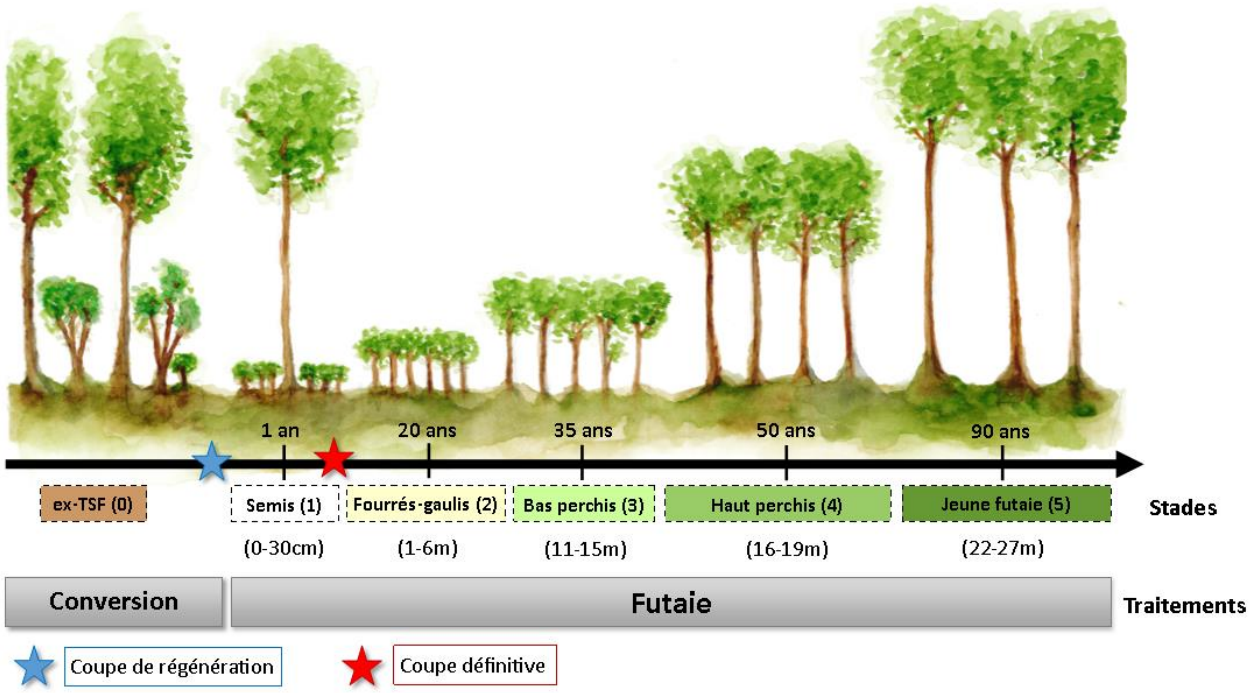


Figure 4.8: Stades et traitements d'une forêt convertie d'une ex-Taillis-sous-futaie (ex-TSF), vers une futaie régulière. D'après une aquarelle réalisée par Laura Chevaux.

Les coupes de conversion, la coupe définitive (marquant la fin de la conversion) ainsi que les coupes de régénération (récolte du bois impliquant un retour au stade semis) peuvent être considérées comme des perturbations écologiques. Ce concept peut être défini comme « un événement ponctuel, localisé et imprévisible, dû à une force physique, un agent ou un processus biotique ou abiotique, qui provoque une perturbation, i.e. qui endommage, déplace ou tue un ou plusieurs individus ou communautés, et permet la colonisation par de nouveaux organismes, d'espèces identiques et différentes, en référence à l'état de l'écosystème avant le dérangement » (d'après Richard 2004, inspiré de Sousa 1984, Rykiel 1985, Blondel 1995 et Begon et al. 1996). L'impact réel de ces dérangements dépend de leur ampleur spatiale, de leur magnitude (intensité et sévérité), de leur fréquence, de leur prédictibilité et de leur période de rotation sur la région (Sousa 1984). Bien que prévisibles, les différentes coupes constituent donc des dérangements

majeurs d'origine anthropique. Ainsi, les conditions locales de l'habitat et des ressources en nourriture changent brutalement lors de la coupe, mais elles changent aussi graduellement au cours du cycle sylvicole. Les successions sylvicoles causées par la coupe et la repousse de la futaie modifient le paysage passant d'un habitat constitué exclusivement d'une strate muscinale et herbacée, à une strate majoritairement arbustive et herbacée, à enfin une strate majoritairement arborée et herbacée. Cependant, la plupart des stades du cycle sylvicole sont représentés si l'on s'intéresse à une zone assez grande (e.g. à l'échelle du paysage).

Les changements brutaux (causés par les coupes) et graduels (causés par les successions sylvicoles) se répercutent sur la flore et la faune locale (Dauffy-Richard et al. 2010) tels que les lépidoptères nocturnes qui sont impactés négativement par les coupes d'encensement (Bonneil 2005). Les insectes, dont les carabidae (famille de grands coléoptères terrestres), sont également fortement impactés. Les carabes n'étant pas des espèces directement exploitées par la coupe, la perturbation est indirecte puisque les ressources (habitat et nourriture) de ces communautés sont modifiées (microclimat du sol, micro-habitats et disponibilité en ressources). Les coupes de régénération ont des impacts variés sur les différentes espèces de carabes puisqu'elles ont tendance à pénaliser les espèces associées aux stades matures de la forêt, et plutôt à favoriser d'autres espèces telles que les espèces généralistes (Dauffy-Richard et al. 2010). À grande échelle spatiale (avec différents stades sylvicoles représentés), les dynamiques de colonisation / recolonisation, entre les patches présentant un stade sylvicole différent, dépendent de la capacité de dispersion des espèces. Pour les espèces carabiques, les capacités de dispersions peuvent être liées à différents traits de l'espèce tels que sa préférence en termes d'habitats, ou des aspects plus morphologiques et fonctionnels tels que la capacité de

déplacement (sur le sol ou en vol en fonction des espèces – Bouget, 2004 ; Richard, 2004). Enfin, à grande échelle temporelle, les effets observés ont tendance à s'inverser, puisqu'à court terme (20 ans après la coupe), les espèces forestières sont plutôt désavantagées au profit des espèces non-forestières par la coupe de régénération, tandis qu'à long terme (100 ans après la coupe), elles sont plutôt avantagées au détriment des espèces non-forestières (Dauffy-Richard et al. 2010).

I. 2. c. Cas spécifique étudié

La forêt domaniale de Montargis, qui couvre une superficie de 4200 hectares, est située dans le département du Loiret, dans la sylvo-écorégion du Pays d'Othe et Gâtinais oriental, elle-même comprise dans la grande région écologique du Centre Nord semi-océanique (IFN 2011). La forêt est caractérisée par un substrat crayeux (craie du Crétacé supérieur) recouvert d'argiles, sables et cailloutis (Denizot 1971) qui forment des sols de types brunisols oligo-saturés et luvisols typiques (Baize and Girard 1992), qui sont des sols plutôt acides. La forêt de Montargis est dominée en majorité par du chêne sessile et pédonculé, et dans une moindre mesure du pin sylvestre, du hêtre et du charme. Le climat est océanique avec des influences continentales, et se caractérise par une température moyenne annuelle de 10.9°C, une variation de température annuelle de 15.3°C (différence entre la température moyenne du mois le plus chaud - 19.0°C en juillet - et celle du mois le plus froid - 3.7°C en janvier) et des précipitations de 647mm (données issues de relevés sur la période 1971-2000, Chevalier 2003).

La forêt de Montargis, anciennement propriété royale, est devenue domaine privé de l'Etat en 1848. Tout comme beaucoup d'autres forêts françaises dont le régime de sylviculture était le

taillis sous futaie, la forêt de Montargis a commencé sa conversion vers la futaie régulière entre 1857 et 1872 (ONF 1971). Les différents stades sylvicoles de la futaie régulière pouvant être définis en fonction de la hauteur dominante (cf. Figure 4.8), les parcelles étudiées en forêt de Montargis ont donc été séparées en cinq stades de sylviculture avec d'un côté le stade d'ex-taillis sous futaie (0, parcelle toujours en conversion), et d'un autre côté les cinq stades (1 à 5) de la futaie régulière (parcelles post conversion) dont les hauteurs dominantes moyennes des ligneux étaient respectivement de 0.0, 4.7, 14.2, 17.8 et 24.9 mètres (Figure 4.9).

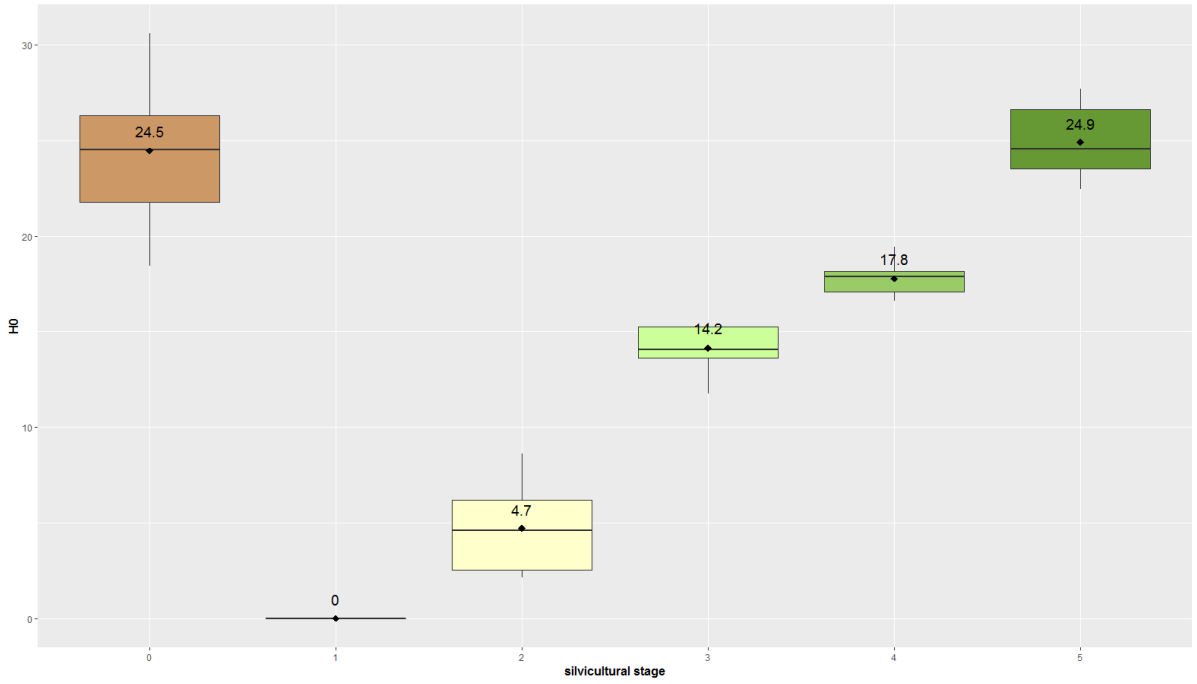


Figure 4.9: Hauteur dominante des ligneux (en m) en fonction du stade sylvicole des parcelles étudiés en forêt de Montargis

Les effets de la conversion et des coupes de régénération sur les communautés de coléoptères carabiques ont déjà été étudiés sur des données d'abondance (cf. Richard et al. 2003, Richard 2004). Le but de notre étude était d'observer si nous obtenions les mêmes résultats sur des données de présence/absence, en utilisant le même type d'analyses (GLM binomial avec prise en

compte des traits des espèces) mais en prenant en compte les asymptotes variables, en appliquant des variables latentes (portant la dépendance entre espèces) et en étudiant de nouveaux traits. Cette nouvelle étude nous permet donc à la fois d'étudier l'intérêt de la prise en compte des asymptotes dans la réciproque de la fonction de lien utilisée dans le modèle multi-espèces, tout en essayant d'affiner la réponse apportée à des problématiques de gestion et de conservation importantes.

Le jeu de données étudié (issu de Richard 2004) se concentre donc sur la relation entre d'une part les coléoptères terrestres (carabidae) dans la forêt française de Montargis sous la région parisienne, et d'autre part les traitements ainsi que stades sylvicoles.

Les carabes ont été collectés dans des pièges à écueil en 1999 (Richard 2004), pendant trois périodes d'une semaine (la deuxième et la cinquième semaine de juin ainsi que la cinquième semaine d'août). Chaque parcelle consistait en un carré de 14 × 14 m et contenait quatre pièges, nommés en fonction de leur quartier cardinal (N, S, E, O) dans lequel ils se trouvaient. Les pièges (8,5 cm de diamètre × 11 cm de profondeur) étaient à moitié remplis d'une solution à 50% d'éthylène glycol et protégés par un toit en plastique transparent afin d'éviter les inondations par la pluie. Les coléoptères terrestres ont été identifiés selon Jeannel (1941), (Lindroth (1974) et Hůrka (1996). La nomenclature a suivi Hůrka (1996). Les larves n'ont pas été identifiées et n'ont donc pas été analysées. Au final, les données d'occurrence (présence / absence) étaient disponibles pour 47 espèces (cf. Table 4.1) sur trois périodes, pour quatre pièges répartis sur 100 parcelles avec un total de 55 037 observations (29 données étaient manquantes pour toutes les espèces en raison de la destruction des pièges). Le nombre d'observations par espèce (Table 4.1) était très variable, avec des espèces rarement capturées (e.g. une seule observation pour huit

espèces dont *Amara lunicollis*) et d'autres espèces plus communément capturées (e.g. 955 observations pour *Abax ovalis*).

Les données sur le stade de régénération et le traitement sylvicole ont été collectées sur les mêmes parcelles. Deux types de traitements ont été définis: futaie régulière (Fut) et conversion de taillis sous futaie en futaie régulière (Conv). Les six stades du cycle sylvicole ont également été précisés (stade 0 d'ex-TSF pour le traitement en conversion et les stades 1 à 5 pour le traitement en futaie régulière, Figure 4.8).

Enfin, nous avons collecté des informations sur plusieurs traits sur les 47 espèces à partir d'autant de sources que possible. Les caractères étudiés ont été choisis en fonction de leur pertinence pour la question écologique et de la disponibilité des données. Lorsque différentes sources étaient disponibles pour un même caractère, nous avons choisi celle contenant des informations pour les 47 espèces étudiées et cela ne semblait pas en contradiction avec d'autres sources (cf. Data IV.S1). Au final (Table 4.1), les traits étudiés dans le cadre de cette étude comprenaient la préférence d'habitat (issue de Coulon et al. 2000) car probablement fortement liée à la capacité de l'espèce à s'établir aux différents stades sylvicoles. La deuxième caractéristique majeure incluse dans l'étude est la morphologie de l'aile, qui nous renseigne sur la capacité d'une espèce à se déplacer et pourrait donc être liée à sa capacité de colonisation. Nous avons également incorporé dans le modèle la taille minimale du corps (de "carabids.org" n.d.) comme indicateur de la taille corporelle qui pourrait également jouer un rôle dans la capacité de l'espèce à s'établir dans un environnement.

Table 4.1: Espèces de carabidés étudiés, avec des informations sur leur préférence d'habitat (Coulon et al. 2000), leur morphologie alaire, leur taille corporelle minimale en millimètres ("carabids.org" n.d.), et leur nombre d'observations dans l'ensemble de données.

Espèce (nom latin)	**Nom de code**	**Préférence d'habitat**	**Morphologie alaire**	**Taille corporelle minimale (mm)**	**Nombre d'observations**
Abax ovalis	abov	forêt	brachyptère/aptère	11	955
Abax parallelepipedus	abat	forêt	brachyptère/aptère	16	429
Abax parallelus	abpa	forêt	brachyptère/aptère	13	284
Amara lunicollis	amlu	prairie	macroptère	6	1
Amara similata	amsi	prairie	macroptère	7	14
Amara equestris	ameq	prairie	macroptère	8	5
Badister bullatus	babu	prairie	macroptère	4	18
Calathus rotundicollis	caro	indifférent	di-polymorphique	8	6
Calathus luctuosus	calu	forêt	brachyptère/aptère	10	522
Carabus nemoralis	cane	indifférent	brachyptère/aptère	18	78
Carabus auronitens	cani	indifférent	brachyptère/aptère	16	6
Carabus violaceus	cavi	indifférent	brachyptère/aptère	22	281
Carabus problematicus	capr	indifférent	brachyptère/aptère	20	98
Carabus monilis	camo	indifférent	brachyptère/aptère	22	595
Carabus auratus	caau	prairie	brachyptère/aptère	20	133
Cicindela campestris	cicm	prairie	macroptère	10	467
Harpalus latus	hala	prairie	macroptère	8	1
Harpalus luteicornis	halu	indifférent	macroptère	6	15
Harpalus rubripes	haru	prairie	macroptère	8	2
Harpalus rufipalpis	harf	prairie	macroptère	8	7
Harpalus tardus	hata	prairie	macroptère	7	6
Lebia chlorocephala	lech	prairie	macroptère	4	1
Leistus rufomarginatus	leru	forêt	di-polymorphique	7	1
Loricera pilicornis	lopi	point d'eau en forêt	macroptère	6	62
Bembidion lampros	bela	indifférent	di-polymorphique	2	1

Microlestes maurus	mima	prairie	di-polymorphique	2	1
Molops piceus	mopi	forêt	brachyptère/aptère	9	7
Nebria brevicollis	nebr	forêt	macroptère	9	5
Notiophilus biguttatus	nobi	indifférent	di-polymorphique	3	4
Notiophilus palustris	nopa	indifférent	di-polymorphique	4	3
Notiophilus rufipes	noru	indifférent	macroptère	4	9
Panagaeus bipustulatus	pabi	prairie	macroptère	6	3
Platyderus depressus	plru	forêt	brachyptère/aptère	5	6
Poecilus kugelanni	poku	prairie	macroptère	10	26
Poecilus cupreus	pocu	prairie	macroptère	9	7
Poecilus versicolor	pove	prairie	macroptère	8	11
Harpalus rufipes	psru	prairie	di-polymorphique	11	1
Pterostichus oblongopunctatus	ptob	forêt	di-polymorphique	9	10
Pterostichus melanarius	ptme	indifférent	di-polymorphique	12	280
Pterostichus niger	ptni	forêt	di-polymorphique	15	4
Pterostichus cristatus	ptcr	forêt	brachyptère/aptère	12	56
Pterostichus madidus	ptma	forêt	di-polymorphique	13	65
Stomis pumicatus	stpu	marais, prairie	di-polymorphique	6	3
Syntomus obscuroguttatus	syob	indifférent	macroptère	2	3
Synuchus vivalis	syvi	indifférent	di-polymorphique	5	41
Trechus obtusus	trob	prairie	di-polymorphique	3	1
Trechus quadristriatus	trqu	indifférent	di-polymorphique	3	4

II. DEVELOPPEMENT DES MODELES ET LIMITES

II. 1. Présentation du modèle général et méthode de comparaison

II. 1. a. Modèle à variables latentes

Nous avons appliqué un modèle synécologique à variables latentes, où la réponse moyenne des espèces aux prédicteurs environnementaux dépend du groupe auquel appartient l'espèce (cf. $\overline{\beta_{1J}}$).

$$f(\mu_{ij}) = \alpha_i + \beta_{0j} + (\overline{x'_\iota} * \overline{\beta_{1J}}) + (\overline{z'_\iota} * \overline{\lambda_J})$$

$$\alpha_i \sim N(0, \sigma_\alpha^2)$$

$$\beta_{0j} \sim N(0,5)$$

$$\overline{\beta_{1J}} \sim N(\overline{m_{gJ}}, \sigma_{\beta 1}^2)$$

$$\overline{z'_\iota} \sim MN(\overline{0}, \overline{\overline{Id}})$$

$$\lambda_{jd} \sim N(0,5)$$

Avec:

- f la fonction moyenne
- i l'indice du site
- j l'indice de l'espèce et gj le groupe auquel appartient l'espèce j

μ_{ij} la réponse moyenne de l'espèce j au site i

$\overline{x'_\iota}$ la ligne (correspondant au site i) de la matrice $\overline{\overline{x'}}$ qui est la transposée de la matrice $\bar{\bar{x}}$, où $\bar{\bar{x}}$ est la matrice des prédicteurs environnementaux

α_i l'effet du site i partagé par toutes les espèces (effet aléatoire), avec

σ_α^2 est la variance de l'effet site (estimée), avec $\ln(\sigma_\alpha^2) \sim N(0,5)$

β_{0j} l'effet additif de l'espèce j (effet fixe), avec

$\overline{\beta_{1J}}$ le vecteur coefficients de régressions décrivant les effets des prédicteurs environnementaux sur l'espèce j, avec

$\overline{m_{gJ}}$ le vecteur de la réponse moyenne attendue pour l'espèce j aux prédicteurs environnementaux compte tenu de ses traits écologiques (estimé), avec $\overline{m_{gJ}} \sim N(0,10)$

$\sigma_{\beta 1}^2$ la variance constante de tous les groupes d'espèces (estimée), avec $\ln(\sigma_{\beta 1}^2) \sim N(0,5)$

$\overline{z'_\iota}$ la ligne (correspondant au site i) de la matrice $\overline{\overline{z'}}$, qui est la transposée de la matrice $\bar{\bar{z}}$, où $\bar{\bar{z}}$ est la matrice des variables latentes qui modélise la corrélation entre les espèces (correspond à la partie synécologique du modèle). $\overline{z'_\iota} = (z_{i1}, \dots, z_{id})$

d est le nombre de variables latentes considérées

$\overline{\overline{Id}}$ est la matrice identité

$\bar{\lambda_j}$ est le vecteur des effets des variables latentes sur l'espèce j. $\bar{\lambda_j} = (\lambda_{j1}, \dots, \lambda_{jd})$

Afin de respecter les contraintes sur la matrice $\bar{\bar{\lambda}}$ (Huber et al. 2004, Niku et al. 2017), nous avons fixé à zéro les éléments triangulaires supérieurs, et forcé les éléments diagonaux à être positifs à l'aide d'une exponentielle.

Nous avons appliqué ce modèle avec sept différents réglages concernant le nombre de variables latentes (d), un modèle sans variable latente (nolv), et six modèles avec un à six variables latentes (lv1sp, lv2sp, lv3sp, lv4sp, lv5sp et lv6sp). Le modèle sans variables latentes (nolv) correspond à un modèle autécologique, car il ne porte pas de dépendance entre les espèces mais permet tout de même du partage d'information au niveau des paramètres α_i et $\overline{\beta_{1j}}$.

II. 1. b. Traitement des asymptotes

Dans le cadre de cette étude, nous avons choisi de tester des modèles avec l'asymptote haute estimée, mais avec l'asymptote basse fixée à zéro. En effet, nous avons imposé une asymptote basse à zéro car il n'y a pas, à notre connaissance, de concept de niche écologique selon lequel une espèce pourrait se rencontrer avec une probabilité positive minimale partout (Huisman et al. 1993 semblent être partis du même postulat).

Nous avons comparé cinq types de modèles différents du point de vue du traitement de l'asymptote haute de la fonction de lien logistique :

- K1 : modèle pour lequel l'asymptote haute est fixée à 1.0 (modèle logistique commun)
- K : modèle pour lequel l'asymptote haute est estimée commune à toutes les espèces et inférieure à 1

- Kgroup : modèle pour lequel l'asymptote haute est estimée, variable et dépendante de la combinaison linéaire des traits des espèces. Chaque groupe d'espèces ayant les mêmes traits qualitatifs (préférence d'habitat et morphologie de l'aile) possèdent une asymptote haute moyenne commune et dévient de cette moyenne à travers leur trait continu (taille minimale du corps).
- Ksp : modèle pour lequel l'asymptote haute est estimée, variable en fonction de chaque espèce
- Ksp.spgroup : modèle pour lequel l'asymptote haute est estimée, variable en fonction de chaque espèce, mais dont la moyenne dépend du groupe auquel appartient l'espèce

II. 1. c. Les niveaux d'inclusion des variables latentes

Nous avons supposé que la corrélation entre espèces pouvait se manifester à différents niveaux hiérarchiques. Nous avons considéré différents niveaux pouvant contenir cette forme de dépendance aléatoire (niveaux auxquels les corrélations entre espèces sont incluses dans $\bar{\bar{z}}$), et avons donc définis les variables latentes aux niveaux :

- de la parcelle (plot) ;
- du plot lié à la saison d'échantillonnage (plot/season) ;
- du plot lié au piège (plot/trap) ;
- de la parcelle lié à la saison d'échantillonnage et au piège (plot/season/trap). Ce niveau d'inclusion des effets aléatoires est celui qui se rapprocherait le plus du niveau observation (mais avec le niveau espèce en moins).

II. 1. d. Méthode d'optimisation et de comparaison des modèles

Afin de s'approcher au plus d'une modélisation Bayésienne, nous avons appliqué des distributions a priori (priors) pour chacun des paramètres (Annexe IV. Table S1.1).

Nous avons conçu une fonction permettant de tirer les valeurs initiales dans des distributions prédéfinies, comprises dans les priors (Annexe IV. Table S1.1), permettant donc d'obtenir des valeurs initiales différentes en fonction de la graine[8] choisie.

Pour chaque modèle, dans le but d'obtenir une mesure de la qualité d'un modèle statistique qui prenne en compte le nombre d'observations, nous avons calculé la version corrigée du critère d'information d'Akaike (AICc). Cette métrique se base sur le nombre de paramètres à estimer dans le modèle (np), l'opposé de la log-vraisemblance (negative-log-likelihood, NLL), les distributions a priori ($priors$) et le nombre d'observation ($nobs$) et se calcule de la manière suivante :

$$AICc = 2 * np + 2 * (NLL - priors) + 2 * np * \frac{np + 1}{nobs - np - 1}$$

[8] Selon Wikipédia, une graine (ou graine aléatoire) est un « nombre utilisé pour l'initialisation d'un générateur de nombres pseudo-aléatoires ».

II. 2. Première étape de modélisation : optimisation simple et instabilité

Lors de la première étape de modélisation, nous avons lancé tous les modèles (avec les différents nombres de variables latentes et les différentes manières d'estimer l'asymptote haute) avec les variables latentes incluses simultanément à tous les niveaux présentés ci-dessus (cf. Chapitre 4 - II. 1. c.), c'est à dire au niveau du plot/season/trap, du plot/season, du plot/trap et du plot.

Dans le logiciel R (R Core Team 2018), nous avons utilisé l'outil Template Model Builder (TMB), au moyen de la librairie R «TMB» (Kristensen et al. 2018). Après des essais de nombreuses autres alternatives, nous avons choisi la méthode d'optimisation « nlminb » disponible dans cette même librairie. Au total, 35 modèles différents ont été appliqués aux donnés et comparés (sept réglages concernant le nombre de variables latentes x cinq traitements concernant l'asymptote haute), et pour chacun de ces modèles, nous avons comparés les résultats obtenus avec les quatre premières graines. Les paramètres α_i, $\overline{\beta_{1j}}$ et $\overline{z'_\iota}$ ont été considérés comme aléatoires et donc marginalisés avec TMB.

Les résultats obtenus avec les différents modèles, à l'aide de l'optimisation « nlminb » étaient très variables entre graines, et donc très instables (Figure 4.10). Seul le modèle le plus simple (modèle K1 sans variable latente) était stable. Il semblerait que l'optimisation s'arrête à un point sans qu'il ne s'agisse pour autant réellement de l'optimum. Ceci peut avoir un impact important lors de la comparaison et sélection de modèle basé sur l'AICc et sur une seule graine. En effet, si

l'on prend par exemple la première graine (s1), le meilleur modèle sélectionné serait le modèle avec une asymptote variable par groupes d'espèces et cinq variables latentes (Kgroup lv5sp), tandis que si l'on prend la troisième graine (s3), le meilleur modèle sélectionné serait le modèle avec une asymptote fixée à 1.0, avec trois variables latentes (K1 lv3sp), aboutissant à des conclusions donc très différentes sur les données (Table 4.2). En effet, dans le premier cas (avec s1) on pourrait conclure que la probabilité de présence maximale des espèces dépendrait de leurs traits. Dans le second cas (avec s3), toutes les espèces de la communauté étudiée seraient obligatoirement présentes lorsque le prédicteur tend vers l'infini. Cette différence de conclusion dépendante de la graine peut être tout aussi délétère à des conclusions d'un point de vue de la recherche en écologie, ici de la connaissance du fonctionnement des communautés, que d'un point de vue de gestion.

Table 4.2: AICc des 35 modèles appliqué aux données pour les quatre premières graines avec une optimisation simple.

Traitement de l'asymptote haute	Graine	nolv	lv1sp	lv2sp	lv3sp	lv4sp	lv5sp	lv6sp
K1	s1	14838.7	14202.2	14209.8	13927.9	13739.4	13641.1	13807.51
	s2	14838.7	NaN	14073.0	13823.9	13784.1	13806.0	NaN
	s3	14838.7	14527.5	14008.1	13977.7	NaN	14639.2	14161.92
	s4	14838.7	14045.9	14183.1	14351.9	14021.4	13944.9	14042.17
K	s1	14841.3	14076.4	14174.1	21723.0	13739.4	14051.0	14318.39
	s2	14840.8	14327.7	14829.6	14061.2	13784.1	NaN	14374.54
	s3	16300.9	18350.3	19730.4	14185.0	NaN	17634.5	14512.33
	s4	18179.4	14936.9	14796.6	16462.8	14021.4	14187.3	14420.76
Kgroup	s1	14854.9	14092.7	14063.1	NaN	13983.1	**13607.6**	13863.22
	s2	14855.8	14119.2	14265.2	13715.1	22646.1	14035.8	17754.07
	s3	14854.9	14890.6	13980.6	14016.2	14054.0	14091.0	14340.04
	s4	14854.9	14150.1	14529.4	14352.4	14162.8	13927.7	14243.94
Ksp	s1	14921.7	14165.1	14369.4	13995.7	14000.4	14758.3	14060.69
	s2	14880.7	14227.6	14188.7	13851.7	14721.4	15097.8	14425.76

	s3	14897.3	14335.3	14248.6	14151.9	NaN	14680.8	14803.52
	s4	14921.4	14403.5	14256.0	14438.6	14285.6	14093.5	14226.84
Kspspgroup	s1	14655.3	14250.0	14670.8	14239.5	14814.0	14732.4	14581.31
	s2	15033.6	14590.2	14469.9	14611.7	NaN	15246.1	14697.51
	s3	15040.7	14048.7	14107.4	14275.9	14487.6	NaN	14795.39
	s4	14642.6	14608.4	14874.5	14726.0	14803.2	14782.0	15932.54

En souligné, la valeur d'AICc du meilleur modèle par graine. En gras et souligné, la valeur d'AICc du meilleur de tous les modèles.

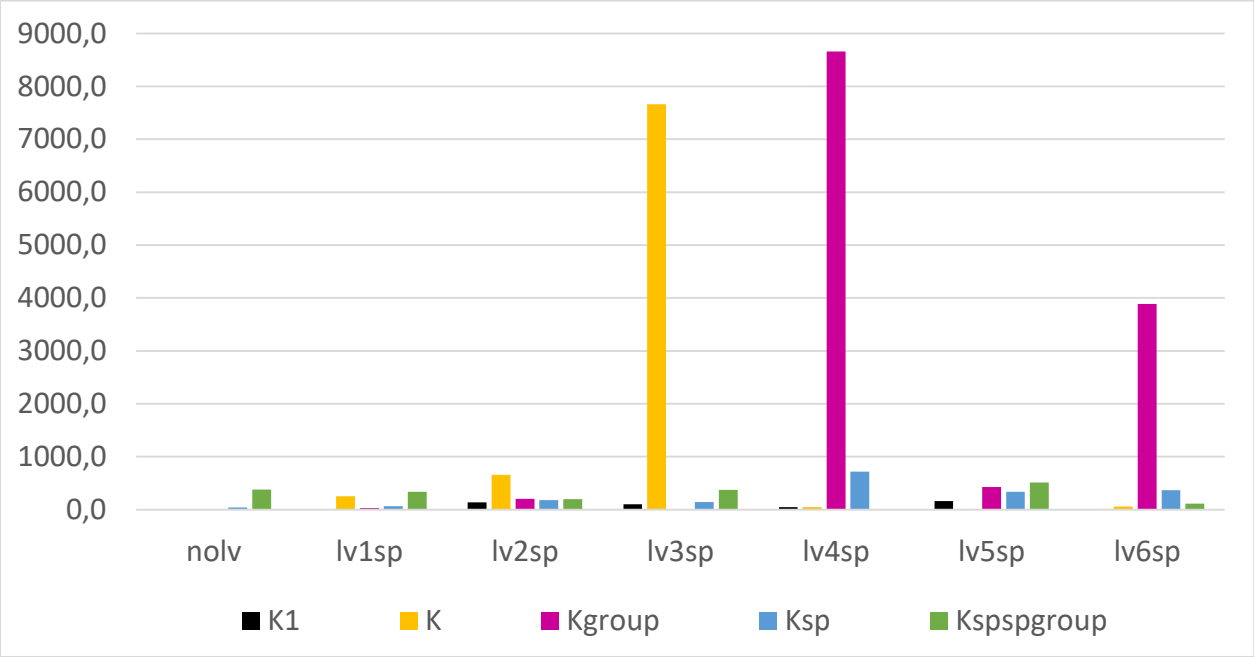


Figure 4.10: Différences entre la valeur maximale et la valeur minimale d'AICc obtenues avec les deux premières graines (afin d'être comparable avec les tests suivants) des 35 modèles appliqués aux données avec de zéro à six variables latentes, avec une optimisation simple.

Afin de pouvoir continuer à utiliser de tels modèles, il est donc nécessaire de stabiliser les résultats obtenus avec ce type d'optimisation. Nous avons émis l'hypothèse qu'une succession d'optimisations, reprenant à chaque étape les paramètres optimisés de sortie de l'étape précédente comme valeurs initiales, pourrait permettre de parvenir à une optimisation beaucoup plus aboutie du modèle.

II. 3. Deuxième étape de modélisation : optimisations multiples et incertitudes persistantes

Dans cette optique, F. Gosselin (cf. Annexe IV. Note S1.1) a développé une librairie, appelé « tioptimTMB », qui a pour but de réaliser cette succession d'optimisations, l'idée étant de continuer les optimisations jusqu'à obtenir une stabilisation de l'optimisation, c'est-à-dire ici des valeurs de « Optimum function value » identiques entre deux étapes successives d'optimisation.

Les paramètres α_i, $\overline{\beta_{1j}}$ et $\overline{z'_i}$ ont été gardés comme aléatoires (marginalisés avec TMB). Pour cette étape, et pour les étapes suivantes, nous avons procédé à un double ajustement des modèles aux données via le package « tioptimTMB » (eux-mêmes constitués de plusieurs itérations): (i) un premier ajustement faisant appel à des arguments de températures afin de tenter d'optimiser au mieux les modèles ; (ii) un second ajustement consistant en une succession d'optimisations, utilisant comme valeurs initiales les valeurs optimisées du premier ajustement, et sans utilisation de température (Annexe IV. Section S1.1).

Nous avons gardé les mêmes niveaux d'inclusion des variables latentes que lors de l'étape précédente (plot/season/trap, plot/season, plot/trap et plot) et avons également testé les modèles avec de zéro à six variables latentes. Pour chaque modèle, en raison d'un temps de calcul beaucoup plus long, nous n'avons comparé les résultats obtenus qu'avec les deux premières graines (s1 et s2). Nous avons lancé une troisième graine (s3), uniquement pour les modèles pour lesquels la différence en termes d'AICc entre les deux premières graines était trop importante.

Nous avons pu constater tout d'abord, que l'utilisation de multiples jeux de valeurs initiales et d'un aplatissement de la vraisemblance permettait d'éliminer les NaN obtenus lors de la phase précédente avec une seule optimisation. Ensuite, les modèles avec les traitements K1, K et Kgroup, et sans variable latente (nolv, soient les modèles les plus simples), étaient les seuls modèles pour lesquels une optimisation simple et une optimisation multiple à l'aide de tioptimTMB arrivaient à la même valeur d'AICc (cf. Table 4.2 et Table 4.3). Autrement, avec la moindre complication (au niveau des asymptotes à estimer ou du nombre de variables latentes), la capacité prédictive des modèles était toujours meilleure (AICc plus faible) avec une optimisation multiple (tioptimTMB) qu'avec une optimisation simple, et semblait donc mieux s'optimiser. Par ailleurs, les résultats des modèles semblaient beaucoup moins instables entre les différents jeux de valeurs initiales (Figure 4.11) avec plusieurs optimisations successives (différence maximum entre deux graines d'un même modèle = 981.9 pour le modèle Ksp.spgroup lv2sp), qu'avec une seule optimisation (différence maximum entre deux graines d'un même modèle = 8663.0 pour le modèle Kgroup lv4sp). Une fois de plus, les modèles sélectionnés différaient en fonction de la graine choisie (Figure 4.12), avec des conclusions écologiques potentiellement très différentes (modèle sélectionné avec la première graine : modèle K avec lv5sp ; modèle sélectionné avec la seconde graine : K1 avec lv4sp).

Table 4.3: AICc des 35 modèles appliqués aux donnés pour chaque graine testée avec une optimisation multiple à l'aide de tioptimTMB.

Traitement de l'asymptote haute	Graine	nolv	lv1sp	lv2sp	lv3sp	lv4sp	lv5sp	lv6sp
K1	s1	14838.7	14005.3	13584.4	13282.1	13087.4	13073.9	13239.1
	s2	14838.7	14014.9	13584.2	13235.7	13087.5	13135.0	13362.7
K	s1	14840.9	14008.0	13569.6	13441.5	13576.7	**13033.2**	13606.8
	s2	14840.9	14010.8	13550.9	13356.5	13686.9	13641.9	13396.7
	s3	-	-	-	-	13544.7	-	-
Kgroup	s1	14933.3	14096.2	13633.9	13341.5	13346.5	13412.9	13624.2
	s2	14933.3	14093.5	13642.3	13376.2	13193.5	13414.6	13624.5
Ksp	s1	14859.2	14042.9	13757.8	13477.0	13499.5	13676.8	13875.2
	s2	14863.9	14088.8	13708.7	13595.3	13618.8	13676.8	14128.2
Kspspgroup	s1	14089.7	13828.8	14208.8	13894.0	13417.0	14040.3	14364.7
	s2	14063.6	13576.8	13226.8	13729.2	13908.4	14258.3	14434.8
	s3	-	-	13398.2	-	13645.8	-	-

En souligné, la valeur d'AICc du meilleur modèle par graine. En gras et souligné, la valeur d'AICc du meilleur de tous les modèles.

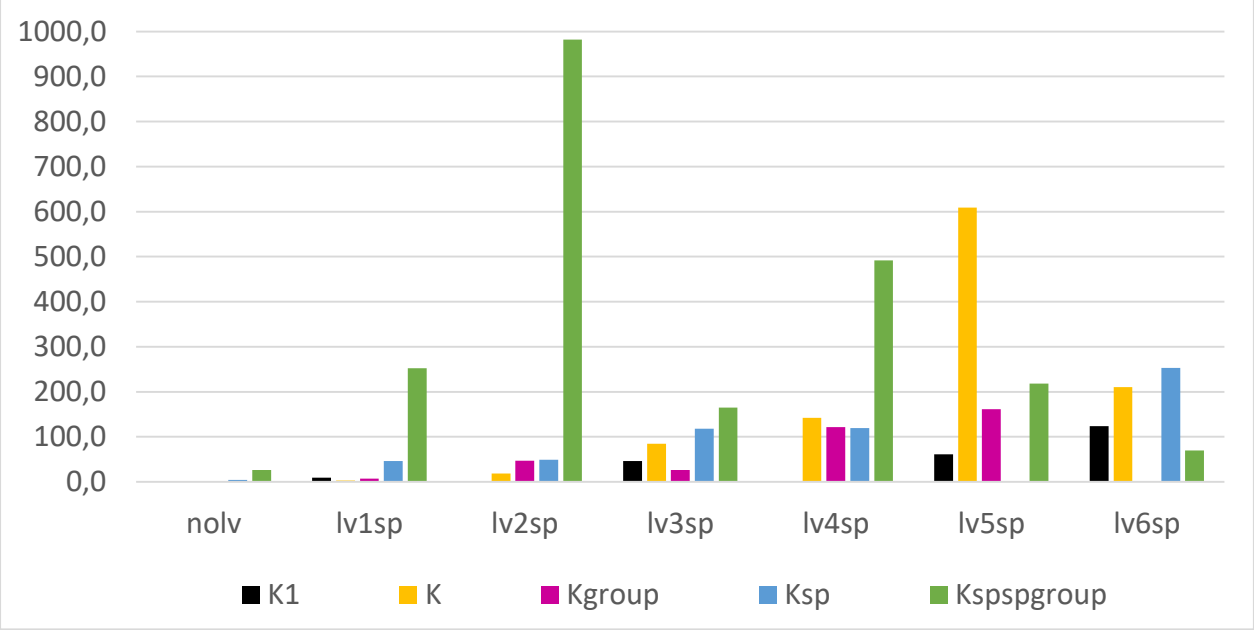


Figure 4.11: Différence entre la valeur maximale et la valeur minimale d'AICc obtenues avec les deux ou trois premières graines des 35 modèles appliqués aux donnés avec de zéro à six variables latentes, avec une optimisation multiple à l'aide de tioptimTMB.

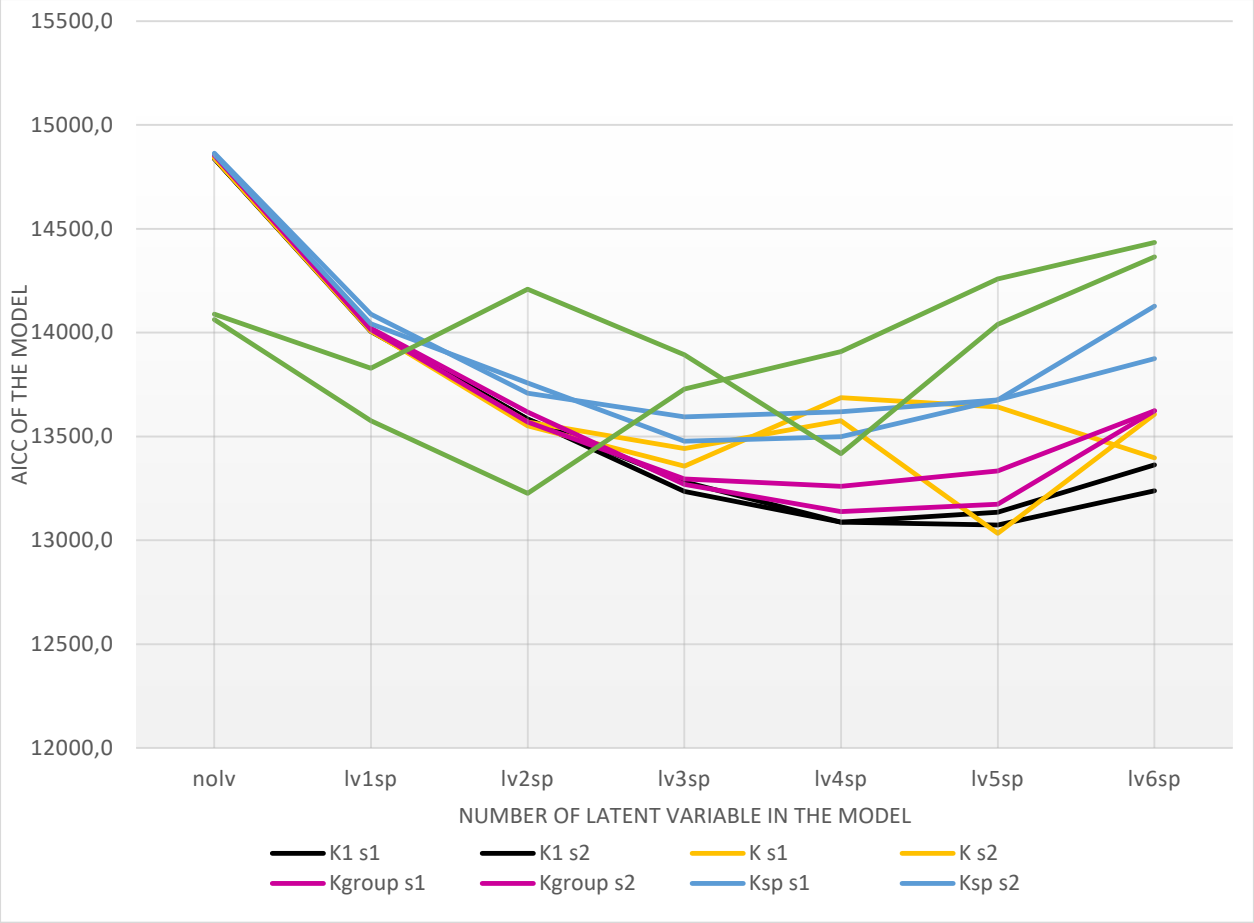


Figure 4.12: Valeur de l'AICc des différents modèles en fonction du nombre de variables latentes pour les deux premières graines uniquement.

Nous avons également comparé les asymptotes hautes estimées par les modèles K et Kgroup afin de constater s'il existait également une instabilité dans cette estimation entre graines et si l'asymptote estimée différait entre les modèles avec différents nombres de variables latentes (nolv à lv6sp). Concernant le modèle K (avec une seule asymptote estimée), l'asymptote estimée par le modèle était très stable entre graines et entre les différents réglages de variables latentes (cf. Annexe IV. Table S1.2). En revanche, concernant le modèle Kgroup, nous avons pu observer que les asymptotes estimées étaient légèrement instables entre les deux graines. Les asymptotes estimées avec les modèles Ksp et Kspspgroup étaient par contre très instables entre les deux

graines (cf.Figure 4.13 ; Annexe IV. Table S1.2 ; Annexe IV. Figure S1.1). De plus, les asymptotes estimées avec le modèle Kgroup variaient légèrement entre les différents réglages de variables latentes, avec notamment une augmentation de la majorité des asymptotes pour un modèle avec trois variables latentes (cf. Figure 4.14 et Annexe IV. Table S1.2).

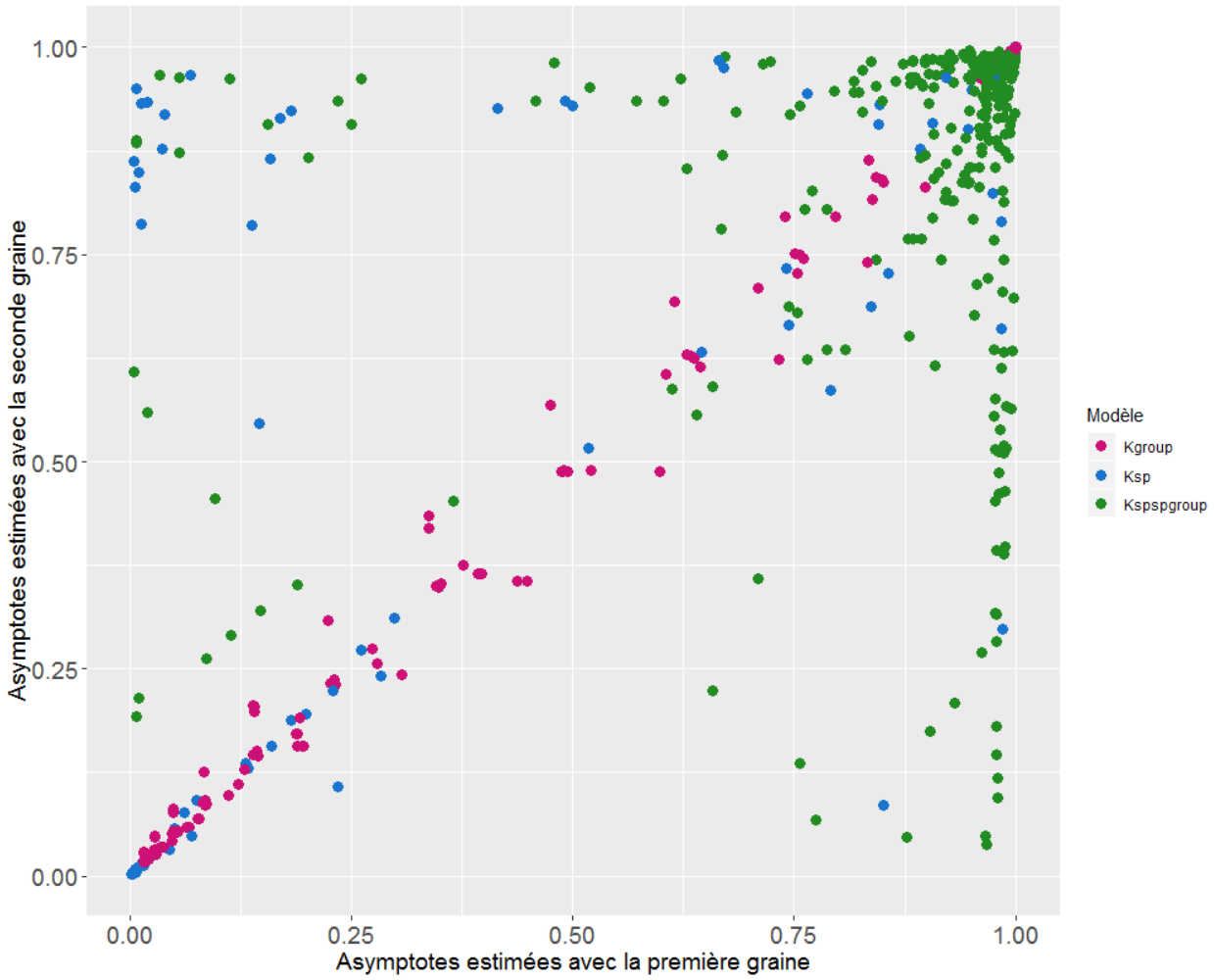


Figure 4.13 : Asymptotes estimées pour les 47 espèces avec la première graine en fonction des asymptotes estimées avec la deuxième graine, pour les trois types de modèles : Kgroup, Ksp, Kspspgroup.

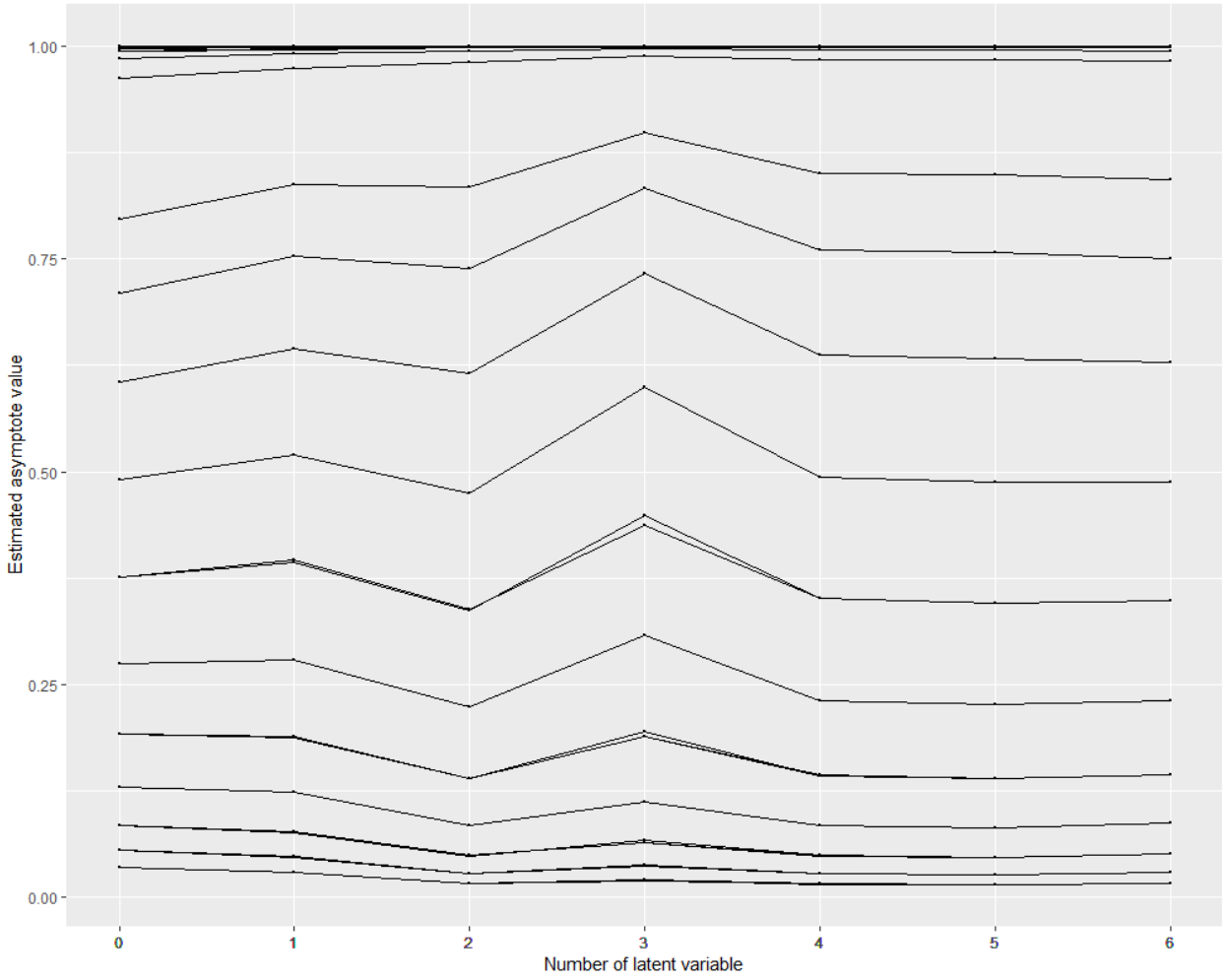


Figure 4.14: Asymptotes estimées des 47 espèces par les modèles avec la première graine, le traitement Kgroup et un nombre variable de variables latentes. 18 espèces possèdent une asymptote supérieure à 0.90 (peu importe le nombre de variables latentes), 10 espèces possèdent une asymptote inférieure à 0.1 (peu importe le nombre de variables latentes) et une espèce possède une asymptote inférieure ou légèrement supérieure à 0.1 (dépendant du nombre de variables latentes).

Nous avons aussi comparé les asymptotes produites par les modèles Kgroup, Ksp et Kspspgroup, afin de constater si les asymptotes estimées (pour un même nombre de variables latentes, une même espèce et une même graine) étaient équivalentes entre ces trois groupes. Nous avons pu observer que les asymptotes estimées par ces trois groupes différaient beaucoup (Figure 4.15).

D’autre part, le modèle Kspspgroup avait tendance à estimer plus d’asymptotes proches de 1.0 que les autres modèles.

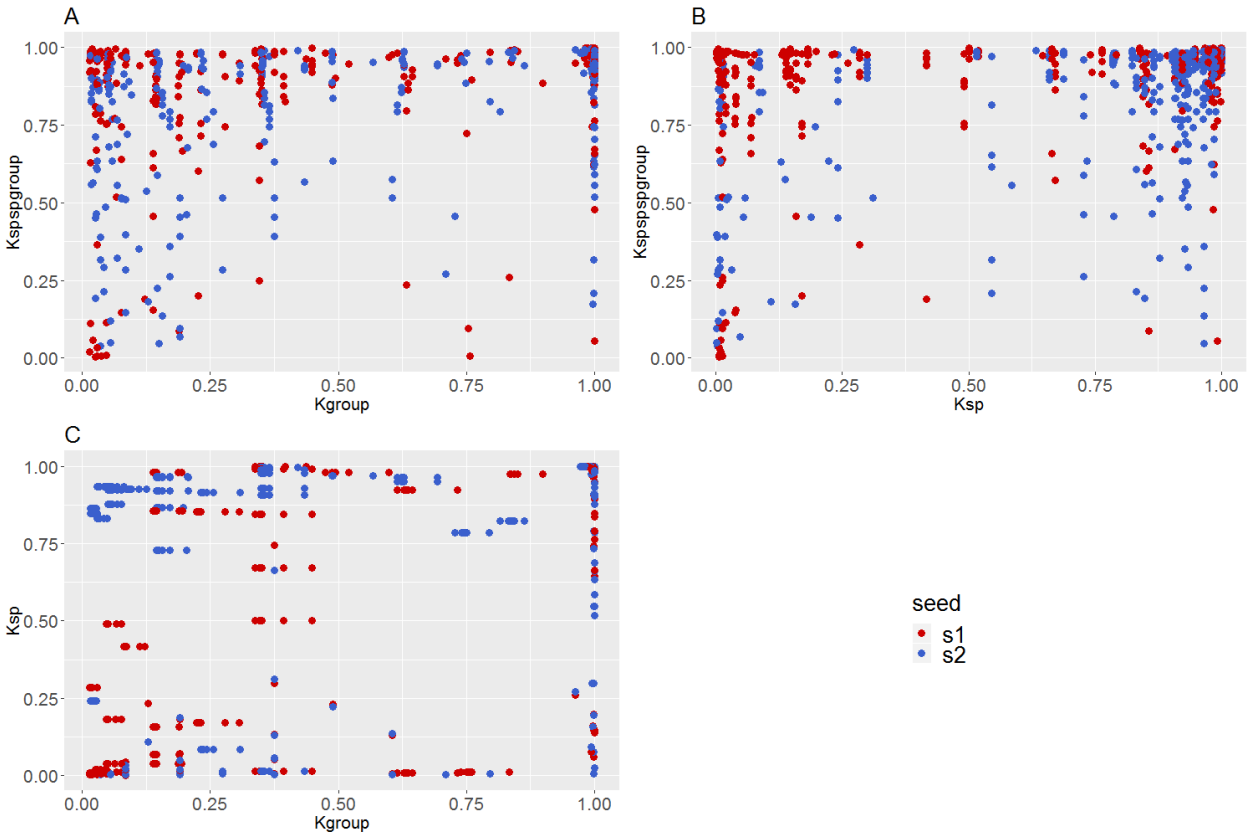


Figure 4.15: Asymptotes estimées par le modèle Kspspgroup en fonction des asymptotes estimées par le modèle Kgroup (A), asymptotes estimées par le modèle Kspspgroup en fonction des asymptotes estimées par le modèle Ksp (B) et asymptotes estimées par le modèle Ksp en fonction des asymptotes estimées par le modèle Kgroup (C).

En conclusion, les modèles restaient très inconstants d’un point de vue de leur capacité prédictive et des paramètres estimés, malgré les efforts mis dans la manière d’optimiser pour tenter de les stabiliser. L’instabilité observée dans les résultats semblait d’autant plus importante que l’asymptote à estimer était complexe (Kspspgroup et Ksp) et que l’on augmentait le nombre de variables latentes. De plus, l’interprétation des asymptotes par espèce était d’autant plus

compliquée que chacun des trois modèles estimaient des asymptotes différentes. En revanche, ils estiment bien tous les trois des asymptotes différentes de 1.0.

Nous avons voulu savoir si cette instabilité dans les résultats pouvait également être due à un autre facteur : les niveaux d'inclusion des variables latentes. Nous avons ici supposé que plus on multiplie et complexifie les niveaux d'inclusions des variables latentes, plus l'optimisation pouvait devenir instable, donnant des résultats variables selon les valeurs initiales.

II. 4. Troisième étape de modélisation : modalités d'inclusion d'effets aléatoires et stabilisation partielle

Nous avons émis l'hypothèse qu'en augmentant le nombre et la complexité des niveaux auxquels une disparité entre les réponses des espèces est supposée apparaître (niveaux d'inclusions des variables latentes), l'estimation de l'asymptote perdrait de l'intérêt car la variabilité de l'asymptote, par rapport à une asymptote fixée à 1.0 et entre groupes ou espèces, pourrait être en partie captée par la variabilité induite par les variables latentes à différents niveaux hiérarchiques. Afin de tester cette hypothèse, nous avons réduit le nombre de modèles à ajuster aux données : (i) en sélectionnant les modèles qui semblaient être les deux meilleurs en termes de traitements d'asymptotes, c'est-à-dire les modèles K et Kgroup, en plus du traitement K1 ; et (ii) en fixant le nombre de variables latentes à quatre, qui semblait donner de bons résultats pour ces trois traitements d'asymptotes. Un nombre de cinq variables latentes aurait également pu être choisi en raison des très bons résultats de ce réglage sur les différents modèles, cependant,

ces deux réglages (lv5sp et lv4sp) donnant des résultats très proches, nous avons choisi celui dont l'optimisation serait la plus rapide.

Sur cette sélection de modèles, nous avons réduit le nombre de niveaux d'inclusion des variables latentes, afin d'observer si cette simplification pouvait donner des résultats différents en termes de choix du meilleur modèle (via la comparaison d'AICc), et réduire l'instabilité entre les graines. Nous avons donc testé six nouvelles modalités d'inclusion des variables latentes en plus des niveaux précédemment testés (plot/season/trap, plot/season, plot/trap et du plot) :

- aux niveaux du plot/season/trap, du plot/season et du plot ;
- aux niveaux du plot/season/trap et du plot ;
- uniquement au niveau du plot/season/trap ;
- aux niveaux du plot/season et du plot ;
- et enfin, uniquement au niveau du plot.

Nous avons donc ajusté un total de 18 modèles pour cette étape (3 traitements d'asymptotes x 1 réglage de variables latentes x 6 modalités d'inclusion des variables latentes). Par ailleurs, les différents $\overline{z'_i}$, correspondant aux différents niveaux d'inclusion des variables latentes, ont tous été désignés en aléatoires (marginalisés avec TMB, en plus des paramètres α_i et $\overline{\beta_{1j}}$).

A l'exception du modèle K1 qui est de manière générale relativement stable, une réduction de la complexité de la modalité d'inclusion des variables latentes (réduction du nombre de niveaux ou de la précision de ces niveaux) avait pour effet de réduire la différence en termes d'AICc entre les graines (Figure 4.16). Seul le modèle avec la modalité plot/season/trap et le traitement K

dérogeait à la règle avec une valeur d'AICc particulièrement haute avec la première graine (Table 4.4). Ainsi, il semblerait qu'une complexification dans l'intégration des variables latentes apportait une instabilité dans l'optimisation du modèle. Pour une modalité d'inclusion des variables latentes donnée, les trois modèles étudiés étaient classés de la même manière selon leur résultats d'AICc peu importe la graine choisie (à condition d'exclure la valeur problématique d'AICc de la première graine du modèle avec la modalité plot/season/trap et le traitement K). En revanche, le meilleur modèle général sélectionné différait toujours en fonction de la graine choisie (le modèle avec la modalité plot/season/trap, plot/season, plot/trap et plot et le traitement K1 avec la première graine contre le modèle avec la modalité plot/season/trap, plot/season et plot, et le traitement Kgroup avec la seconde graine). Enfin, bien qu'une simplification au niveau de l'inclusion des variables latentes apporte plus de stabilité dans l'optimisation, elle s'accompagne aussi globalement d'une diminution de la capacité prédictive du modèle (augmentation de la valeur d'AICc), surtout lorsque l'on retire le niveau d'inclusion le plus complexe : plot/season/trap. En effet, la valeur d'AICc augmente de plus de 1140 points (graine1, modèles K1, K et Kgroup) en passant d'un modèle avec la modalité plot/season/trap, plot/season et plot vers la modalité plot/season et plot. De la même manière, la valeur d'AICc augmente de plus de 1080 points (graine2, modèles K1, K et Kgroup) lorsque l'on passe d'un modèle avec la modalité plot/season/trap et plot vers un modèle avec uniquement plot comme niveau d'inclusion des variables latentes.

Table 4.4: AICc des 18 modèles appliqués aux donnés pour chaque graine testée avec une optimisation multiple à l'aide de tioptimTMB

Traitement de l'asymptote haute	Seed	plot/season/trap pot/season plot/trap plot	plot/season/trap plot/season plot	plot/season/trap plot	plot/season/trap	plot/season plot	plot
K1	s1	13087.4	13106.3	NaN	13381.8	14248.8	14217.9
	s2	13087.5	13107.3	13120.4	13385.4	14248.8	14217.9
	s3	-	-	13111.7	-	-	-
K	s1	13576.7	12994.6	13121.0	16123.1	14253.0	14220.3
	s2	13686.9	13089.3	13137.1	13479.8	14251.1	14220.3
	s3	-	-	-	13464.2	-	-
Kgroup	s1	13265.4	12988.8	13142.5	13415.7	14267.2	14234.2
	s2	13224.6	13199.9	13098.1	13407.3	14267.2	14234.2

En souligné, la valeur d'AICc du meilleur modèle par graine. En gras et souligné, la valeur d'AICc du meilleur de tous les modèles.

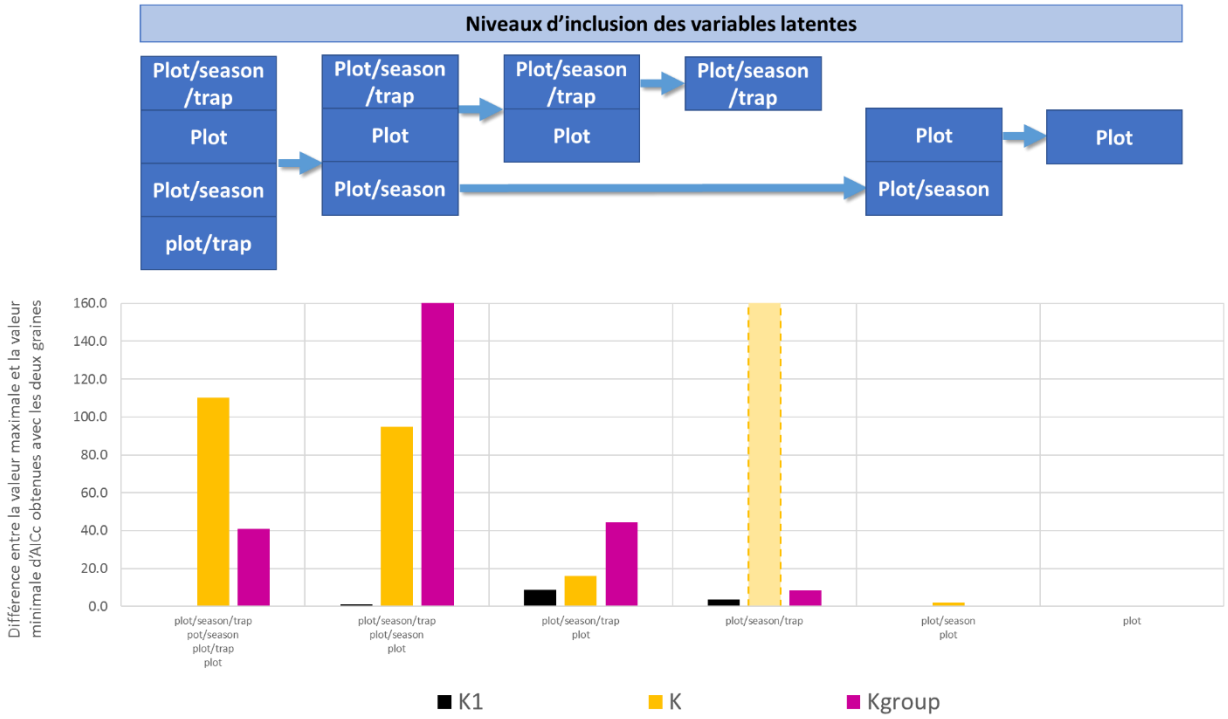


Figure 4.16: Graphique reprenant, pour chacun des 18 modèles, leurs niveaux d'inclusions des variables latentes et un histogramme représentant la différence entre la valeur maximale et la valeur minimale d'AICc obtenues avec les deux graines. Ces 18 modèles ont été appliqués aux données avec une optimisation multiple à l'aide de tioptimTMB. La barre du modèle avec le traitement K et la modalité plot/season//trap n'est pas représentée entièrement afin de faciliter la lecture du graphique, et atteint normalement une valeur de 2658.9.

En explorant les asymptotes estimées par les différents modèles en fonction de la modalité d'inclusion des variables latentes (Annexe IV. Table S1.2), nous avons pu remarquer que celles-ci étaient relativement stables, mais fluctuaient tout de même un petit peu. Pour un modèle avec des variables latentes uniquement au niveau plot/season/trap, les asymptotes avaient généralement tendance à augmenter, sauf les asymptotes déjà au-dessus de 0.75 qui avaient elles plutôt tendance à diminuer (Figure 4.17).

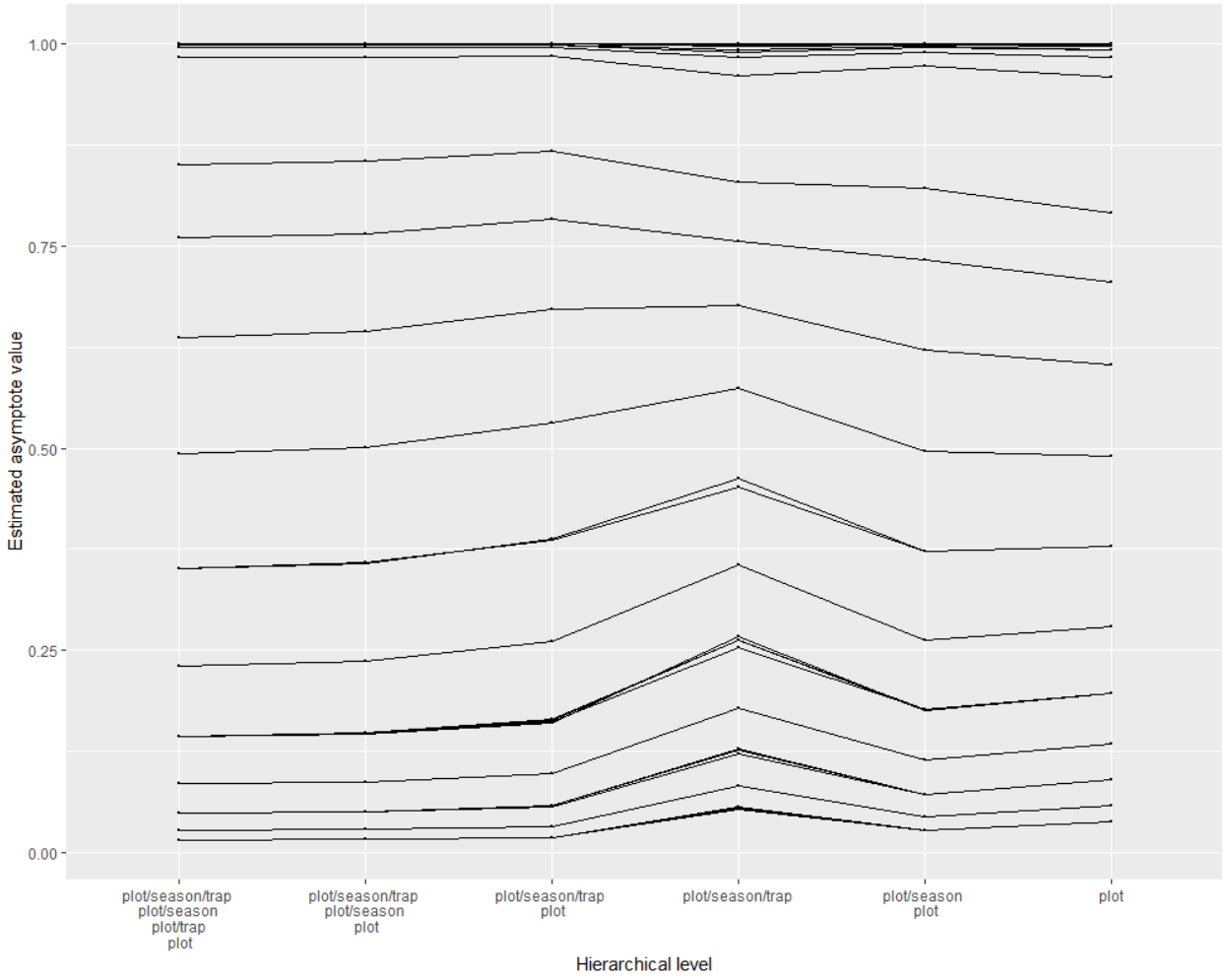


Figure 4.17: Asymptotes estimées des 47 espèces par les modèles avec le traitement Kgroup, la première graine, 4 variables latentes et différentes modalités d'inclusion variables latentes. 18 espèces possèdent une asymptote supérieure à 0.90 (peu importe le niveau d'inclusion des variables latentes), 6 espèces possèdent une asymptote inférieure à 0.1 (peu importe le niveau d'inclusions des variables latentes) et 5 espèces possèdent une asymptote inférieure ou légèrement supérieure à 0.1 (dépendant du niveau d'inclusion des variables latentes).

II. 5. Quatrième étape de modélisation : exclusion des espèces rares et incidence sur la capacité prédictive

Enfin, notre dernière hypothèse était que les espèces rares incluses dans le jeu de données pouvaient éventuellement rendre instable l'optimisation des modèles. En effet, en raison d'une quantité d'information très limitée contenue dans le jeu de données sur ces espèces, le modèle pourrait optimiser différents jeux de paramètres pour celles-ci.

Nous avons donc testé de nouveaux les deux premières graines des modèles avec les trois traitements K1, K et Kgroup et quatre variables latentes, et avons fixé la modalité d'inclusion des effets aléatoires qui semblait donner les meilleurs résultats sur les trois traitements : plot/season/trap, plot/season et plot. Nous avons comparé la différence d'AICc entre les deux graines des modèles appliqués aux données avec toutes les espèces (soit 47 espèces, i.e. résultats précédent), aux modèles appliqués aux données pour lesquelles les espèces avec moins de 10 observations ont été exclues (soit uniquement 21 espèces).

L'exclusion des espèces rares des données n'a pas eu pour effet de réduire l'instabilité entre les graines (Table 4.5 et Figure 4.18). En effet, l'instabilité entre les deux graines a diminué pour les modèles K et Kgroup mais a fortement augmenté pour le modèle K1. Ainsi, la suppression de ces 26 espèces rares ne résout pas le problème d'optimisation.

L'inclusion des espèces rares dans les jeux de données semble être l'un des moteurs du choix vers un modèle avec asymptotes estimées et variables entre groupes. En effet, sans les espèces rares,

le modèle sélectionné est le modèle avec K1, tandis qu'avec les espèces rares le meilleur modèle sélectionné est le modèle avec Kgroup. Les espèces rares pourraient donc potentiellement être les espèces incluses dans les groupes avec une asymptote différente de 1.0.

Table 4.5: AICc des 3 modèles appliqués aux donnés avec et sans les espèces rares (<10 observations) pour chaque graine testée avec une optimisation multiple à l'aide de tioptimTMB

Traitement de l'asymptote haute	Graine	Données avec les espèces rares	Données sans les espèces rares
K1	s1	13106.3	11617.8
	s2	13107.3	**11508.7**
K	s1	12994.6	11651.5
	s2	13089.3	11678.3
Kgroup	s1	**12988.8**	11659.4
	s2	13199.9	11575.4

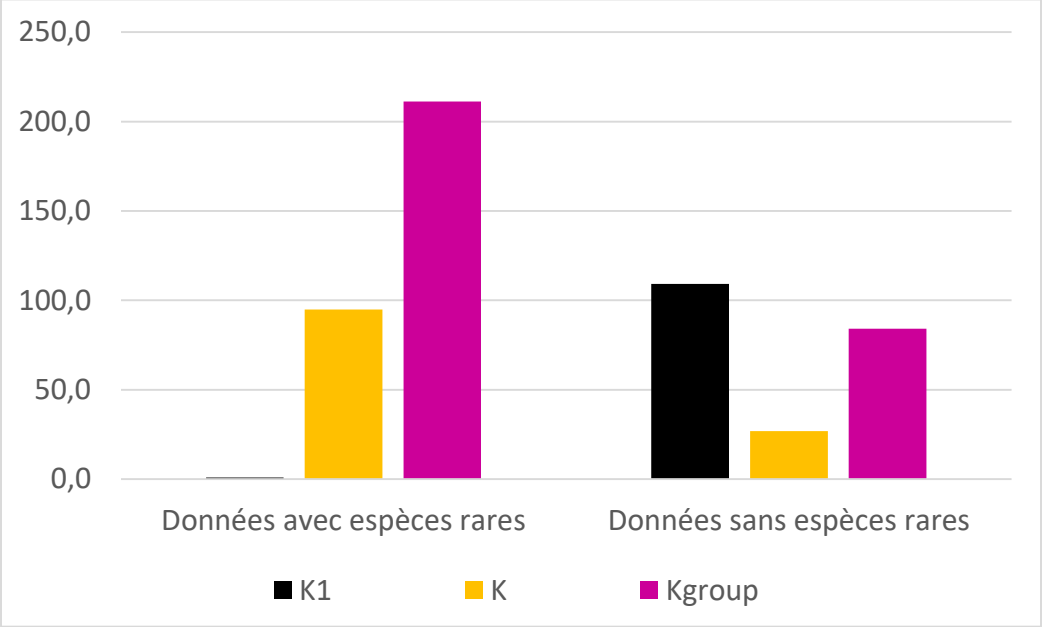


Figure 4.18: Différence entre la valeur maximale et la valeur minimale d'AICc obtenues avec les deux ou trois premières graines des 3 modèles appliqués aux donnés avec et sans les espèces rares, avec une optimisation multiple à l'aide de tioptimTMB.

De manière générale, l'exclusion des espèces rares dans le jeu de données avait pour effet de diminuer notablement les asymptotes estimées des autres espèces non exclues (sauf les espèces dont l'asymptote était proche de 1.0 - cf. Figure 4.19, et Annexe IV. Table S1.4).

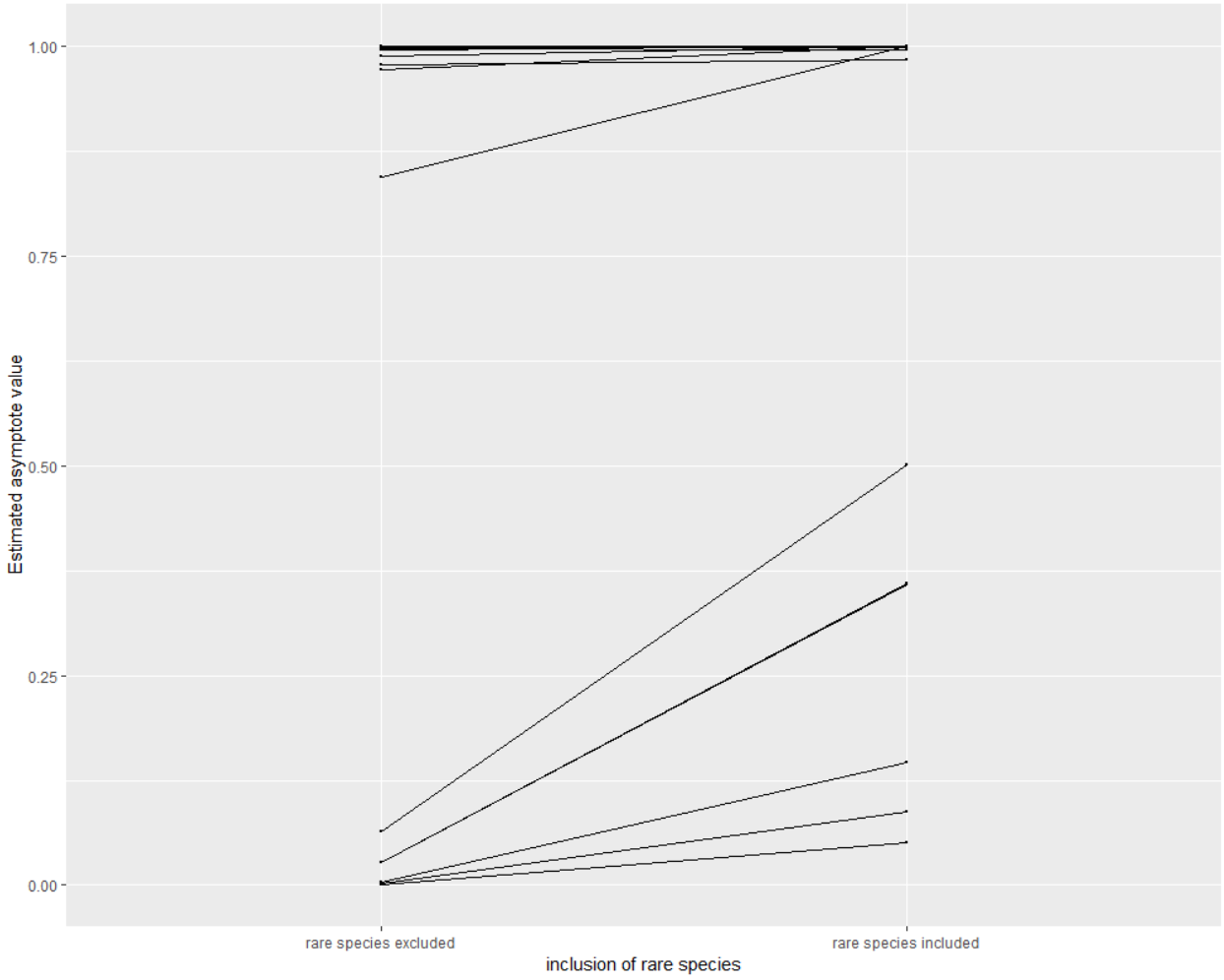


Figure 4.19 : Asymptotes estimées des 21 espèces non rares par les modèles avec le traitement Kgroup, la première graine, 4 variables latentes, un niveau d'inclusion des variables latentes fixes (plot, plot/season, plot/season/trap), et incluant ou non les espèces rares dans les données. 13 espèces possèdent une asymptote supérieure à 0.90 (en incluant ou non les espèces rares), 2 espèces possèdent une asymptote inférieure à 0.1 (en incluant ou non les espèces rares) et 5 espèces possèdent une asymptote inférieure, légèrement supérieure ou fortement supérieure à 0.1 (en fonction de si les données incluent ou non les espèces rares).

II. DISCUSSION

Les modèles de type Joint Species distribution Models (JSDMs) sont de plus en plus populaires car permettent de modéliser la relation des espèces à leur environnement tout en prenant en compte les corrélations entre ces espèces (corrélations dues à leurs interactions ou une réponse commune à une variable environnementale non prise en compte). Mais leur utilisation demeure malgré tout un parcours parsemé d'épreuves et de difficultés.

Le premier défi consiste à choisir le type de modèle multi-espèces que l'on va appliquer aux données (cf. Chapitre 4 - I.2.). Ce choix va dépendre de ce que l'on désire extraire du modèle et des données et ressources à disposition. Par exemple, les modèles synécologiques aléatoires peuvent apporter de nombreuses réponses intéressantes, mais ils sont très gourmands en ressources de calcul et demandent des jeux de données conséquents.

Ensuite, vient le choix de l'approche statistique (fréquentiste ou bayésienne) ainsi que de la plateforme de modélisation avec laquelle les modèles seront optimisés sur les données (librairie déjà écrite, TMB…). Les librairies déjà écrites permettent notamment d'économiser du temps sur la rédaction du modèle en soi, mais il est compliqué de changer certaines caractéristiques imposées. Par exemple, la librairie « HMSC » (permettant de lancer des modèles hiérarchiques de communautés d'espèces ; Tikhonov, 2019) a certes une approche assez intuitive et offre beaucoup de fonctionnalités, mais reste contrainte notamment concernant les fonctions de liens qui sont imposées et peu flexibles. Il n'est par exemple pas possible d'utiliser cette librairie pour caler un modèle avec asymptotes estimées comme nous l'avons fait dans ce chapitre.

Enfin, une fois ces étapes passées, le travail n'est pas encore terminé car il subsiste quelques obstacles. Dans notre cas, nous avons été confrontés à une instabilité des optimisations dès lors que nous tentions de complexifier le modèle (en termes de traitement de l'asymptote, du nombre de variables latentes, ou de la modalité d'intégration des effets aléatoires). Nos résultats semblaient donc orienter vers une simplification des modèles pour obtenir des résultats plus fiables. Or, notre objectif était justement de complexifier les modèles dans le but de les rendre plus précis et réalistes, notamment parce que dans notre cas il s'agissait d'étudier l'impact de la gestion sylvicole sur des espèces d'insectes de grand intérêt, les carabes, afin d'émettre des recommandations de gestion sylvicole qui favoriseraient la diversité de ces espèces et la présence d'espèces rares.

A l'aide d'une librairie développée pour surmonter cette complication (tioptimTMB), nous avons réussi à améliorer légèrement la stabilité de l'optimisation des modèles, mais celle-ci demeure insuffisante. Une réflexion sur les niveaux d'inclusion des variables latentes nous a également permis de stabiliser les modèles, mais il apparaît dommage d'être contraint par de telles considérations lors de la mise en place de modèle, surtout lorsque nous avons de bonnes raisons de penser que la dépendance entre les espèces se joue à différents niveaux hiérarchiques. En revanche, la suppression des espèces rares du jeu de données n'a pas permis de stabiliser les modèles, et leur exclusion diminue l'intérêt des modèles hiérarchiques et JSDMs (Ovaskainen and Soininnen 2011).

Malgré nos efforts, les résultats des modèles ne peuvent pas encore être interprétés d'un point de vue écologique car trop instables d'un point de vue de leur capacité prédictive et des paramètres estimés, et ils sont donc peu fiables. En revanche, ils nous ont permis de constater que les asymptotes des espèces pouvaient bien différer de 1.0 et être variables entre espèces. Les estimations de l'asymptote ne variant que légèrement entre les graines (contrairement aux $\overline{\overline{\beta_1}}$ et $\bar{\lambda}$, cf. Annexe IV. S3), lorsqu'elles dépendent des traits de l'espèce (modèle Kgroup), ce résultat semble relativement robuste et non le produit d'un artefact. De nouvelles tentatives sont toujours en cours, à la fois avec TMB et aussi en utilisant des approches bayésiennes, et nous espérons, à terme, être en mesure de proposer un modèle avec des asymptotes hautes variables et estimées, qui soit stable pour étudier les communautés de carabes. Nous envisageons également de développer nos modèles pour proposer aussi une estimation des asymptotes basses afin de prendre en compte d'autres effets. Par exemple, dans un modèle ou la ressource est incluse sous forme logarithmique (e.g. Williams et al. 2009), alors des processus tels que la « dette à l'extinction », qui décrit le processus temporel de disparition différée d'une espèce lorsque ses ressources disparaissent, pourraient expliquer une probabilité de présence d'une espèce positive à un instant t bien que la ressource soit nulle.

Dans leur article de 2015 et 2016, Thorson et al. utilisent des méthodes comparables à la nôtre pour modéliser la distribution jointe d'espèces. En effet, ils utilisent également un modèle multi-espèces à variables latentes. La librairie « TMB » leur permet aussi d'obtenir les fonctions de vraisemblance via des techniques de différenciation automatique grâce à la fonction « MakeADFun ». Puis ils utilisent l'optimisateur non-linéaire « nlimb » issue de « R core » (R Core

Team 2018), pour maximiser la fonction de vraisemblance marginale. En revanche, Thorson et al. (2015) ont fait un choix différent sur les paramètres mis en aléatoire : le β_{0j} (noté α_i dans leur article) et le $\overline{z'_i}$ (noté Ω dans leur article), en plus d'un paramètre de sur-dispersion (ε_i). La question est donc de savoir si les auteurs ont également une incertitude dans leur modèle, causée soit par la marginalisation de TMB ou l'optimisation de « nlimb », ou bien si les choix qui différent par rapport aux nôtres ont permis d'éviter cette incertitude, tels que le choix des effets aléatoires, ou l'utilisation préalable de INLA pour calculer les valeurs initiales.

En conclusion les JSDMs sont d'un intérêt majeur dans la modélisation des communautés, mais nécessitent encore beaucoup d'améliorations autant pour perfectionner les modèles en soi, en prenant mieux en compte des situations variées s'écartant des situations simples (travail sur les fonctions de liens, la forme des niches…), que pour les rendre efficients et accessibles.

DISCUSSION GENERALE

"*Remember that all models are wrong ; the practical question is how wrong do they have to be to not be useful*."(Box and Draper 1987, p.74).

S'il est admis que tous les modèles sont faux, l'enjeu de la modélisation est d'améliorer ces modèles dans le but de les rendre plus utiles et d'en proposer de nouveaux. Appliqué au domaine de l'écologie, il s'agit de modéliser les écosystèmes afin d'en comprendre le fonctionnement. Les modèles peuvent également être utilisés à des fins plus appliqués telles que la conservation des espèces ou la gestion des ressources. La question du degré d'inexactitude des modèles est donc omniprésente en écologie, et le secteur est en perpétuelle évolution à travers de grands bouleversements ainsi que des améliorations plus discrètes. Afin d'apporter une pierre à l'édifice et d'améliorer la pertinence et l'efficacité des modèles statistiques en écologie, nous avons fixé notre attention sur l'utilisation des fonctions sigmoïdales, en particulier lors de l'étude de la distribution de la biodiversité. En écologie, cette fonction possède un intérêt particulier car elle permet d'ajuster des formes de relation monotone et dont la magnitude n'est pas constante. Lomolino (2000) justifie d'ailleurs son utilisation dans un contexte de relation aire-espèce (SAR) : "with richness remaining relatively low and apparently independent of area for the smaller islands, increasing rapidly to rise through an inflection point for islands of intermediate size, and then asymptotically approaching, or leveling off at the richness of the species pool for the largest islands". L'objectif était donc d'estimer si une extension de ce type de fonction pouvait permettre de rendre les modèles moins faux et plus utiles (au sens de Box and Draper 1987 et Legay 1997).

Face à la diversité des fonctions de forme sigmoïdale, il a été important de définir le terme de sigmoïde à travers une description générale, abordable pour les écologues et qui en facilite leur

usage. Définir les objets permet d'initier des raisonnements autour des répercussions que peuvent avoir les hypothèses auxiliaires, induites par les propriétés de l'objet, sur les résultats des modèles. Parmi les propriétés des fonctions sigmoïdales introduisant des hypothèses auxiliaires, figurent notamment : (i) le caractère fixé, ou variable et estimé, des asymptotes et leurs positions ; (ii) la symétrie forcée autour du point d'inflexion ou l'asymétrie estimée ; (iii) la flexibilité des formes de courbes permise par des variations autour d'une forme moyenne, de la magnitude de la relation, ainsi que des positions des asymptotes et l'asymétrie. Ainsi, nous avons mené ce type de réflexion en intégrant des formes sigmoïdales étendues dans divers types de modèles.

Dans un premier temps, nous avons appliqué une fonction avec une forme sigmoïdale dans un contexte de *Species-Ressource-Relationship* (SReRs). Nous avons étudié la relation entre le volume de bois mort et la communauté de coléoptères saproxyliques, à travers la richesse spécifique, dans divers massifs forestiers français. Nous avons pour cela comparé des fonctions courantes pour ce type d'étude (linéaire, exponentielle, logistique commune) et deux fonctions sigmoïdales, dérivées de la fonction logistique, mais pour lesquelles les positions des asymptotes (logis4p) et l'asymétrie (logis5p) sont estimées à partir des données. Nous avons également testé une grande variété de manières d'intégrer des effets aléatoires, sur les paramètres des fonctions, dans le but de faire varier la forme de la relation entre les massifs forestiers. Les fonctions logistiques avec estimation des asymptotes et de l'asymétrie ont donné des résultats prometteurs incitant à considérer davantage ce type de fonctions dans les études SReRs. Par ailleurs, la

flexibilité des fonctions sigmoïdales et leur capacité à adopter des formes variées en fonction des données, a été d'un intérêt majeur pour l'étude de la relation entre les coléoptères saproxyliques et le volume de bois mort. En effet, nous avons pu mettre en évidence une variation aléatoire dans la forme de la relation entre les différents massifs forestiers, sous-entendant probablement des variations de la relation en fonction de conditions écologiques non prises en compte dans le modèle. Au final, ces modèles ont permis de mettre en place une structure aléatoire de variations entre massifs forestiers, qu'il conviendrait, par la suite, d'expliquer en prenant en compte les variables écologiques pouvant être à l'origine de ces variations.

Fort de cette expérience positive, nous avons ensuite continué notre réflexion sur la position des asymptotes dans des modèles de régression logistique (qui constituent un cas particulier de modèle linéaire généralisé pour lequel la variable à expliquer est binomiale). Ces modèles reposent sur une fonction logistique classique, et n'ont, à notre connaissance, pas bénéficié de réflexions sur l'impact d'asymptotes fixées à zéro et un. À l'aide de données simulées, nous avons étudié le comportement du modèle basique (avec la fonction logistique commune dont les asymptotes sont fixées à 0.0 et 1.0), et de modèles intégrants des fonctions logistiques dont les asymptotes sont estimées à partir des données (notamment le modèle LestKest qui intègre la fonction ELUA logistic – « Estimated Lower & Upper Asymptotes »). Le modèle avec les deux asymptotes estimées a constitué un progrès significatif par rapport au modèle basique, appliqué à des régressions univariées et bivariées, en améliorant à la fois la capacité prédictive, l'adéquation aux données et l'estimation des paramètres, surtout la pente. Nous avons aussi testé le modèle avec la fonction ELUA sur des données réelles et les résultats ont également été

encourageants, puisque la fonction a permis, dans certains cas, d'améliorer la capacité prédictive tout en estimant des asymptotes réellement différentes de zéro et un. Parmi les jeux de données sur lesquels notre modèle avait de meilleurs résultats, comparé au modèle de régression logistique basique, figuraient deux jeux de données appartenant au domaine de l'écologie et plus précisément de l'étude de la distribution des espèces (SDM). Les données étudiées dans le cadre de SDMs constituent des données de présence/absence des espèces en fonction d'un gradient (telle qu'une ressource). Appliquée à ce type de modèle, la fonction ELUA s'ajuste mieux aux cas où la probabilité de présence de l'espèce ne tendrait pas vers 0.0 (pour une ressource tendant vers moins l'infini) et 1.0 (pour une ressource tendant vers plus l'infini).

Au demeurant, d'après Austin (2007), des conclusions sur la courbe de réponse des espèces ne peuvent être clairement déterminées que si le gradient environnemental échantillonné dépasse clairement les limites supérieures et inférieures de la présence de l'espèce. Ainsi, une réponse peut avoir un aspect exponentiel si le gradient étudié est incomplet et que la réponse est uniquement étudiée sur la partie croissance de la courbe sigmoïdale. Les formes de courbes sigmoïdales flexibles, et plus généralement les courbes issues de fonctions non-linéaires, ont donc l'avantage de rendre compte de divers patterns de réponses des espèces en diminuant la perte d'information. Pour autant, de manière générale, lors de la modélisation de la distribution des espèces, peu de modèles non-linéaires paramétriques semblent être appliqués. Malgré l'intérêt indéniable que pourraient présenter les modèles développés par Huisman et al. (1993), notamment pour modéliser les niches écologiques, ceux-ci n'ont pas rencontré le succès attendu. Les résultats obtenus incitent à accorder plus de considérations aux modèles paramétriques et à

la fonction de lien. En effets, nous avons trouvé que l'utilisation des splines dans le cadre d'un GAM (approche semi-paramétrique) utilisant la fonction de lien logistique classique, ne permettait pas systématiquement de compenser le mauvais choix de la fonction de lien.

La modélisation de la distribution d'une espèce est une étape majeure pour étudier une population, mais elle ne permet pas de rendre compte de la communauté. D'autre part, la richesse spécifique (étudiée dans le cadre du deuxième chapitre), est certes une bonne mesure résumant de la diversité d'une communauté, mais elle n'est valide que si la communauté répond de manière emboitée au gradient. Elle ne permet pas de dépeindre les différents patrons sous-jacents possibles et donc de distinguer des compositions des communautés fondamentalement différentes, telles que:

1) une quasi disparition en terme d'abondance de certaines espèces composant la communauté (e.g. Figure 5.1);

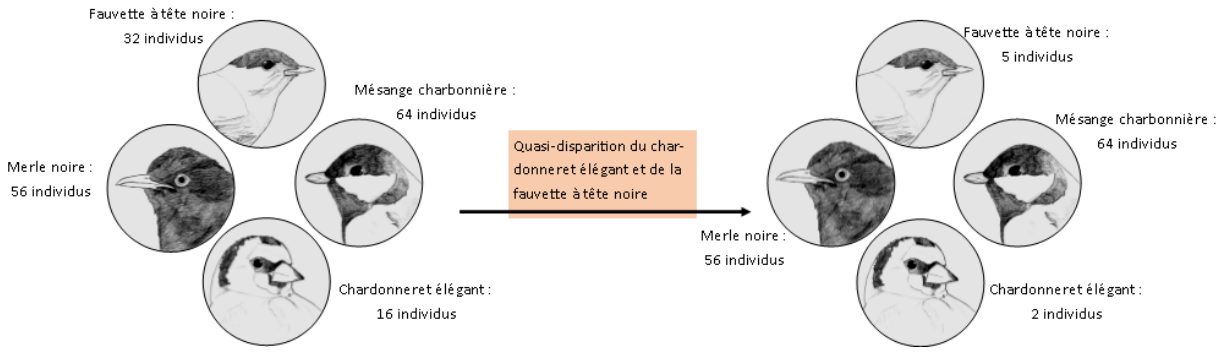


Figure 5.1: Représentation d'une communauté théorique de passereaux, pour laquelle un effondrement de la population de fauvettes à tête noir et de chardonnerets élégant n'est pas observable à l'aide de la métrique de richesse spécifique.

2) un changement dans les espèces composants la communauté (certaines espèces étant remplacées par d'autres), ou deux communautés différentes composée du même nombre d'espèces (e.g. Figure 5.2).

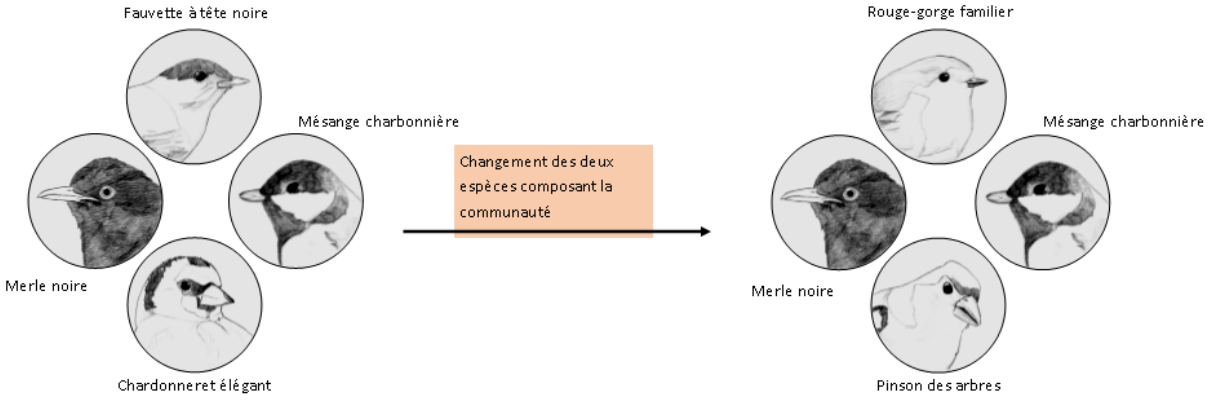


Figure 5.2 : Représentation de deux communautés théoriques de passereaux, que l'étude de la richesse spécifique ne permet pas de distinguer.

Au cours du quatrième chapitre, nous avons donc choisi de traiter une communauté d'espèces en abordant un point de vue véritablement multi-spécifique, permettant, *a minima*, de rendre compte d'une éventuelle permutation d'espèces (Figure 5.2). Il s'agissait donc d'appliquer des modèles de distributions d'espèces à plusieurs espèces en parallèle, à l'aide de modèles dits de distributions d'espèces joints (JSDMs). Cependant, nous avons étudié ces communautés en termes de présence/absence et non d'abondance, ne permettant donc pas de rendre compte de changements tels que des effondrements de populations au sein de la communauté (Figure 5.1).

De manière générale, les modèles JSDMs, tout comme les modèles SDMs, reposent sur des modèles linéaires généralisés, et les données en présence/absence sont étudiées avec des régressions logistiques. Ainsi, dans notre approche de la modélisation multi-espèces, nous avons

également tenté d'aborder les questions liées aux hypothèses auxiliaires, c'est à dire de savoir si une fonction sigmoïde avec asymptotes estimées, remplaçant la fonction logistique dans la loi de probabilité, pouvait également améliorer les résultats du modèle. Malgré un potentiel intérêt de l'asymptote basse, pour cette étape de développement des modèles, nous avons porté notre attention uniquement sur l'asymptote haute. De surcroît, nous nous sommes demandés si l'asymptote haute de cette fonction logistique pouvait être variable entre les espèces composant la communauté, et à quel niveau jouait cette variabilité (au niveau de l'espèce ou du groupe). Pour cela nous avons exploré la relation entre la présence de coléoptères carabiques et les traitements et stades sylvicoles dans la forêt française de Montargis. Contrairement à l'étude menée sur les GLMs (exposée dans le troisième chapitre), les covariables explorées étaient de nature discrète (à l'exception de l'effet site et des variables latentes). L'application du modèle multi-espèces synécologique à variables latentes développé a malheureusement été plus compliqué que prévu et les difficultés rencontrées ne nous ont pas permis de produire des conclusions d'un point de vue écologique sur la communauté étudiée. Cependant, nous avons tout de même pu mettre en évidence que la relation des espèces carabiques avec l'environnement sylvicole étudié semblaient être mieux représentée par un modèle pour lequel l'asymptote haute est estimée et dépend des traits des espèces. Une fois de plus, la forme de la relation est donc bien variable au sein d'un même jeu de données et la flexibilité de la fonction non-linéaire sigmoïdale utilisée nous a permis de mettre cette variabilité en évidence.

D’autres améliorations des JSDMs peuvent être envisagées en menant une réflexion poussée sur des formes non-linéaires des niches écologiques, et sur leur intégration dans les modèles. Ce travail a notamment été initié par les travaux de Huisman et collègues (1993), puis Oksanen et Minchin (2002), qui ont respectivement développé et perfectionné des fonctions qui peuvent être intégrés aux modèles pour mieux s’ajuster aux données et mieux rendre compte de la réalité écologique sous-jacente (e.g. des courbes en cloche asymétrique comme le modèle V proposé par Huisman et al., 1993). Ce travail pourrait être approfondi en explorant davantage les formes de courbes non-linéaires pour représenter les niches écologiques. A titre d’exemple, une forme de niche écologique bimodale semble très peu utilisée, mais pourrait pourtant tout à fait être pertinente (Michaelis and Diekmann 2017). En effet, on définit une communauté composée en partie d’une espèce A qui possède une gamme écologique très grande sur le gradient environnemental considéré. Dans cette même communauté, l’espèce B est compétitrice (ou prédatrice) de l’espèce A, mais a une gamme beaucoup plus resserrée sur la variable environnementale considérée, et centrée sur la moyenne de celle de l’espèce A. Dans de telles conditions, l’espèce A serait abondante au début de sa gamme, puis moins abondante au centre de sa gamme, et enfin de nouveau abondante à la fin de sa gamme. Dans le cas où l’espèce B n’aurait pas été prise en compte dans le modèle (en raison d’un manque de connaissance ou de données par exemple), alors on pourrait observer une réponse bimodale de l’espèce A à la variable environnementale étudiée (Figure 5.3). D’autres formes complexes peuvent être envisagées pour représenter la niche écologique des espèces telles que, par exemple, une courbe de réponse en cloche étendue (Figure 5.4) avec plateau au milieu (la fonction utilisée pourrait être le produit de deux logistiques de sens inverse).

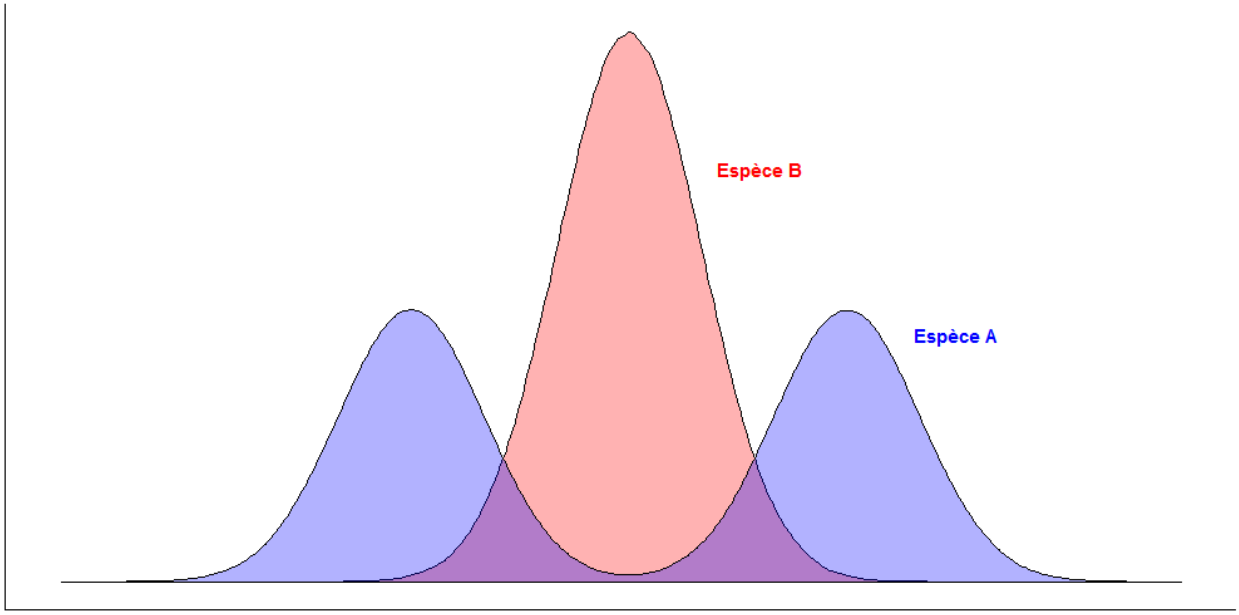


Figure 5.3 : Exemple de courbe bimodale pour l'espèce A étudiée et gaussienne pour l'espèce B non prise en compte

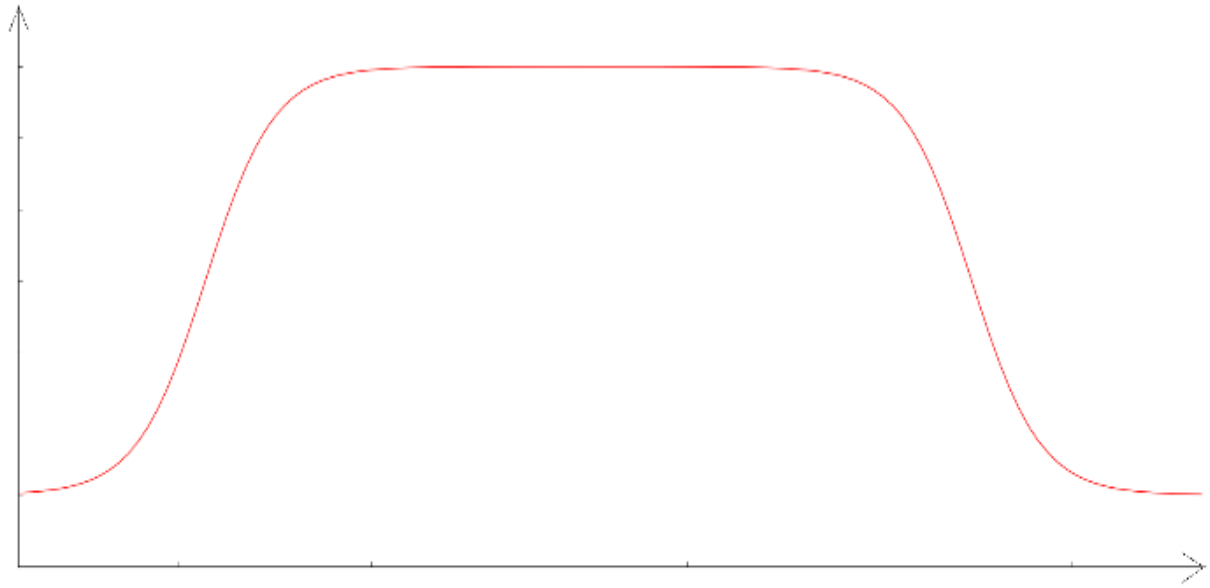


Figure 5.4 : Exemple de courbe en cloche étendue pouvant être obtenue par une combinaison de courbes sigmoïdales.

Initialement le projet de thèse devait s'articuler en trois axes majeurs :

(i) développer une structure de modèle probabiliste qui améliore l'étude des données issues de comptage en incorporant de nouvelles familles de distributions de probabilité de comptage adaptées aux données sous-dispersées.

(ii) proposer et tester d'autres formes de relations non-linéaires pour décrire la relation entre la présence/absence ou l'abondance des espèces et un gradient environnemental.

(iii) permettre d'inclure une forme d'incertitude sur les traits des espèces dans les JSDMs.

Dans le but d'aborder ces divers points, nous avions espéré développer tous ces modèles dans le cadre d'inférence bayésienne. Or, en raison de la longueur de cette méthode d'inférence et de nos ressources de calcul limitées, nous n'avons pu utiliser l'inférence bayésienne que pour développer les modèles les moins complexes et nécessitant le moins de ressources de calcul : les modèles liant la richesse spécifique des coléoptères saproxyliques au volume de bois mort dans les forêts françaises (cf. Chapitre 2). Concernant les deux autres projets (traités dans les Chapitres 3 et 4), nous n'avons eu d'autres choix que de passer à une méthode d'inférence fréquentiste. L'expérience liée à ces deux méthodes d'inférence a plutôt été en faveur des méthodes bayésiennes. En effet, bien que plus longs à converger, les modèles bayésiens étaient beaucoup moins instables et semblaient être mieux optimisés que les modèles fréquentistes, dont les sorties constituaient parfois des valeurs extrêmes et aberrantes pour les GLMs (cf. Chapitre 3) et une instabilité manifeste pour les JSDMs (cf. Chapitre 4). Un accès à des ressources de calculs

infiniment plus importantes permettrait de tester ces modèles avec des méthodes d'inférence bayésienne.

"The history of science, like the history of all human ideas, is a history of irresponsible dreams, of obstinacy, and of error ... This is why we can say that, in science, we often learn from our mistakes, and why we can speak clearly and sensibly about making progress there." (Popper 1963). Cette thèse fut également composée de rêves et d'obstinations ainsi que d'erreurs. Les objectifs initiaux n'ont d'ailleurs pas pu être atteints, mais les échecs les entourant ont permis de soulever d'autres questions et d'approfondir d'autres aspects de la modélisation au profit de l'écologie.

Au final, faute d'avoir pu aborder les points espérés susmentionnés, nous avons plutôt contribué à faire avancer la connaissance sur :

- la définition d'un terme utilisé en écologie, base nécessaire pour faire progresser la connaissance ;
- la réflexion autour des hypothèses auxiliaires des fonctions sigmoïdales utilisées en écologie, et le moyen de s'en affranchir ;
- de potentielles améliorations significatives d'une famille de modèles en particulier : les modèles de régressions logistiques, et ce dans un contexte très large mais aussi lors de leur utilisation en écologie des communautés ;

- le processus d'évaluation des modèles. En effet, en écologie, la méthode généralement employée dans le but de choisir et valider un modèle est la comparaison de modèles à travers la confrontation de la capacité prédictive des modèles. Le Goodness-of-fit est plutôt utilisé pour évaluer un modèle déjà sélectionné, et l'étude de la relation en elle-même (via la significativité et la magnitude des effets) est plutôt utilisée pour traiter *in fine* les sorties d'un modèle. Or, nous avons pu mettre en évidence que ces outils d'analyse peuvent également servir lors de l'étape de comparaison des modèles pour aider à choisir le meilleur modèle. De manière générale, chacun des outils abordés (AIC, GOF, significativité et magnitude) sont complémentaires dans leur recours et devrait se faire conjointement.

D'autres questions ont été soulevées au cours de cette thèse mais nous n'avons pas eu le temps nécessaire pour les traiter, notamment autour de l'application de fonctions non-linéaires dans des modèles étudiant des facteurs limitants (Paris 1992b) et leur application en écologie des communautés (Danger et al. 2008), ou encore un approfondissement des questions autour des SEMs.

BIBLIOGRAPHIE

Adams, D. C., M. S. D. Bitetti, C. H. Janson, L. B. Slobodkin, and N. Valenzuela. 1997. An “Audience Effect” for Ecological Terminology: Use and Misuse of Jargon. Oikos 80:632–636.

Agresti, A. 1990. Categorical Data Analysis. Wiley.

Amrhein, V., S. Greenland, and B. McShane. 2019. Scientists rise up against statistical significance. Nature 567:305–307.

Araújo, M. B., and A. Guisan. 2006a. Five (or so) challenges for species distribution modelling. Journal of Biogeography 33:1677–1688.

Araújo, M. B., and A. Guisan. 2006b. Five (or so) challenges for species distribution modelling. Journal of Biogeography 33:1677–1688.

Araújo, M. B., and M. Luoto. 2007. The importance of biotic interactions for modelling species distributions under climate change. Global Ecology and Biogeography 16:743–753.

Austin, M. P. 1976. On Non-Linear Species Response Models in Ordination. Vegetatio 33:33–41.

Austin, M. P. 1999. A silent clash of paradigms: some inconsistencies in community ecology. Oikos 86:170–178.

Austin, M. P. 2002. Spatial prediction of species distribution: An interface between ecological theory and statistical modelling. Ecological Modelling 157:101–118.

Austin, M. P. 2007. Species distribution models and ecological theory: A critical assessment and some possible new approaches. Ecological Modelling 200:1–19.

Baize, D., and M.-C. Girard. 1992. Référentiel pédologique. Principaux sols d’Europe. INRA Éditions, Versailles.

Begon, M., J. L. Harper, and C. R. Townsend. 1996. Ecology: Individuals, Populations and Communities. Blackwell Science.

Bersier, L.-F., and D. R. Meyer. 1994. Bird assemblages in mosaic forests : the relative importance of vegetation structure and floristic composition along the successional gradient. Acta Oecologica 15:561–576.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. 2014. Semiparametrics. Page *in* N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels, editors. Wiley StatsRef: Statistics Reference Online.

Biggs, R., S. R. Carpenter, and W. A. Brock. 2009. Spurious Certainty: How Ignoring Measurement Error and Environmental Heterogeneity May Contribute to Environmental Controversies. BioScience 59:65–76.

Birch, C. 1999. A New Generalized Logistic Sigmoid Growth Equation Compared with the Richards Growth Equation. Annals of Botany 83:713–723.

Blackburn, T. M., P. Pyšek, S. Bacher, J. T. Carlton, R. P. Duncan, V. Jarošík, J. R. U. Wilson, and D. M. Richardson. 2011. A proposed unified framework for biological invasions. Trends in Ecology & Evolution 26:333–339.

Blondel, J. 1995. Biogéographie: approche écologique et évolutive. Masson.

Bolker, B. M. 2008. Ecological Models and Data in R. Princeton University Press, New Jersey, United States.

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

Bolker, B. M., B. Gardner, M. Maunder, C. W. Berg, M. E. Brooks, L. Comita, E. Crone, S. Cubaynes, T. Davies, P. de Valpine, J. Ford, O. Gimenez, M. Kéry, E. J. Kim, C. Lennert-Cody, A.

Magnusson, S. Martell, J. Nash, A. Nielsen, J. Regetz, H. Skaug, and E. Zipkin. 2013. Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS. Methods in Ecology and Evolution.

Bonneil, P. 2005. Diversité et structure des communautés de Lépidoptères nocturnes en chênaie de plaine dans un contexte de conversion vers la futaie régulière. Thèse de doctorat en Écologie, Muséum national d'histoire naturelle, Paris.

Bouget, C. 2004. Chablis et diversité des coléoptères en forêt feuillue de plaine : impact à court terme de la trouée, de sa surface et de son contexte paysager. Thèse de doctorat en Écologie, Muséum national d'histoire naturelle, Paris.

Bouget, C., and M. Gosselin. 2017. 5- Effet des caractéristiques de peuplement et de naturalité biolgique sur la biodiversité - Quelles implications possibles pour les stratégies de gestion ? Rendez-vous techniques ONF 56:44–50.

Bouget, C., L. Larrieu, and A. Brin. 2014. Key features for saproxylic beetle diversity derived from rapid habitat assessment in temperate forests. Ecological Indicators 36:656–664.

Bouleau, N. 1999. Philosophies des mathématiques et de la modélisation : Du chercheur à l'ingénieur. L'Harmattan, Paris (France).

Box, G. E. P., and N. R. Draper. 1987. Empirical model-building and response surfaces. Wiley.

Brown, A. M., D. I. Warton, N. R. Andrew, M. Binns, G. Cassis, and H. Gibb. 2014. The fourth-corner solution – using predictive models to understand how species traits interact with the environment. Methods in Ecology and Evolution 5:344–352.

carabids.org. (n.d.). . https://www.carabids.org/portal/en-us/home/.

Caradima, B., N. Schuwirth, and P. Reichert. 2019. From individual to joint species distribution models: A comparison of model complexity and predictive performance. Journal of Biogeography:jbi.13668.

Chase, J. M., M. A. Leibold, and U. of C. Press. 2003. Ecological Niches: Linking Classical and Contemporary Approaches. University of Chicago Press.

Cherruault, Y. 1998. Modèles et méthodes mathématiques pour les sciences du vivant. Presses universitaires de France, Paris.

Chevalier, R. 2003. Sylviculture du chêne et biodiversité végétale spécifique : Étude d'une forêt en conversion vers la futaie régulière : la forêt domaniale de Montargis (45). École pratique des hautes études, Paris (France).

Clark, J. S., and A. E. Gelfand. 2006. A future for models and data in environmental science. Trends in Ecology and Evolution 21:375–380.

Codd, C. L., and R. Cudek. 2014. Nonlinear random-effects mixture models for repeated measures. Psychometrika 79:60–83.

Coelho, M. T. P., J. A. Diniz-Filho, and T. F. Rangel. 2019. A parsimonious view of the parsimony principle in ecology and evolution. Ecography 42:968–976.

Collett, D. 2002. Modelling Binary Data. Second Ed. CRC Press.

Colwell, R. K. . 2009. III.1 Biodiversity: Concepts, Patterns, and Measurement. Pages 257–263 Princeton Guide to Ecology. Princeton University Press.

Connor, E. F., and E. D. McCoy. 2001. Species-Area Relationships. Ecyclopedia of Biodiversity 5:397–411.

Coulon, J., P. Marchal, R. Pupier, P. Richoux, R. Allemand, L. C. Genest, and J. Clary. 2000. Coléoptères de Rhône-Alpes. Carabiques et cicindèles. Museum d'Histoire Naturelle de Lyon, Société Linnéenne de Lyon.

Cox, D. R. 1990. Role of Models in Statistical Analysis. Statistical Science 5:169–174.

Cox, D. R. 2018. Analysis of Binary Data. Second Ed. Routledge.

Cressie, N., C. A. Calder, J. S. Clark, J. M. Ver Hoef, and C. K. Wikle. 2009. Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecological Applications 19:553–570.

Cripps, E., and M. Pecht. 2017. A Bayesian nonlinear random effects model for identification of defective batteries from lot samples. Journal of Power Sources 342:342–350.

Danger, M., T. Daufresne, F. Lucas, S. Pissard, and G. Lacroix. 2008. Does Liebig's law of the minimum scale up from species to communities? Oikos.

Dauffy-Richard, E., L. Bergès, P. Bonneil, R. Chevalier, and F. Gosselin. 2010. Conversion de chênaies en futaie régulière : quel impact sur la biodiversité ? Illustration en forêt domaniale de Montargis. Rendez-vous techniques ONF hors-série 5:36–44.

Dauvin, J.-C., G. Bellan, and D. Bellan-Santini. 2008. The need for clear and comparable terminology in benthic ecology. Part I. Ecological concepts. Aquatic Conservation: Marine and Freshwater Ecosystems 18:432–445.

Dengler, J. 2009a. Which function describes the species-area relationship best? A review and empirical evaluation. Journal of Biogeography 36:728–744.

Dengler, J. 2009b. Which function describes the species-area relationship best? A review and empirical evaluation. Journal of Biogeography 36:728–744.

Denizot, G. 1971. Carte géologique au 1/50 000. Montargis XXIV - 19. Carte + Notice, BRGM, Orléans.

Dickey-Collas, M., M. R. Payne, V. M. Trenkel, and R. D. M. Nash. 2014. Hazard warning: model misuse ahead. ICES Journal of Marine Science 71:2300–2306.

Drakare, S., J. J. Lennon, and H. Hillebrand. 2006. The imprint of the geographical, evolutionary and ecological context on species-area relationships. Ecology Letters 9:215–227.

Drikvandi, R. 2017. Nonlinear mixed-effects models for pharmacokinetic data analysis: assessment of the random-effects distribution. Journal of Pharmacokinetics and Pharmacodynamics 44:223–232.

Driscoll, D. A., and D. B. Lindenmayer. 2012. Framework to improve the application of theory in ecology and conservation. Ecological Monographs 82:129–147.

Dubbeldam, J. L. 2007. An annotated bibliography of C.J. van der Klaauw with notes on the impact of his work. Acta Biotheoretica 55:1–22.

Duhem, P. M. M. 1981. La théorie physique. Son objet, sa structure. Second Ed. Vrin, Paris (France).

Fattorini, S., E. Maurizi, and A. D. Giulio. 2012. Tackling the taxonomic impediment: A global assessment for ant-nest beetle diversity (Coleoptera: Carabidae: Paussini). Biological Journal of the Linnean Society 105:330–339.

Fauth, J. E., J. Bernardo, M. Camara, W. J. Resetarits, J. Van Buskirk, and S. A. McCollum. 1996. Simplifying the Jargon of Community Ecology: A Conceptual Approach. The American Naturalist 147:282–286.

Fisher, B., R. Costanza, R. K. Turner, and P. Morling. 2007. Defining and classifying ecosystem services for decision making. CSERGE Working Paper EDM, No. 07-04, University of East Anglia, The Centre for Social and Economic Research on the Global Environment (CSERGE), Norwich.

Fortin, M. 2013. Population-averaged predictions with generalized linear mixed-effects models in forestry: an estimator based on Gauss−Hermite quadrature. Canadian Journal of Forest Research 43:129–138.

Fox, J. 2011a. Contrarian ecology and why we need it. http://www.oikosjournal.org/blog/contrarian-ecology-and-why-we-need-it.

Fox, J. 2011b. Contrarian ecology and why we need it. http://www.oikosjournal.org/blog/contrarian-ecology-and-why-we-need-it.

Gelfand, A. E., A. M. Schmidt, S. Wu, J. A. Silander, A. Latimer, and A. G. Rebelo. 2005. Modelling species diversity through species level hierarchical modelling. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54:1–20.

Gelman, A. 2004. Parameterization and Bayesian Modeling. Journal of American Statistical Association 99:537–545.

Gelman, A. 2011, Summer. Economic Divisions and Political Polarization in Red and Blue America. Pathways:3–6.

Gelman, A., D. Park, B. Shor, and J. Cortina. 2010. Red State, Blue State, Rich State, Poor State. Princeton University Press.

Gentile, G., and R. Argano. 2005. Island biogeography of the Mediterranean sea: The species-area relationship for terrestrial isopods. Journal of Biogeography 32:1715–1726.

Gimenez, O., S. T. Buckland, B. J. T. Morgan, N. Bez, S. Bertrand, R. Choquet, S. Dray, M.-P. Etienne, R. Fewster, F. Gosselin, B. Mérigot, P. Monestiez, J. M. Morales, F. Mortier, F. Munoz, O. Ovaskainen, S. Pavoine, R. Pradel, F. M. Schurr, L. Thomas, W. Thuiller, V. Trenkel, P. de Valpine, and E. Rexstad. 2014. Statistical ecology comes of age. Biology letters 10:20140698.

Godeau, U., C. Bouget, J. Piffady, T. Pozzi, and F. Gosselin. (n.d.). The importance of being random! Taking full account of random effects in nonlinear sigmoid hierarchical Bayesian models.

Godeau, U., C. Bouget, J. Piffady, T. Pozzi, and F. Gosselin. (n.d.). The importance of being random! Taking full account of random effects in nonlinear sigmoid hierarchical Bayesian models.

Godeau, U., C. Bouget, J. Piffady, T. Pozzi, and F. Gosselin. (n.d.). Lack of definition of mathematical terms in ecology: the case of the sigmoid class of functions in biogeography and community ecology.

Gosselin, F. 1997. Modèles stochastiques d'extinction de population : propriétés mathématiques et leurs applications. Université Paris 6.

Gosselin, F. 2001. Lorenz partial order: the best known logical framework to define evenness indices. Community Ecology 2:197–207.

Gosselin, F. 2008. Redefining ecological engineering to promote its integration with sustainable development and. Ecological Engineering 32:199–205.

Gosselin, F. 2009. Management on the basis of the best scientific data or integration of ecological research within management? Lessons learned from the Northern spotted owl saga on the connection between research and management in conservation biology. Biodiversity and Conservation.

Gosselin, F. 2011. A New Calibrated Bayesian Internal Goodness-of-Fit Method: Sampled Posterior p-Values as Simple and General p-Values That Allow Double Use of the Data. PLoS ONE 6.

Grêt-Regamey, A., S. E. Rabe, R. Crespo, S. Lautenbach, A. Ryffel, and B. Schlup. 2014. On the importance of non-linear relationships between landscape patterns and the sustainable provision of ecosystem services. Landscape Ecology 29:201–212.

Grove, S. J. 2002. Tree basal area and dead wood as surrogate indicators of saproxylic insect faunal integrity: a case study from the Australian lowland tropics. Ecological Indicators 1:171–188.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: Offering more than simple habitat models. Ecology Letters.

Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecological Modelling 135:147–186.

Hachich, N. F., M. B. Bonsall, E. M. Arraut, D. R. Barneche, T. M. Lewinsohn, and S. R. Floeter. 2015. Island biogeography: patterns of marine shallow-water organisms in the Atlantic Ocean. Journal of Biogeography 42:1871–1882.

Hand, D. J. 1998. Data Mining: Statistics and More? The American Statistician 52:112–118.

Hanski, I. A., and M. E. Gilpin. 1997. Metapopulation biology : ecology, genetics and evolution. Page (I. A. Hanski and M. E. Gilpin, Eds.). Academic Press, San Diego, CA.

Harrell, F. E. Jr. 2001. Regression Modeling Strategies. First Ed. Springer-Verlag New York, New York.

Harris, D. J. 2015. Generating realistic assemblages with a joint species distribution model. Methods Ecol. Evol 6:465–473.

Hastie, T. J. 2017. Generalized Additive Models. Routledge.

Hastie, T., and R. Tibshirani. 1986. General Additive Models. Statistical Science 1:297–318.

He, F., and P. Legendre. 1996. On species-area relations. American Naturalist 148:719–737.

He, F., and P. Legendre. 2002. Species diversity patterns derived from species-area models. Ecology 83:1185–1198.

Heams, T. 2009. Expression stochastique des gènes et différenciation cellulaire. Page 192 Le hasard au cœur de la cellule. Probabilités, déterminisme, génétique. Syllepse.

Herrando-Pérez, S., B. W. Brook, and C. J. A. Bradshaw. 2014. Ecology Needs a Convention of Nomenclature. BioScience 64:311–321.

Herrando-Pérez, S., B. W. Brook, and C. J. A. Bradshaw. 2017. Ecology Needs a Convention of Nomenclature. BioScience 64:311–321.

Herrando-Pérez, S., S. Delean, B. W. Brook, and C. J. A. Bradshaw. 2012. Density dependence: an ecological Tower of Babel. Oecologia 170:585–603.

Hill, T., and P. Lewicki. 2006. Statistics: Methods and Applications: A Comprehensive Reference for Science, Industry, and Data Mining. StatSoft, Inc., Tulsa.

Hodges, K. E. 2008. Defining the problem: terminology and progress in ecology. Frontiers in Ecology and the Environment 6:35–42.

Huber, P., E. Ronchetti, and M.-P. Victoria-Feser. 2004. Estimation of generalized linear latent variable models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66:893–908.

Hui, F. K. C. 2016. BORAL - Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in R. Methods in Ecology and Evolution 7:744–750.

Huisman, J., H. Olff, and L. F. M. Fresco. 1993a. A Hierarchical Set of Models for Species Response Analysis. Source Journal of Vegetation Science 4:37–46.

Huisman, J., H. Olff, and L. F. M. Fresco. 1993b. A Hierarchical Set of Models for Species Response Analysis. Source Journal of Vegetation Science 4:37–46.

Hunsicker, M. E., C. V. Kappel, K. A. Selkoe, B. S. Halpern, C. Scarborough, L. Mease, and A. Amrhein. 2015. Characterizing driver-response relationships in marine pelagic ecosystems for improved ocean management. Ecological Applications 26:651–663.

Hůrka, K. 1996. Carabidae of the Czech and Slovak Republics. Kabourek, Zlin.

Hurlbert, S. H., R. A. Levine, and J. Utts. 2019. Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires. The American Statistician 73:352–357.

Hutchinson, G. E. 1957. Concluding remarks. Pages 415–427.

IFN. 2011. GRECO B Centre Nord semi-océanique. https://inventaire-forestier.ign.fr/IMG/pdf/Tome_B.pdf.

Ishibashi, T., Y. Yano, and T. Oguma. 2003. Population pharmacokinetics of platinum after nedaplatin administration and model validation in adult patients. British Journal of Clinical Pharmacology 56:205–213.

Jamil, T., W. A. Ozinga, M. Kleyer, and C. J. F. ter Braak. 2013. Selecting traits that explain species-environment relationships: a generalized linear mixed model approach. Journal of Vegetation Science 24:988–1000.

Jax, K. 2005a. Function and '" functioning "' in ecology : what does it mean ? Oikos 111:641–648.

Jax, K. 2005b. Function and ‘“ functioning ”’ in ecology : what does it mean ? Oikos 111:641–648.

Jax, K., and K. E. Hodges. 2008. Concepts, not terms. Frontiers in Ecology and the Environment 6:178–179.

Jeannel, R. 1941. Faune de France. Coléoptères carabiques. Première partie. Office central de Faunistique, Paris.

Kirk, D. A., A. C. Park, A. C. Smith, B. J. Howes, B. K. Prouse, N. G. Kyssa, E. N. Fairhurst, and K. A. Prior. 2018. Our use, misuse, and abandonment of a concept: Whither habitat? Ecology and Evolution 8:4197–4208.

Kobayashi, S. 1976. The Species-area relation - III a third model for delimited community. Researches on Population Ecology 17:243–254.

Komori, O., S. Eguchi, S. Ikeda, H. Okamura, M. Ichinokawa, and S. Nakayama. 2016. An asymmetric logistic regression model for ecological data. Methods in Ecology and Evolution 7:249–260.

Kristensen, K., B. Bell, H. Skaug, A. Magnusson, C. Berg, A. Nielson, M. Maechler, T. Michelot, M. Brooks, A. Forrence, C. M. Albertsen, and C. Monnahan. 2018. TMB: Template Model Builder: A General Random Effect Tool Inspired by “ADMB.” R package version 1.7.15.

Lakatos, I., J. Worrall, and G. Currie. 1980. The Methodology of Scientific Research Programmes: Volume 1: Philosophical Papers. Cambridge University Press.

Lande, R., S. Engen, and S. Bernt-Erik. 2003. Stochastic Population Dynamics in Ecology and Conservation. Oxford University Press.

Lassauce, A., Y. Paillet, H. Jactel, and C. Bouget. 2011. Deadwood as a surrogate for forest biodiversity: Meta-analysis of correlations between deadwood volume and species richness of saproxylic organisms. Ecological Indicators 11:1027–1039.

Le Petit Robert : Sigmoïde. 2017. . https://pr.bvdep.com/robert.asp.

Lebreton, J.-D., F. Gosselin, and C. Niel. 2007. Extinction and viability of populations: Paradigms and concepts of extinction models. Ecoscience 14:472–481.

Legay, J.-M. 1997. L'expérience et le modèle: Un discours sur la méthode. Quae.

Legendre, P., and L. Legendre. 1998. Numerical Ecology. 2nd edition. Elsevier Science.

Levin, S. A. 1992. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. Ecology 73:1943–1967.

Levins, R. 1966. The strategy of model building in population biology. American Scientist 54:421–431.

Levins, R. 1993. A Response to Orzack and Sober: Formal Analysis and the Fluidity of Science. The Quarterly Review of Biology 68:547–555.

Lim, C. C., R. Arora, and E. C. Townsend. 1998. Comparing Gompertz and Richards Functions to Estimate Freezing Injury in Rhododendron Using Electrolyte Leakage. Journal of the American Society for Horticultural Science 123:246–252.

Lindroth, C. H. 1974. Handbook for the identification of British insects. Coleoptera Carabidae. Society of London, London.

Liu, J.-H., Y. Yan, A. Ali, M.-F. Yu, Q.-J. Xu, P.-J. Shi, and L. Chen. 2018. Simulation of crop growth, time to maturity and yield by an improved sigmoidal model. Scientific Reports 8:7030.

Loehle, C. 1983. Evaluation of theories and calculation tools in ecology. Ecological Modelling 19:239–247.

Lomolino, M. V. 2000a. Ecology's most general, yet protean pattern: The species-area relationship. Journal of Biogeography 27:17–26.

Lomolino, M. V. 2000b. Ecology's most general, yet protean pattern: The species-area relationship. Journal of Biogeography 27:17–26.

Maaf, and IGN. 2016. Indicateurs de gestion durable des forêts françaises métropolitaines, édition 2015, Résultats. Page 343. Maaf-IGN, Paris.

MacGregor-Fors, I. 2011. Misconceptions or misunderstandings? On the standardization of basic terms and definitions in urban ecology. Landscape and Urban Planning 100:347–349.

Madin, J. S., S. Bowers, M. P. Schildhauer, and M. B. Jones. 2008. Advancing ecological research with ontologies. Trends in Ecology & Evolution 23:159–168.

Martikainen, P., J. Siitonen, P. Punttila, L. Kaila, and J. Rauh. 2000. Species richness of Coleoptera in mature managed and old-growth boreal forests in southern Finland. Biological Conservation 94:199–209.

Mashayekhi, M., B. MacPherson, and R. Gras. 2014. Species-area relationship and a tentative interpretation of the function coefficients in an ecosystem simulation. Ecological Complexity 19:84–95.

McCarthy, M. A., D. C. Franklin, and M. A. Burgman. 1994. The importance of demographic uncertainty: An example from the helmeted honeyeater Lichenostomus melanops cassidix. Biological Conservation 67:135–142.

McCullagh, P., and J. A. Nelder. 1989. Generalized Linear Models. Second Ed. Chapman & Hall.

McShane, B. B., D. Gal, A. Gelman, C. Robert, and J. L. Tackett. 2019. Abandon Statistical Significance. The American Statistician 73:235–245.

Medellín, R. A., and J. Soberón. 1999. Predictions of Mammal Diversity on Four Land Masses. Conservation Biology 13:143–149.

Menon, A., K. Mehrotra, C. K. Mohan, and S. Ranka. 1996. Characterization of a class of sigmoid functions with applications to neural networks. Neural Networks 9:819–835.

Michaelis, J., and M. R. Diekmann. 2017. Biased niches – Species response curves and niche attributes from Huisman-Olff-Fresco models change with differing species prevalence and frequency. PLOS ONE 12:e0183152.

Middleton, A. La., and E. Prigoda. 2008. What does 'fledging' mean? Ibis 143:296–298.

Miller, G. A. 2003. The cognitive revolution: a historical perspective. Trends in Cognitive Sciences 7:141–144.

Morante-Filho, J. C., D. Faria, E. Mariano-Neto, and J. Rhodes. 2015. Birds in Anthropogenic Landscapes : The Responses of Ecological Groups to Forest Loss in the Brazilian Atlantic Forest. PLoS ONE 10:e0128923.

Moreau, C., C. Barnaud, and R. Mathevet. 2019. Conciliate agriculture with landscape and biodiversity conservation: A role-playing game to explore trade-offs among ecosystem services through social learning. Sustainability, MPDI 11:20.

Muggeo, V. M. R. 2003. Estimating regression models with unknown break-points. Statistics in Medicine 22:3055–3071.

Niku, J., D. I. Warton, F. K. C. Hui, and S. Taskinen. 2017. Generalized Linear Latent Variable Models for Multivariate Count and Biomass Data in Ecology. Journal of Agricultural, Biological and Environmental Statistics 22:498–522.

Oakes, D. 2014. Semi-Parametric Models. Page *in* N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels, editors. Wiley StatsRef: Statistics Reference Online.

Oksanen, J., and P. R. Minchin. 2002. Continuum theory revisited: What shape are species responses along ecological gradients? Ecological Modelling 157:119–129.

ONF. 1971. Forêt domaniale de Montargis. Résumé des aménagements antérieurs. Page 7 p. Document interne ONF.

Ovaskainen, O., N. Abrego, P. Halme, and D. Dunson. 2016a. Using latent variable models to identify large networks of species-to-species associations at different spatial scales. Methods in Ecology and Evolution 7:549–555.

Ovaskainen, O., Gleb Tikhonov, A. Norberg, F. G. Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecology Letters:1–16.

Ovaskainen, O., D. B. Roy, R. Fox, and B. J. Anderson. 2016b. Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. Methods in Ecology and Evolution 7:428–436.

Ovaskainen, O., and J. Soininnen. 2011. Making more out of sparse data : hierarchical modeling of species communities Reports. Ecology 92:289–295.

Padilla, F. M., and F. I. Pugnaire. 2006. The role of nurse plants in the restoration of degraded environments. Frontiers in Ecology and the Environment 4:196–202.

Paillet, Y., F. Archaux, S. du Puy, C. Bouget, V. Boulanger, N. Debaive, O. Gilg, F. Gosselin, and E. Guilbert. 2018. The indicator side of tree microhabitats: A multi-taxon approach based on bats, birds and saproxylic beetles. Journal of Applied Ecology 55:2147–2159.

Paine, C. E. T., T. R. Marthews, D. R. Vogt, D. Purves, M. Rees, A. Hector, and L. A. Turnbull. 2012a. How to fit nonlinear plant growth models and calculate growth rates: An update for ecologists. Methods in Ecology and Evolution 3:245–256.

Paine, C. E. T., T. R. Marthews, D. R. Vogt, D. Purves, M. Rees, A. Hector, and L. A. Turnbull. 2012b. How to fit nonlinear plant growth models and calculate growth rates: An update for ecologists. Methods in Ecology and Evolution 3:245–256.

Paris, Q. 1992a. The von Liebig Hypothesis. American Journal of Agricultural Economics 74:1020–1028.

Paris, Q. 1992b. The von Liebig Hypothesis. American Journal of Agricultural Economics 74:1020–1028.

Pausas, J. G., and M. P. Austin. 2001. Patterns of plant species richness in relation to different environments: An appraisal. Journal of Vegetation Science 12:153–166.

Peters, R. H. 1991. A Critique for Ecology. Cambridge University Press.

Pickett, S. T. A., J. Kolasa, and C. G. Jones. 2007a. Ecological Understanding: The Nature of Theory and the Theory of Nature. Second Edi. Academic Press, San Diego.

Pickett, S. T. A., J. Kolasa, and C. G. Jones. 2007b. Ecological Understanding: The Nature of Theory and the Theory of Nature. Second Edi. Academic Press, San Diego.

Pierson, J. C., P. S. Barton, P. W. Lane, and D. B. Lindenmayer. 2015. Can habitat surrogates predict the response of target species to landscape change? Biological Conservation 184:1–10.

Pillai, G. (Colin), F. Mentré, and J.-L. Steimer. 2005. Non-Linear Mixed Effects Modeling – From Methodology and Software Development to Driving Implementation in Drug Development Science. Journal of Pharmacokinetics and Pharmacodynamics 32:161–183.

Pollock, L. J., R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. Mccarthy. 2014. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). Methods in Ecology and Evolution 5:397–406.

Popper, K. R. 1963. Conjectures And Refutations: The Growth Of Scientific Knowledge. Harper & Row.

Preston, F. W. 1962a. The Canonical Distribution of Commonness and Rarety: Part I. Ecology 43:185–215.

Preston, F. W. 1962b. The Canonical Distribution of Commonness and Rarety: Part I. Ecology 43:185–215.

Quandt, R. E. 1958. The Estimation of the Parameters of a Linear Regression System Obeying Two Separate Regimes. Journal of the American Statistical Association 53:873–880.

Quine, W. V. O. 1951. Two Dogmas of Empiricism. The Philosophical Review 60:20–43.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Radanielson, A. M., O. Angeles, T. Li, A. M. Ismail, and D. S. Gaydon. 2018. Describing the physiological responses of different rice genotypes to salt stress using sigmoid and piecewise linear functions. Field Crops Research 220:46–56.

Ranius, T., and M. Jonsson. 2007a. Theoretical expectations for thresholds in the relationship between number of wood-living species and amount of coarse woody debris: A study case in spruce forests. Journal for Nature Conservation 15:120–130.

Ranius, T., and M. Jonsson. 2007b. Theoretical expectations for thresholds in the relationship between number of wood-living species and amount of coarse woody debris: A study case in spruce forests. Journal for Nature Conservation 15:120–130.

Richard, E. 2004. Réponse des communautés de coléoptères carabiques à la conservation en futaie régulière de chêne : aspects écologique et méthodologiques.

Richard, E., F. Gosselin, and J. Lhonoré. 2003. Short-term and Mid-term Response of Ground Beetle Communities (Coleoptera, Carabidae) to Disturbance by Regeneration Felling. Pages 179–192 Forest biodiversity: lessons from history for conservation. Louvain, BELGIUM.

le Roux, P. C., L. Pellissier, M. S. Wisz, and M. Luoto. 2014. Incorporating dominant species as proxies for biotic interactions strengthens plant community models. Journal of Ecology 102:767–775.

Ruxton, G. D., and H. M. Schaefer. 2011. Resolving current disagreements and ambiguities in the terminology of animal communication: Definitions in animal communication. Journal of Evolutionary Biology 24:2574–2585.

Rykiel, E. J. 1985. Towards a definition of ecological disturbance. Austral Ecology 10:361–365.

Saas, Y., and F. Gosselin. 2014. Comparison of regression methods for spatially-autocorrelated count data on regularly- and irregularly-spaced locations. Ecography 37:476–789.

Sasieni, P. 2014. Semiparametric Regression. Page *in* N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, and J. L. Teugels, editors. Wiley StatsRef: Statistics Reference Online.

Schielzeth, H., and W. Forstmeier. 2009. Conclusions beyond support: Overconfident estimates in mixed models. Behavioral Ecology 20:416–420.

Seibold, S., C. Bässler, R. Brandl, B. Büche, A. Szallies, S. Thorn, M. D. Ulyshen, and J. Müller. 2016. Microclimate and habitat heterogeneity as the major drivers of beetle diversity in dead wood. Journal of Applied Ecology 53:934–943.

Shaw, D. J., and A. P. Dobson. 1995. Patterns of macroparasite abundance and aggregation in wildlife populations: a quantitative review. Parasitology 111:S111–S133.

Shmueli, G. 2010. To Explain or to Predict? Statistical Science 25:289–310.

Slisko, J., and D. I. Dykstra. 1997. The role of scientific terminology in research and teaching: Is something important missing? Journal of Research in Science Teaching 34:655–660.

Sousa, W. P. 1984. The Role of Disturbance in Natural Communities. Annals of Ecology and Systematics 15:353–391.

Starfield, A. M. 1997. A pragmatic approach to modeling for wildlife management. Journal of Wildlife Management 61:261–270.

Thorson, J. T., J. N. Ianelli, E. A. Larsen, L. Ries, M. D. Scheuerell, C. Szuwalski, and E. F. Zipkin. 2016. Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring: Joint dynamic species distribution models. Global Ecology and Biogeography 25:1144–1158.

Thorson, J. T., M. D. Scheuerell, A. O. Shelton, K. E. See, H. J. Skaug, and K. Kristensen. 2015. Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. Methods in Ecology and Evolution 6:627–637.

Tikhonov, G. 2019. Hmsc: Demonstrates the potential user interface and distribution of HMSC package. R package version 0.2.11.3.

Tikhonov, G., N. Abrego, D. Dunson, and O. Ovaskainen. 2017. Using joint species distribution models for evaluating hows pecies-to-species associations depend on the environmental context. Methods in Ecology and Evolution:443–452.

Tjørve, E. 2003a. Shapes and functions of species-area curves: A review of possible models. Journal of Biogeography 30:827–835.

Tjørve, E. 2003b. Shapes and functions of species-area curves: A review of possible models. Journal of Biogeography 30:827–835.

Tjørve, E. 2009a. Shapes and functions of species-area curves (II): A review of new models and parameterizations. Journal of Biogeography 36:1435–1445.

Tjørve, E. 2009b. Shapes and functions of species-area curves (II): A review of new models and parameterizations. Journal of Biogeography 36:1435–1445.

Tjørve, E. 2012. Arrhenius and Gleason revisited: New hybrid models resolve an old controversy. Journal of Biogeography 39:629–639.

Underwood, A. J. 1995. Ecological Research and (and Research into) Environmental Management. Ecological Applications 5:232–247.

Veech, J. A. 2000. Choice of species-area function affects identification of hotspots. Conservation Biology 14:140–147.

Verwijst, T., and H. A. V. Fircks. 1994. Plant response to temperature stress is characterized by an asymmetric sigmoid function. Environmental and Experimental Botany 34:69–74.

Vetter, D., G. Rücker, and I. Storch. 2013. Meta-analysis: A need for well-defined usage in ecology and conservation biology. Ecosphere 4:1–24.

Warton, D. I., F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. C. Hui. 2015. So Many Variables: Joint Modeling in Community Ecology. Trends in Ecology and Evolution.

Wasserstein, R. L., and N. A. Lazar. 2016. The ASA Statement on p-Values: Context, Process, and Purpose. The American Statistician 70:129–133.

Wasserstein, R. L., A. L. Schirm, and N. A. Lazar. 2019. Moving to a World Beyond "p<0.05." The American Statistician 73:1–19.

Wikipedia. (n.d.). Sigmoid function. https://en.wikipedia.org/wiki/Sigmoid_function.

Williams, M. R., B. B. Lamont, and J. D. Henstridge. 2009a. Species-area functions revisited. Journal of Biogeography 36:1994–2004.

Williams, M. R., B. B. Lamont, and J. D. Henstridge. 2009b. Species-area functions revisited. Journal of Biogeography 36:1994–2004.

Yin, X., J. Goudriaan, E. A. Lantinga, J. Vos, and H. J. Spiertz. 2003. A flexible sigmoid function of determinate growth. Annals of Botany 91:361–371.

Yoccoz, N. G. 1999. Evolution de l'utilisation des statistiques : quelques réflexions sur le rôle des modèles. Nature Sciences Sociétés 7:14–18.

Zilliox, C., and F. Gosselin. 2014. Tree species diversity and abundance as indicators of understory diversity in French mountain forests: Variations of the relationship in geographical and ecological space. Forest Ecology and Management 321:105–116.

Ugoline GODEAU

Améliorer la pertinence et l'efficacité des modèles statistiques en écologie : extension des fonctions sigmoïdes dans le cadre de l'étude de la distribution de la biodiversité

Résumé : La modélisation est un outil majeur en écologie pour décrire et comprendre les écosystèmes ou prédire leur réponse. Nous nous sommes intéressés aux modèles non-linéaires de forme sigmoïdale en macro-écologie avec pour objectif de mieux les définir, d'en comprendre les limites et de proposer des améliorations. Nous les avons d'abord étudiés dans des modèles de biodiversité hiérarchiques Bayésiens. Nous avons démontré que la prise en compte de variations aléatoires de différents paramètres de fonctions sigmoïdales avait un impact sur l'estimation des effets. Nous nous sommes ensuite intéressés aux modèles linéaires généralisés binomiaux binaires pour lesquels nous avons comparé la fonction classique logistique à d'autres fonctions sigmoïdales dont les asymptotes étaient estimées. Cela a permis de mettre en évidence les erreurs d'estimation induites par l'utilisation de la fonction logistique classique si les données ne sont pas cohérentes avec ce modèle. Enfin, nous avons appliqué ces fonctions logistiques avec asymptotes estimées dans le cadre de modèles d'occurrence hiérarchiques multi-espèces, grâce auxquels nous avons pu établir un intérêt probable de l'estimation des asymptotes. Les résultats instables ne nous ont pas permis de développer des conclusions écologiques. Lors de ces différents travaux, nous avons utilisé différents outils d'évaluation et interprétation des modèles, et prôné leur utilisation conjointe. En conclusion, nous avons développé de nouveaux modèles statistiques non-linéaires sigmoïdaux, qui sont de nouveaux outils pour l'écologue permettant d'enrichir sa palette pour mieux estimer les relations entre des variables et des données de biodiversité.

Mots clés : statistiques, modèles hiérarchiques, sigmoïde, logistique, modèles linéaires généralisés binomiaux, biodiversité

Improving relevance and efficiency of statistical models in ecology: extension of sigmoid functions in the context of the study of the distribution of biodiversity

Summary: Modeling is a major tool in ecology to describe and understand ecosystems or predict their response. We here focused our attention on non-linear sigmoidal models in macroecology, in order to better define them, understand their limitations and suggest improvements. We first studied them in hierarchical Bayesian biodiversity models. We found that taking into account random variations of different parameters of sigmoidal functions has an impact on the estimation of the effects. We then turned our attention to binary binomial generalized linear models for which we compared the classical logistic function to other sigmoidal functions whose asymptotes were estimated. We found strong estimation errors induced by the use of the classical logistic function if the data are not consistent with this model. Finally, we applied these logistic functions with estimated asymptotes in the context of hierarchical joint species occurrence models, thanks to which we were able to demonstrate the usefulness of considering the estimation of asymptotes. However, the unstable results did not allow us to develop ecological conclusions. Throughout, we have used various tools to better apprehend model evaluation and proposed that they should be used jointly. In conclusion, we have developed new forms of non-linear sigmoidal statistical models, which are new tools for the ecologist allowing to enrich his/her toolbox to better estimate the relationships between ecological variables and biodiversity data.

Keywords: statistics, hierarchical models, sigmoid, logistic, binomial generalized linear models, biodiversity

INRAE
Domaine des Barres
45290 Nogent-sur-Vernisson

UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE Santé, Sciences Biologiques et Chimie du Vivant

Unité de Recherche Ecosystèmes Forestiers - INRAE

THÈSE présentée par :

Ugoline GODEAU

soutenue le : **15 juin 2020**

pour obtenir le grade de : **Docteur de l'Université d'Orléans**

Discipline/ Spécialité : Écologie

Améliorer la pertinence et l'efficacité des modèles statistiques en écologie : extension des fonctions sigmoïdes dans le cadre de l'étude de la distribution de la biodiversité

THÈSE dirigée par :

M. GOSSELIN Frédéric, Ingénieur en Chef des Ponts, des Eaux et des Forêts, INRAE Nogent-sur-Vernisson

RAPPORTEURS :

Mme DELIGNETTE-MULLER Marie-Laure, Professeur des Universités, VetAgro Sup

M. GIMENEZ Olivier, Directeur de recherches, CNRS

JURY :

Mme DELIGNETTE-MULLER Marie-Laure, Professeur des Universités, VetAgro Sup

M. GIMENEZ Olivier, Directeur de recherches, CNRS

Mme PEYRARD Nathalie, Directeur de recherches, INRAE Toulouse

M. REINEKING Björn, Directeur de recherches, INRAE Grenoble

M. GOSSELIN Frédéric, Ingénieur en Chef des Ponts, des Eaux et des Forêts, INRAE Nogent-sur-Vernisson

VOLUME II : Annexes

Sommaire

I. ANNEXE I (CHAPITRE 1)

Annexes associées à l'article : **Lack of definition of mathematical terms in ecology: the case of the sigmoid class of functions in macro-ecology**

Annexe I. S1 : Méthodes

Annexe I. Table S1.1: Papers in the SARs and SReRs domains that use functions with a sigmoidal form or that discuss about sigmoidal relationships, with precision about their use of a term.

Article reference	Use "sigmoid" or "sigmoidal" word in the article	Define or describe sigmoid
Bolgovics et al. 2016	YES	NO
Boomsma et al. 1987	YES	NO
Burbidge et al. 1996	NO	NO
Connor & McCoy 2001	YES	NO
Dengler 2009	YES	NO
Fattorini 2006a	YES	NO
Fattorini 2006b	YES	NO
Fattorini et al. 2012	YES	NO
Gao et al. 2016	YES	NO
Gentile et al. 2005	YES	NO
Hachich et al. 2015	NO	NO
He & Lengendre 1996	NO	NO
He & Lengendre 2002	YES	NO
Huisman et al. 1993	NO	NO
Kilburn 1963	YES	NO
Kobayashi 1976	YES	NO
Lomolino 2000a	YES	NO
Mashayekhi et al. 2014	YES	NO
Monteil et al. 2004	YES	NO
Natuhara and Imai 1999	YES	NO
Oksanen & Michin 2002	NO	NO
Panitsa et al. 2006	YES	NO
Preston 1962	YES	YES

Simaiakis et al. 2012	YES	NO
Stiles et al. 2007	YES	PARTLY
Tjørve 2003	YES	YES
Tjørve 2009	YES	YES
Tjørve 2012	YES	NO
Tjørve and Tjørve 2011	YES	NO
Tjørve and Turner 2009	YES	NO
Tjørve et al. 2008	YES	NO
Triantis et al. 2012	YES	NO
Turner & Tjorve 2005	YES	NO
Veech 2000	YES	PARTLY
Williams 1995	YES	NO
Williams et al. 2009	YES	NO
Total number : 36	Number of YES : 31	Number of YES or PARTLY : 5

"NO" in the second column is for papers that did not use a sigmoid family term. "YES" in the second column and "NO" in the third is for papers that used a sigmoid family term but did not define. "YES" in the second column and "YES" or "PARTLY" in the third is for papers that either entirely or partly defined the sigmoid.

Bolgovics, Á. et al., 2016. Species area relationship (SAR) for benthic diatoms: a study on aquatic islands. *Hydrobiologia*, 764(1), pp.91–102.

Boomsma, J.J. et al., 1987. Insular biogeography and distribution ecology of ants on the Frisian islands. *Journal of Biogeography*, 14(1), pp.21–37.

Burbidge, A., Williams, M. & Abbott, I., 1997. Mammals of Australian islands: factors influencing species richness. *Journal of Biogeography*, 24(6), pp.703–715.

Connor, E.F. & McCoy, E.D., 2001. Species-Area Relationships. *Ecyclopedia of Biodiversity*, 5, pp.397–411.

Dengler, J., 2009. Which function describes the species-area relationship best? A review and empirical evaluation. *Journal of Biogeography*, 36(4), pp.728–744.

Fattorini, S., 2006a. Detecting biodiversity hotspots by species-area relationships: A case study of mediterranean beetles. *Conservation Biology*, 20(4), pp.1169–1180.

Fattorini, S., 2006b. Testing the latitudinal gradient: a narrow-scale analysis of tenebrionid richness (Coleoptera, Tenebrionidae) in the Aegean archipelago (Greece). *Italian Journal of Zoology*, 73(3), pp.203–211.

Fattorini, S., Maurizi, E. & Giulio, A. Di, 2012. Tackling the taxonomic impediment: A global assessment for ant-nest beetle diversity (Coleoptera: Carabidae: Paussini). *Biological Journal of the Linnean Society*, 105(2), pp.330–339.

Gao, D. & Perry, G., 2016. Species–area relationships and additive partitioning of diversity of native and nonnative herpetofauna of the West Indies. *Ecology and Evolution*, 6(21), pp.7742–7762.

Gentile, G. & Argano, R., 2005. Island biogeography of the Mediterranean sea: The species-area relationship for terrestrial isopods. *Journal of Biogeography*, 32(10), pp.1715–1726.

Hachich, N.F. et al., 2015. Island biogeography: patterns of marine shallow-water organisms in the Atlantic Ocean. *Journal of Biogeography*, 42(10), pp.1871–1882.

He, F. & Legendre, P., 1996. On species-area relations. *American Naturalist*, 148(4), pp.719–737.

He, F. & Legendre, P., 2002. Species diversity patterns derived from species-area models. *Ecology*, 83(5), pp.1185–1198.

Huisman, J., Olff, H. & Fresco, L.F.M., 1993. A Hierarchical Set of Models for Species Response Analysis. *Source Journal of Vegetation Science*, 4(1), pp.37–46.

Kilburn, P.D., 1963. Exponential Values for the Species-Area Relation. *Science, New series*, 141(3587), p.1276.

Kobayashi, S., 1976. The Species-area aelation - III a third model for delimited community. *Researches on Population Ecology*, 17(2), pp.243–254.

Lomolino, M. V., 2000. Ecology's most general, yet protean pattern: The species-area relationship. *Journal of Biogeography*, 27(1), pp.17–26.

Mashayekhi, M., MacPherson, B. & Gras, R., 2014. Species-area relationship and a tentative interpretation of the function coefficients in an ecosystem simulation. *Ecological Complexity*, 19, pp.84–95.

Monteil, C., Deconchat, M. & Balent, G., 2005. Simple neural network reveals unexpected patterns of bird species richness in forest fragments. *Landscape Ecology*, 20(5), pp.513–527.

Natuhara, Y. & Imai, C., 1999. Prediction of species richness of breeding birds by landscape-level factors of urban woods in Osaka Prefecture, Japan. *Biodiversity and Conservation*, 8(2), pp.239–253.

Oksanen, J. & Minchin, P.R., 2002. Continuum theory revisited: What shape are species responses along ecological gradients? *Ecological Modelling*, 157(2), pp.119–129.

Panitsa, M. et al., 2006. Patterns of species richness on very small islands: The plants of the Aegean archipelago. *Journal of Biogeography*, 33(7), pp.1223–1234.

Preston, F.W., 1962. The Canonical Distribution of Commonness and Rarety: Part I. *Ecology*, 43(2), pp.185–215.

Simaiakis, S.M. et al., 2012. The species-area relationship in centipedes (Myriapoda: Chilopoda): A comparison between Mediterranean island groups. *Biological Journal of the Linnean Society*, 105(1), pp.146–159.

Stiles, A. & M. Scheiner, S., 2007. Evaluation of species-area functions using Sonoran Desert plant data: Not all species-area curves are power functions. *Oikos*, 116(11), pp.1930–1940.

Tjørve, E., 2012. Arrhenius and Gleason revisited: New hybrid models resolve an old controversy. *Journal of Biogeography*, 39(4), pp.629–639.

Tjørve, E., 2003. Shapes and functions of species-area curves: A review of possible models. *Journal of Biogeography*, 30(6), pp.827–835.

Tjørve, E., 2009. Shapes and functions of species-area curves (II): A review of new models and parameterizations. *Journal of Biogeography*, 36(8), pp.1435–1445.

Tjørve, E. et al., 2008. Species-area relationship: Separating the effects of species abundance and spatial distribution. *Journal of Ecology*, 96(6), pp.1141–1151.

Tjørve, E. & Tjørve, K., 2011. Subjecting the theory of the small-island effect to Ockham's razor. *Journal of Biogeography*, 38(9), pp.1834–1839.

Tjørve, E. & Turner, W.R., 2009. The importance of samples and isolates for species-area relationships. *Ecography*, 32(3), pp.391–400.

Triantis, K.A., Guilhaumon, F. & Whittaker, R.J., 2012. The island species-area relationship: biology and statistics. *Journal of Biogeography*, 39(2), pp.215–231.

Turner, W.R. & Tjørve, E., 2005. Scale-dependence in species-area relationships. *Ecography*, 6(28), pp.721–730.

Veech, J.A., 2000. Choice of species-area function affects identification of hotspots. *Conservation Biology*, 14(1), pp.140–147.

Williams, M.R., 1995. An Extreme-Value Function Model of the Species Incidence and Species-Area Relations. *Ecology*, 76(88), pp.2607–2616.

Williams, M.R., Lamont, B.B. & Henstridge, J.D., 2009. Species-area functions revisited. *Journal of Biogeography*, 36(10), pp.1994–2004.

II. ANNEXE II. (CHAPITRE 2)

Annexes associées à l'article : **The importance of being random! Taking full account of random effects in nonlinear sigmoid hierarchical Bayesian models reveals the relationship between deadwood and the species richness of saproxylic beetles**

Annexe II. S1 : Méthodes

Annexe II. Table S1.1: Name of the project and number of plots and traps per plot for each sites (forest districts) in which the observations were made

Site (forest district)	Associated project	Reference	Total Number of plots	Number of traps/plot	Sampling year
Aigoual	GNB		8	2	2015
Allier	GP2013	(Parmain & Bouget 2018)	26	1	2013
Alsace	TailliAL	(Lassauce et al. 2012)	6	2	2009
Anost	GNB		8	2	2014
Auberive	GNB		24	2	2009
Bauges	JanssenPhD	(Janssen & Bouget 2016)	15	3	2014
Ballons-Comtois	GNB		15	2	2010
Bois-du-Parc	GNB		10	1	2011
Chartreuse	JanssenPhD	(Janssen & Bouget 2016)	18	3	2014
Chatillon	GNB		8	2	2013
Chize	GNB		23	2	2010
Citeaux	GNB		12	2	2010
Combe-Lavaux	GNB		8	2	2010
Gascogne	Distrafor	(Brin et al. 2016)	45	3	2012
Gâtinais	Distrafor	(Brin et al. 2016)	43	3	2012

Engins	GNB		10	2	2011
Fontainebleau	GNB		25	2	2008
Haut-Tuileau	GNB		14	1	2011
Haute Chaîne du Jura	GNB		16	2	2013
Landes	RESINE	(Bouget et al. 2014)	50	2	2006
lozere	GNB		12	2	2015
Lure	GNB		8	2	2011
Nievre	TailliAL	(Lassauce et al. 2012)	11	2	2009
Parroy	GNB		8	2	2013
Rambouillet	RESINE & TailliAL	(Bouget et al. 2014) (Lassauce et al. 2012)	63	2	2006
Rambouillet-GNB	GNB		16	2	2012
Troncais	LassaucePhD	(Lassauce et al. 2013)	31	2	2009
Ventoux	GNB		10	2	2011
Vercors	JanssenPhD	(Janssen & Bouget 2016)	7	3	2014
Ventron	GNB		8	2	2009
Yonne	TailliAL	(Lassauce et al. 2012)	5	2	2009
Yvelines	GP2013	(Parmain & Bouget 2018)	18	1	2013
Verrières	GNB		8	2	2012
Total			**589**		

Bouget, C., Larrieu, L. & Brin, A., 2014. Key features for saproxylic beetle diversity derived from rapid habitat assessment in temperate forests. Ecological Indicators, 36, pp.656–664. Available at: http://dx.doi.org/10.1016/j.ecolind.2013.09.031.

Brin, A., Valladares, L. & Ladet, S., 2016. Effects of forest continuity on flying saproxylic beetle assemblages in small woodlots embedded in agricultural landscapes. Biodiversity and

Conservation, 25, pp.587–602.

Janssen, P. & Bouget, C., 2016. Are biodiversity patterns of saproxylic beetles shaped by habitat limitation or dispersal limitation ? A case study in unfragmented montane forests. Biodiversity and Conservation, 25, pp.1167–1185.

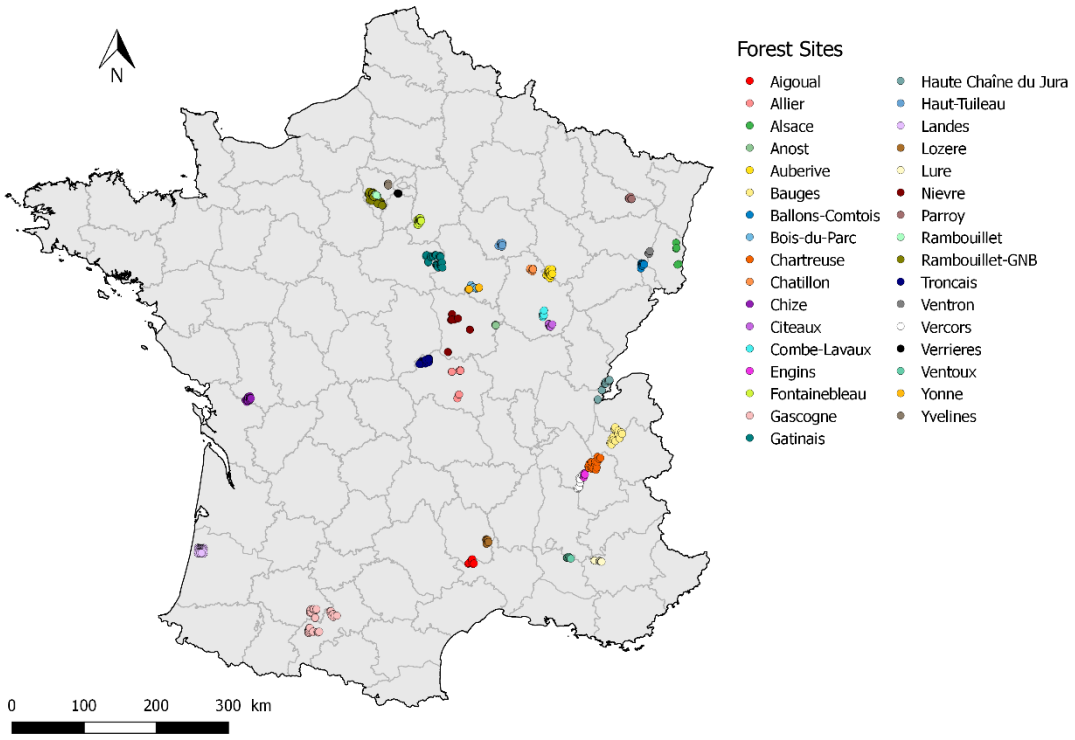
Lassauce, A. et al., 2012. Overmature coppices enhance saproxylic beetle biodiversity: a case study in French deciduous forests. Forest Ecology and Management, 266, pp.273–285.

Lassauce, A. et al., 2013. The effects of forest age on saproxylic beetle biodiversity : implications of shortened and extended rotation lengths in a French oak high forest. Insect Conservation and Diversity, 6, pp.396–410.

Paine, C.E.T. et al., 2012. How to fit nonlinear plant growth models and calculate growth rates: An update for ecologists. Methods in Ecology and Evolution, 3(2), pp.245–256.

Parmain, G. & Bouget, C., 2018. Large solitary oaks as keystone structures for saproxylic beetles in European agricultural landscapes. Insect Conservation and Diversity, 11, pp.100–115.

Annexe II. Figure S1.1: Distribution of the plots in France and colored by site (forest districts). Some sites (Rambouillet & Rambouillet-GNB; and Yonne & Bois du Parc) were spatially close but were kept as separate sites because they corresponded to different collection years.



Annexe II. Section S1.1: Dendrometric survey protocol

Snags greater than 30cm in diameter at breast height (dbh) were inventoried within a 20m radius disk whereas those between 10 and 30cm in dbh were inventoried in a 10m radius disk. For downed wood pieces, those with diameter greater than 30cm were inventoried in the 20m disk as well, while those between 2.5 and 30cm in diameter were inventoried along three 20m linear transects (except for projects GNB and JanssenPhD, where minimum diameter for downed wood pieces were respectively 5 and 7.5 cm instead of 2.5 cm). Deadwood volume

per ha was then calculated using formulas adapted to each sampling scheme, before summing snags and downed deadwood.

Annexe II. Equation S1.1: Paine et al. (2012) five-parameters logistic function parametrization

$$f(X) = L + \frac{M_0\,(K-L)}{(M_0^{\frac{1}{\theta}} + \left((K-L)^{\frac{1}{\theta}} - M_0^{\frac{1}{\theta}}\right) e^{-(r/\theta)X})^{\theta}}$$

Where :

L	the lower asymptote
K	the upper asymptote
M0	the intercept minus L (ie. The intercept equals L+M0)
r	Parameter proportional to the slope at X=0
θ	the asymetry parameter

Annexe II. Table S1.2: Parametrization of the functions compared

Function name	**Function reparametrized**	**Number of parameters**	**Reference**
Linear function	$F(\tilde{x}_{i,j}, \widetilde{lNp}_{i,j}, \theta_i, \varphi) = \exp(\rho\ \tilde{lp}_{i,j} + h_i)\ \ (a_i + r_i\ \tilde{x}_{i,j})$	2	Revised parametrization from Paine et al. (2012)
Exponential function	$F(\tilde{x}_{i,j}, \widetilde{lNp}_{i,j}, \theta_i, \varphi) = \exp(\rho\ \tilde{lp}_{i,j} + h_i + r_i\ \tilde{x}_{i,j})$	2	Revised parametrization from Paine et al. (2012)
Logistic function with 3 parameters (logis3p)	$F(\tilde{x}_{i,j}, \widetilde{lNp}_{i,j}, \theta_i, \varphi) = \exp(\rho\ \tilde{lp}_{i,j} + h_i) \left(\frac{M_0\ K}{\left(M_0 + (K - M_0) * \left(\exp(-(4\ r_i\ /\ K)\ \tilde{x}_{i,j} + a_i)\right)\right)} \right)$	3	Revised parametrization from Paine et al. (2012)
Extrem value function (EVF)	$F(\tilde{x}_{i,j}, \widetilde{lNp}_{i,j}, \theta_i, \varphi) = \exp(\rho\ \tilde{lp}_{i,j} + h_i)\ K\ \left\{1 - [(K - M_0)/K]^{\exp(-(\exp(1)\ r_i\ /\ K)\ \tilde{x}_{i,j} + a_i)}\right\}$	3	Revised parametrization from Godeau et al. (submitted)
Gompertz function	$F(\tilde{x}_{i,j}, \widetilde{lNp}_{i,j}, \theta_i, \varphi) = \exp(\rho\ \tilde{lp}_{i,j} + h_i)\ K\ (M_0/K)^{\exp(-(\exp(1)\ r_i\ /\ K)\tilde{x}_{i,j} + a_i)}$	3	Revised parametrization from Godeau et al. (submitted)
Logistic function with 4 parameters (logis4p)		4	Revised parametrization from Paine et al. (2012)

	$F(\tilde{x}_{i,j}, l\widetilde{N}p_{i,j}, \theta_i, \varphi) = \exp(\rho\, \widetilde{lp}_{i,j} + h_i) \left(\mathrm{L} + \frac{(\mathrm{M}_0 - \mathrm{L})\,(\mathrm{K} - \mathrm{L})}{\left((\mathrm{M}_0 - \mathrm{L}) + (K - \mathrm{M}_0)\left(\exp\left(-(4\, r_i/(\mathrm{K} - \mathrm{L}))\, \tilde{x}_{i,j} + a_i\right)\right)\right)} \right)$		
Logistic function with 5 parameters (logis5p)	cf. equation 3 in main text	5	Revised parametrization from Paine et al. (2012)

Annexe II. Table S1.3: Prior probability distribution of the parameters of the functions and probability distributions from which initial values were drawn.

Functions concerned	**Parameter**	**Prior distribution**	**Initial values**
All functions	ρ	U(0,10)	exp(N(0,0.3))
	α	U(0,1)	U(0,1)
	λ	U(0,1000)	$\delta_{0.01}$
	K	U(0,300)	$\delta_{\max(\min(\exp(\exp(a)),295),\min(\exp(\exp(d)),295))}$
	$\log(\sigma_v)$ ($v \in \{1,2,3,4\}$)	T(0.04,10)	U(-6,-5)
	$corr_{vw}$ (v and $w \in \{1,2,3,4\}$)	U(-1,1)	δ_0
All except constant	r	T(0.04,10)	U(2/10,5/10)*sign(U(-0.5,0.5))
All except constant and linear	M_0	U(0,K)	U(0.3,0.5)
logis4p and logis5p	L	U(0,300)	$\delta_{\min(\min(\exp(\exp(a)),250),\min(\exp(\exp(d)),250))}$
	M_0	U(L,K)	$\delta_{(L+M)/2}$
Only logis5p	ε	U(-3,3)	U(-0.01,0.01)

δ = Dirac distribution; T = Student distribution; U = uniform distribution. The initial values of K and L are based on two initial draws, for parameters called a and d, from the distribution N(log(3.4),0.3).

Annexe II. Table S1.4: X-axis starting (Xinit) points to which we added the different Deltax that were considered for the magnitude analyses. The levels are rounded values stemming from considerations of empirical quantiles (in parentheses)

	Quantile 0.0	Quantile 0.25	Quantile 0.9
Deadwood (m^3/ha)	0 (0)	10 (9.27)	70 (67.85)

Annexe II. S2 : Résultats

Annexe II. Table S2.1: Score of conditional version of classical WAIC (Hooten & Hobbs 2015) and WAICbis (Gelman et al. 2014) of all models tested, including models with no correlations between random effects (nocorr).

Function	Constant		Linear		Exp		Gompertz		EVF		Logis3p		Logis4p		Logis5p	
WAIC type	WAICbis	WAIC	WAICbis	WAIC	WAICbis	WAIC	WAICbis	WAIC	WAICbis	WAIC	WAICbis	WAIC	WAICbis	WAIC	WAICbis	WAIC
no.re	4798.1	4798.1	4799.7	4799.8	4799.8	4799.9	4799.5	4799.6	4799.5	4799.6	4799.6	4799.7	4799.3	4799.3	4802.7	4802.7
re.1r			4773.8	4776.4	4769.9	4772.5	4752.7	4755.5	4728.4	4730.6	4744.5	4747.3	4647.4	4652.4	4651.9	4657.3
re.1h	4329.5	4333.6	4329.0	4333.3	4329.1	4333.4	4324.1	4328.6	4319.4	4324.0	4322.2	4326.8	4319.7	4324.3	4319.4	4324.0
re.1a			4325.8	4330.2	4329.1	4333.4	4326.1	4330.7	4326.9	4331.4	4325.8	4330.3	4326.2	4330.7	4326.4	4331.0
nocorr.re.2ra			4321.9	4327.5			4322.2	4327.8	4321.6	4327.5	4321.9	4327.7	4322.0	4327.8	4322.5	4328.4
nocorr.re.2hr			4321.3	4327.5	4320.9	4327.3	4321.3	4326.9	4319.9	4325.2	4320.9	4326.2	4314.5	4320.94	4313.3	4320.2
nocorr.re.2ha							4324.9	4329.5	4321.9	4327.3	4324.0	4328.8	4318.9	4324.2	4319.1	4324.5
re.2hr			4321.6	4327.5	4321.5	4327.5	4321.2	4326.6	4319.3	4324.6	4321.0	4326.3	4311.4	4318.3	4312.0	4318.8
re.2ra			4323.3	4328.5			4323.3	4328.7	4322.8	4328.5	4323.2	4328.6	4323.0	4328.5	4323.2	4328.7
re.2ha							4325.0	4329.5	4321.8	4327.0	4324.1	4328.9	4320.4	4325.8	4319.4	4324.8
nocorr.re.3hra							4322.3	4327.7	4320.9	4326.8	4322.1	4327.6	4315.2	4322.2	4315.1	4322.1
re.3hra							4322.9	4328.2	4321.1	4326.8	4322.4	4327.8	4314.1	4320.8	4314.2	4320.8
nocorr.re.4hrae															4315.2	4322.3
re.4hrae															4314.2	4321.1

Blue is for models with NBLM distribution used for convergence reasons.

Annexe II. Table S2.2: Difference in terms of LOOic (and standard error) between models with no correlation between random effect and models with correlation between random effects for logis4p and logis5p link functions.

	Logis4p nocorr/corr	Logis5p nocorr/corr
re.2ra	-0.3 (0.4)	0.0 (0.5)
re.2hr	1.3 (1.1)	0.8 (0.8)
re.2ha	-0.2 (0.3)	-0.1 (0.3)
re.3hra	0.8 (0.6)	0.7 (0.6)
re.4hrae		0.6 (0.5)

** is for models significantly different at the 5% significance level. A positive value indicates that the best model is the model with correlations. A negative value indicates that the best model is the model without correlations.*

Annexe II. Table S2.3: Difference in terms of LOOic (and standard error) between models re.1a and models re.1h.

Loo.diff (se)	Linear	Exp	Gompertz	EVF	Logis3p)	Logis4p	Logis5p
re.1h / re.1a	1.5 (1.1)	0.0 (0.1)	-1.0 (1.9)	-3.7 (2.8)	-1.8 (2.2)	-3.2 (2.8)	-3.5 (3.0)

** is for models significantly different at the 5% significance level. A positive value indicates that the best model is the model with random effects on the intercept. A negative value indicates that the best model is the model with homothetic random effects.*

Annexe II. Table S2.4: Difference in terms of LOOic (and standard error) between the model logis4p with re.2hr and other models with re.2hr and a different link function, or other models with logis4p function and other random effects settings.

link function / random effect	logis4p re.2hr	link function / random effect	logis4p re.2hr
logis4p no.re	-240.1 (18.0)*	**linear re.2hr**	-4.8 (2.4)*
logis4p re.1r	-183.5 (18.9)*	**Exp re.2hr**	-4.7 (2.5)

logis4p re.1h	-2.9 (2.9)	**Gompertz re.2hr**	-4.3 (2.2)
logis4p re.1a	-6.1 (3.7)	**EVF re.2hr**	-3.2 (2.1)
logis4p re.2ra	-5.3 (3.0)	**logis3p re.2hr**	-4.2 (2.2)
logis4p re.2ha	-3.7 (2.4)	**logis5pV4 re.2hr**	-0.2 (0.2)
logis4p re.3hra	-1.3 (0.6)*		

** is for models significantly different at the 5% significance level. A positive value indicates that the best model is the one represented in row, and conversely a negative value indicates that the best model is the model logis4p with re.2hr.*

Annexe II. Table S2.5: Parameter estimates of the best model overall (with logis4p function and re.2hr random effects settings) and discussion around these estimates

Parameters	**Mean**	**SD**	**Credibility interval between 2.5 and 97.5%**
L	22.333279	5.01795768	[8.71235; 30.81386]
M0	31.3207381	2.33336941	[27.007748; 36.13795]
r	4.7859172	2.42487192	[1.11090; 10.28674]
K	37.756073	21.4198949	[29.24737; 50.23030]
λ	1.4171996	0.21820345	[0.99628; 1.84930]
α	0.2916648	0.1897457	[0.01552; 0.71797]
ρ	0.2434446	0.03121617	[0.18265; 0.30503]
$\log(\sigma_3)$	-0.9205117	0.13781265	[-1.17716; -0.64088]
$\log(\sigma_2)$	0.5347816	2.68178063	[-7.72541; 2.31553]
$corr_{23}$	-0.4000944	0.42370922	[-0.95430; 0.70754]

Annexe II. Section S2.1: Results discussion of estimated parameters for the best model

The estimated parameters for the best model indicated that the ordinate at the Media M0 was more precisely estimated (31.32 ± 2.33) than either the lower asymptote L (22.33 ± 5.02) or the upper asymptote K (37.76 ± 21.42). They also indicated a clear effect of the number of sampling periods since parameter ρ was precisely estimated (0.24 ± 0.03). Homothetic variation among sites implied a lognormal variation among sites with a lognormal standard deviation that varied between 0.3 and 0.5, which was a rather strong variation. The mean slope at the inflexion point r

indicated a positive, strong but uncertain relationship with x at the inflexion point (a 10m^3/ha increase in x would imply a 47.8 increase of species with the mean estimate). Yet, this happens only for negative values of x and therefore does not occur for positive x values (for magnitude analyses in the observed set of x values, see Results). The slope also varied strongly with considerable imprecision among sites ($\log(\sigma_2)$ was estimated at 0.53 ± 2.68). The correlation between the two random effects was estimated to be negative but with a high level of uncertainty. Finally, the parameters of the Negative Binomial distribution indicated a moderate level of overdispersion (λ = 1.42 ± 0.22), showing similarities to the first type of Negative Binomial in which variance is related to mean rather than to the second type of Negative binomial (α = 0.29 ± 0.19).

Annexe II. Table S2.6: Score of marginal (marg) and conditional versions of classical WAIC (Hooten & Hobbs 2015) and WAICbis (Gelman et al. 2014) of models with logis4p function and with all different random effects. Marginal versions were based on importance sampling estimates of marginal likelihoods (using 2,000 importance samples).

random effetcs	WAICbis	WAICbismarg	WAICvar	WAICvarmarg
re.1r	4647.4	4718.0	4652.4	4724.8
re.1h	4319.7	4404.3	4324.3	4406.3
re.1a	4326.2	4405.7	4330.7	4407.7
re.2hr	4311.4	4399.3	4318.3	4402.9
re.2ra	4323.0	4408.8	4328.5	4411.2
re.2ha	4320.4	4404.3	4325.8	4406.3
re.3hra	4314.1	4398.6	4320.8	4402.0

Annexe II. Table S2.7: Score of marginal (marg) and conditional versions of classical WAIC (Hooten & Hobbs 2015) and WAICbis (Gelman et al. 2014) of models with random effect re.2hr with all different link functions. Marginal versions were based on importance sampling estimates of marginal likelihoods (using 2,000 importance samples).

Link Function	**WAICbis**	**WAICbismarg**	**WAICvar**	**WAICvarmarg**
Linear	4321.6	4408.9	4327.5	4410.9
Exp	4321.5	4408.9	4327.5	4410.9
Gompertz	4321.2	4409.3	4326.6	4411.0
EVF	4319.3	4405.9	4324.6	4407.7
Logis3p	4321.0	4407.5	4326.3	4409.3
Logis4p	4311.4	4399.3	4318.3	4402.9
Logis5p	4312.0	4399.8	4318.8	4403.2

Annexe II. Section S2.2: Results discussion of marginal WAIC

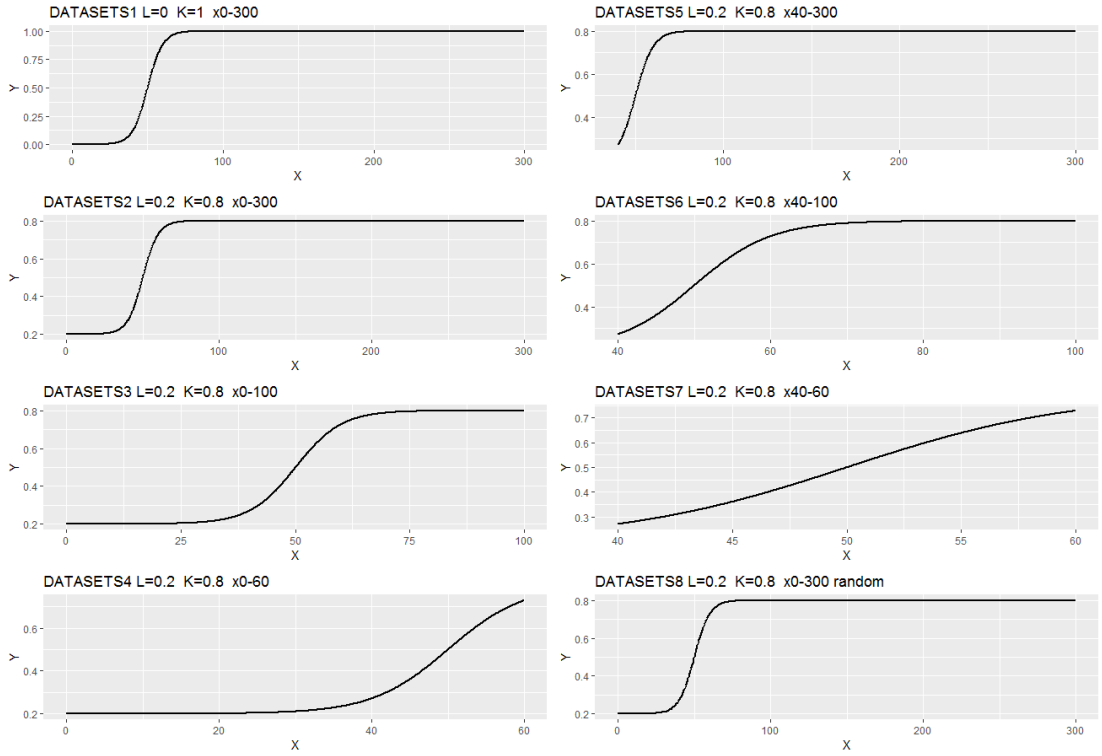
Marginal WAIC results were very similar to those for conditional WAIC (Table S3.S1 and S3.S2) but with a lag of approximatively 83 units. Overall, the order of the models remained the same, expect in some cases where the WAIC score difference was very small and models are not considered different from each other (for example when comparing the model with linear function and the model with Gompertz function with re.2hr random effects settings – cf. Table S3.S2). We, therefore, considered that in our study conditions (about 20 replications per level of random effects), the conditional and marginal WAICs were equivalent. Thus, we did not recalculate marginal WAIC for all models.

III. ANNEXE III. (CHAPITRE 3)

Annexes associées à l'article : **Generalized Linear Misleading: need for a logistic function with estimated asymptotes to complement canonical link functions for binomial GLMs**

Annexe III. S1 : Méthodes

Annexe III. Figure S1.1: Curve shapes and observed gradients of the simulation scenarios. The name of the simulation scenario, the values of L and K and the gradient are specified above each curve. The mention "random" indicates that observations were randomly drawn unlike other simulation scenarios where observations were equally distributed. Slope and inflexion point were respectively fixed to 0.2 and 50.0.



Annexe III. Table S1.1: Initial value for the models applied to univariate datasets.

Parameter	First set of initial values	Second set of initial values	Third set of initial values	Fourth set of initial values	Fifth set of initial values
Lp	U(-5,-4)	U(0,1)	U(-5,-4)	U(-5,-4)	U(-5,-4)
Kp	U(4,5)	U(0,1)	U(4,5)	U(4,5)	U(4,5)
sl	0	0	0	0	½: U(0.1,0.2) ½: -U(0.1,0.2)
ip	U(X_{min},X_{max})	U(X_{min},X_{max})	X_{min}	X_{max}	U(X_{min},X_{max})

U = uniform distribution; X_{min} = minimum value of gradient X ; X_{max} = maximum value of gradient X.

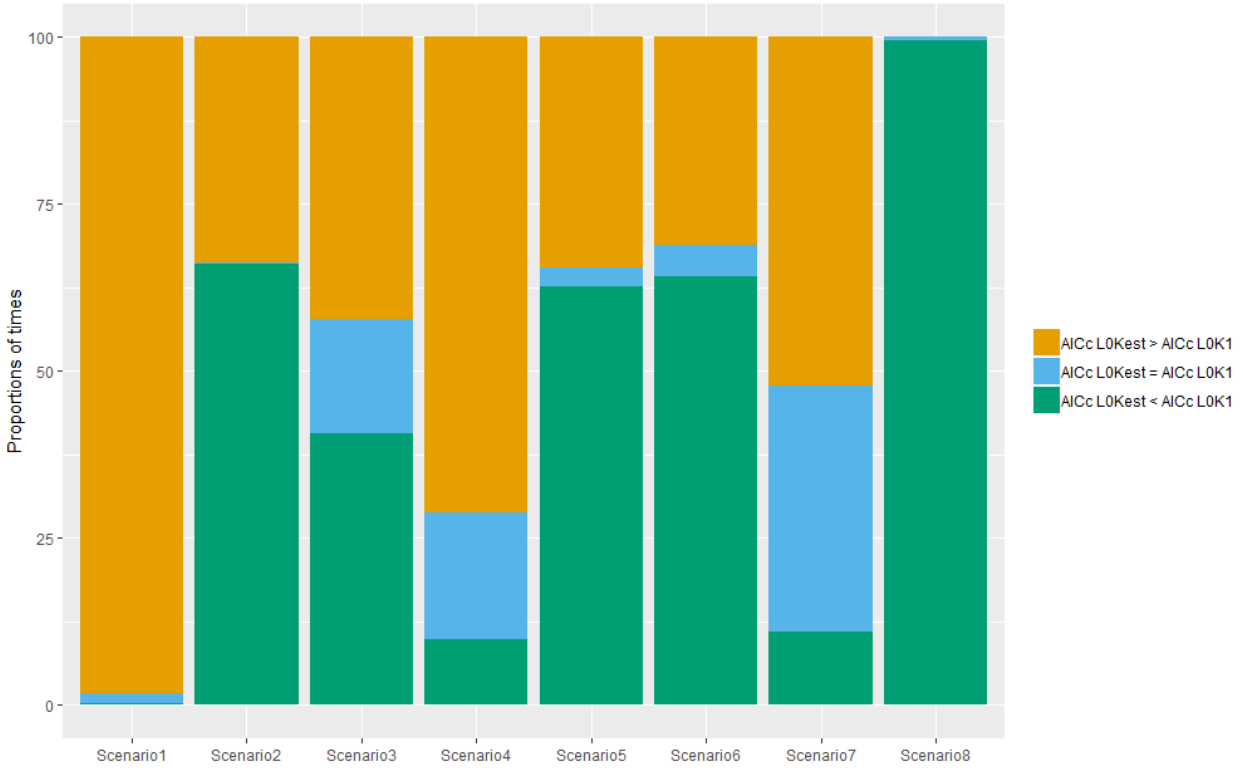
Annexe III. S2 : Résultats

Annexe III. Table S2.1: AICc and estimators of *sl* and *ip* for models GLM and L0K1 for 10 datasets randomly drawn from simulation Scenario1 and 10 datasets randomly drawn from simulation Scenario2.

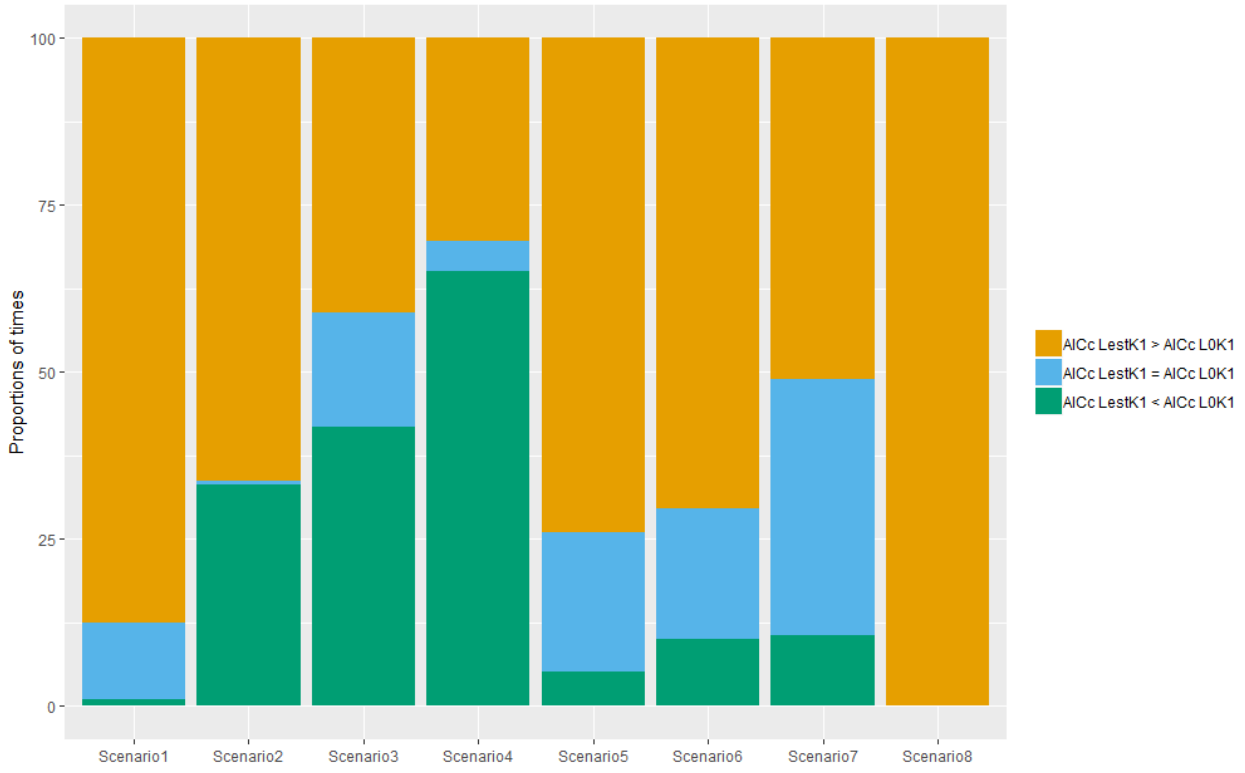
Randomly drawn dataset from Scenario1 and Scenario2	**AICc GLM**	**AICc L0K1**	**estimated sl GLM**	**estimated sl L0K1**	**estimated ip GLM**	**estimated ip L0K1**
Scenario1 dataset 1	248.9036	248.9036	0.175095	0.175095	50.6894	50.6894
Scenario1 dataset 2	50.4705	50.4705	0.215333	0.215333	54.9342	54.9342
Scenario1 dataset 3	95.9225	95.9225	0.180593	0.180593	51.3205	51.3205
Scenario1 dataset 4	43.0823	43.0823	0.261123	0.261123	51.9355	51.9355
Scenario1 dataset 5	173.8372	173.8372	0.177224	0.177224	49.2708	49.2708
Scenario1 dataset 6	54.5374	54.5374	0.240090	0.240090	49.0958	49.0958
Scenario1 dataset 7	150.7294	150.7294	0.190540	0.190540	49.2937	49.2937
Scenario1 dataset 8	82.6536	82.6536	0.199677	0.199677	52.1647	52.1647
Scenario1 dataset 9	291.5185	291.5185	0.198224	0.198224	49.9228	49.9228
Scenario1 dataset 10	110.6349	110.6349	0.181584	0.181584	51.4718	51.4718
Scenario2 dataset 1	1059.3891	1059.3891	0.005462	0.005462	-14.9528	-14.9528
Scenario2 dataset 2	1428.2215	1428.2215	0.015552	0.015552	58.4690	58.4690
Scenario2 dataset 3	962.9000	962.9000	0.006459	0.006459	125.8438	125.8438

Scenario2 dataset 4	330.9468	330.9468	0.088661	0.088661	40.7800	40.7800
Scenario2 dataset 5	622.7524	622.7524	0.086224	0.086224	41.9025	41.9025
Scenario2 dataset 6	775.9819	775.9819	0.006317	0.006317	93.8066	93.8066
Scenario2 dataset 7	699.1639	699.1639	0.009545	0.009545	42.9438	42.9438
Scenario2 dataset 8	2378.2791	2378.2791	0.005519	0.005519	74.6897	74.6897
Scenario2 dataset 9	573.3892	573.3892	0.067208	0.067208	35.4539	35.4539
Scenario2 dataset 10	523.0958	523.0958	0.080506	0.080506	37.9944	37.9944

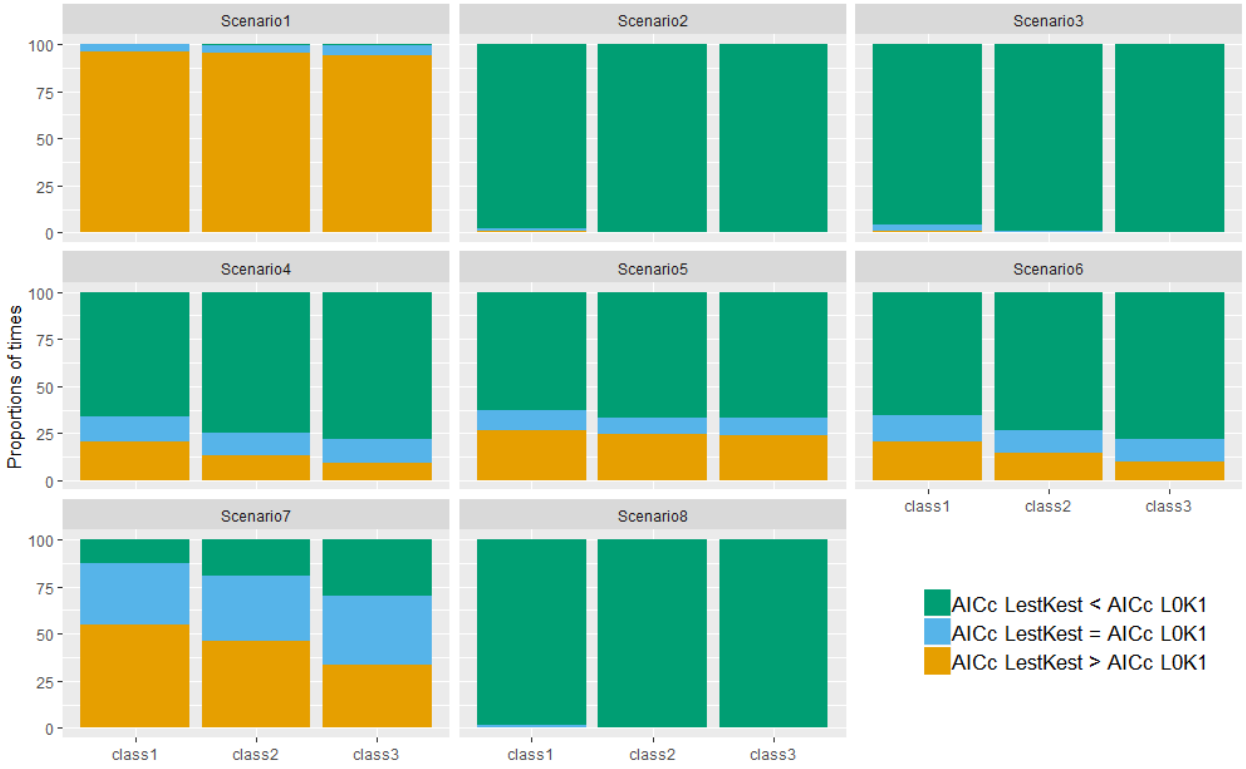
Annexe III. Figure S2.1: For each univariate simulation scenario, the proportion of times when: i) L0Kest had less predictive capacity than L0K1 (an AICc greater by at least 2.0 points, in orange), ii) the predictive capacity of models L0Kest and L0K1 was equivalent (AICc within 2.0 points, in blue); and iii) L0Kest had better predictive capacity than L0K1 (an AICc lower by at least 2.0 points, in green).



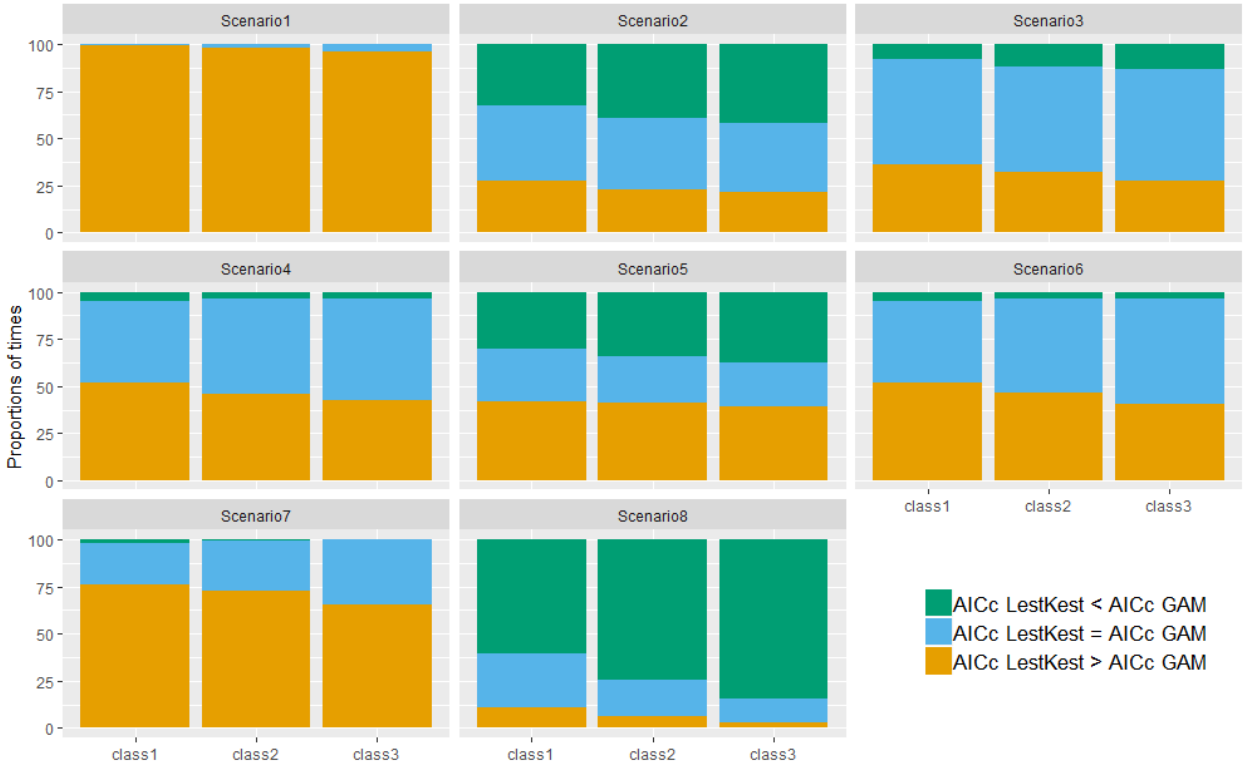
Annexe III. Figure S2.2: For each univariate simulation scenario, the proportion of times when: i) LestK1 had less predictive capacity than L0K1 (an AICc greater by at least 2.0 points, in orange); ii) predictive capacity of models LestK1 and L0K1 was equivalent (AICc within 2.0 points, in blue); and LestK1 had better predictive capacity than L0K1 (an AICc lower by at least 2.0 points, in green).



Annexe III. Figure S2.3: For each univariate simulation scenario, the proportion of times when: i) LestKest had less predictive capacity than L0K1 (an AICc greater by at least 2.0 points, in orange); ii) the predictive capacity of models LestKest and L0K1 was equivalent (AICc within 2.0 points, in blue); and iii) LestKest had better predictive capacity than L0K1 (an AICc lower by at least 2.0 points, in green). Three different classes of number of observations are shown - class1: nobs Є [147;1092], class2: nobs Є]1092;2037] and class3: nobs Є]2037;2982].



Annexe III. Figure S2.4: For each univariate simulation scenario, the proportion of times when: i) LestKest had less predictive capacity than GAM (an AICc greater by at least 2.0 points, in orange); ii) predictive capacity of models LestKest and GAM was equivalent (AICc within 2.0 points, in blue); and iii) LestKest had better predictive capacity than GAM (an AICc lower by at least 2.0 points, in green). Three different classes of number of observations are shown - class1: nobs Є [147;1092], class2: nobs Є]1092;2037] and class3: nobs Є]2037;2982].



Annexe III. Table S2.2: Type-I errors (in percentage of times that the true value fell outside the confidence interval) at the 1% level for of each parameter and each univariate simulation scenario. The underlined values 0.00% were expected because the true value was by nature outside the confidence interval due to the construction with Lp and Kp. * indicates values significantly different from 1.0 at the 5% level. Grey highlights values not significantly different from 1.0 at the level 5%.

Simulation scenario	Parameter	L0K1	L0Kest	LestK1	LestKest
Datasets1	L	NA	NA	0.00 *	0.00 *
Datasets1	K	NA	0.00 *	NA	0.00 *
Datasets1	sl	0.70 *	0.60 *	0.50 *	0.40 *
Datasets1	ip	1.60 *	1.70 *	1.60 *	1.70 *
Datasets1	NSL	0.70 *	0.60 *	0.50 *	0.40 *
Datasets2	L	NA	NA	33.83 *	1.88 *
Datasets2	K	NA	1.30 *	NA	0.78 *
Datasets2	sl	99.97 *	67.40 *	66.91 *	3.74 *
Datasets2	ip	69.00 *	60.59 *	37.51 *	2.98 *
Datasets2	NSL	99.88 *	66.85 *	66.77 *	3.37 *
Datasets3	L	NA	NA	14.66 *	0.82
Datasets3	K	NA	15.30 *	NA	1.02
Datasets3	sl	100.00 *	67.37 *	66.60 *	1.98 *
Datasets3	ip	85.56 *	44.73 *	43.60 *	1.82 *
Datasets3	NSL	99.99 *	67.38 *	66.63 *	2.01 *
Datasets4	L	NA	NA	0.75 *	0.80 *
Datasets4	K	NA	37.19 *	NA	15.20 *
Datasets4	sl	80.11 *	67.53 *	24.01 *	2.40 *
Datasets4	ip	63.47 *	37.61 *	54.86 *	10.23 *
Datasets4	NSL	71.07 *	67.32 *	5.16 *	0.91
Datasets5	L	NA	NA	28.14 *	25.69 *
Datasets5	K	NA	1.00	NA	0.98
Datasets5	sl	69.94 *	11.79 *	62.84 *	3.50 *
Datasets5	ip	49.59 *	23.35 *	36.54 *	9.56 *
Datasets5	NSL	66.49 *	4.42 *	63.01 *	2.86 *
Datasets6	L	NA	NA	37.35 *	15.36 *
Datasets6	K	NA	0.80 *	NA	0.75 *
Datasets6	sl	80.51 *	26.20 *	66.95 *	2.78 *
Datasets6	ip	63.74 *	55.56 *	36.74 *	10.06 *
Datasets6	NSL	70.67 *	5.18 *	66.52 *	0.81 *

Datasets7	L	NA	NA	18.27 *	16.72 *
Datasets7	K	NA	18.49 *	NA	16.56 *
Datasets7	sl	83.80 *	42.82 *	41.12 *	1.46 *
Datasets7	ip	82.75 *	47.10 *	45.84 *	8.27 *
Datasets7	NSL	3.15 *	1.78 *	1.69 *	0.17 *
Datasets8	L	NA	NA	99.7 *	2.82 *
Datasets8	K	NA	2.60 *	NA	1.11
Datasets8	sl	100.00 *	99.90 *	100.00 *	6.75 *
Datasets8	ip	33.50 *	82.50 *	33.40 *	6.14 *
Datasets8	NSL	100.00 *	98.10 *	100.00 *	6.04 *

Annexe III. Table S2.3: Type-I errors (in percent of times that the true value fell outside the confidence interval) at the 5% level for each parameters and each univariate simulation scenario. The underlined values 0.00% were expected because the true value was by nature outside the confidence interval due to the construction with Lp and Kp. * indicates values significantly different from 5.0 at the 5% level. Grey is to highlight values not significantly different from 5.0 at the 5% level.

Simulation scenario	**Parameter**	**L0K1**	**L0Kest**	**LestK1**	**LestKest**
Datasets1	L	NA	NA	0.00 *	0.00 *
Datasets1	K	NA	0.00 *	NA	0.00 *
Datasets1	sl	3.80 *	3.60 *	3.40 *	3.20 *
Datasets1	ip	5.20	5.30	5.00	5.10
Datasets1	NSL	3.80 *	3.60 *	3.40 *	3.20 *
Datasets2	L	NA	NA	35.35 *	4.84
Datasets2	K	NA	5.01	NA	3.61 *
Datasets2	sl	99.98 *	68.26 *	67.96 *	6.55 *
Datasets2	ip	76.97 *	65.20 *	46.08 *	7.23 *
Datasets2	NSL	99.94 *	68.09 *	67.87 *	6.26 *
Datasets3	L	NA	NA	21.57 *	3.68 *
Datasets3	K	NA	22.20 *	NA	3.77 *
Datasets3	sl	100.00 *	68.46 *	67.81 *	5.16
Datasets3	ip	89.09 *	50.04 *	48.93 *	6.18 *
Datasets3	NSL	100.00 *	68.64 *	67.81 *	5.27
Datasets4	L	NA	NA	3.68 *	3.86 *
Datasets4	K	NA	38.75 *	NA	20.35 *
Datasets4	sl	85.69 *	68.79 *	36.50 *	5.37

Datasets4	ip	72.57 *	46.36 *	62.06 *	16.84 *
Datasets4	NSL	76.09 *	68.79 *	12.62 *	3.86 *
Datasets5	L	NA	NA	33.37 *	31.90 *
Datasets5	K	NA	3.76 *	NA	4.07 *
Datasets5	sl	73.95 *	19.93 *	63.75 *	5.52 *
Datasets5	ip	67.83 *	38.46 *	50.13 *	16.88 *
Datasets5	NSL	68.48 *	8.28 *	64.06 *	4.92
Datasets6	L	NA	NA	38.90 *	20.69 *
Datasets6	K	NA	3.60 *	NA	3.80 *
Datasets6	sl	85.47 *	37.56 *	67.96 *	5.40
Datasets6	ip	73.44 *	63.19 *	45.91 *	16.60 *
Datasets6	NSL	75.75 *	12.85 *	68.25 *	3.86 *
Datasets7	L	NA	NA	20.63 *	22.95 *
Datasets7	K	NA	21.06 *	NA	22.64 *
Datasets7	sl	90.82 *	49.99 *	48.40 *	4.21 *
Datasets7	ip	87.06 *	52.48 *	51.42 *	12.84 *
Datasets7	NSL	10.33 *	6.20 *	6.29 *	1.54 *
Datasets8	L	NA	NA	99.70 *	7.25 *
Datasets8	K	NA	7.30 *	NA	4.93
Datasets8	sl	100.00 *	100.00 *	100.00 *	10.47 *
Datasets8	ip	45.70 *	89.60 *	45.30 *	10.57 *
Datasets8	NSL	100.00 *	99.10 *	100.00 *	9.67 *

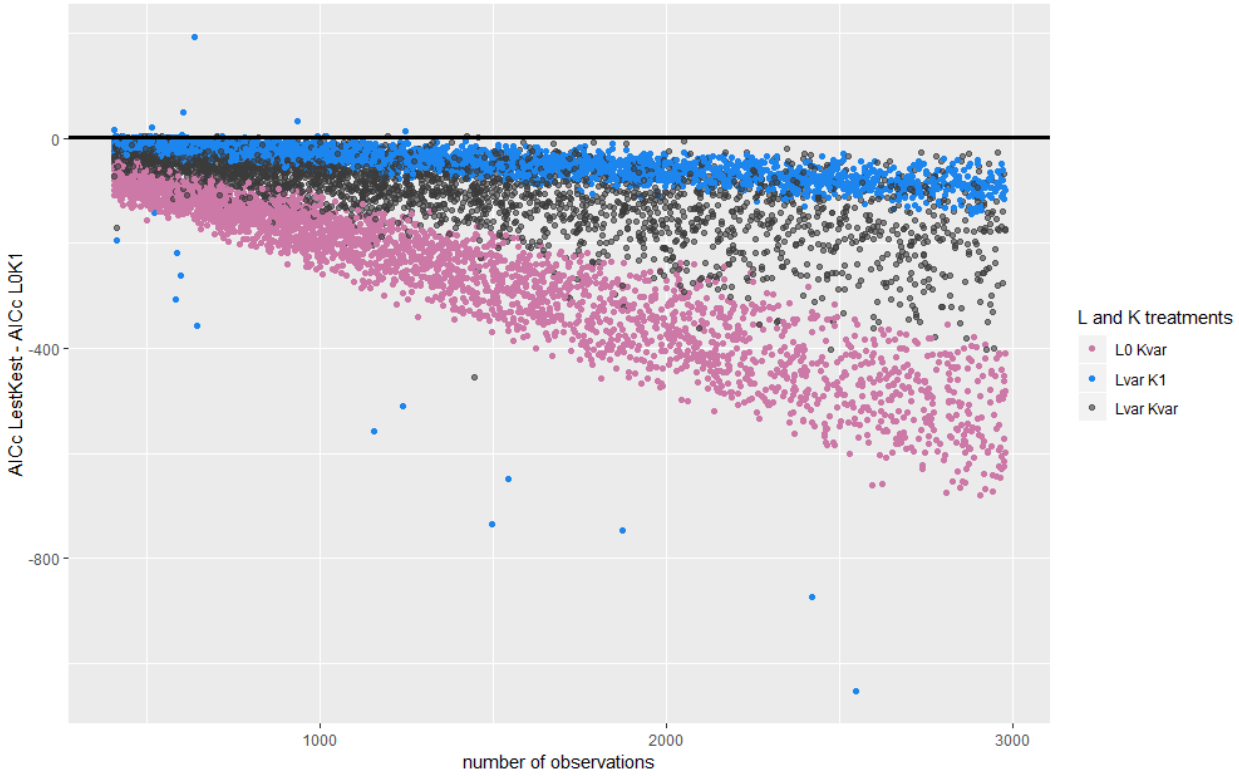
Annexe III. Table S2.4: Plugin p-value results at the 5% level for each univariate simulation scenario and each model. Yellow highlights cases with a departure from a uniform distribution at the 5% level; red highlights cases with a departure from a uniform distribution at the 1% level.

Simulation scenario	Model	mean Ynorm	var Ynorm	skew Ynorm	kur Ynorm	corYnorm link	corYnorm x	meanYnorm lower	corYnorm linklower	meanYnorm central	corYnorm linkcentral	meanYnorm upper	corYnorm linkupper
Scenario1	L0K1	0.060	0.050	0.030	0.090	0.040	0.050	0.050	0.060	0.050	0.050	0.040	0.040
Scenario2	L0K1	0.020	0.040	0.130	0.040	0.510	0.520	0.680	0.170	0.400	0.500	0.300	0.040
Scenario3	L0K1	0.050	0.040	0.420	0.070	0.490	0.490	0.380	0.120	0.030	0.620	0.300	0.040
Scenario4	L0K1	0.000	0.020	0.050	0.040	0.390	0.390	0.280	0.100	0.160	0.250	0.510	0.050
Scenario5	L0K1	0.030	0.040	0.030	0.010	0.100	0.080	0.490	0.430	0.110	0.210	0.040	0.060
Scenario6	L0K1	0.020	0.040	0.040	0.050	0.320	0.340	0.510	0.160	0.150	0.240	0.200	0.050
Scenario7	L0K1	0.000	0.020	0.060	0.030	0.050	0.050	0.150	0.140	0.020	0.060	0.000	0.050
Scenario8	L0K1	0.010	0.030	0.110	0.090	0.680	0.680	0.860	0.050	0.380	0.540	0.300	0.040
Scenario1	L0Kest	0.060	0.050	0.030	0.080	0.040	0.050	0.050	0.060	0.050	0.050	0.050	0.040
Scenario2	L0Kest	0.030	0.040	0.050	0.040	0.060	0.080	0.110	0.230	0.020	0.060	0.010	0.030
Scenario3	L0Kest	0.050	0.040	0.260	0.040	0.290	0.290	0.400	0.110	0.020	0.430	0.020	0.040
Scenario4	L0Kest	0.010	0.020	0.030	0.040	0.390	0.390	0.280	0.100	0.160	0.260	0.510	0.030
Scenario5	L0Kest	0.030	0.040	0.030	0.020	0.080	0.100	0.010	0.070	0.030	0.050	0.030	0.030
Scenario6	L0Kest	0.030	0.040	0.020	0.060	0.110	0.110	0.080	0.160	0.020	0.080	0.010	0.040
Scenario7	L0Kest	0.000	0.020	0.050	0.030	0.060	0.060	0.130	0.150	0.020	0.080	0.000	0.050
Scenario8	L0Kest	0.010	0.030	0.070	0.090	0.010	0.010	0.000	0.170	0.010	0.030	0.030	0.040
Scenario1	LestK1	0.060	0.050	0.030	0.080	0.040	0.050	0.060	0.060	0.050	0.050	0.030	0.040
Scenario2	LestK1	0.030	0.040	0.131	0.030	0.515	0.515	0.606	0.121	0.414	0.505	0.303	0.061
Scenario3	LestK1	0.020	0.040	0.200	0.080	0.220	0.230	0.140	0.140	0.030	0.320	0.350	0.040
Scenario4	LestK1	0.000	0.020	0.040	0.040	0.010	0.010	0.100	0.130	0.020	0.040	0.000	0.070
Scenario5	LestK1	0.031	0.031	0.031	0.010	0.103	0.093	0.515	0.454	0.103	0.237	0.031	0.062
Scenario6	LestK1	0.020	0.030	0.030	0.060	0.330	0.350	0.470	0.150	0.150	0.210	0.200	0.050

Scenario7	LestK1	0.000	0.020	0.060	0.040	0.090	0.090	0.100	0.120	0.020	0.060	0.000	0.050
Scenario8	LestK1	0.010	0.030	0.110	0.090	0.680	0.680	0.860	0.050	0.380	0.540	0.300	0.040
Scenario1	LestKest	0.060	0.050	0.030	0.070	0.040	0.050	0.060	0.060	0.050	0.050	0.040	0.040
Scenario2	LestKest	0.041	0.041	0.041	0.031	0.072	0.082	0.021	0.113	0.031	0.072	0.010	0.041
Scenario3	LestKest	0.020	0.041	0.051	0.051	0.051	0.051	0.112	0.133	0.020	0.031	0.010	0.041
Scenario4	LestKest	0.010	0.021	0.021	0.042	0.010	0.010	0.125	0.135	0.021	0.063	0.000	0.021
Scenario5	LestKest	0.036	0.036	0.024	0.024	0.072	0.108	0.024	0.072	0.024	0.096	0.036	0.048
Scenario6	LestKest	0.020	0.030	0.010	0.051	0.131	0.121	0.020	0.131	0.020	0.061	0.010	0.030
Scenario7	LestKest	0.000	0.020	0.051	0.041	0.082	0.082	0.061	0.122	0.020	0.061	0.000	0.041
Scenario8	LestKest	0.011	0.033	0.066	0.077	0.033	0.000	0.000	0.055	0.033	0.044	0.033	0.022
Scenario1	GAM	0.060	0.050	0.030	0.080	0.020	0.050	0.060	0.040	0.020	0.040	0.010	0.040
Scenario2	GAM	0.040	0.030	0.030	0.040	0.040	0.120	0.000	0.110	0.020	0.100	0.020	0.020
Scenario3	GAM	0.030	0.040	0.050	0.050	0.050	0.070	0.030	0.070	0.020	0.040	0.010	0.050
Scenario4	GAM	0.010	0.030	0.030	0.040	0.050	0.040	0.060	0.060	0.010	0.030	0.000	0.060
Scenario5	GAM	0.030	0.040	0.030	0.020	0.050	0.120	0.020	0.010	0.010	0.010	0.030	0.050
Scenario6	GAM	0.030	0.030	0.010	0.060	0.100	0.110	0.020	0.110	0.020	0.090	0.010	0.030
Scenario7	GAM	0.000	0.020	0.050	0.040	0.110	0.110	0.040	0.110	0.000	0.090	0.000	0.030
Scenario8	GAM	0.010	0.030	0.060	0.090	0.080	0.110	0.010	0.110	0.010	0.020	0.040	0.110

meanYnorm, varYnorm, skewYnorm, kurtYnorm = mean, variance, skewness and kurtosis of normalized random quantile residuals (Dunn and Smyth 1996); corYnormlink, corYnormx = Hoeffding's D statistics measure of independence between normalized random quantile residuals and the fitted mean value or the explanatory variable x. The last six metrics are the same as meanYnorm and corYnormlink but restricted either on the smallest 10% (lower), intermediate between 10% and 90% (central), and largest 10% (upper) fitted mean values.

Annexe III. Figure S2.5: Differences in AICc between the LestKest and L0K1 models as a function of the number of observations for the datasets in the ninth simulation scenario (Datasets9). Three cases are presented: i) L is fixed at zero and K varied during data generation (L0 Kvar in pink); ii) L varied and K was fixed at one during data generation (Lvar K1, in yellow); iii) and L and K both varied during data generation (Lvar Kvar, in grey). Points above lines represent cases where L0K1 was better than LestKest; points between the two lines (merged in this graph due to the Y-scale) represent cases where L0K1 and LestKest were equivalent (a difference in AICc of less than two units); points below lines represent cases where LestKest was better than L0K1.



Annexe III. Figure S2.6: For each multivariate simulation scenario, the proportion of times when: i) LestK1 had less predictive capacity than L0K1 or GAM (an AICc greater by at least 2.0 points, in orange); ii) predictive capacity of models LestK1 and L0K1 (or GAM) was equivalent (AICc within 2.0 points, in blue); and iii) LestK1 had better predictive capacity than L0K1 or GAM (an AICc lower by at least 2.0 points, in green). Three different classes of number of observations are shown - class1: nobs Є [147;1092], class2: nobs Є]1092;2037] and class3 nobs Є]2037;2982].

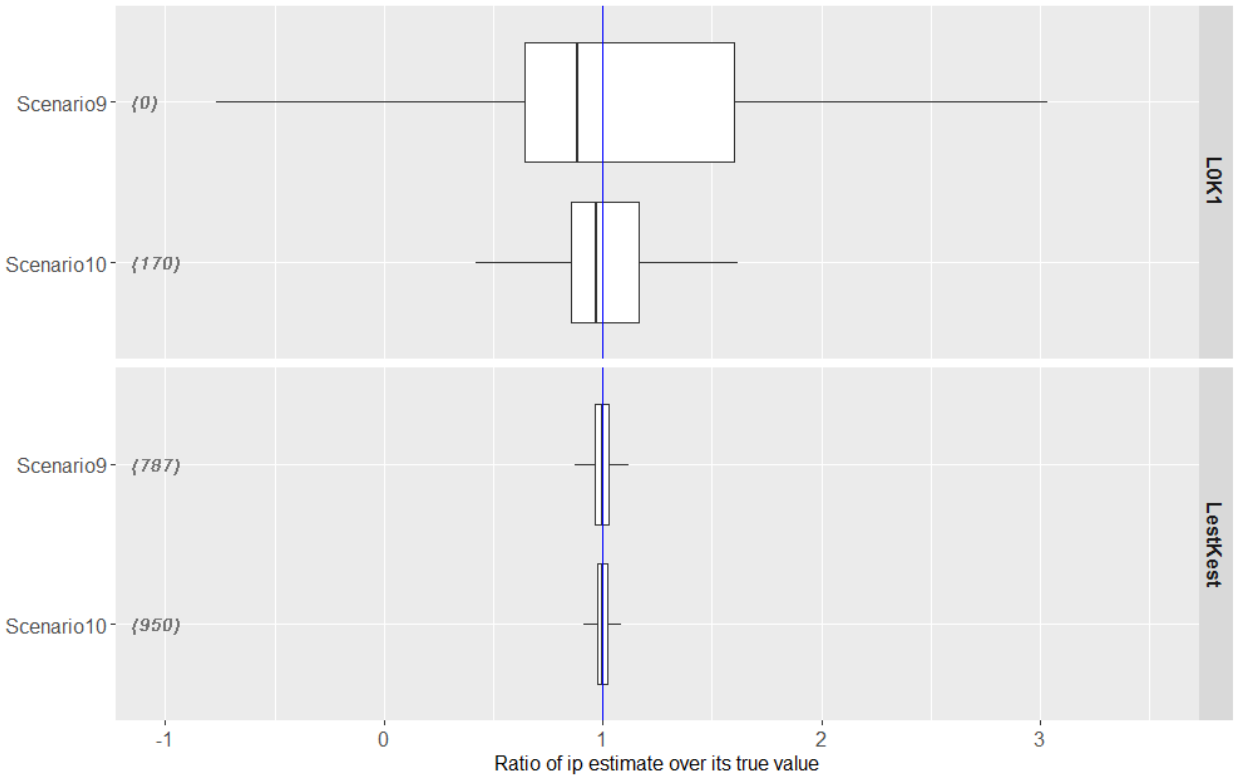
Annexe III. Table S2.5: Type-I errors (in percentage of times that the true value fell outside the confidence interval) at the 1% level for each parameter and each multivariate simulation scenario. * indicates values significantly different from 1.0 at the 5% level. Grey is to highlight values not significantly different from 1.0 at the 5% level.

Simulation scenario	Parameter	L0K1	LestKest
Datasets9	L	NA	2.39 *
Datasets9	K	NA	0.76
Datasets9	NSLX1	99.86 *	3.18
Datasets9	NSLX2	99.57 *	2.97 *
Datasets9	ip	68.68 *	3.55 *
Datasets10	L	NA	2.34 *
Datasets10	K	NA	1.03
Datasets10	NSLX1	77.28 *	1.65 *
Datasets10	NSLX2	78.17 *	1.61 *
Datasets10	ip	70.79 *	2.67 *

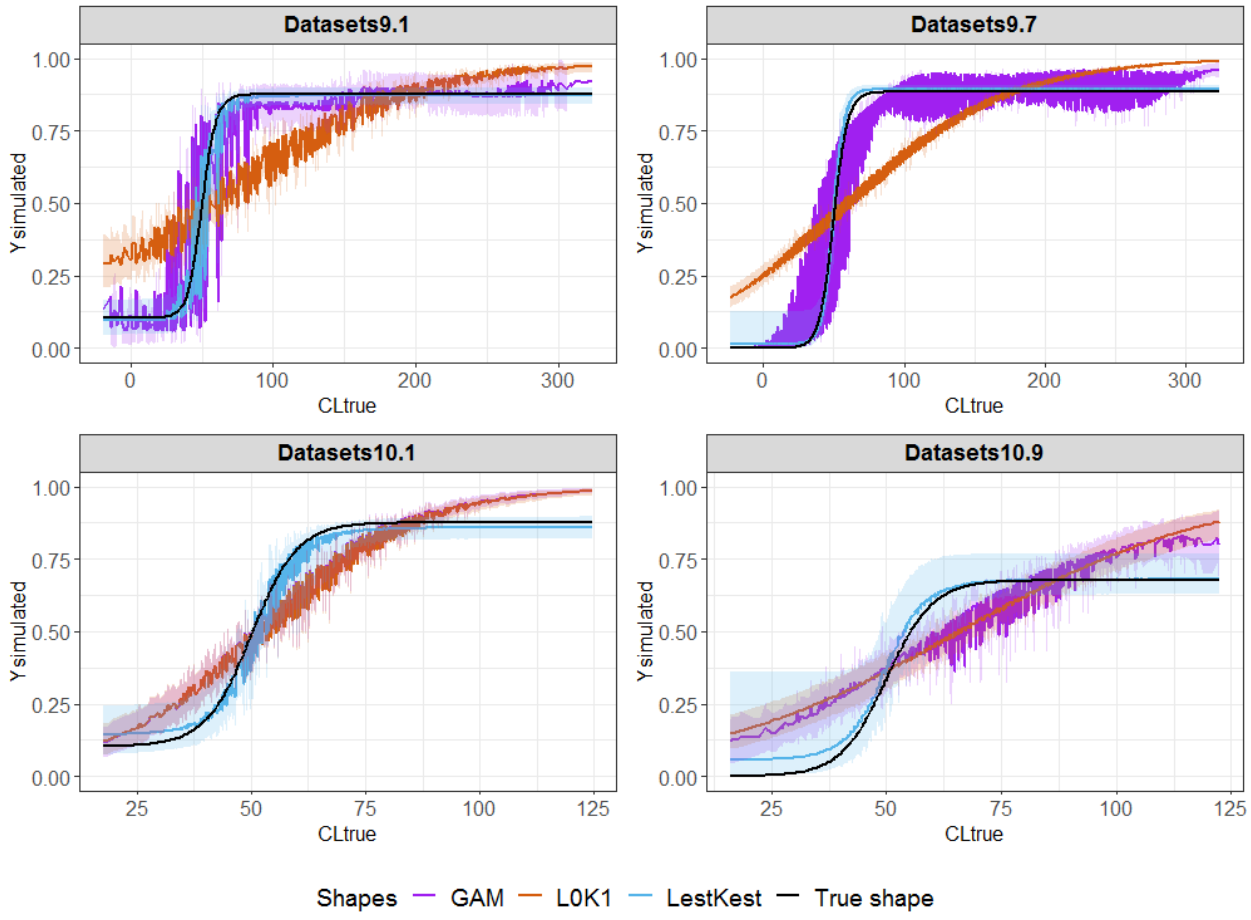
Annexe III. Table S2.6: Type-I errors (in percentage of times that the true value fell outside the confidence interval) at the 5% level for each parameter and each multivariate simulation scenario. * indicates values significantly different from 5.0 at the 5% level. Grey is to highlight values not significantly different from 5.0 at the 5% level.

Simulation scenario	Parameter	L0K1	LestKest
Datasets9	L	NA	5.80 *
Datasets9	K	NA	3.61
Datasets9	NSLX1	99.95 *	6.05 *
Datasets9	NSLX2	99.83 *	5.88 *
Datasets9	Ip	76.88 *	8.16 *
Datasets10	L	NA	5.59 *
Datasets10	K	NA	3.81
Datasets10	NSLX1	82.99 *	4.82
Datasets10	NSLX2	83.83 *	4.58
Datasets10	ip	78.58 *	7.13 *

Annexe III. Figure S2.7: Boxplot showing the ratio of optimum estimates of the inflexion point (ip) over its true value for models L0K1 and LestKest. The blue line represents the target (= 1.0 when estimates were equal to the true value). Outliers were excluded for ease of reading. The number of excluded values is specified in parentheses.



Annexe III. Figure S2.8: Y predicted by models L0K1, LestKest and GAM as a function of true CL, for four multivariate datasets. Confidence intervals of the curve shapes are between quantiles 0.025 and 0.975. Plots A and B are respectively based on the first and seventh dataset in Scenario9; plots C and D are respectively based on the first and ninth dataset in Scenario10. In plot C, the GAM and L0K1 curve shapes are almost superimposed so GAM is not visible.



Annexe III. Table S2.7: Plugin p-value results at the 5% level for each multivariate simulation scenario and each model. Yellow highlights cases with a departure from a uniform distribution at the 5% level; red highlights cases with a departure from a uniform distribution at the 1% level.

Simulation scenario	Model	mean Ynorm	var Ynorm	skew Ynorm	kur Ynorm	corYnorm link	corYnorm x1	corYnorm x2	meanYnorm lower	corYnorm linklower	meanYnorm central	corYnorm linkcentral	meanYnorm upper	corYnorm linkupper
Scenario9	L0K1	0.0200	0.040	0.140	0.040	0.570	0.560	0.020	0.730	0.070	0.370	0.570	0.390	0.000
Scenario10	L0K1	0.0000	0.060	0.150	0.060	0.480	0.160	0.090	0.520	0.110	0.240	0.310	0.460	0.040
Scenario9	LestKest	0.0227	0.057	0.034	0.034	0.034	0.034	0.045	0.023	0.080	0.023	0.068	0.034	0.045
Scenario10	LestKest	0.0108	0.054	0.065	0.054	0.086	0.022	0.032	0.022	0.032	0.043	0.011	0.054	0.075
Scenario9	GAM	0.0100	0.040	0.050	0.050	0.090	0.050	0.030	0.040	0.070	0.030	0.090	0.040	0.100
Scenario10	GAM	0.0000	0.050	0.130	0.060	0.390	0.040	0.060	0.440	0.090	0.170	0.290	0.320	0.050

meanYnorm, varYnorm, skewYnorm, kurtYnorm = mean, variance, skewness and kurtosis of normalized random quantile residuals (Dunn and Smyth 1996); corYnormlink, corYnormx = Hoeffding's D statistics measure of independence between normalized random quantile residuals and the fitted mean value or the explanatory variable x. The last six metrics are the same as meanYnorm and corYnormlink but restricted either on the smallest 10% (lower), intermediate between 10% and 90% (central), and largest 10% (upper) fitted mean values.

Annexe III. Table S2.8: Summary of univariate real dataset results (variable selected by the different models; AICc of the different models; estimated values of parameters L and K with their standard deviations for the LestKest model). Bold AICc values indicate the best model for the real dataset in the row and models with a difference in AICc of less than 2.0 points. Underlined variable names are for cases where the model chose a different variable from the other models.

Dataset	Best variable selected				AICc				Estimated asymptotes			
	L0K1	GAM free	GAM monotone	LestKest	L0K1	GAM free	GAM monotone	LestKest	L (mean)	L (sd)	K (mean)	K (sd)
RealData1	V5	V5	V5	V5	**84.02**	**83.18**	**84.05**	87.20	0.16	0.1030	1.00	0.0127
RealData2	V2	V2	V2	V2	1212.89	**1202.88**	1208.71	1210.03	0.54	0.0402	1.00	0.0014
RealData3	V2	V2	V2	V2	**440.97**	**440.97**	**440.98**	444.42	0.09	0.0667	0.70	0.1396
RealData4	V3	V3	V3	V3	1008.02	**841.81**	895.52	896.86	0.43	0.0241	1.00	0.0014
RealData5	V2	V2	V2	V2	4012.68	3996.33	**3992.75**	4002.37	0.00	0.0089	0.71	0.0331
RealData6	V3	V3	V3	V3	1341.75	**1312.63**	1316.49	1323.74	0.01	0.0057	0.90	0.0200
RealData7	V5	V2	V5	V5	244.80	**218.49**	235.05	235.15	0.16	0.0361	0.70	0.0854
RealData8	V3	V3	V3	V3	281.16	**265.98**	**266.51**	**266.93**	0.15	0.0350	0.97	0.0401
RealData9	V3	V3	V3	V3	44.57	**43.27**	**42.00**	44.14	0.09	0.0451	1.00	0.0028
RealData10	V2	V2	V2	V2	1432.34	1432.34	1432.35	**1430.05**	0.15	0.0188	0.29	0.0154
RealData11	V2	V2	V2	V2	**8592.31**	**8591.06**	**8591.37**	**8591.53**	0.28	0.0315	0.58	0.0553
RealData12	V3	V3	V3	V4	4846.68	4517.20	4847.08	**4472.35**	0.00	0.0004	0.80	0.0139
RealData13	V2	V2	V2	V2	320.22	320.22	**318.20**	324.33	0.33	0.2115	0.66	19.6838
RealData14	V5	V5	V5	V5	**1610.13**	**1608.72**	**1608.73**	1611.91	0.00	0.0027	0.36	0.0886
RealData15	V5	V3	V3	V3	1179.74	**1156.42**	1173.15	1174.32	0.16	0.1630	0.51	0.0205
RealData16	V2	V2	V2	V2	185.76	**160.72**	173.20	170.62	0.06	0.0208	0.57	0.0969
RealData17	V2	V2	V2	V2	**225.20**	**225.20**	**225.21**	227.90	0.10	0.0658	0.61	0.0847
RealData18	V6	V6	V6	V6	924.58	**922.08**	924.59	928.61	0.00	0.0020	1.00	0.0081
RealData19	V3	V3	V3	V2	**121.12**	**121.11**	**121.14**	123.43	0.52	0.0548	1.00	0.0039

RealData20	V3	V3	V3	V3	**330.03**	**330.03**	**330.03**	333.96	0.04	0.0854	1.00	0.0106
RealData21	V3	V3	V3	V3	**211.43**	**211.43**	**211.44**	215.00	0.00	0.0121	0.86	0.1566
RealData22	V2	V2	V2	V2	328.53	**311.60**	325.57	327.78	0.29	0.1551	0.61	0.0580
RealData23	V2	V2	V2	V2	328.53	**311.60**	325.57	327.78	0.29	0.1551	0.61	0.0580
RealData24	V2	V2	V2	V2	317.50	**302.56**	315.09	317.30	0.32	0.1279	0.62	0.0609
RealData25	V2	V2	V2	V2	90.92	**88.12**	89.79	**87.69**	0.50	0.0791	0.95	0.0281
RealData26	V2	V2	V2	V2	**170.72**	**170.72**	**170.74**	174.88	0.04	0.1498	1.00	0.0425
RealData27	V2	V2	V2	V2	**126.19**	**126.19**	**126.21**	129.06	0.23	0.1030	0.50	0.1296
RealData28	V3	V3	V3	V3	90.95	**75.18**	79.97	81.46	0.05	0.0565	0.84	0.1589
RealData29	V3	V3	V3	V3	**100.94**	**100.94**	**100.96**	104.50	0.11	0.1299	0.95	0.1495
RealData30	V4	V4	V4	V4	**31.42**	**31.41**	**31.46**	35.81	0.00	0.0020	0.96	0.0673
RealData31	V2	V2	V3	V3	115.67	**108.14**	112.60	115.06	0.00	0.0015	0.40	0.0593
RealData32	V3	V3	V3	V3	**373.03**	**371.48**	**371.43**	374.79	0.04	0.0323	1.00	0.0017
RealData33	V3	V3	V3	V3	**318.92**	**318.89**	**318.93**	321.86	0.00	0.0000	0.29	0.1760
RealData34	V3	V2	V3	V3	575.67	**571.19**	573.11	574.19	0.39	0.0479	0.91	0.1071
RealData35	V2	V2	V2	V2	**36.09**	**36.09**	**36.13**	40.39	0.00	0.0037	0.33	0.3029
RealData36	V3	V3	V3	V3	**61.37**	**61.24**	**61.38**	**61.40**	0.17	0.0761	0.74	0.0843
RealData37	V2	V2	V2	V2	**129.51**	**128.19**	**127.96**	130.43	0.00	0.0002	0.99	0.0043
RealData38	V5	V5	V5	V5	**57.48**	**57.48**	**57.52**	59.67	0.11	0.1144	0.88	0.0861
RealData39	V3	V3	V3	V3	519.28	507.39	**504.29**	517.81	0.01	0.0043	1.00	0.0059
RealData40	V2	V2	V2	V2	1079.21	1075.96	**1072.52**	1075.28	0.11	0.0097	0.22	0.0423
RealData41	V3	V4	V3	V3	**259.15**	263.92	**259.16**	262.88	0.50	0.2300	1.00	0.0492
RealData42	V4	V4	V4	V4	2493.92	**2421.09**	2425.66	2430.10	0.14	0.0817	0.56	0.0139
RealData43	V2	V2	V2	V2	5947.00	**5926.00**	**5926.37**	5930.82	0.05	0.0099	1.00	0.0007
RealData44	V2	V2	V2	V2	**60.83**	**60.83**	**60.86**	64.06	0.07	0.0380	0.35	0.0933
RealData45	V2	V2	V2	V2	117.90	**109.25**	116.42	118.57	0.49	0.1603	1.00	0.0071
RealData46	V2	V2	V2	V2	79.40	**74.32**	**74.59**	**74.93**	0.00	0.0019	0.59	0.0702
RealData47	V2	V2	V2	V2	**81.25**	**81.12**	**81.28**	83.27	0.71	NA	1.00	NA
RealData48	V2	V2	V2	V2	236.58	**234.42**	**233.82**	**235.48**	0.49	0.0630	1.00	0.0047
RealData49	V3	V2	V3	V2	**66.31**	**64.59**	**66.33**	67.57	0.88	0.0402	0.99	0.0140

RealData50	V4	V4	V4	V4	3234.23	**2753.47**	2761.79	2807.68	0.14	0.0088	1.00	0.0003
RealData51	V2	V2	V2	V2	2643.51	2482.66	2499.69	**2516.30**	0.08	0.0068	1.00	0.0066
RealData52	V4	V3	V3	V3	2154.08	**2142.58**	**2142.47**	**2143.80**	0.56	0.0422	0.94	0.0431

Annexe III. Table S2.9: Summary of multivariate real datasets results (variable selected by the different models; AICc of the different models; estimated values of parameters L and K with their standard deviation for model LestKest). Bold AICc values indicate the best model for the real dataset in the row, and models with a difference in AICc of less than 2.0 points. Underlined variable names are for cases where the model chose a different variable from the other models.

Dataset	Best variables selected				AICc				Estimated asymptotes			
	L0K1	GAM free	GAM monotone	LestKest	L0K1	GAM free	GAM monotone	LestKest	L (mean)	L (sd)	K (mean)	K (sd)
RealData1	V5 V2	V5 V4	V5 V2	V5 V2	83.88	**68.11**	83.91	76.11	0.19	NA	0.92	NA
RealData4	V3 V4	V3 V4	V3 V2	V3 V4	994.44	**820.51**	895.52	884.47	0.44	0.0223	0.94	0.0210
RealData5	V2 V3	V2 V3	V2 V4	V2 V3	3936.68	**3896.31**	3949.63	3929.05	0.18	0.1294	0.80	0.0408
RealData6	V3 V2	V3 V2	V3 V2	V3 V2	1343.75	**1286.19**	1307.86	1325.69	0.01	0.0060	0.90	0.0199
RealData7	V5 V2	V2 V5	V5 V2	V5 V8	226.93	**208.32**	223.78	211.71	0.15	0.0316	0.76	0.0510
RealData8	V3 V2	V3 V2	V3 V2	V3 V2	267.32	**254.24**	266.51	265.89	0.10	0.0377	0.98	0.0548
RealData9	V3 V2	V3 V2	V3 V2	V3 V2	43.42	**23.97**	34.02	46.52	0.09	0.0452	1.00	0.0028
RealData10	V2 V3	V2 V3	V2 V3	V2 V3	**1424.21**	**1423.02**	1432.35	1426.51	0.13	0.0528	0.34	0.1088
RealData11	V2 V3	V2 V3	V2 V3	V2 V3	**8562.67**	**8560.85**	8591.37	8562.90	0.24	0.0479	0.76	0.1412
RealData12	V3 V4	V3 V4	V3 V4	V4 V3	4411.27	**3495.87**	4441.52	3715.12	0.00	0.0002	0.91	0.0101
RealData14	V5 V2	V5 V2	V5 V2	V5 V2	1538.76	**1498.34**	1528.34	1533.02	0.03	0.0109	0.71	0.1310
RealData15	V5 V3	V3 V2	V3 V5	V3 V5	1139.25	**1108.96**	1116.20	1120.28	0.27	NA	0.64	NA
RealData16	V2 V3	V2 V3	V2 V3	V2 V3	180.33	**157.70**	166.95	166.43	0.05	0.0189	0.55	0.0613

RealData18	V6 V9	V6 V9	V6 V9	V6 V9	846.37	**841.99**	846.38	**842.82**	0.02	0.0136	0.93	0.0272
RealData19	V3 V2	V3 V2	V3 V2	V2 V3	**118.61**	**117.31**	121.14	121.76	0.17	0.1349	0.71	0.0949
RealData20	V3 V2	V3 V8	V3 V8	V3 V2	317.09	**303.87**	**304.62**	321.20	0.00	0.0404	1.00	0.0115
RealData21	V3 V8	V3 V8	V3 V8	V3 V8	203.23	**197.13**	200.38	201.54	0.02	0.0578	0.71	0.0864
RealData28	V3 V4	V3 V5	V3 V5	V3 V5	77.82	**61.59**	65.48	68.81	0.00	0.0009	0.67	0.0857
RealData29	V3 V2	V3 V2	V3 V2	V3 V2	**102.88**	**102.88**	134.87	106.77	0.01	0.0048	0.89	0.0509
RealData30	V4 V2	V4 V2	V4 V3	V4 V2	19.43	**13.37**	18.35	24.26	0.00	0.0012	1.00	0.0010
RealData31	V2 V3	V2 V3	V3 V2	V3 V2	115.72	**105.44**	**103.75**	120.11	0.00	0.0110	1.00	0.1303
RealData32	V3 V4	V3 V4	V3 V4	V3 V4	351.83	**345.77**	371.43	355.88	0.00	0.0261	1.00	0.0029
RealData33	V3 V5	V3 V5	V3 V2	V3 V5	314.05	**309.32**	318.30	312.06	0.00	0.0000	0.21	0.0625
RealData34	V3 V2	V2 V3	V3 V2	V3 V2	576.43	**550.51**	561.18	575.43	0.43	0.0505	0.92	0.0945
RealData35	V2 V3	V2 V3	V2 V3	V2 V3	38.29	**32.48**	39.15	42.79	0.00	0.0042	0.36	0.3611
RealData36	V3 V2	V3 V2	V3 V2	V3 V2	57.90	**55.32**	61.38	62.05	0.00	0.0043	0.82	0.1641
RealData37	V2 V3	V2 V3	V2 V3	V2 V3	**6.02**	**6.02**	129.51	10.06	0.00	0.0001	1.00	0.0001
RealData38	V5 V3	V5 V3	V5 V3	V5 V3	53.03	**50.25**	53.07	57.78	0.00	0.0099	1.00	0.0060
RealData39	V3 V4	V3 V4	V3 V4	V3 V4	512.11	501.39	**498.46**	512.03	0.01	0.0041	1.00	0.0029
RealData40	V2 V3	V2 V3	V2 V3	V2 V3	1079.68	1076.51	1073.13	**1065.56**	0.11	NA	0.30	NA
RealData41	V3 V2	V4 V3	V3 V2	V3 V2	**258.30**	261.20	**258.59**	**259.48**	0.00	0.0429	0.82	0.0518
RealData42	V4 V3	V4 V3	V4 V3	V4 V2	2492.58	**2417.66**	2422.47	2428.71	0.10	0.0810	0.56	0.0138
RealData43	V2 V4	V2 V4	V2 V4	V2 V4	5784.97	**5751.51**	5759.39	5771.67	0.04	0.0098	0.98	0.0152
RealData46	V2 V3	V2 V3	V2 V3	V2 V3	79.02	**67.95**	**68.51**	79.44	0.00	0.0047	0.74	0.1140
RealData47	V2 V4	V2 V4	V2 V4	V2 V4	71.70	**69.28**	71.73	75.42	0.37	0.2308	1.00	0.0022
RealData49	V3 V2	V2 V3	V3 V2	V2 V5	67.45	**63.85**	66.97	72.87	0.87	0.0478	0.97	0.0191
RealData50	V4 V2	V4 V3	V4 V3	V4 V3	3217.11	**2736.23**	2757.31	2796.85	0.14	0.0088	1.00	0.0003
RealData51	V2 V4	V2 V4	V2 V4	V2 V4	2555.02	**2278.55**	2303.43	2483.84	0.08	0.0066	1.00	0.0397
RealData52	V4 V3	V3 V2	V3 V2	V3 V2	2110.37	**2048.27**	2052.00	2082.18	0.45	0.0942	0.97	0.0403

Annexe III. S3 : RMSE et biais Binomial.

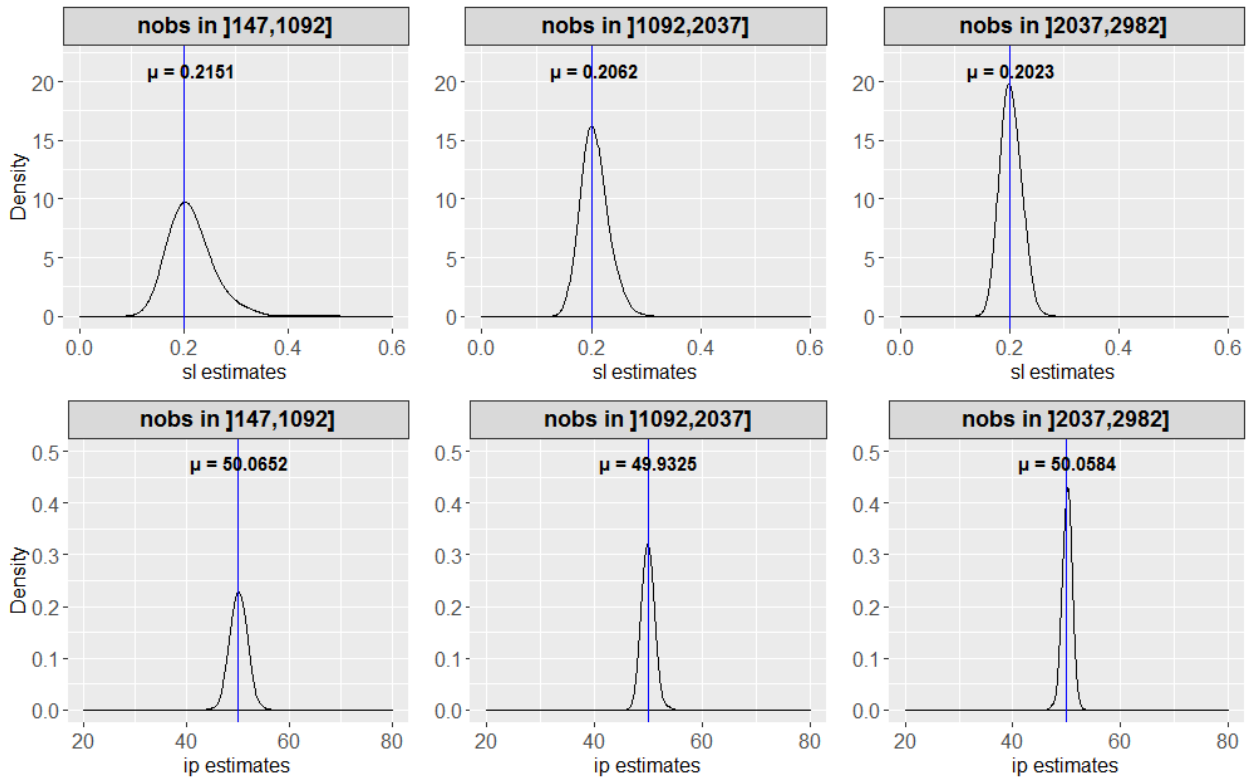
Annexe III. Section S3.1: Method description

We tried to estimate the true global variability of each estimator for the whole series with the Root Mean Square Error (RMSE), which is accessible only when applying models to simulated data. RMSE is based on the square root of the mean squared difference between the mean of the estimate and the true values:

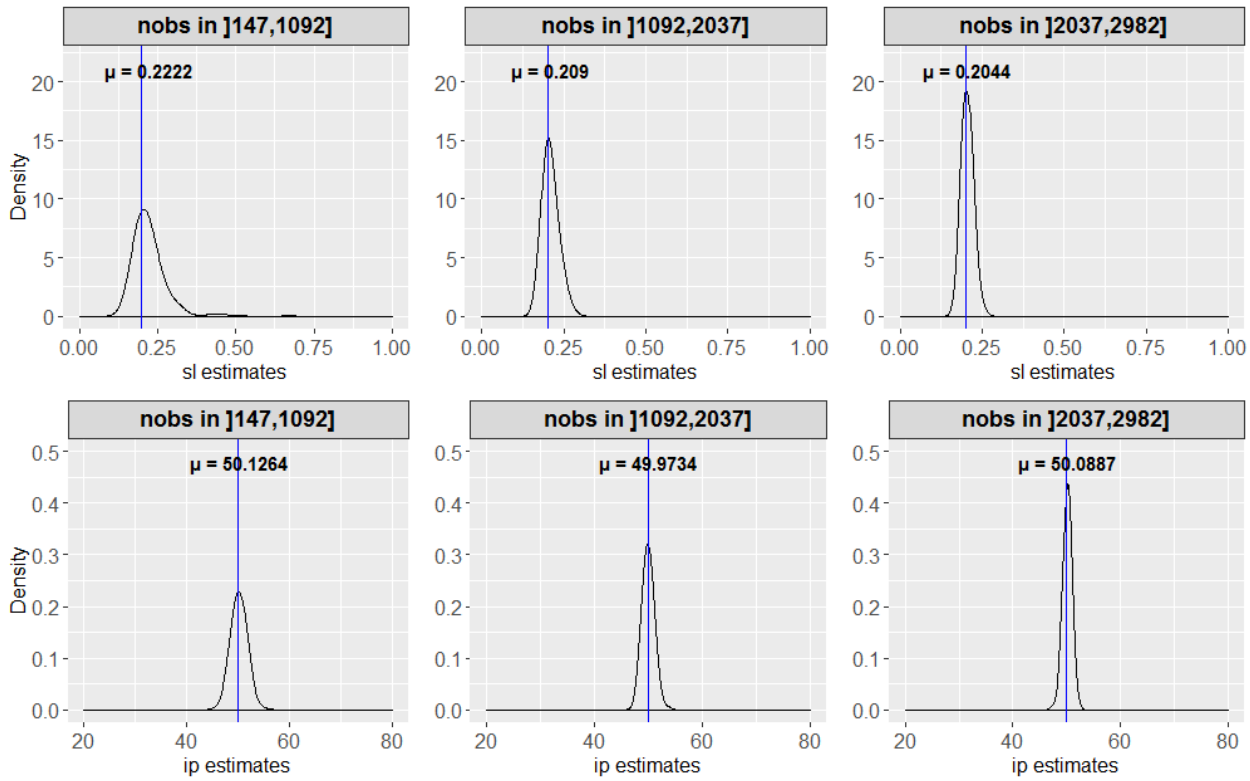
$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Estimated_i - True_i)^2}{N}}$$

Logistic models yield biased odds ratios, especially with small sample sizes; they generally slightly overestimate the parameter and therefore positively skew the distribution function of the odds ratio. This bias decreases as sample size increases, and the distribution converges to a normal distribution centered on the estimated effect (Nemes et al. 2009). We wanted to get a glimpse of this binomial bias for the L0K1 and LestKest models, and compare them to see if LestKest could bring the distribution closer to a normal distribution. For that purpose, we used the first and second simulated series of datasets (Scenario1 and Scenario2) to graphically compare the sampling distribution of the inflection point (*ip*) and the slope (*sl*) in the L0K1 and LestKest models, for different sample size classes (to confirm the impact of the number of observations on the bias).

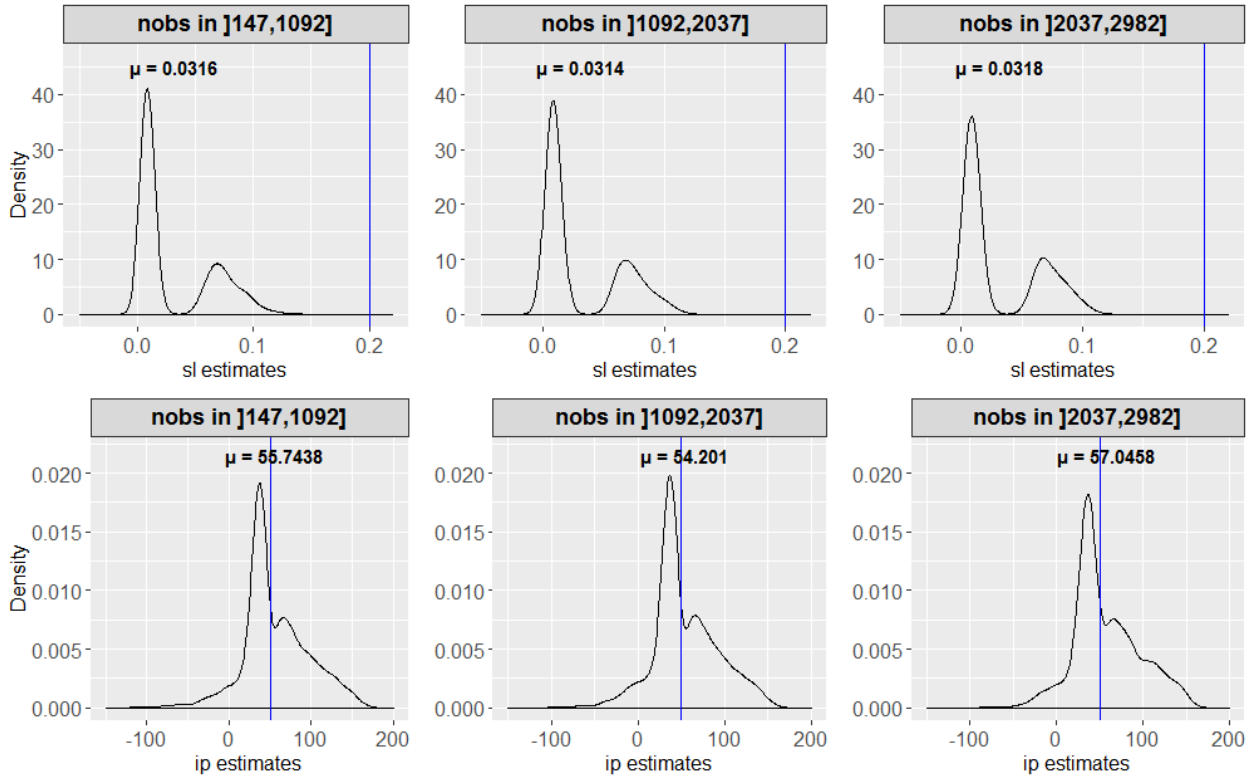
Annexe III. Figure S3.1: Sampling distribution of optimal logistic regression coefficient estimates (with model L0K1 on Scenario1) for different sample size classes. nobs in = the number of observations is included in the range.



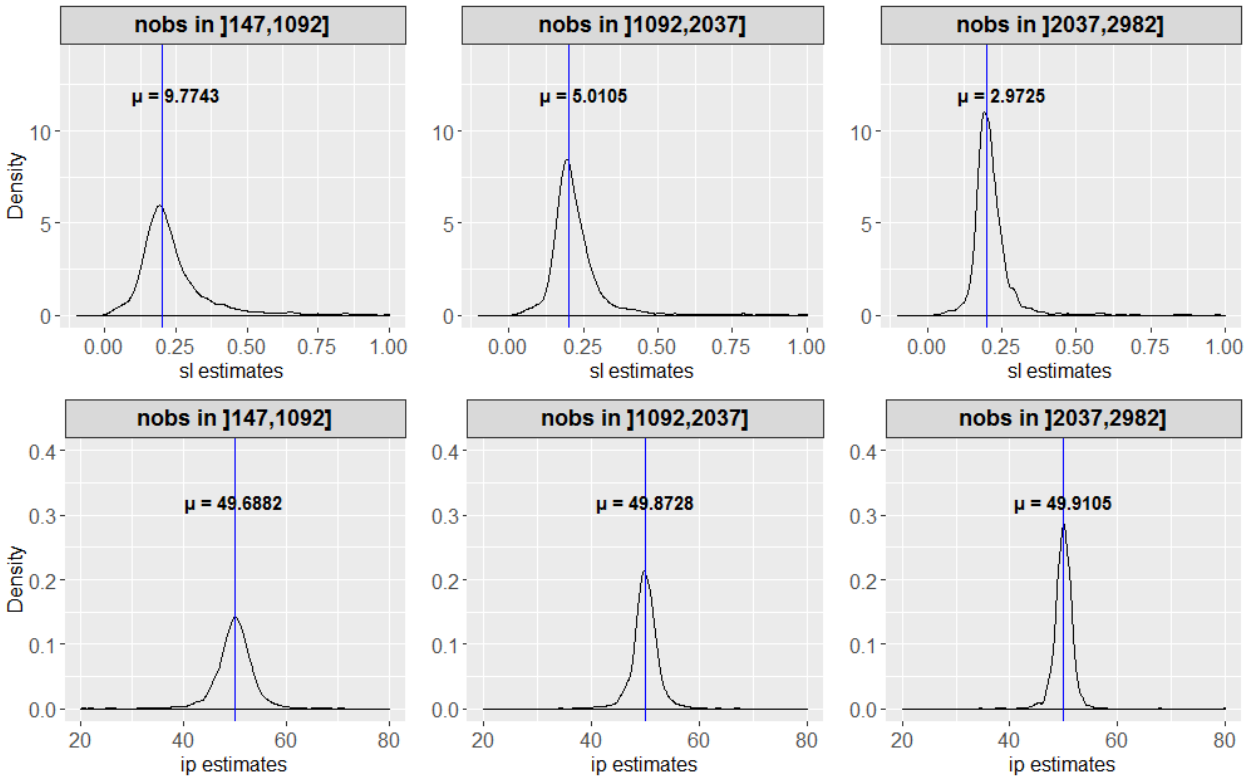
Annexe III. Figure S3.2: Sampling distribution of optimal logistic regression coefficient estimates (with model LestKest on Scenario1) for different sample size classes. nobs in = the number of observations is included in the range.



Annexe III. Figure S3.3: Sampling distribution of optimal logistic regression coefficient estimates (with model L0K1 on Scenario2) for different sample size classes. nobs in = the number of observations is included in the range. Seven observations were removed from nobs in]147,1092] for ease of reading.



Annexe III. Figure S3.4: Sampling distribution of optimal logistic regression coefficient estimates (with model LestKest on Scenario2) for different sample size classes. Seven observations were removed from nobs in]147,1092] for ease of reading. For ease of reading, 495, 91, 19, 20, and 3 observations were respectively removed from the first five figures for ease of reading).



Annexe III. Section S3.2: Results and Discussion

When applied on datasets where the true link function is the classical logistic function (Scenario1), the distribution function of the odds ratio with a logistic regression (model L0K1 and LestKest) was a bit skewed in smaller samples sizes and became more symmetric with increasing sample size (Annexe III. Section S3.1 and Annexe III. Section S3.2). Similarly, the

mean of the estimators (μ in Appendix figures), overestimated on small sample sizes, approached the real value (blue line in Appendix figures) with larger sample sizes.

With model L0K1, the sampling distribution of the slope (*sl*) was dramatically underestimated for all sample size classes (approximately 0.03 instead of 0.2). Furthermore, the sampling distribution of both slope and inflection point (*sl* and *ip*) was bimodal when applied to datasets where the asymptotes of the link function were derived from zero and one (Scenario2) and did not improve with increasing sample size class (Annexe III. Section S3.3).

With the model LestKest and the same simulation scenario (Scenario2), the sampling distribution of the parameters was no longer bimodal, although it was slightly positively skewed in smaller sample size classes. Symmetry was partially recovered with increasing sample size (Annexe III. Section S3.4). However, it seems obvious that extreme values for slope completely distort the average, even though graphically the distribution seems well centered on the real value. This deviation between the estimated mean and the real value decreases as sample size increases (going from a multiplying factor of 4.89 to 1.49 between the smallest sample size class to the largest sample size class).

To conclude, in this study, we first have shown that binomial bias decreases considerably with an increasing number of samples, in line with Nemes et al. (2009). We have also highlighted the fact that the basic logistic regression model (L0K1), when applied to datasets whose asymptotes are different from zero and one, tends to present a bimodality in the estimation of its parameters; a bimodality, which does not appear with the LestKest model. However, extreme values for the slope parameter appeared in the LestKest model, making it impossible for us to calculate the RMSE as we had first hoped.

IV. ANNEXE IV (CHAPITRE 4)

Annexes associées au Chapitre 4 du manuscrit de thèse.

Annexe IV. S1 : Développement des modèles et limites

Annexe IV. Table S1.1: Distributions a priori et valeurs initiales utilisées dans les modèles

Traitement de l'asymptote haute	**Paramètres**	**Distribution a priori**	**Valeurs initiales**
Tous les modèles (K1, K, Kgroup, Ksp and Kspspgroup)	β_0	$N(0.0,5.0)$	$N(0,0.3)$
	α	$N(0.0, exp(\sigma_\alpha^2))$	$\alpha \sim N(0,0.3)$ $\sigma_\alpha^2 \sim log\ (U(0.0,0.2))$
	$\overline{\beta_{1j}}$	$N(m_{gj}, \sigma_{\beta 1}^2)$, with $m_{gj} = envgrouppar(i,g) * spgroup(j,g)$	$\overline{\beta_{1j}} \sim N(0,0.3)$ $\sigma_{\beta 1}^2 \sim log(U(0.0,0.2))$
	$envgrouppar(i,g)$	$N(0.0,10.0)$	$envgrouppar \sim N(0,0.3)$
Traitement K	K, with $K = \frac{1}{1+exp(-K_{logit})}$		$K_{logit} \sim N(0.0,1.0)$
Traitement Kgroup	σ_α^2	$N(0.0,5.0)$	$log\ (U(0.0,0.2))$
	$\sigma_{\beta 1}^2$	$N(0.0,5.0))$	$log(U(0.0,0.2))$
	K_{group}, with $K_{group} = \frac{1}{1+exp(-K_{group.logit})}$	$K_{group.logit} \sim N(0.0,10.0)$	$K_{group.logit} \sim N(4.0,0.5))$
Traitement Ksp	K_{sp}, with $K_{sp} = \frac{1}{1+exp(-K_{sp.logit})}$	$K_{sp.logit} \sim N(0.0,10.0)$	$K_{sp.logit} \sim N(4.0,0.5))$
Traitement Kspspgroup	$K_{spspgroup}$, with $K_{sp}spgroup = \frac{1}{1+exp(-K_{spspgroup.logit})}$	$K_{spspgroup.logit} \sim N(m_{gj}, \exp(\sigma_K^2)$	$K_{spspgroup.logit} \sim N(4.0,0.5))$
Tous les modèles avec au moins une variables latente	λ	above the diagonal = 0 the diagonal = exp(N(0.0,5.0)) below the diagonal = N(0.0,5.0)	$N(0,0.3)$
	z'	$N(0.0,1.0)$	$N(0,0.3)$

Annexe IV. Note S1.1: Note de F. Gosselin sur l'utilisation de la librairie « tioptimTMB ».

Iterated Tempered Optimization: an interesting improvement over simple optimization tools to reach more stable and optimal results, with a focus on TMB

Frédéric Gosselin

INRAE, UR EFNO, Domaine des Barres, 45290 Nogent-sur-Vernisson

Introduction

Current practice in parametric statistical model fitting is often based either on optimization procedures for frequentist models or on Bayesian fitting (often of Monte Carlo Markov Chain type; MCMC) for Bayesian models. We will here restrict our attention to the first kind of tools, especially tools of the gradient descent type. The aim of these tools is to reach the global optimum of the statistical model, i.e. the set of parameters that correspond to the global maximum likelihood of the model. Yet, the practice around traditional statistical model fitting tools (e.g. the tools found in the software R such as glm, glmm, glmmTMB…) often rely on a simple optimization tool, such as nlminb. This seems to be also the standard practice proposed with TMB (Template Model Builder; (Kristensen *et al.*)), a numerically very efficient model fitting framework based on model marginalization (over random effects) that relies on automatic differentiation and Laplace approximation. Yet, it is a standard observation with such tools that the "optimal" result found depends on the initial values of the parameters. The user is therefore not very confident that the result found is indeed the global maximum likelihood.

We here propose a strategy based on classical gradient-descent optimization tools to circumvent this problem. It is based on two ideas:

(i) tempering the objective function, i.e. putting it at higher temperatures to smooth it and render movements between local and global optima easier. The idea is more precisely to start with a first optimization with a tempered function and then (gradually or not) "cool" the function, until it reaches the untransformed objective function. This idea comes from the stochastic optimization literature (e.g. simulated annealing) or from some Bayesian techniques based e.g. on parallel tempering (Swendsen & Wang, 1986, Geyer, 1991, Earl & Deem, 2005);
(ii) repeating optimization over the same objective function from different starting values until the function value stabilizes.

We describe the associated algorithm and apply it to one case study.

Material and Methods

Description of the algorithm

The algorithm we propose is quite simple. Its main arguments are:

– the arguments used to build the TMB functions (TMB_args), provided as a list with components such as *data*, *random* and *DLL* (and possibly other arguments). The *DLL* is the dll file associated with the TMB code of the function, involving a data element called temperature used to temper the function. The objective function is obtained when temperature equals 1. The *parameters* argument is not provided here as it is managed independently in tioptimTMB.
– the argument that contains a list of initial values for consideration, called *init_pars*;
– the arguments for the tempering part of the algorithm (called temper_args): it can consist of either a named list in, which case, there will be only one tempering scheme, or, a list of named lists (or a list of lists of named lists), in which case, there will be as many tempering schemes as the number of lists, and the best result of the different tempering schemes will be kept. The unit tempering list has either:
- a temperature numeric vector component giving the vector of temperatures to be considered in the tempering scheme (the last temperature should be 1.0, otherwise a 1.0 will be added to it)
- a Textr (positive double argument, the extremum temperature being considered) value and a nTa1 (positive integer argument, the number of temperatures different from 1.0) value in which case temperatures will be calculated based on control$tempering_scheme
- or a Textr vector with N values (N>1) and a nTa1 value (or vector of length N) in which case N lists of temperatures will be created based on control$tempering_scheme, corresponding to each Textr [and nTa1 if of length N] value in turn
- or refers to random tempering scheme and has two components:
  - a first component called fT ghat is an expression of a random function to draw a temperature
  - a second component called control which arguments are listed in Appendix 1.

– the argument to control the overall process (argument called *control*), which components are listed in Appendix 2. It especially has a component called *epsilon* which is a positive, typically small, number that specifies when the iteration stops once out of the tempering phase (in terms of difference of two successive functions. It also has a component called *nOptMax* which specifies the maximum number of successive optimizations once in the untempered phase.
– the arguments specifying the optimization methods: first, the argument called optim_type specifying which function to use for optimization (current possibilities include: 'optimr:optimr', 'optimr::polyopt', 'optimx::optimx', 'nlminb'); second, the argument called optim_args which is a named list of arguments (or list of such lists) to be passed to the optimization function being selected in optim_type. It is indeed possible to use different optimization procedures, but we will here only describe the

use of the nlminb function with arguments chosen to render it more stringent than default arguments (for example: rel.tol=1e-15,eval.max=1000,iter.max=200).
– an argument called AIC_fun, which is an expression that provides the possibility to use a formula to specify which transformation of the TMB function is to be used to compare multiple fits. Default to NULL, in which case the TMB function value is used to compare fits.

The pseudo-code for the tempered optimization algorithm for TMB is the following one for the first two possibilities of temper_args:

- if *temperatures* is NULL, calculate *temperatures* of the tempered phase from *Tmax*, *nTa1*, *control$templad*
- #tempering phase
- for (i in 1:length(init_pars))
- choose init_pars[[i]] as initial parameter values
  - for (i In 1:length(temperatures))
    - calculate the data for the TMB function, including temperature=temperatures[i]
    - form the TMB function TMB.fn with these data and temperature with the function MakeADfun
    - Optimize TMB.fn, using current values of parameters as starting values, method optim_type and potentially other arguments to specify the optimization method
    - Update parameters based on the optimization results
  - Store values of TMB.fn$fn() and parameters
- #iteration phase
- Put iteration number to 0.
- Take as parameters the ones corresponding to the lowest TMB.fn$fn() out of the tempering phase
- while (abs(obj.old-obj.new)>control$epsilon) and (iteration number<control$nOptMax)
  - if first iteration:
    - calculate the data for the TMB function, including temperature=1
    - form the TMB function TMB.fn with these data and temperature with the function MakeADfun
  - calculate obj.old=TMB.fn() twice
  - optimize TMB.fn, using current values of parameters as starting values, method optim_type and potentially other arguments to specify the optimization method
  - Update parameters based on the optimization results and iteration number
  - Update parameters based on the optimization results
  - calculate obj.new=TMB.fn(parameters) twice

- Put the results of the tempered optimization (parameters, TMB.fn()) in final list object.

For the last possibility of temper_args, based on random temperatures generation – when argument fT is not null –, the pseudo-code is the following one (with temper_args$control$mixIV=FALSE and temper_args$control$bestIV=TRUE to simplify the presentation):

- # initial tempering phase
- for (i in 1:temper_args$control$Ninit)
  - draw temperature[i] at random – can have if done at the beginning in case temper_args$control$regularize is TRUE to ensure a minimal coverage of the temperature interval: [temper_args$control$Tmin; temper_args$control$Tmax]
  - Do the tempering phase of the previous algorithm with temperatures created from Textr= temperature[i] and nta1= temper_args$control$nTa1 with a geometric ladder
  - If (in case temper_args$control$nIVs.init>1)
- for (j in 1: (case temper_args$control$nIVs.init-1))
  - Use independent starting values
  - Do the tempering phase of the previous algorithm with temperatures created from Textr= temperature[i] and nta1= temper_args$control$nTa1 with a geometric ladder and these new starting values
    - If TMB.fn$fn() for the current IV is better than before the j loop of at least temper_args$control$fntol and if it is not the best overall result
- Test the optimization with the best overall temperature and the stating value associated with j
- Keep as starting value for the first part of next step the one with lowest TMB.fn$fn() value (including for previous i's)
- Store the best TMB.fn$fn() value for i

- # final tempering phase
- count_temper= temper_args$control$Ninit
- while (count_temper< temper_args$control$Nmax) & ((count_temper< temper_args$control$Nmin) | (# of optimised TMB.fn$fn() within temper_args$control$AICtol of the best optimized TMB.fn$fn() < temper_args$control$Nconv))
  - Draw if in case 1, 2 or 3 based on probabilities in temper_args$control$probs.sampling
  - If in case 1: draw a random initial value for parameters and a random temperature corresponding to past TMB.fn$fn() values within temper_args$control$AICtol of the best optimised TMB.fn$fn()
  - If in case 2: draw a random temperature corresponding to past TMB.fn$fn() values within temper_args$control$AICtol of the best optimised TMB.fn$fn(); take its corresponding starting value; disturb the temperature using parameters temper_args$control$sigma0, temper_args$control$mult.sigma0 and temper_args$control$power

- If in case 3: draw a random temperature corresponding to past TMB.fn$fn() values within temper_args$control$AICtol of the best optimized TMB.fn$fn(); take its corresponding starting value; draw another temperature from temper_args$fT
- Do the tempering phase of the previous algorithm with temperatures created from Textr= temperature and nta1= temper_args$control$nTa1 with a geometric ladder
- If (in case temper_args$control$nIVs.init>1)

• for (j In 1: (case temper_args$control$nIVs.init-1))

- Use independent starting values associated with j
- Do the tempering phase of the previous algorithm with temperatures created from Textr= temperature[i] and nta1= temper_args$control$nTa1 with a geometric ladder and these new starting values

▪ If TMB.fn$fn() for the current IV is better than before the j loop of at least temper_args$control$fntol and if it is not the best overall result

• Test the optimization with the best overall temperature and the stating value associated with j

• Keep as starting value for the first part of next step the one with lowest TMB.fn$fn() value (including for previous i's)

• Store the best TMB.fn$fn() value for i

• #iteration phase

• Put iteration number to 0.

• Take as parameters the ones corresponding to the lowest TMB.fn$fn() out of the tempering phase

• while (abs(obj.old-obj.new)>control$epsilon) and (iteration number<control$nOptMax)

- if first iteration:

▪ calculate the data for the TMB function, including temperature=1

▪ form the TMB function TMB.fn with these data and temperature with the function MakeADfun

- calculate obj.old=TMB.fn() twice
- optimize TMB.fn, using current values of parameters as starting values and nopt, method and potentially other arguments to specify the optimization method
- Update parameters based on the optimization results and iteration number
- Update parameters based on the optimization results
- calculate obj.new=TMB.fn(parameters) twice

• Put the results of the tempered optimization (parameters, TMB.fn()) in final list object.

Case study: data and model

The data analyzed correspond to the species richness (i.e. number of species) of saproxylic beetles caught with window traps in forests throughout France, denoted as Y. We wished to explain the variations of species richness by the volume of deadwood measured at a local scale (rough area of measurement: 0.2ha), denoted as x (for more details cf. Godeau et al. Accepted or Chapter 2 of Ugoline Godeau's PhD).

The model we here chose is one related to those proposed by Godeau et al. (Accepted). It is a sigmoid model with four parameters and with:

$$\mu[\mathrm{n}] = \exp\left[\mu_a + \sigma_a * z_a\left[\mathrm{forest}[\mathrm{n}]\right] + \rho * (\log(\mathrm{N}[\mathrm{n}]) - \mathrm{mN})/\sqrt{\mathrm{vN}}\right] \times \ldots$$

$$\ldots \left(1 + \frac{1}{1+\exp(-d)}\left(1 - \frac{2}{\left(1 + \exp\left(c * \frac{\mathrm{x}[\mathrm{n}] - x_{05} - b * (x_{95} - x_{05})}{sd_x}\right)\right)}\right)\right)$$

where:

- n is the index of the observation
- $\mu[\mathrm{n}]$ is the modelled mean of $Y[\mathrm{n}]$
- $\mathrm{forest}[\mathrm{n}]$ denotes the forest in which the n-th observation occurred
- $\mu_a + \sigma_a * z_a\left[\mathrm{forest}[\mathrm{n}]\right]$ allows a different mean level by forest based on a Gaussian standard distribution for z_a
- $\rho * (\log(\mathrm{N}[\mathrm{n}]) - \mathrm{mN})/\sqrt{\mathrm{vN}}$ models the effect of the – standardized – number of actve traps $\mathrm{N}[\mathrm{n}]$ (with a power model)
- b controls the position of the inflexion point
- c controls the slope at the inflexion point
- d controls the inter-asymptotes variation in the sigmoid function

In the model, $Y[\mathrm{n}]$ is assumed to follow a negative binomial distribution with mean $\mu[\mathrm{n}]$ and variance $disp\ \mu[\mathrm{n}]$, where disp is an estimated parameter greater than 1.0. The model includes priors for all parameters, that are non-informative and that have either Gaussian distributions (μ_a, b, c, d) or t-distributions with order 4 (ρ, log(σ_a) and log(*disp*-1)). The model is therefore a Bayesian non-linear hierarchical model linking species richness to deadwood volume.

Initial values were drawn from random functions that were relatively informative for μ_a, b, c, log(disp) and ρ (respectively Gaussian with mean log(3.4) and standard deviation 0.3; uniform between 0.2 and 0.3; Gaussian with mean 10 and standard deviation 0.3, uniform draw between 2 and 4 and exponential of a Gaussian random draw with mean 0 and standard deviation 0.3), and less informative for d, log(σ_a) and z_a (respecively: uniform draw between 0 and 4, uniform draw between -2 and -1, and standard Gaussian).

The simple optimization of the associated TMB model did not meet convergence issues in that we reached similar function values and parameter values from different initial values. Yet, as we reached such instability problems for a similar model fitted with Bayesian tool, we wished to study the profiled optimum log-posterior density value according to the value of parameter μ_a, to study if the profile was flat around the optimum or not. We then reached an apparently incoherent behaviour with different initial values for each value of μ_a (see Figure 1). The behavior was incoherent because we could hardly explain why the resulting graph had such a non-continuous, erratic behaviour for high values of μ_a.

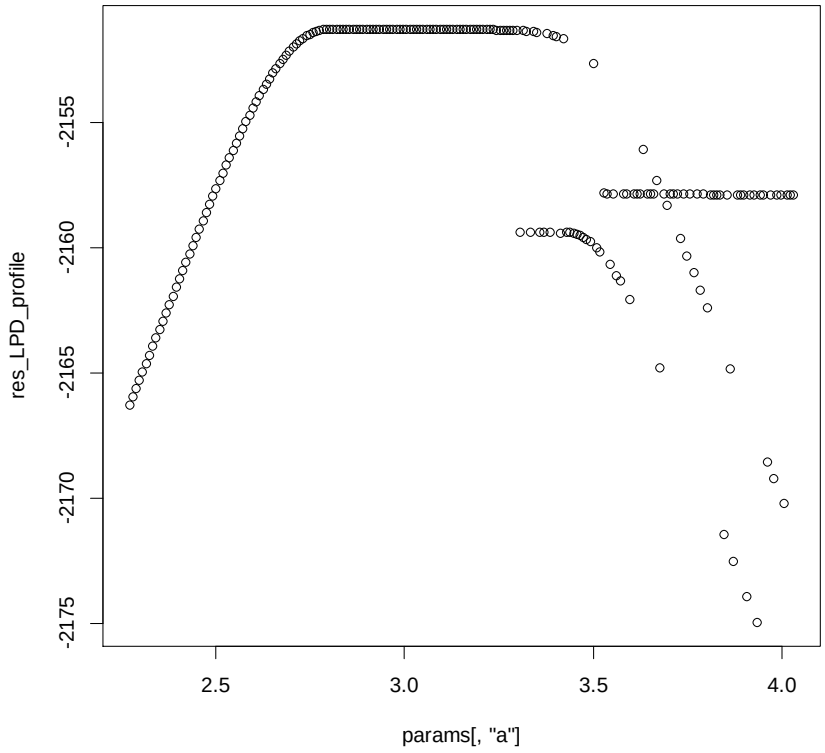


Figure 1. Profiled optimum log-posterior density functions for the model with TMB and successive, fixed values of μ_a (here denoted as a), based on simple optimizations from a unique random initial value for each fixed values of μ_a.

We therefore tested the tioptim.TMB function on this case, with different settings. At the end, we reached a stable behavior with:

- Optimization 1: A first tioptim.TMB use with the random tempering algorithm using the following settings:
temper_args=list(fT=expression(exp(runif(1,0,log(40000)))),control=list(Nmax=50,AICtol=0.2,Ninit=20,nIVs.init=2, fntol=5e-5,
mixIV=FALSE,bestIV=TRUE,regularize=TRUE,probs.sampling=c(0,1,0),Tmin=1,Tmax=40000,log=TRUE)), but without a developed iteration scheme (nOptMax=1),
followed by a second tioptim.TMB with the parameter values coming from the first optimization, but with no tempering and now with a developed iteration scheme (default nOptMax).

The good behaviour of this scheme could be simply due to the fact of iterating the optimization from different initial values of parameters, and not really to the tempering scheme itself. To have an element of comparison with only different initial values, we compared this first optimization with two other optimizations:

- Optimization 2: The first one similar as above except that the first optimization had no tempering but 20 different initial values (the number usually used in the first

optimization since in most cases the tempering ended after the initial tempering phase and since the parameter nIVs.init was equal to 20);
- Optimization 3: The second one similar as above except that the first optimization had no tempering but 80 different initial values (the number of function evaluations in the initial phase of the first model, or something close to the maximum number of maximum values used in the first model – in the few cases where the optimization lasted after the initial tempering phase).

This comparison of the three Optimization schemes was repeated 10 times, with 10 different sets of random initial values. We just compared the resulting Profiled optimum log-posterior density functions graphs of the different schemes.

Results

The tempered Optimization 1 invariably yielded the same Profiled optimum log-posterior density functions graphs, shown in Figure 2. Each replicated Profile likelihood optimization took around five hours of calculus.

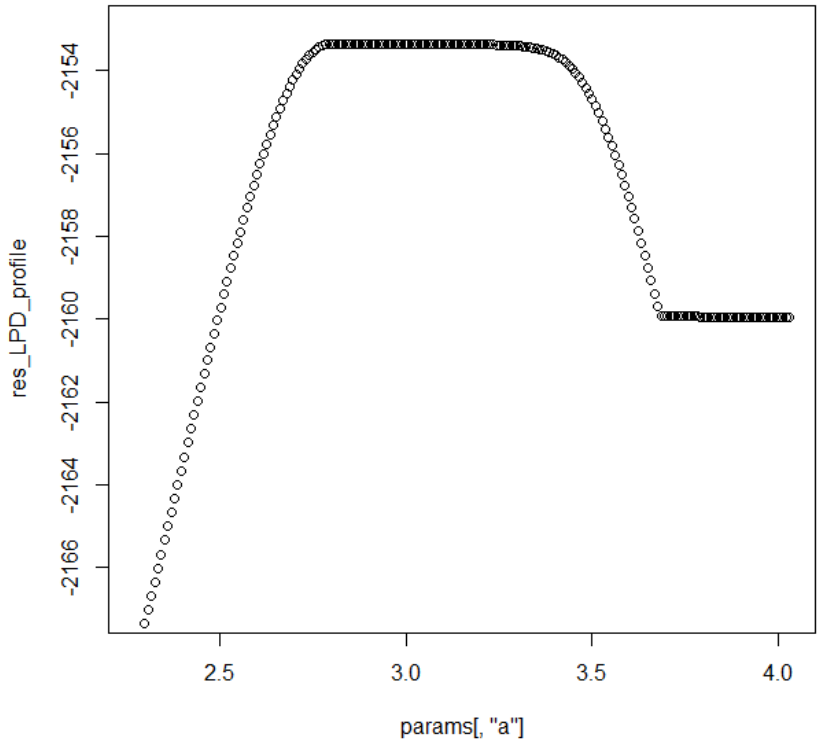


Figure 2. Profiled optimum log-posterior density functions for the model with TMB and successive, fixed values of μ_a (here denoted as a), based on tempered and iterated Optimization 1. We reached this same result for 10 out of 10 different sets of starting values.

Optimization2 was much faster (around one hour and a half by replicated profiled likelihood) but each replicated profile failed (Figure 3)

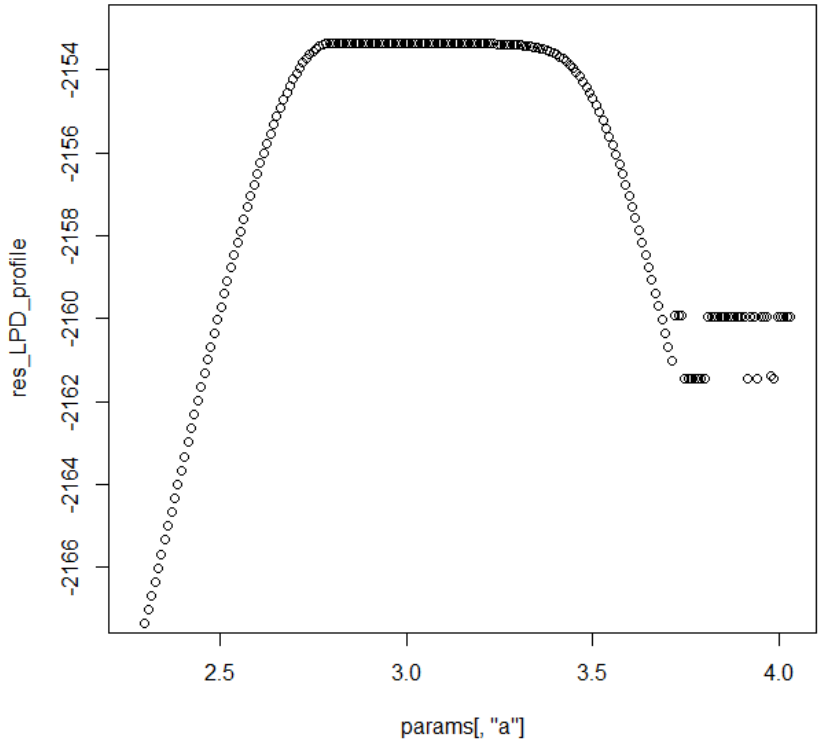


Figure 3. Profiled optimum log-posterior density functions for the model with TMB and successive, fixed values of μ_a (here denoted as a), based on tempered and iterated Optimization 2. We reached incoherent results for every replicated profile.

Optimization 3 was slower than the first two ones (around six hours per replication). Among the first five replicates, three gave coherent Profiled optimum log-posterior density (as in Figure 2), but two gave slightly incoherent ones (cf. Figure 4 for one of the replicates).

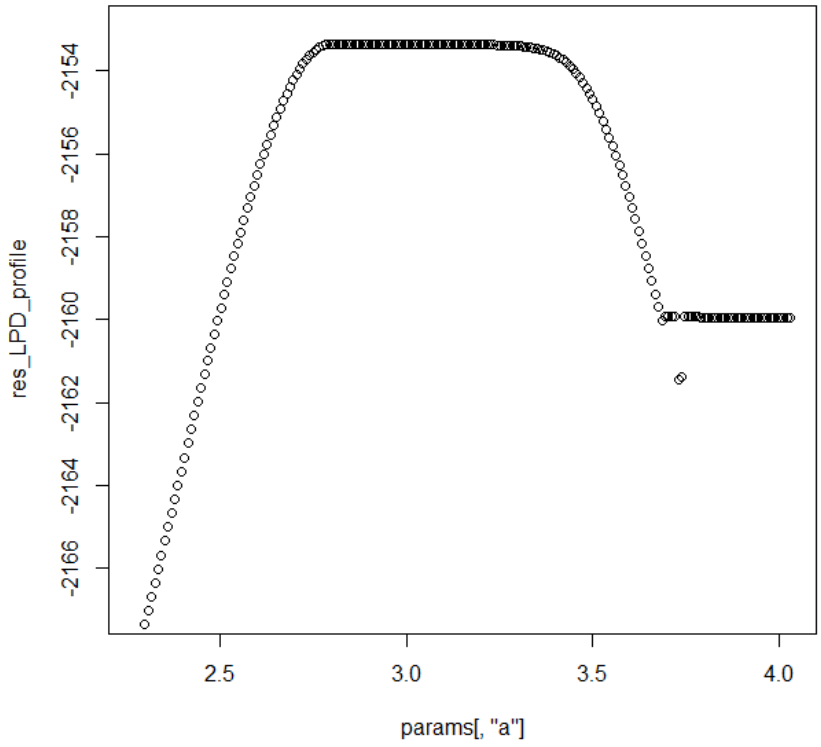


Figure 4. Second replicate Profiled optimum log-posterior density functions for the model with TMB and successive, fixed values of μ_a (here denoted as a), based on tempered and iterated Optimization3. We reached incoherent results for the third replicate as well, but not for the first, fourth and fifth ones.

Discussion

We have highlighted in this note that TMB had optimization problems, even for models of moderate complexity. This is no surprise, since techniques such as Expectation Maximization (EM), were in particular developed based on the limitation of more classical gradient descent optimization procedures ((Dempster *et al.*, 1977)). The use of the function tioptim.TMB highlights the possibility that using repeated initial values, tempering of the objective function and iteration of the optimization procedure may solve some problems met with classical gradient descent optimization procedures. Whether the procedure is efficient in a broader context is unclear. Anyway, the message from his note is that models fit by TMB may be hard to optimize and that care should be given to whether the optimization is really optimal, at the very least by trying various optimizations from different starting values.

References

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Earl, D.J. & Deem, M.W. (2005) Parallel tempering: Theory, applications, and new perspectives. *Physical Chemistry Chemical Physics*, **7**, 3910-3916.

Geyer, C.J. (1991) Markov Chain Monte Carlo Maximum Likelihood. In: *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface*, pp. 156-163. Interface Foundation of North America

Godeau, U., Bouget, C., Piffady, J., Pozzi, T., Gosselin, F. (In Press) The importance of being random! Taking full account of random effects in nonlinear sigmoid hierarchical Bayesian models reveals the relationship between deadwood and the species richness of saproxylic beetles. *Forest Ecology and Management.*

Kristensen, K., Nielsen, A. & et al. TMB: automatic differentiation and Laplace aproximation.

Swendsen, R.H. & Wang, J.S. (1986) Replica Monte Carlo simulation of spin-glasses. *Physical Review Letters*, **57**, 2607-2609.

Appendix 1 : Arguments of the control argument in temper_args :

– Nmax: maximum number of temperatures for which optimization is performed (one with the temperature and one with temperature=1.0 based on the optimized values obtained? with the temperature) (Default: 50).

– Nmin: minimum number of temperatures for which optimization is performed (one with the temperature and one with temperature=1.0 based on the optimized values obtained? with the temperature) (Default: 5).

– Ninit: number of initial random draws of temperature with fT. Should be less than Nmax (Default: 5).

– Nconv: number of replicated optimizations within AICtol of the best optimization at which we declare that the model has converged (Default: 5).

– nTa1: number of temperatures above 1 including the first one (Default: 2).

– nIVs.init: number of Initial Values (IVs) considered for each initial Temperature (including the past best-; Default: 1).

– fntol: tolerance for two function values calculated for a same initial temperature to be considered as equivalent (and not provoke new calculations in best temperatures – Default: 0). Should be nonnegative. Could be a good idea to consider a small positive value (e.g. between 1e-6 and 1 e-4) for speeding the algorithm.

– bestOptOnly.init: logical: should we recalculate with only the best optimization method on the best temperature in case an initial value did best that the best initial value? (Default: TRUE).

– mixIV: logical: should we mix initial values at the beginning? (Default: FALSE, meaning that new temperatures all start with the first initial value, except if nIVs.init is greater than 1 or in the terminal tempering phase in case we are in the first configuration/case of the terminal phase).

– bestIV: logical: should we use the initial values related to the "best" (random within AICtol interval of best) initial value (only valid for phases 2 and 3, and only if mixIV=FALSE in the completely new temperature phase) (Default: FALSE).
– keepbestSOV: logical: should we keep the best set of values (either an initial value or a final value if it has a lower fn value (Default: FALSE)
– AICtol: positive value specifying the tolerance for AIC of optimization to be considered as good as the best optimization (Default: 2).
– probs.sampling: vector with 3 values giving the probability (after the Nmin initial phase) to sample Temperature according to one of the following schemes: (i)-sampling an exact past value among the best ones (with tolerance AICtol); (ii) sampling a past value among the best ones (with tolerance AICtol) and multiplying by a lognormal noise of sd: sigma0*iterations^(-power); (iii) sampling with fT. (Default: c(0.35,0.35,0.3))
– sigma0: scaling parameter for the terminal tempering phase. (Default: 0.3)
– mult.sigma0: multiplier of the scaling parameter for the terminal tempering phase. (Default: 1)
– power: power parameter for the terminal tempering phase. (Default: 1)
– regularize: logical: should we regularize the Nmin first temperature draws, so that the temperature gradient is covered by these draws? (Default: FALSE ; forced to be FALSE if either Tlow or Thigh is NA).
– Tmin: minimum temperature for initial regularization (Default: NA).
– Tmax: maximum temperature for initial regularization (Default: NA).
– log: logical: should we use log scale for initial regularization? (Default: TRUE).

Appendix 2 : Arguments of the control argument in *control* :

- tempering_scheme: a character chain among 'arithmetic', 'geometric', 'doublelog' to specify how temperatures are calculated if argument temperatures of temper_args is NULL. (Default: 'geometric').
- epsilon: minimum difference between successive TMB function values at which iteration is stopped. (Default: 1e-6).
- nOptMax: maximum number of optimization phases in the untempered (Temperature==1) phase. (Deafult: Inf)
- doubleFnCall: logical: should two calls to obj_TMB$fn be made inside the function being optimized? This slows optimization but makes it more rigorous due to the tendency of TMB to give different values on the first call of fn() on new parameters compared to the next calls. (Default: FALSE, but TRUE might be tried as well).
- reMakeADF: logical: should we remake the AD Function at each optimization step and at the end? Should have an impact via the value of random effects. If FALSE, optimizations should be strictly decreasing while if TRUE this is not guaranteed (due to re-initialization of random effects). (Default: TRUE; strongly recommended).
- keepBestParams: logical: should we keep he best overall parameters (TRUE) or the ones associated to the last optimization (even if not the best). (Default: TRUE).
- keepGrad: logical: should we keep the gradient calculated by TMB in the optimization list? (Default: TRUE).

- keepBestParTa1: logical: should we give the last best parameters of TMB function for temperature>1 or not? If not, only nonrandom parameters are changed. (Default: FALSE).
- keepAllIters: logical: should we keep the info of all iterations (except the best one) in the component named "archive" of the output? (Default: TRUE).
- printParamChanges: logical: should we print changes in parameters during optimization? (Default: FALSE).

Annexe IV. Section S1.1 : Arguments et réglages de « TMB » et « tioptimTMB » des modèles pour les étapes de modélisation 2, 3 et 4 (Chapitre 4 – II.3, II.4 et II.5)

Réglages concernant les arguments généraux de « TMB » pour les deux ajustements :

(i) Une méthode d'optimisation nlminb : `method = c("nlminb")`
(ii) Un nombre maximal d'itération de 150 : `maxit = c(150)` (dans optim_args)
(iii) Un nombre limite sur le nombre d'évaluations de fonctions utilisées dans la recherche de 200 : `maxfeval = c(200)` (dans optim_args)

Réglages concernant les températures des modèles pour le premier ajustement (appelé firstrun), dans la librairie « tioptimTMB » (dans l'argument temper_args):

(i) nous avons choisi des températures comme étant l'exponentielle d'une valeur tirée au hasard entre 0 et log(4000) : `fT=expression(exp(runif(1,0,log(40000))))`
(ii) nombre minimum et maximum de températures pour lesquelles l'optimisation est effectuée respectivement de 30 et 40 : `Nmin=30, Nmax=40`
(iii) un nombre de tirages aléatoires initiaux de la température de 30 : `Ninit=30`.
(iv) une tolérance d'AIC pour comparaison avec la meilleure optimisation de 2.0 : AICtol=2.
(v) un paramètre d'échelle pour la phase finale de tempering de 0.1 : `sigma0=0.1`
(vi) un tirage des valeurs initiales dans les meilleurs valeurs initiales (dans une tolérance d'AIC de 2.0 prévu ci-dessus) : `bestIV=TRUE`.
(vii) le gradient de température est couvert par les tirages du nombre minimum de températures : `regularize=TRUE`.
(viii) Une probabilité de 100% de tirer la valeur de température passée parmi les meilleures (avec tolérance AICtol) et multiplié par un bruit lognormal de sd : `probs.sampling=c(0,1.00,0)`.
(ix) Une température minimale et maximale pour la régularisation initiale respectivement de 1.0 et 40000 : `Tmin=1, Tmax=40000`.
(x) Une utilisation de l'échelle logarithmique pour la régularisation initiale : `log=TRUE`.

Autres réglages des modèles pour le premier ajustement (firstrun), dans la librairie « tioptimTMB » (dans l'argument control):

(i) Un nombre maximum de phases d'optimisation dans la phase non tempérée (température == 1) de 1.0 : `nOptMax=1`.
(ii) Une différence minimale entre les valeurs de fonction TMB successives auxquelles l'itération est arrêtée de 0.01 : `epsilon=1e-2`.

(iii) les derniers meilleurs paramètres de la fonction TMB sont donnés pour une température>1 : `keepBestParTa1=TRUE`.

Réglages des modèles pour le second ajustement (secondrun), dans la librairie « tioptimTMB » :

(i) Pas d'utilisation de temperature : `temper_args=list(list(Textr=1,nTa1=1))`,
(ii) Utilisation des valeurs optimisées des paramètres du premier ajustement comme valeurs initiales : `init_pars=firstrun$parameters.TMBformat`.
(iii) Un nombre maximum de phases d'optimisation dans la phase non tempérée de 500 : `control=list(nOptMax=500`.
(iv) Une différence minimale entre les valeurs de fonction TMB successives auxquelles l'itération est arrêtée inchangée : `epsilon=1e-2`.
(v) les derniers meilleurs paramètres de la fonction TMB sont toujours donnés pour une température>1 : `keepBestParTa1=TRUE`.

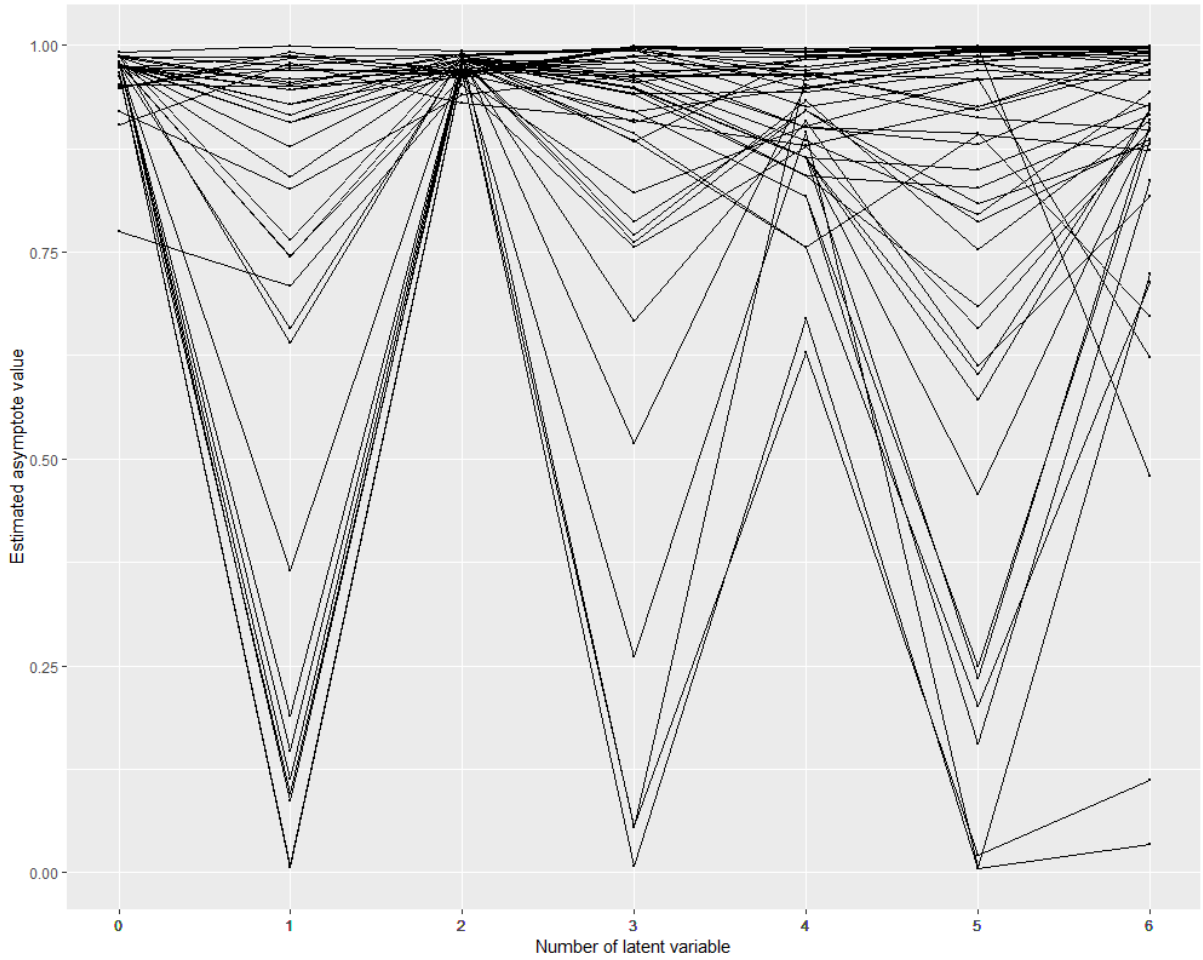
Annexe IV. Table S1.2: Valeurs des asymptotes hautes estimées par les modèles avec les traitements K et Kgroup, la modalité d'inclusion des effets aléatoires plot/season/trap, pot/season, plot/trap et plot, et de zéro à six variables latentes. Les modèles ont été optimisés à l'aide de la librairie tioptimTMB.

Traitement de l'asymptote haute	Groupe	Seed	nolv	lv1sp	lv2sp	lv3sp	lv4sp	lv5sp	lv6sp
K	-	s1	0.9996	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999
		s2	0.9996	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999
Kgroup	abat	s1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		s2	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	abov	s1	0.9996	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999
		s2	0.9996	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999
	abpa	s1	0.9998	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
		s2	0.9998	0.9999	0.9999	1.0000	1.0000	1.0000	1.0000
	ameq	s1	0.3759	0.3930	0.3374	0.4486	0.3512	0.3456	0.3487
		s2	0.3757	0.3648	0.4340	0.3559	0.3525	0.3499	0.3487
	amlu	s1	0.1913	0.1877	0.1394	0.1947	0.1432	0.1393	0.1447
		s2	0.1912	0.1713	0.2063	0.1562	0.1508	0.1466	0.1447

	amsi	s1	0.2740	0.2789	0.2231	0.3073	0.2312	0.2262	0.2313
		s2	0.2739	0.2562	0.3087	0.2423	0.2372	0.2332	0.2314
	babu	s1	0.0850	0.0762	0.0490	0.0670	0.0491	0.0472	0.0508
		s2	0.0850	0.0692	0.0810	0.0584	0.0547	0.0520	0.0508
	bela	s1	0.0352	0.0291	0.0162	0.0200	0.0157	0.0150	0.0166
		s2	0.0352	0.0261	0.0275	0.0203	0.0185	0.0172	0.0166
	caau	s1	0.9939	0.9966	0.9980	0.9989	0.9984	0.9984	0.9981
		s2	0.9939	0.9962	0.9977	0.9975	0.9978	0.9980	0.9981
	calu	s1	0.9994	0.9997	0.9999	0.9999	0.9999	0.9999	0.9999
		s2	0.9994	0.9997	0.9997	0.9998	0.9999	0.9999	0.9999
	camo	s1	0.9976	0.9988	0.9994	0.9997	0.9995	0.9995	0.9994
		s2	0.9976	0.9986	0.9992	0.9991	0.9993	0.9994	0.9994
	cane	s1	0.9847	0.9906	0.9938	0.9963	0.9948	0.9949	0.9941
		s2	0.9847	0.9896	0.9928	0.9924	0.9933	0.9939	0.9941
	cani	s1	0.9620	0.9742	0.9808	0.9876	0.9833	0.9836	0.9817
		s2	0.9620	0.9716	0.9790	0.9777	0.9796	0.9811	0.9817
	capr	s1	0.9939	0.9966	0.9980	0.9989	0.9984	0.9984	0.9981
		s2	0.9939	0.9962	0.9975	0.9975	0.9978	0.9980	0.9981
	caro	s1	0.3759	0.3967	0.3380	0.4375	0.3518	0.3457	0.3487
		s2	0.3757	0.3648	0.4205	0.3559	0.3525	0.3499	0.3487
	cavi	s1	0.9976	0.9988	0.9994	0.9997	0.9995	0.9995	0.9994
		s2	0.9976	0.9986	0.9992	0.9991	0.9993	0.9994	0.9994
	cicm	s1	0.6053	0.6445	0.6153	0.7325	0.6368	0.6328	0.6288
		s2	0.6051	0.6148	0.6934	0.6227	0.6254	0.6277	0.6287
	hala	s1	0.3759	0.3930	0.3374	0.4486	0.3512	0.3456	0.3487
		s2	0.3757	0.3648	0.4340	0.3559	0.3525	0.3499	0.3487
	halu	s1	0.1913	0.1892	0.1399	0.1884	0.1430	0.1393	0.1447
		s2	0.1912	0.1713	0.1978	0.1562	0.1508	0.1466	0.1447
	harf	s1	0.3759	0.3930	0.3374	0.4486	0.3512	0.3456	0.3487
		s2	0.3757	0.3648	0.4340	0.3559	0.3525	0.3499	0.3487
	haru	s1	0.3759	0.3930	0.3374	0.4486	0.3512	0.3456	0.3487
		s2	0.3757	0.3648	0.4340	0.3559	0.3525	0.3499	0.3487
	hata	s1	0.2740	0.2789	0.2231	0.3073	0.2312	0.2262	0.2313
		s2	0.2739	0.2562	0.3087	0.2423	0.2372	0.2332	0.2314
	lech	s1	0.0850	0.0762	0.0490	0.0670	0.0491	0.0472	0.0508
		s2	0.0850	0.0692	0.0810	0.0584	0.0547	0.0520	0.0508
	leru	s1	0.9974	0.9985	0.9993	0.9995	0.9994	0.9995	0.9993
		s2	0.9974	0.9985	0.9989	0.9991	0.9992	0.9993	0.9993
	lopi	s1	0.1913	0.1886	0.1398	0.1954	0.1429	0.1393	0.1447
		s2	0.1912	0.1713	0.2044	0.1562	0.1508	0.1466	0.1447
	mima	s1	0.0352	0.0288	0.0161	0.0208	0.0157	0.0150	0.0166
		s2	0.0352	0.0261	0.0289	0.0203	0.0185	0.0172	0.0166
	mopi	s1	0.9990	0.9994	0.9998	0.9998	0.9998	0.9998	0.9998
		s2	0.9990	0.9995	0.9996	0.9997	0.9997	0.9998	0.9998
	nebr	s1	0.9990	0.9995	0.9998	0.9998	0.9998	0.9998	0.9998
		s2	0.9990	0.9995	0.9996	0.9997	0.9997	0.9998	0.9998

	nobi	s1	0.0550	0.0477	0.0283	0.0361	0.0279	0.0267	0.0292
		s2	0.0550	0.0427	0.0463	0.0346	0.0320	0.0300	0.0292
	nopa	s1	0.0850	0.0773	0.0492	0.0643	0.0492	0.0473	0.0508
		s2	0.0850	0.0692	0.0770	0.0584	0.0547	0.0520	0.0508
	noru	s1	0.0850	0.0769	0.0492	0.0645	0.0490	0.0473	0.0508
		s2	0.0850	0.0692	0.0771	0.0584	0.0547	0.0520	0.0508
	pabi	s1	0.1913	0.1877	0.1394	0.1947	0.1432	0.1393	0.1447
		s2	0.1912	0.1713	0.2063	0.1562	0.1508	0.1466	0.1447
	plru	s1	0.9934	0.9955	0.9979	0.9977	0.9982	0.9983	0.9979
		s2	0.9934	0.9959	0.9962	0.9972	0.9976	0.9978	0.9979
	pocu	s1	0.4900	0.5200	0.4744	0.5988	0.4935	0.4882	0.4877
		s2	0.4899	0.4891	0.5684	0.4885	0.4881	0.4879	0.4878
	poku	s1	0.6053	0.6445	0.6153	0.7325	0.6368	0.6328	0.6288
		s2	0.6051	0.6148	0.6934	0.6227	0.6254	0.6277	0.6287
	pove	s1	0.3759	0.3930	0.3374	0.4486	0.3512	0.3456	0.3487
		s2	0.3757	0.3648	0.4340	0.3559	0.3525	0.3499	0.3487
	psru	s1	0.7099	0.7533	0.7391	0.8333	0.7601	0.7569	0.7508
		s2	0.7097	0.7268	0.7949	0.7404	0.7451	0.7490	0.7507
	ptcr	s1	0.9997	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
		s2	0.9997	0.9999	0.9999	0.9999	1.0000	1.0000	1.0000
	ptma	s1	0.9998	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
		s2	0.9998	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000
	ptme	s1	0.7961	0.8376	0.8344	0.8980	0.8506	0.8492	0.8427
		s2	0.7960	0.8160	0.8633	0.8313	0.8366	0.8408	0.8427
	ptni	s1	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
		s2	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	ptob	s1	0.9990	0.9995	0.9998	0.9998	0.9998	0.9998	0.9998
		s2	0.9990	0.9995	0.9996	0.9997	0.9997	0.9998	0.9998
	stpu	s1	0.1913	0.1888	0.1391	0.1891	0.1435	0.1393	0.1447
		s2	0.1912	0.1713	0.2058	0.1562	0.1508	0.1466	0.1447
	syob	s1	0.0352	0.0289	0.0162	0.0201	0.0157	0.0150	0.0166
		s2	0.0352	0.0261	0.0276	0.0203	0.0185	0.0172	0.0166
	syvi	s1	0.1291	0.1230	0.0840	0.1119	0.0851	0.0822	0.0869
		s2	0.1290	0.1103	0.1253	0.0967	0.0921	0.0885	0.0869
	trob	s1	0.0550	0.0473	0.0282	0.0375	0.0280	0.0267	0.0292
		s2	0.0550	0.0427	0.0487	0.0346	0.0320	0.0300	0.0292
	trqu	s1	0.0550	0.0477	0.0283	0.0361	0.0279	0.0267	0.0292
		s2	0.0550	0.0427	0.0463	0.0346	0.0320	0.0300	0.0292

Annexe IV. Figure S1.1: Asymptotes estimées des 47 espèces par les modèles avec la première graine, le traitement Kspspgroup et un nombre variable de variables latentes.



Annexe IV. Table S1.3: Valeurs des asymptotes hautes estimées par les modèles avec les traitements K et Kgroup, un nombre de variable latente fixé à 4, et différentes modalités d'inclusion des effets aléatoires. Les modèles ont été optimisés à l'aide de la librairie tioptimTMB.

Traitement de l'asymptote haute	Groupe	Seed	plot/season/trap pot/season plot/trap plot	plot/season/trap plot/season plot	plot/season/trap plot	plot/season/trap	plot/season plot	plot
K	-	s1	0.9999	0.9999	0.9999	0.6532	0.9998	0.9995
		s2	0.9999	0.9999	0.9992	0.9992	0.9998	0.9995
		s3	-	-	-	0.9985	-	-
Kgroup	abat	s1	1.0000	1.0000	1.0000	0.9999	1.0000	1.0000
		s2	1.0000	1.0000	1.0000	0.9999	1.0000	1.0000
	abov	s1	0.9999	0.9999	0.9999	0.9993	0.9998	0.9995
		s2	0.9999	0.9999	0.9999	0.9993	0.9998	0.9995
	abpa	s1	1.0000	1.0000	1.0000	0.9997	0.9999	0.9998
		s2	1.0000	1.0000	1.0000	0.9997	0.9999	0.9998
	ameq	s1	0.3512	0.3578	0.3875	0.4629	0.3721	0.3787
		s2	0.3525	0.3556	0.3400	0.6366	0.4024	0.3787
	amlu	s1	0.1432	0.1467	0.1639	0.2618	0.1759	0.1967
		s2	0.1508	0.1558	0.1397	0.4124	0.1929	0.1966
	amsi	s1	0.2312	0.2363	0.2604	0.3560	0.2624	0.2787
		s2	0.2372	0.2419	0.2244	0.5258	0.2864	0.2786
	babu	s1	0.0491	0.0504	0.0573	0.1274	0.0714	0.0895
		s2	0.0547	0.0581	0.0487	0.2194	0.0782	0.0895
	bela	s1	0.0157	0.0162	0.0184	0.0542	0.0271	0.0380
		s2	0.0185	0.0202	0.0157	0.1010	0.0294	0.0380
	caau	s1	0.9984	0.9984	0.9985	0.9931	0.9963	0.9932
		s2	0.9978	0.9975	0.9981	0.9963	0.9968	0.9932
	calu	s1	0.9999	0.9999	0.9999	0.9988	0.9997	0.9992

		s2	0.9999	0.9998	0.9999	0.9988	0.9997	0.9992
	camo	s1	0.9995	0.9995	0.9995	0.9972	0.9987	0.9972
		s2	0.9993	0.9992	0.9994	0.9984	0.9988	0.9972
	cane	s1	0.9948	0.9949	0.9951	0.9835	0.9897	0.9832
		s2	0.9933	0.9925	0.9941	0.9900	0.9909	0.9832
	cani	s1	0.9833	0.9837	0.9843	0.9609	0.9720	0.9591
		s2	0.9796	0.9779	0.9814	0.9755	0.9749	0.9591
	capr	s1	0.9984	0.9984	0.9985	0.9932	0.9963	0.9932
		s2	0.9978	0.9975	0.9981	0.9960	0.9968	0.9932
	caro	s1	0.3518	0.3595	0.3865	0.4515	0.3731	0.3787
		s2	0.3525	0.3556	0.3376	0.6361	0.4035	0.3787
	cavi	s1	0.9995	0.9995	0.9995	0.9972	0.9987	0.9972
		s2	0.9993	0.9992	0.9994	0.9984	0.9988	0.9972
	cicm	s1	0.6368	0.6436	0.6712	0.6769	0.6219	0.6028
		s2	0.6254	0.6227	0.6204	0.8139	0.6548	0.6029
	hala	s1	0.3512	0.3578	0.3875	0.4629	0.3721	0.3787
		s2	0.3525	0.3556	0.3400	0.6366	0.4024	0.3787
	halu	s1	0.1430	0.1476	0.1649	0.2626	0.1766	0.1967
		s2	0.1508	0.1558	0.1388	0.3892	0.1913	0.1966
	harf	s1	0.3512	0.3578	0.3875	0.4629	0.3721	0.3787
		s2	0.3525	0.3556	0.3400	0.6366	0.4024	0.3787
	haru	s1	0.3512	0.3578	0.3875	0.4629	0.3721	0.3787
		s2	0.3525	0.3556	0.3400	0.6366	0.4024	0.3787
	hata	s1	0.2312	0.2363	0.2604	0.3560	0.2624	0.2787
		s2	0.2372	0.2419	0.2244	0.5258	0.2864	0.2786
	lech	s1	0.0491	0.0504	0.0573	0.1274	0.0714	0.0895
		s2	0.0547	0.0581	0.0487	0.2194	0.0782	0.0895
	leru	s1	0.9994	0.9994	0.9994	0.9962	0.9985	0.9970
		s2	0.9992	0.9991	0.9994	0.9973	0.9986	0.9970
	lopi	s1	0.1429	0.1464	0.1607	0.2668	0.1757	0.1967

		s2	0.1508	0.1558	0.1390	0.3900	0.1924	0.1966
	mima	s1	0.0157	0.0161	0.0183	0.0540	0.0270	0.0380
		s2	0.0185	0.0202	0.0158	0.1101	0.0297	0.0380
	mopi	s1	0.9998	0.9998	0.9998	0.9982	0.9994	0.9988
		s2	0.9997	0.9997	0.9998	0.9981	0.9994	0.9988
	nebr	s1	0.9998	0.9998	0.9998	0.9985	0.9995	0.9988
		s2	0.9997	0.9997	0.9998	0.9988	0.9995	0.9988
	nobi	s1	0.0279	0.0288	0.0326	0.0821	0.0443	0.0586
		s2	0.0320	0.0344	0.0277	0.1508	0.0483	0.0586
	nopa	s1	0.0492	0.0507	0.0570	0.1223	0.0717	0.0895
		s2	0.0547	0.0581	0.0482	0.2191	0.0786	0.0895
	noru	s1	0.0490	0.0507	0.0577	0.1278	0.0717	0.0895
		s2	0.0547	0.0581	0.0483	0.2034	0.0775	0.0895
	pabi	s1	0.1432	0.1467	0.1639	0.2618	0.1759	0.1967
		s2	0.1508	0.1558	0.1397	0.4124	0.1929	0.1966
	plru	s1	0.9982	0.9981	0.9979	0.9894	0.9958	0.9926
		s2	0.9976	0.9972	0.9980	0.9886	0.9955	0.9926
	pocu	s1	0.4935	0.5008	0.5319	0.5733	0.4968	0.4903
		s2	0.4881	0.4883	0.4785	0.7346	0.5306	0.4903
	poku	s1	0.6368	0.6436	0.6712	0.6769	0.6219	0.6028
		s2	0.6254	0.6227	0.6204	0.8139	0.6548	0.6029
	pove	s1	0.3512	0.3578	0.3875	0.4629	0.3721	0.3787
		s2	0.3525	0.3556	0.3400	0.6366	0.4024	0.3787
	psru	s1	0.7601	0.7648	0.7837	0.7565	0.7327	0.7055
		s2	0.7451	0.7406	0.7438	0.8836	0.7636	0.7056
	ptcr	s1	1.0000	1.0000	1.0000	0.9995	0.9999	0.9997
		s2	1.0000	0.9999	1.0000	0.9995	0.9999	0.9997
	ptma	s1	1.0000	1.0000	1.0000	0.9997	0.9999	0.9998
		s2	1.0000	1.0000	1.0000	0.9998	0.9999	0.9998
	ptme	s1	0.8506	0.8550	0.8677	0.8294	0.8210	0.7908

		s2	0.8366	0.8316	0.8368	0.9159	0.8429	0.7909
	ptni	s1	1.0000	1.0000	1.0000	0.9999	1.0000	0.9999
		s2	1.0000	1.0000	1.0000	0.9999	1.0000	0.9999
	ptob	s1	0.9998	0.9998	0.9998	0.9985	0.9995	0.9988
		s2	0.9997	0.9997	0.9998	0.9989	0.9995	0.9988
	stpu	s1	0.1435	0.1466	0.1615	0.2536	0.1764	0.1967
		s2	0.1508	0.1558	0.1393	0.4035	0.1956	0.1966
	syob	s1	0.0157	0.0162	0.0186	0.0569	0.0271	0.0380
		s2	0.0185	0.0202	0.0158	0.0928	0.0290	0.0380
	syvi	s1	0.0851	0.0877	0.0980	0.1785	0.1141	0.1343
		s2	0.0921	0.0964	0.0827	0.3071	0.1252	0.1343
	trob	s1	0.0280	0.0286	0.0323	0.0818	0.0442	0.0586
		s2	0.0320	0.0344	0.0279	0.1635	0.0488	0.0586
	trqu	s1	0.0279	0.0288	0.0326	0.0821	0.0443	0.0586
		s2	0.0320	0.0344	0.0277	0.1508	0.0483	0.0586

Annexe IV. Table S1.4: Valeurs des asymptotes hautes estimées par les modèles avec les traitements K et Kgroup, un nombre de variable latente fixé à 4, la modalité d'inclusion des effets aléatoires plot/season/trap, pot/season, plot/trap et plot et intégrant ou excluant les espèces rares (moins de 10 observations). Les modèles ont été optimisés à l'aide de la librairie tioptimTMB.

Traitement de l'asymptote haute	Groupe	Seed	with rare species	without rare species
K	-	s1	0.9999	0.9999
		s2	0.9999	0.9999
Kgroup	abat	s1	1.0000	1.0000
		s2	1.0000	1.0000
	abov	s1	0.9999	0.9954
		s2	0.9999	0.9956
	abpa	s1	1.0000	0.9993
		s2	1.0000	0.9993
	amlu	s1	0.1467	0.0043
		s2	0.1558	0.0042
	babu	s1	0.0504	0.0007
		s2	0.0581	0.0007
	caau	s1	0.9984	0.9994
		s2	0.9975	0.9995
	calu	s1	0.9999	0.9885
		s2	0.9998	0.9888
	cane	s1	0.9949	0.9964
		s2	0.9925	0.9965
	cani	s1	0.9837	0.9776
		s2	0.9779	0.9782
	capr	s1	0.9984	0.9994
		s2	0.9975	0.9995
	caro	s1	0.3595	0.0266
		s2	0.3556	0.0261
	cavi	s1	0.9995	0.9999
		s2	0.9992	0.9999
	hala	s1	0.3578	0.0266
		s2	0.3556	0.0261
	leru	s1	0.9994	0.8441
		s2	0.9991	0.8457
	pocu	s1	0.5008	0.0643
		s2	0.4883	0.0634

	pove	s1	0.3578	0.0266
		s2	0.3556	0.0261
	ptcr	s1	1.0000	0.9982
		s2	0.9999	0.9982
	ptma	s1	1.0000	0.9993
		s2	1.0000	0.9993
	ptni	s1	1.0000	0.9999
		s2	1.0000	0.9999
	ptob	s1	0.9998	0.9716
		s2	0.9997	0.9723
	syvi	s1	0.0877	0.0017
		s2	0.0964	0.0016

Annexe IV. S3 : Analyse complémentaire sur l'instabilité des paramètres

Annexe IV. Section S3.1 : Méthodes et résultats

Les asymptotes estimées avec les modèles Kgroup étant relativement stables entre deux graines, nous avons voulu regarder s'il en était de même pour deux autres paramètres d'intérêt :

(i) les matrices $\overline{\overline{\beta_1}}$ décrivant les effets des prédicteurs environnementaux sur les espèces (de dimension : nombre d'espèces x nombre de conditions environnementales = 47 x 7).

(ii) et les $\bar{\lambda}$ décrivant les effets des variables latentes sur les espèces (dans le code, sous la forme d'un vecteur de dimension :

$$\text{nombre de variables latentes} * \left(\text{nombre d'espèces} - \frac{\text{nombre de variables latentes} - 1}{2}\right)).$$

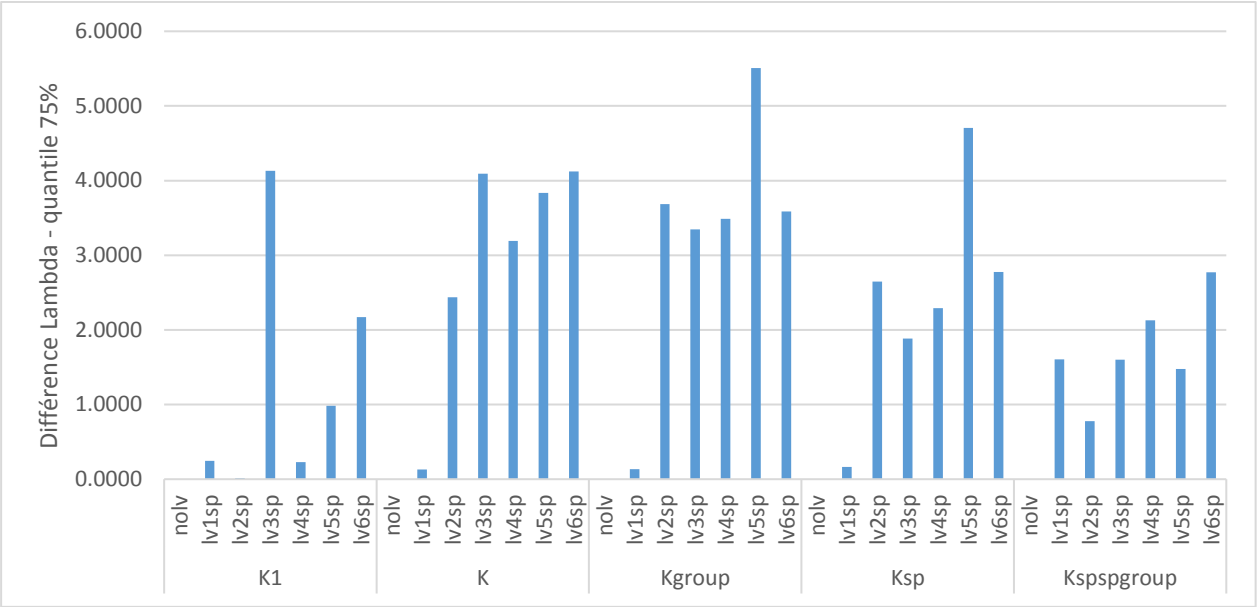
Pour cela, pour chacun des modèles des étapes 2 et 3 avec le traitement Kgroup, nous avons calculé la valeur absolue de la différence entre la première graine et la seconde, pour chaque

valeur de la matrice $\overline{\overline{\beta_1}}$, soit 47*7=329 valeurs. Nous avons fait de même pour le vecteur $\bar{\lambda}$ (dont le nombre de valeurs est variable et dépend du nombre de variables latentes incluses dans le modèle). Puis, nous avons observé, sur ces valeurs absolues des différences entre les deux graines, les quantiles et plus précisément le quantile 75%.

Nous avons pu observer que, peu importe le modèle, les paramètres $\bar{\lambda}$ étaient très instables (le quantile 75% allant jusqu'à une différence de plus de 5.0, pour une distribution a priori normale ou pour l'exponentielle d'une distribution a priori normale dont l'écart type est de 5.0 dans les deux cas – cf. Annexe IV. Figure S3.1 et Annexe IV. Figure S3.2). En revanche, l'instabilité de l'AICc observée ne semblait pas liée à l'instabilité de ce paramètre, pour les modèles de la seconde étape (étape de comparaison des modèles avec un niveau d'inclusion des variables latentes fixé, et un nombre variable de variables latentes, située dans le Chapitre 4 - II.3. - cf. Annexe IV. Figure S3.3). Le lien entre l'instabilité de l'AICc et du paramètre semblait plus probable (bien qu'incertain) pour les modèles de la troisième étape (étape de comparaison des modèles avec le nombre de variables latentes fixés, et des niveaux variables d'inclusions des variables latentes, située dans le Chapitre 4 - II.4.), puisque la relation entre ces deux instabilités n'apparaissait qu'à condition de s'affranchir de la valeur extrême de différence d'AICc (cf. Annexe IV. Figure S3.4).

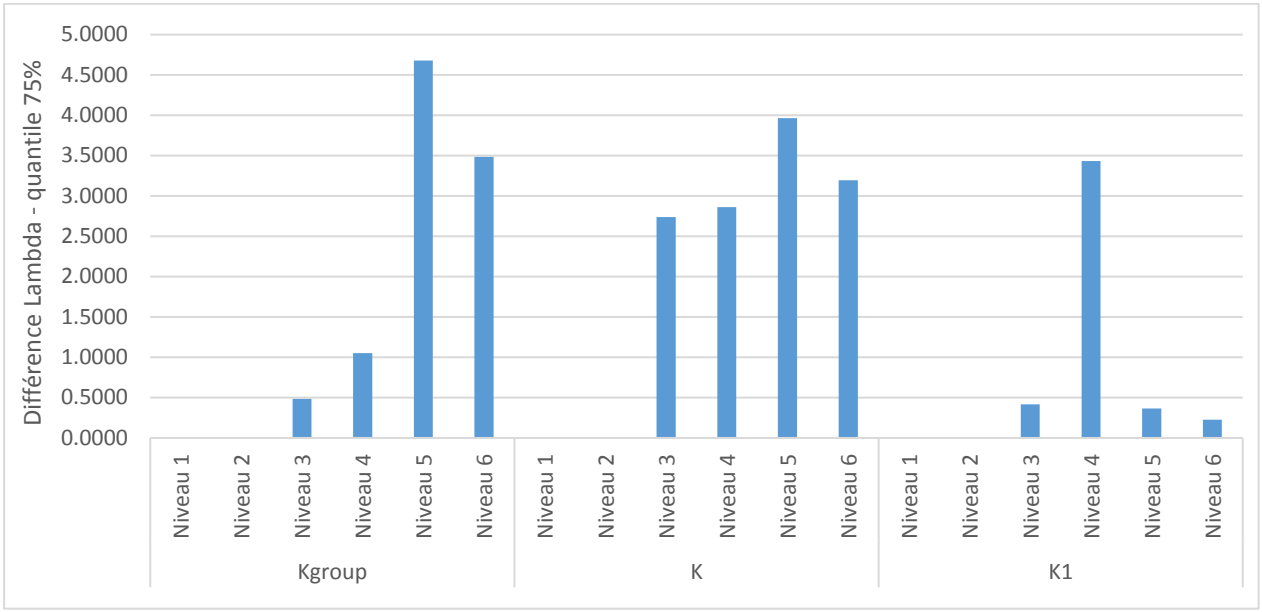
Par ailleurs, le paramètre $\overline{\overline{\beta_1}}$ semblait moins variable (le quantile 75% allant jusqu'à une différence de plus de 1.0 , pour une distribution a priori normale dont l'écart type est tiré d'une distribution $N(0.0,5.0))$ - Annexe IV. Figure S3.5 et Annexe IV. Figure S3.6), mais l'instabilité de l'AICc observée au cours de nos tentatives de modélisation semblait au moins pour parti liée à l'instabilité de ce paramètre (Annexe IV. Figure S3.7 et Annexe IV. Figure S3.8).

Annexe IV. Figure S3.1 : Quantile 75% des différentes valeurs absolues (nombre variable) de la différence entre la valeur estimée avec la première graine et la seconde, pour le paramètre $\bar{\lambda}$. Les modèles étudiés sont les modèles Kgroup de la seconde étape de modélisation (avec des niveaux d'inclusion des variables latentes fixés à plot/season/trap, plot/season, plot/trap et plot ; et un nombre variable de variables latentes).



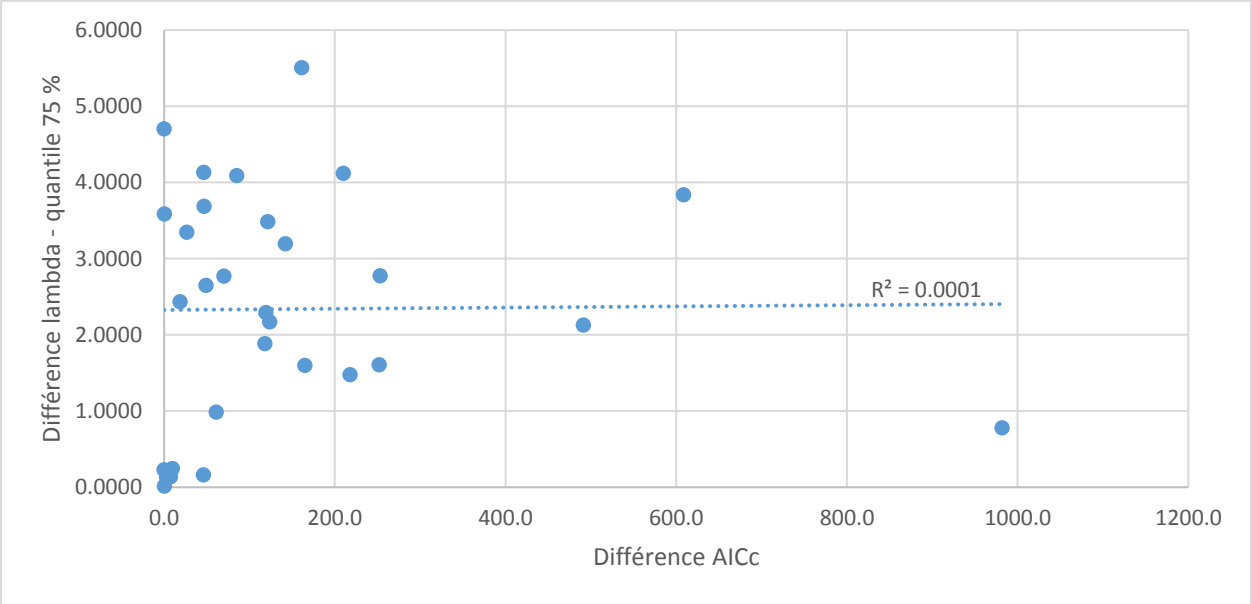
Annexe IV. Figure S3.2 : Quantile 75% des différentes valeurs absolues (nombre variable) de la différence entre la valeur estimée avec la première graine et la seconde, pour le paramètre $\bar{\lambda}$. Les modèles étudiés sont les modèles Kgroup de la troisième étape de modélisation (avec des niveaux variables d'inclusion des variables latentes, et un nombre de variables latentes fixé à 4). Pour faciliter la lecture du graphique nous avons abrégé les niveaux comme suit : Niveau 1= plot ; Niveau 2 = plot/season et plot ; Niveau 3 = plot/season/trap ; Niveau 4 =

plot/season/trap et plot ; Niveau 5 = plot/season/trap ; plot/season et plot ; Niveau 6 = plot/season/trap, plot/season, plot/trap et plot.



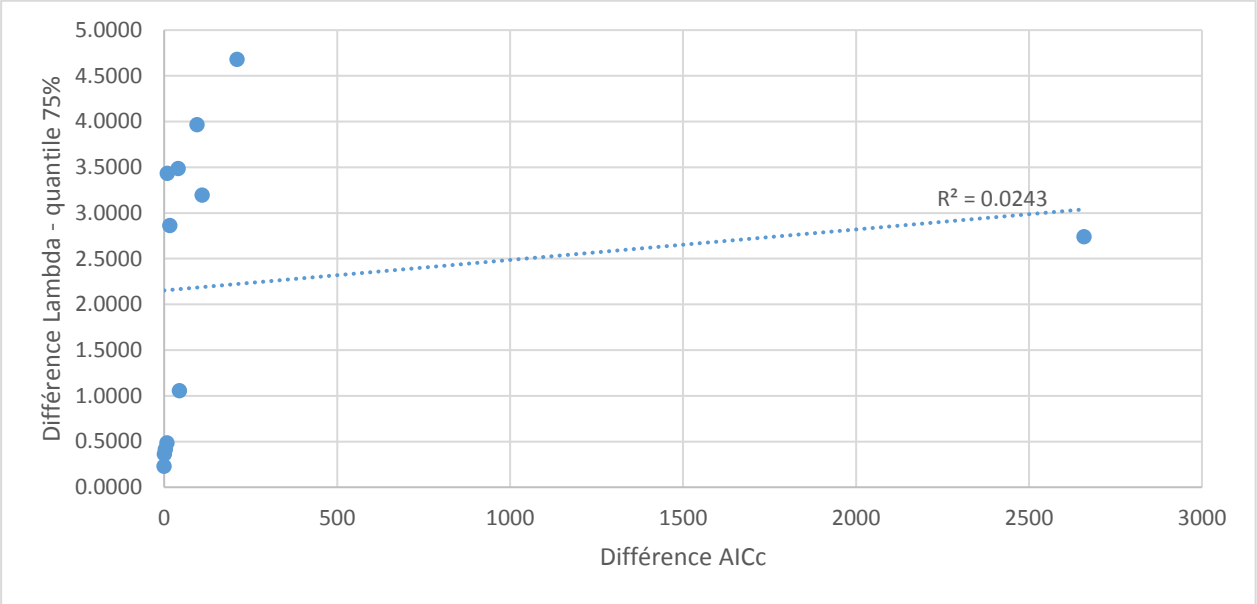
Annexe IV. Figure S3.3 : Quantile 75% des différentes valeurs absolues (nombre variable) de la différence entre la valeur estimée avec la première graine et la seconde, pour le paramètre $\bar{\lambda}$, en fonction de la valeur absolue de la différence d'AICc obtenue entre les deux graines. Les modèles étudiés sont les modèles Kgroup de la seconde étape de modélisation (avec des

niveaux d'inclusion des variables latentes fixés à plot/season/trap, plot/season, plot/trap, et plot et un nombre variable de variables latentes).



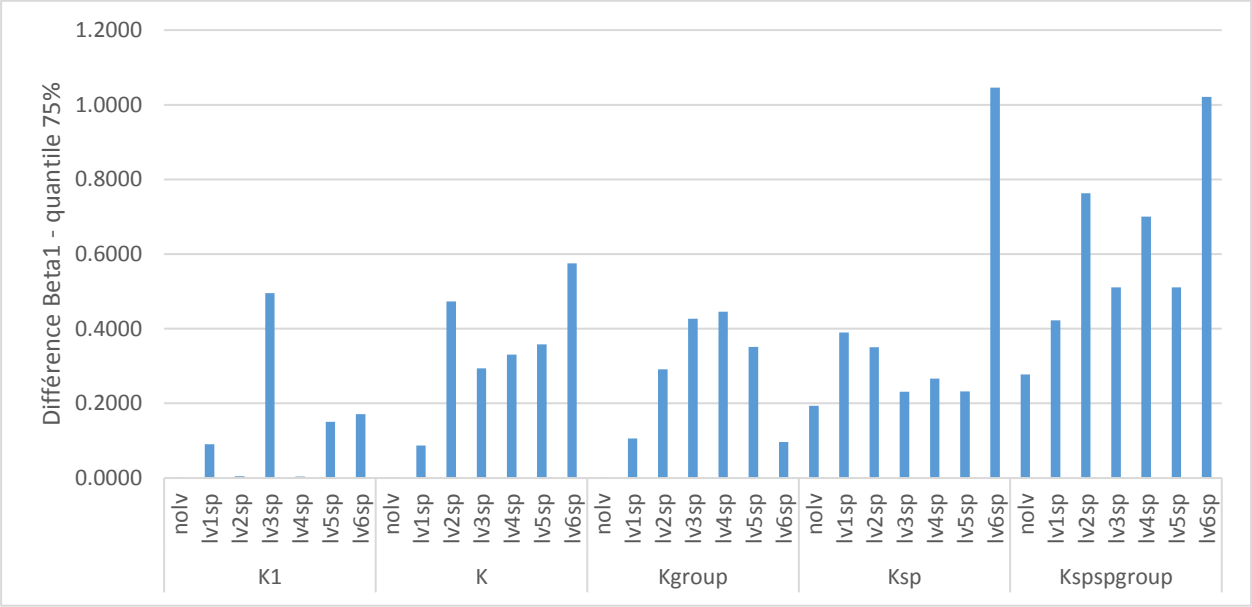
Annexe IV. Figure S3.4 : Quantile 75% des différentes valeurs absolues (nombre variable) de la différence entre la valeur estimée avec la première graine et la seconde, pour le paramètre $\bar{\lambda}$, en fonction de la valeur absolue de la différence d'AICc obtenue entre les deux graines. Les modèles étudiés sont les modèles Kgroup de la troisième étape de modélisation (avec des

niveaux variables d'inclusion des variables latentes, et un nombre de variables latentes fixé à 4). Sans le point extrême : $R^2 = 0.4459$.



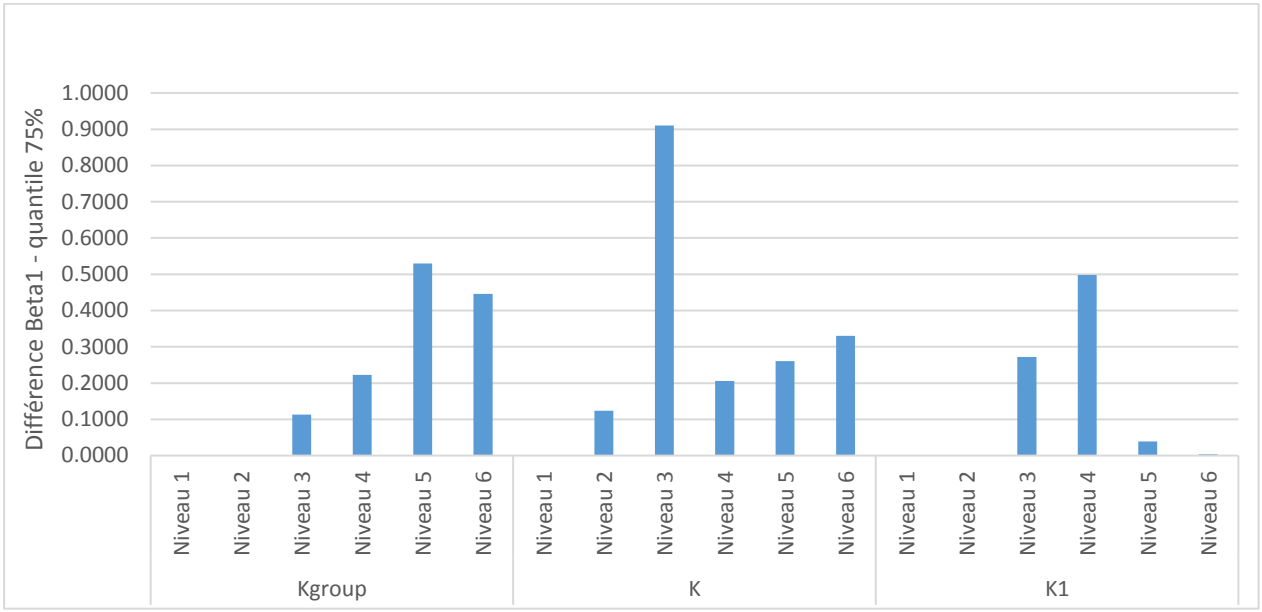
Annexe IV. Figure S3.5 : Quantile 75% des différentes valeurs absolues (329 valeurs) de la différence entre la valeur estimée avec la première graine et la seconde, pour le paramètre $\overline{\overline{\beta_1}}$. Les modèles étudiés sont les modèles Kgroup de la seconde étape de modélisation (avec

des niveaux d'inclusion des variables latentes fixé à plot/season/trap, plot/season, plot/trap et plot et un nombre variable de variables latentes).



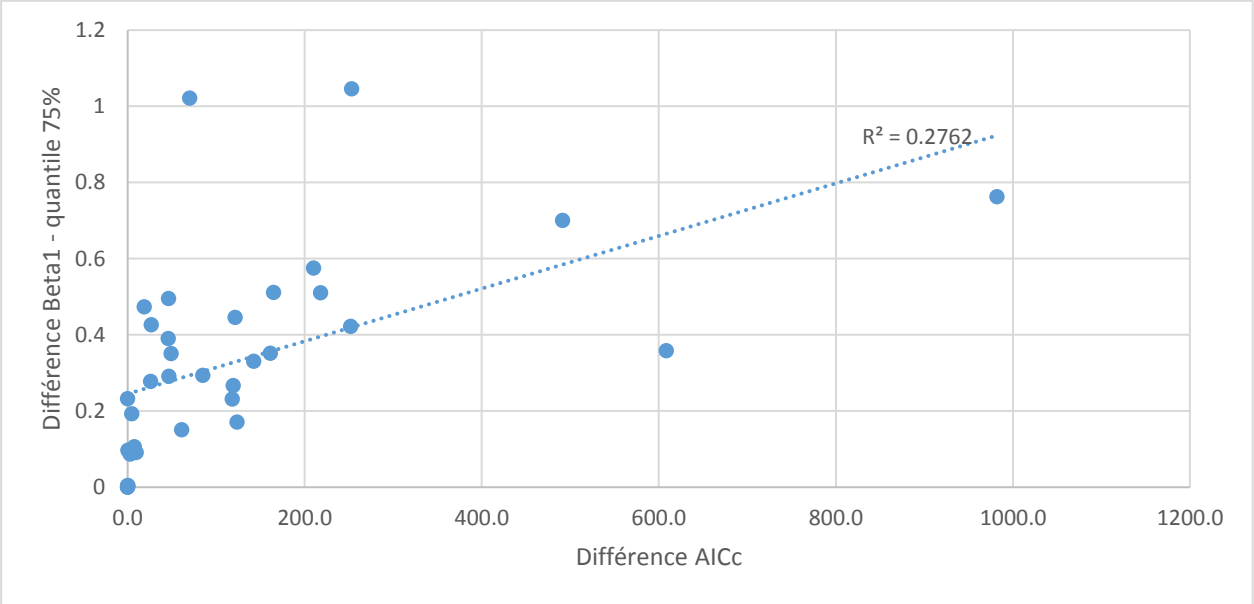
Annexe IV. Figure S3.6 : Quantile 75% des différentes valeurs absolues (329 valeurs) de la différence entre la valeur estimée avec la première graine et la seconde, pour le paramètre $\overline{\overline{\beta_1}}$. Les modèles étudiés sont les modèles Kgroup de la troisième étape de modélisation (avec des niveaux variables d'inclusion des variables latentes, et un nombre de variables latentes fixé à 4). Pour faciliter la lecture du graphique nous avons abrégé les niveaux comme suit : Niveau 1= plot ; Niveau 2 = plot/season et plot ; Niveau 3 = plot/season/trap ; Niveau 4 =

plot/season/trap et plot ; Niveau 5 = plot/season/trap ; plot/season et plot ; Niveau 6 = plot/season/trap, plot/season, plot/trap et plot.



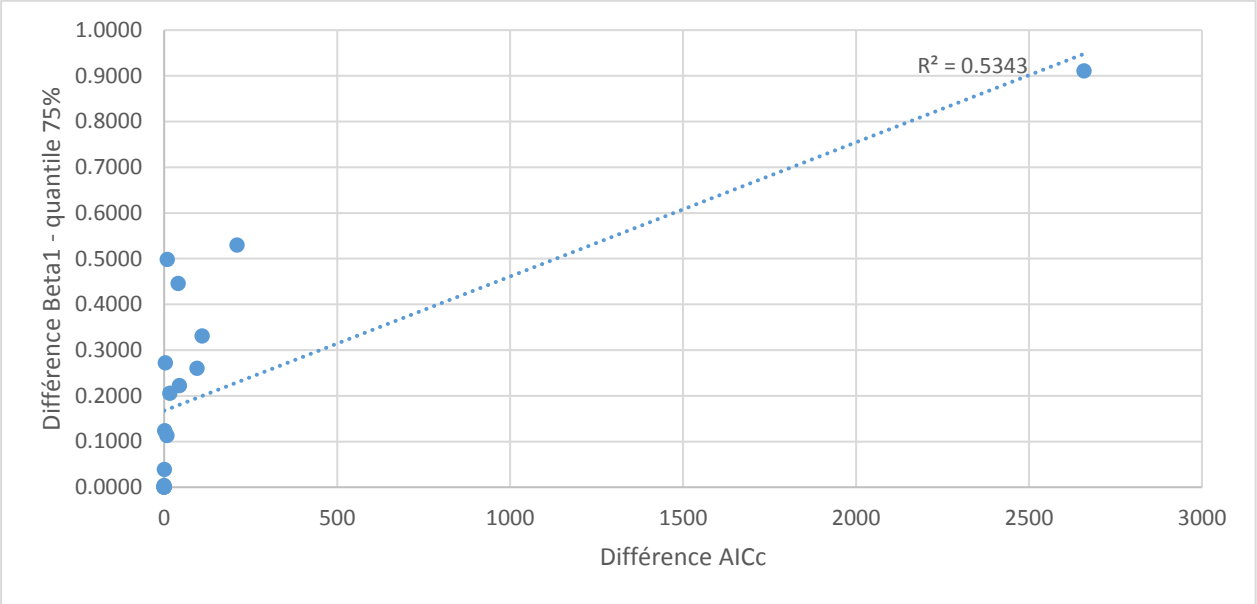
Annexe IV. Figure S3.7 : Quantile 75% des différentes valeurs absolues (329 valeurs) de la différence entre la valeur estimée avec la première graine et la seconde, pour le paramètre $\overline{\overline{\beta_1}}$, en fonction de la valeur absolue de la différence d'AICc obtenue entre les deux graines. Les modèles étudiés sont les modèles Kgroup de la seconde étape de modélisation (avec des

niveaux d'inclusion des variables latentes fixé à plot/season/trap, plot/season, plot/trap et plot et un nombre variable de variables latentes).



Annexe IV. Figure S3.8 : Quantile 75% des différentes valeurs absolues (329 valeurs) de la différence entre la valeur estimée avec la première graine et la seconde, pour le paramètre $\overline{\overline{\beta_1}}$, en fonction de la valeur absolue de la différence d'AICc obtenue entre les deux graines. Les modèles étudiés sont les modèles Kgroup de la troisième étape de modélisation (avec des

niveaux variables d'inclusion des variables latentes, et un nombre de variables latentes fixé à 4).



V. METADATA « *SUPPLEMENTARY MATERIALS* »

METADATA II. Manuscrit 2

Liste des « supplementary materials » associés au deuxième manuscrit : **The importance of being random! Taking full account of random effects in nonlinear sigmoid hierarchical Bayesian models reveals the relationship between deadwood and the species richness of saproxylic beetles.**

Data II.S1 : GOF p-values results (2 fichiers)

- `GOF-pvalues_results_01_pvalue.xls` – Complete results for GOF p-values. Yellow indicates problematic values (under 0.01) and red very problematic values (under 0.001).
- `GOF-pvalues_results_02_005.xls` – Complete results for GOF p-value percentage of significant probabilities at a significance level of 0.005. Yellow indicates problematic values (under 0.01) and red very problematic values (under 0.001).

Data II.S2 : Magnitude results (3 fichiers)

- `Magnitude_results_01_summary.csv` - Summary of the multiplicative effect of adding deltax=10m3/ha (mean and standard error in parentheses), shown with the significance of the relationship between deadwood and species richness, and the magnitude of the relationship for the global metrics. The codes for magnitude analyses corresponded to the following intervals for the logarithm of the multiplicative effect: [-0.05;0.05] (denoted as 000 and qualified as a strongly negligible relationship), [-0.1;0.1] (denoted as 00 and qualified as a moderately negligible relationship), [-0.2;0.2] (denoted as 0 and qualified as a weakly negligible relationship) and [0.05, Inf] (denoted as + and qualified as a weakly positive relationship). We ranked levels of significance as follows: *** corresponding to $p \leq 0.001$; ** to $0.001 \leq p \leq 0.01$; and * to $0.01 < p \leq 0.05$. If there is no symbol then the estimator is not

significantly different from 0.0 at the 5% level 0.05 < p. For random effect abbreviations, see Methods.

- `Magnitude_results_02_inits3.csv` - Summary of the multiplicative effect of adding deltax=10m3/ha (mean and standard error in parentheses) to Xinit=10m3/ha, shown with the significance of the relationship between deadwood and species richness, and the magnitude of the relationship for each forest and for the global metrics. The codes for magnitude analyses corresponded to the following intervals for the logarithm of the multiplicative effect: [-0.05;0.05] (denoted as 000 and qualified as a strongly negligible relationship), [-0.1;0.1] (denoted as 00 and qualified as a moderately negligible relationship), [-0.2;0.2] (denoted as 0 and qualified as a weakly negligible relationship) and [0.05, Inf] (denoted as + and qualified as a weakly positive relationship). We ranked levels of significance as follows: *** corresponding to $p \leq 0.001$; ** to $0.001 \leq p \leq 0.01$; and * to $0.01 < p \leq 0.05$. If there is no symbol then the estimator is not significantly different from 0.0 at the 5% level 0.05 < p. For random effect abbreviations, see Methods.
- `Magnitude_results_03_Xobs.csv` - Summary of the multiplicative effect of adding deltax=10m3/ha (mean and standard error in parentheses) to observed initial X (Xobs), shown with the significance of the relationship between deadwood and species richness, and the magnitude of the relationship for each forest and for the global metrics. The codes for magnitude analyses corresponded to the following intervals for the logarithm of the multiplicative effect: [-0.05;0.05] (denoted as 000 and qualified as a strongly negligible relationship), [-0.1;0.1] (denoted as 00 and qualified as a moderately

negligible relationship), [-0.2;0.2] (denoted as 0 and qualified as a weakly negligible relationship) and [0.05, Inf] (denoted as + and qualified as a weakly positive relationship). We ranked levels of significance as follows: *** corresponding to $p \leq 0.001$; ** to $0.001 \leq p \leq 0.01$; and * to $0.01 < p \leq 0.05$. If there is no symbol then the estimator is not significantly different from 0.0 at the 5% level $0.05 < p$. For random effect abbreviations, see Methods.

Data II.S3 : R codes (3 fichiers)

- `R_code_01_model_logis5p_fullre.R` - Code for model logis5p full.re. Other models can be obtained by simplifying this model.
- `R_code_02_ICs_calculations.R` - Code for ICs calculations based on the outputs of the models.
- `R_code_03_GOFs_calculations.R` - Code for GOF-pvalues calculations based on the outputs of the models.

Data II.S4 : Dataset. (1 fichier)

- `Dataset.csv` – Data used for analyses.

METADATA III. Manuscrit 3

Liste des « supplementary materials » associés au troisième manuscrit : **Generalized Linear Misleading: the need for a logistic function with estimated asymptotes to complement the classical logistic function for binomial GLMs.**

Data III.S1 : R codes (3 fichiers)

- `Model_univar_GLM_GAM.R` - Code for GLM and GAM on univariate simulated datasets.
- `Model_univar_TMBall.R` - Code for TMB models (L0Kest, LestK1 and LestKest) on univariate simulated datasets.
- `Model_multivar_GAM_L0K1_LestKest.R` - Code for GAM, L0K1 and LestKest models on multivariate simulated datasets.

METADATA IV. Chapitre 4

Liste des « supplementary materials » associés au quatrième chapitre de thèse.

Data IV.S1 : Introduction du chapitre 4 (1 fichier)

- `CARAB_traits.xlsx` – Tableau reprenant tous les traits associés au carabes recueillis dans la littérature.

Ugoline GODEAU

Améliorer la pertinence et l'efficacité des modèles statistiques en écologie : extension des fonctions sigmoïdes dans le cadre de l'étude de la distribution de la biodiversité

Résumé : La modélisation est un outil majeur en écologie pour décrire et comprendre les écosystèmes ou prédire leur réponse. Nous nous sommes intéressés aux modèles non-linéaires de forme sigmoïdale en macro-écologie avec pour objectif de mieux les définir, d'en comprendre les limites et de proposer des améliorations. Nous les avons d'abord étudiés dans des modèles de biodiversité hiérarchiques Bayésiens. Nous avons démontré que la prise en compte de variations aléatoires de différents paramètres de fonctions sigmoïdales avait un impact sur l'estimation des effets. Nous nous sommes ensuite intéressés aux modèles linéaires généralisés binomiaux binaires pour lesquels nous avons comparé la fonction classique logistique à d'autres fonctions sigmoïdales dont les asymptotes étaient estimées. Cela a permis de mettre en évidence les erreurs d'estimation induites par l'utilisation de la fonction logistique classique si les données ne sont pas cohérentes avec ce modèle. Enfin, nous avons appliqué ces fonctions logistiques avec asymptotes estimées dans le cadre de modèles d'occurrence hiérarchiques multi-espèces, grâce auxquels nous avons pu établir un intérêt probable de l'estimation des asymptotes. Les résultats instables ne nous ont pas permis de développer des conclusions écologiques. Lors de ces différents travaux, nous avons utilisé différents outils d'évaluation et interprétation des modèles, et prôné leur utilisation conjointe. En conclusion, nous avons développé de nouveaux modèles statistiques non-linéaires sigmoïdaux, qui sont de nouveaux outils pour l'écologue permettant d'enrichir sa palette pour mieux estimer les relations entre des variables et des données de biodiversité.

Mots clés : statistiques, modèles hiérarchiques, sigmoïde, logistique, modèles linéaires généralisés binomiaux, biodiversité

Improving relevance and efficiency of statistical models in ecology: extension of sigmoid functions in the context of the study of the distribution of biodiversity

Summary: Modeling is a major tool in ecology to describe and understand ecosystems or predict their response. We here focused our attention on non-linear sigmoidal models in macroecology, in order to better define them, understand their limitations and suggest improvements. We first studied them in hierarchical Bayesian biodiversity models. We found that taking into account random variations of different parameters of sigmoidal functions has an impact on the estimation of the effects. We then turned our attention to binary binomial generalized linear models for which we compared the classical logistic function to other sigmoidal functions whose asymptotes were estimated. We found strong estimation errors induced by the use of the classical logistic function if the data are not consistent with this model. Finally, we applied these logistic functions with estimated asymptotes in the context of hierarchical joint species occurrence models, thanks to which we were able to demonstrate the usefulness of considering the estimation of asymptotes. However, the unstable results did not allow us to develop ecological conclusions. Throughout, we have used various tools to better apprehend model evaluation and proposed that they should be used jointly. In conclusion, we have developed new forms of non-linear sigmoidal statistical models, which are new tools for the ecologist allowing to enrich his/her toolbox to better estimate the relationships between ecological variables and biodiversity data.

Keywords: statistics, hierarchical models, sigmoid, logistic, binomial generalized linear models, biodiversity

INRAE
Domaine des Barres
45290 Nogent-sur-Vernisson