

**Rapport de stage**  
Master II Mathématiques et applications  
Parcours : Statistique

---

## Titre : Développement d'indicateurs de tendances temporelles de la flore forestière

---

**ETABLISSEMENT :**  
INRAE, UR EFNO, Nogent-sur-Vernisson (45)

21 février 2022-19 août 2022

Soutenu par Ndongo Seye

le 01 septembre 2022

**Tuteurs :**

**M. Frédéric Gosselin<sup>†</sup>**, Ingénieur (ICPEF, HDR) en écologie forestière et biométrie au sein de l'UR EFNO

**M. Jean-Luc Dupouey**, Directeur de recherche à INRAE UMR Silva, Nancy

**M. Frédéric Archaux**, Directeur de l'unité ECODIV, ingénieur/chercheur en écologie des communautés et écologie du paysage

**Jury :**

**M. Laurent Gardes**, Professeur de Mathématiques appliquées à l'Université de Strasbourg, membre de l'équipe statistique UMR 7501, IRMA

**M. Etienne Birmele**, Professeur de Mathématiques appliquées à l'Université de Strasbourg

**M. Nicolas Poulin**, Professeur à l'Université de Strasbourg, ingénieur statisticien à l'IRMA

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	INRAE . . . . .	5
1.2	PASSIFOR2 . . . . .	6
<b>2</b>	<b>Objectifs et Problématique</b>	<b>6</b>
<b>3</b>	<b>Matériels et méthodes</b>	<b>9</b>
3.1	Le modèle hiérarchique sur grille latente . . . . .	9
3.2	Modèle de Gompertz . . . . .	11
3.2.1	Modèle de Rhodes et Jonzén . . . . .	11
3.2.2	Le modèle de Thorson . . . . .	12
3.3	Modèle d'observation . . . . .	17
3.4	Inférence statistique . . . . .	18
3.5	Présentation des données de l'inventaire forestier . . . . .	19
3.6	Phénologie . . . . .	21
3.6.1	Fonction sigmoïde à double asymétrie . . . . .	21
3.6.2	Fonction à double logistique . . . . .	21
<b>4</b>	<b>Simulation et résultats</b>	<b>23</b>
<b>5</b>	<b>Ajustement sur les données IFN</b>	<b>24</b>
<b>6</b>	<b>Discussion</b>	<b>25</b>
<b>7</b>	<b>Appendices</b>	<b>27</b>
7.1	Rhodes et Jonsén . . . . .	27
7.2	Thorson modifié . . . . .	27

## Liste des tableaux

1	Extrait des 3 premiers lignes du jeu de données IFN . . . . .	20
2	Tableau des vraies valeurs, des valeurs estimées, leurs erreurs-types (sd) et leurs intervalles de confiance à 95% (Conf.Int(95%)) . . . . .	24

3	Tableau des estimations pour les 3 espèces. . . . .	25
4	Intervalle de confiance des paramètres estimés pour les 3 espèces. . . . .	25

## Remerciements

Mes remerciements vont tout d'abord à mes tuteurs Frédéric Gosselin, Frédéric Archaux et Jean-Luc Dupouey qui m'ont guidé et soutenu tout au long de ce stage en expliquant, à moi qui débutait ses premiers pas dans l'écologie, son contexte environnemental et écologique, m'apportant articles et en me poussant à poser les bonnes questions pour ma compréhension. Je remercie particulièrement Frédéric Gosselin, qui a toujours été disponible durant ce stage même certains week-ends, de ses enseignements statistiques, code et debugging.

Je remercie tous les membres du personnel de l'INRAE à Nogent, des échanges que nous avons eu à faire pendant les pauses-café et qui ont été très instructifs et divertissants à la fois, sans oublier le personnel de l'Université de Strasbourg et en particulier tous les professeurs que j'y ai rencontrés.

Je remercie El hadj Cissé Faye, étudiant à l'université d'Orléans, qui m'a aidé dans ce stage sur la partie simulation.

Je ne pourrais terminer sans parler de ma famille, mon père, ma mère et mon grand frère Mohamed, que je n'ai pas vu depuis que je suis arrivé étudier en France, merci pour votre amour aussi tendre que doux, à mon frère Cheikh qui a été un guide depuis Marseille et qui m'a soutenu dans les moments les plus durs.

Je remercie Djiby Touré et Allé Kassé de m'avoir accueilli en France et tous mes amis d'ici et d'ailleurs.

Je ne pourrai certainement vous remercier assez dans ces quelques lignes. Alors merci.

**Mots clés :** *modélisation hiérarchique, modèle state-space, auto-corrélation spatiale, auto-corrélation temporelle, carrying capacity, modèle auto-régressif.*

# 1 Introduction

## 1.1 INRAE

J'ai réalisé mon stage de fin d'études de Master à l'INRAE du 21 février au 19 août 2022 sous la direction de Frédéric Gosselin, ingénieur-chercheur en écologie forestière et biométrie.

L'institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement a été créé en 2020, résultant de la fusion entre Inra et Irstea. INRAE est l'un des premiers organismes de recherche mondial spécialisé en agriculture, alimentation, en sciences du végétal et de l'animal et en écologie-environnement. L'institut a pour but de trouver des solutions face au changement climatique, à la raréfaction des ressources et à la dégradation de la biodiversité. Il constitue une communauté de 12000 personnes et 273 unités de recherches implantées dans 18 centres en France dont celle où j'ai effectué mon stage, l'unité EFNO, dans l'équipe "Biodiversité". Elle se situe au domaine des Barres à Nogent-sur-Vernisson et compte une quarantaine de chercheurs, ingénieurs et techniciens. L'unité comporte 4 équipes : Biodiversité, Forhet, Fona et l'équipe Geedaf.

- L'équipe Biodiversité vise à anticiper les conséquences sur la biodiversité de l'évolution de la gestion forestière, à comprendre l'influence du paysage actuel et passé sur la répartition et la dynamique de la biodiversité forestière, à développer des méthodes d'inventaires standardisées et des méthodes d'analyses adaptées aux jeux de données de biodiversité et à synthétiser et diffuser des connaissances scientifiques sur les pratiques de gestion favorables à la biodiversité.
- L'équipe Forhet a pour principal objectif l'étude des différentes interactions dans les peuplements mélangés (interactions entre les individus d'une même espèce ou d'espèces différentes, dans la strate arborée, la végétation du sous-bois, ou entre les différentes strates de l'écosystème) via deux aspects : la modélisation de la croissance et de la dynamique des arbres d'une part et d'autre part le partage des ressources entre les individus (lumière, eau, nutriments).
- L'équipe Fona étudie les effets directs et indirects des ongulés sauvages et des rongeurs sur les écosystèmes forestiers et périforestiers sous 4 thèmes :
  - renouvellent forestier : adapter le renouvellement de la forêt en présence d'ongulés sauvages en prenant en compte le changement climatique ;
  - biodiversité : étudier les effets des ongulés sauvages sur le fonctionnement et la biodiversité des écosystèmes ;
  - rôle fonctionnel : dégager les caractéristiques des ongulés sauvages qui modifient la dynamique des écosystèmes ;
  - gestion intégrée : proposer des outils qui facilitent la gestion des populations d'ongulés sauvages et qui permettent de modéliser leurs effets sur leur milieu.
- L'équipe Geedaf, Groupe d'étude et d'expertise sur la Diversité Adaptative des Arbres Forestiers dont la mission principale est l'appui au Ministère en charge de la forêt (MAA) pour la conservation, l'évaluation et la valorisation des ressources génétiques forestières (RGF) sur le territoire métropolitain, en tenant compte du changement climatique.

## 1.2 PASSIFOR2

Le Sommet Planète Terre du 3 au 14 juin 1992 à Rio de Janeiro qui a réuni des dirigeants politiques, diplomates et scientifiques ainsi que des ONG, a visé à réconcilier l’impact des activités socio-économiques humaines et l’environnement. Cette rencontre a mis en exergue l’interdépendance des facteurs économiques, sociaux et environnementaux en soulignant les changements qui devront s’opérer dans les différents secteurs et en définissant les grands principes du développement durable (y compris la biodiversité). Cette conférence a abouti au lancement du processus paneuropéen de gestion forestière durable d’Helsinki (1993) et a mené :

- à une forte coordination inter-gouvernementale européenne (“Forest Europe”), favorisant des exercices nationaux, européens et mondiaux.
- à des recherches sur le lien entre gestion forestière et biodiversité entre 1996 et 2018 dans le cadre du programme Biodiversité et gestion forestière (BGF) soutenu par le ministère de la transition écologique, notamment sur les indicateurs de biodiversité forestière au sein du GIP Ecofor [Nivet et al., 2012].
- à la reconnaissance de la biodiversité comme un des 6 critères de gestion forestière durable.

En conséquence, le projet PASSIFOR2, financé par le ministère de l’écologie, a pour objet de proposer un programme d’expertise et de recherche pour un système de suivi continu de la biodiversité en forêt à l’échelle métropolitaine.

## 2 Objectifs et Problématique

La biodiversité se définit comme étant l’ensemble des êtres vivants ainsi que les écosystèmes dans lesquels ils vivent. Cela inclut les interactions des espèces entre elles et leurs milieux de vie. Elle est l’observation de la variabilité de l’aspect et des comportements dynamiques des organismes sur terre [Chave, J., 2000]. L’étude de leurs populations a été longtemps axée sur un modèle de distribution des espèces qui ne tient pas compte de la dynamique spatiale des populations, ce qui induit une estimation biaisée des paramètres car ce modèle suppose que toutes les espèces soient en équilibre malgré les variations environnementales, ce qui n’est pas toujours le cas [Pagel and Schurr, 2012].

Lorsqu’on considère certains modèles de régression usuelle en introduisant de l’auto-corrélation spatiale (ACS) définie comme étant l’interdépendance dans l’espace entre observations, on constate que l’hypothèse d’indépendance des résidus est violée (Saas and Gosselin [2014]) et on arrive à une sous-estimation des erreurs-types des estimateurs des effets fixes lorsque l’ACS est positive [Saas and Gosselin, 2014].

Les phénomènes écologiques sont complexes à modéliser [Cressie et al., 2009] car ils sont souvent aléatoires, spatialement hiérarchisés et mélangent des effets de multiples variables aléatoires ou non. L’une des approches les plus utilisées de nos jours pour analyser le système écologique est le modèle state-space (linéaire ou non) qui s’adapte bien aux variations environnementales comme la dynamique des populations, le comportement des animaux ...

[Pedersen et al., 2011] et donne à spécifier des modèles avec une structure de dépendance non stationnaire [Higdon, 1998].

Un modèle state-space est semblable aux séries temporelles et, qui plus, permet de considérer les vrais états du phénomène non observés et d’aboutir à inférer sur ces états à partir des données observées en tenant compte des erreurs qui viennent de ces observations et du processus stochastique du phénomène étudié (1). Jonsen et al. [2003] utilisèrent le même schéma pour modéliser le vrai mouvement non observé, dans le temps, des tortues de mer avec des données simulées. L’erreur qui en résulte de l’ajustement constitue l’erreur de mesure de la position de la tortue dans le temps (erreur d’observations).

Notre objectif sera donc de modéliser la dynamique temporelle de populations d’espèces spatialement structurées, en présence-absence, par une approche (fréquentiste ou bayésienne) hiérarchique consistant en une construction brique par brique de cette dynamique partant d’un processus interne latent régi par des hyper-paramètres jusqu’au modèle d’observations comme l’illustre la figure (1) [Ancelet, 2008]. L’approche hiérarchique nous permet de mieux tenir compte des corrélations qui existent au niveau spatial et au niveau temporel et d’estimer la tendance temporelle de la dynamique des populations d’espèces pour identifier les réponses des populations au changement climatique, aux mesures prises à l’égard des espèces (espèces protégées ou non protégées) ...

Comprendre l’effet du climat sur la flore peut parfois être un défi très difficile à relever en dépit de la longue histoire de la saisonnalité en biologie [Inouye et al., 2019]. Néanmoins des efforts pour appréhender le comportement des espèces face à ses effets ont été déployés ces 2 dernières décennies. Des questions que l’on se pose, notamment à quel moment débute l’activité de vol des insectes [Roy and Sparks, 2000], le passage des oiseaux migratoires [Marra et al., 2005] mais plus encore des questions liées à l’agriculture suscitent l’intérêt de préciser les premières dates de floraison des plantes, des récoltes après la senescence ..., et donc de connaître la phénologie des espèces [Edwards and Crone, 2020, Cleland et al., 2007].

Dans ce travail, notre objectif sera aussi d’étudier la phénologie de chaque espèce et ainsi déterminer leurs différentes phases périodiques face aux effets du climat.

Légitimement des questions que l’on pourrait se poser pour étudier la présence-absence des espèces sont : quelle est la probabilité d’observation d’une plante conditionnellement à sa saisonnalité observée, l’altitude ou la région de relevé?, y a-t-il un effet équipe sur l’observation d’une espèce? Évidemment, une équipe peut passer outre à une espèce et la juger inintéressante ou simplement ne pas la voir à cause des conditions météorologiques du moment (fortes pluies, chaleur ou froid ...) ou de son état d’esprit. Pour mieux prendre en compte le phénomène dans le modèle, nous allons choisir un modèle plus cohérent par rapport aux données pour mieux estimer les autres paramètres. Par exemple, si  $p(s)$  est la probabilité de présence de l’espèce au site  $s$ ,  $p(s) = p_s D_s$  où  $p_s$  est la probabilité que l’espèce soit présente et  $D_s$  la probabilité que l’espèce soit détectée.

## Définitions :

- Soit  $(\Omega, \mathbb{F}, \mathbb{P})$  un espace probabilisé,  $E \subset \mathbb{R}$  un ensemble. Pour  $t \in E$ , soit  $Z_t : (\Omega, \mathbb{F}, \mathbb{P}) \rightarrow (\mathbb{R}, B(\mathbb{R}))$  une variable aléatoire. On appelle processus stochastique d'espaces d'états  $E$  la famille  $(Z_t)_{t \in E}$ .
- $Z$  est dit stationnaire au second ordre si  $Z_t \in L^2(\Omega, \mathbb{F}, \mathbb{P})$  et si :
  - $\mathbb{E}(Z_t)$  est fini,
  - $Cov(Z_t, Z_{t+h})$  ne dépend pas de  $t$ .
- $Z$  est un processus autorégressif d'ordre  $p$  si elle est stationnaire et est solution d'une équation de récurrence de la forme :  $Z_t + \sum_{k=1}^p \phi_k Z_{t-k} = \eta_t$  avec  $\eta_t$  un bruit blanc et  $\phi_p \neq 0$
- Soit  $X_s$  un processus stochastique. On dit que  $X$  est markovien (ou est un processus de Markov) si  $\mathbb{P}(X_s | X_{-s}) = \mathbb{P}(X_s | X_{s-1})$  où  $X_{-s}$  représente toutes les réalisations de  $X$  sauf  $X_s$ .

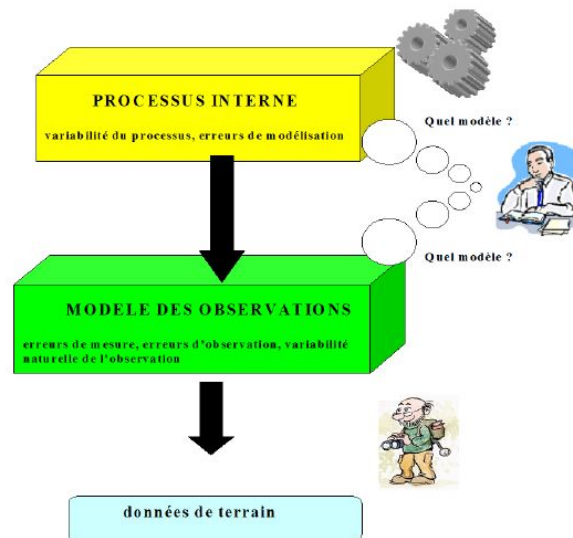


FIGURE 1 – Modélisation statistique hiérarchique Ancelet(2008). On voit la structure du modèle qui montre le processus réel (en jaune) qui se déroule et qui permet de prendre en compte l'incertitude du modèle, la variabilité et les erreurs de processus. Ensuite vient le modèle d'observations (en vert) qui se base sur l'étape précédente et qui tient compte des erreurs d'observation Cressie(2009).



### 3 Matériels et méthodes

Dans cette section, nous nous intéressons aux matériels utilisés et aux méthodes statistiques que nous avons employées pour aboutir à nos objectifs.

Dans la plupart des phénomènes écologiques étudiés, d’importantes quantités d’informations sont souvent disponibles à plusieurs niveaux d’unités d’observation. Nous avons utilisé un modèle hiérarchique dans ce travail qui permet de prendre en considération ces informations. En plus, prendre un modèle non hiérarchique ne serait pas cohérent dans notre cas, car avec peu de paramètres, il ne permet pas d’ajuster avec précision une quantité de données (dans notre cas) et avec beaucoup de paramètres, on arrive à un sur-ajustement de ces données [Gelman et al., 2004, Chap. 5].

La dynamique spatio-temporelle de populations d’espèces, en écologie, indique l’étude de l’évolution dans le temps et dans l’espace du nombre d’individus au sein de la population et vise à comprendre les influences environnementales sur ses effectifs. Un modèle à densité-dépendance dit de *Gompertz* sera principalement l’objet de ce travail pour étudier la dynamique des populations par espèce.

Nous avons ensuite parlé de la méthode statistique utilisée pour inférer sur les paramètres du modèle, l’approximation de Laplace par le biais des outils dans R : [TMB](#) avant d’évoquer quelques utilisés pour caractérisés la phénologie des espèces.

Les données utilisées dans ce travail sont issues de l’Inventaire forestier national (IFN).

#### 3.1 Le modèle hiérarchique sur grille latente

Selon [Ancelet \[2008\]](#), la démarche hiérarchique consiste “à décomposer un phénomène aléatoire complexe en processus aléatoires élémentaires qui peuvent être modélisés avec des structures probabilistes standard”. Le modèle que nous étudions ici est un modèle spatial qui permet de modéliser la dynamique temporelle de la population d’espèces sur la France métropolitaine  $D$  discrétisé pour faire ressortir des ressemblances ou dissemblances de cette dynamique pour des valeurs obtenues dans des zones voisines. Nous construisons ainsi une grille latente où l’on simule l’abondance de la dynamique temporelle sur chaque nœuds de celle-ci pour modéliser cette structure complexe. La figure (2) montre cette grille latente avec 2548 nœuds.

Considérons par exemple le modèle de [Rhodes and Jonzén \[2011\]](#). Soient  $\theta = \{\sigma, \gamma, \rho, r, \alpha, \beta, \sigma_o\}$  les paramètres du modèle,  $Z_i^t$  la variable aléatoire représentant l’abondance de la population des espèces. Soit  $Y$  la variable représentant les relevés de terrain (i.e. données de fréquence ou d’abondance). On suppose que  $(Y_k)_{k \in \{1, \dots, n_{obs}\}}$  sont indépendantes. La variable  $Z_i^t$  n’est pas observée mais constitue une étape importante du processus et permet de conceptualiser le modèle [[Ancelet, 2008](#)]. On suppose que les variations environnementales  $E$ , et  $Z$  sont les effets aléatoires du modèle. La figure (3) illustre la structuration du modèle avec les hyper-paramètres, la couche latente et les observations.

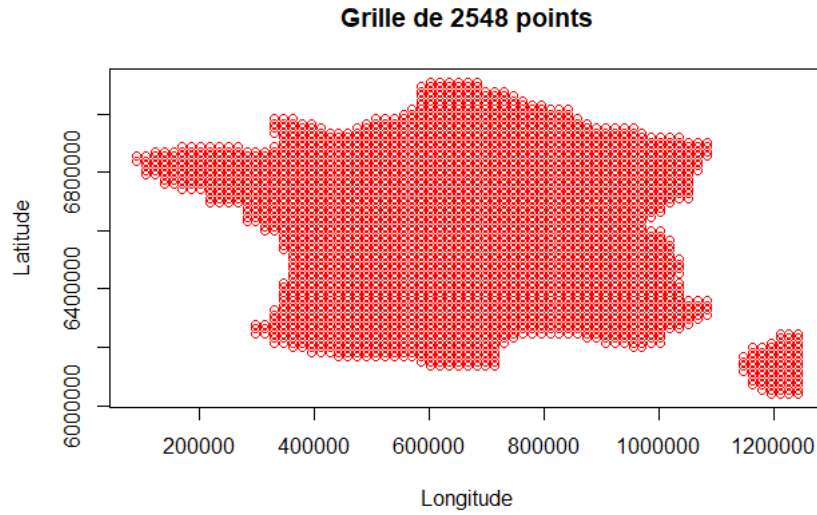


FIGURE 2 – Grille de 2548 points sur la France métropolitaine. On simule sur cette grille la log-abondance ( $\log Z$ ) de la dynamique temporelle de populations d'espèces

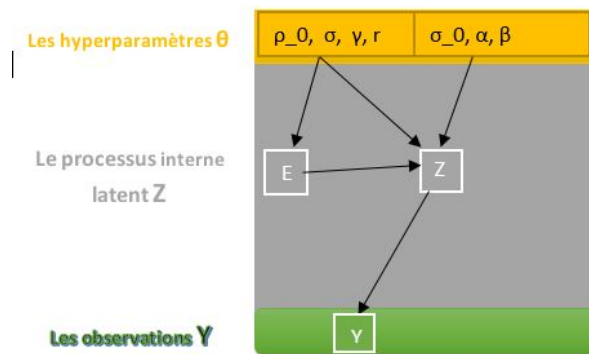


FIGURE 3 – Graphe acyclique hiérarchique du modèle hiérarchique de Rhodes et Jonzèn

Notre travail de modélisation et d’inférence est effectué sous R version 4.2 (R Core team) avec le package **TMB** [Kristensen et al., 2015] codé en *C++*. Une partie des calculs dans R a été effectuée à l’aide du package **INLA** pour obtenir des matrices dont leurs inverses sont creuses et qui approximent les matrices de variances-covariances des champs gaussiens (cf 3.2.2.3).

## 3.2 Modèle de Gompertz

Il existe différentes méthodes pour représenter la dynamique de la population d’espèces (modèle de Ricker [Pagel and Schurr, 2012], modèle de Gompertz ...). Dans ce stage, nous nous sommes limités au modèle à densité dépendante de **Gompertz** pour des raisons computationnelles mais aussi car ce modèle est souvent utilisé en écologie et représente la dynamique de beaucoup d’espèces [Rhodes and Jonzén, 2011]. Le modèle state-space de Gompertz est un processus stochastique, auto-régressif à l’échelle log, utilisé pour représenter des séries temporelles qui modélisent l’abondance de populations d’espèces en tenant compte de leurs structures spatiales comme décrit dans Thorson et al. [2015].

### 3.2.1 Modèle de Rhodes et Jonzén

L’article de Rhodes and Jonzén [2011] met en exergue la bonne stratégie de répartir l’effort d’échantillonnage entre le nombre de sites et le nombre de passages dans chaque site pour estimer la tendance temporelle. En particulier, le mélange optimal du nombre total de sites et du nombre de passages par site est susceptible de dépendre des corrélations spatiales et temporelles dans la dynamique des populations. Il en résulte que :

- lorsque la corrélation spatiale est faible et la corrélation temporelle est élevée, la meilleure option est d’augmenter le nombres de sites en particulier lorsque les erreurs d’observations et/ou les variations spatiales de la tendance temporelle sont élevées.
- lorsque la corrélation temporelle est faible et la corrélation spatiale est élevée, la meilleure stratégie est d’échantillonner fréquemment peu de sites, en particulier si les erreurs d’observation et/ou les variations spatiales de la tendance temporelle sont faible.
- quand les abondances sont spatialement indépendantes, le mélange optimal s’obtient en maximisant le nombre de sites.

Pour un site  $i$  dans le domaine  $D$  et une date  $t$  donnée, la dynamique dans Rhodes and Jonzén [2011] est donnée par :

$$Z_i^t = Z_i^{t-1} \exp(-0.5\sigma^2 + \rho(\ln K_i^{t-1} - \ln Z_i^{t-1}) + E_i^t) \quad (1)$$

avec  $Z_i^t$  l’abondance au site  $i$  au temps  $t$ ,  $1 - \rho$  la force de densité-dépendance, aussi appelée auto-corrélation temporelle,  $E_i^t$  la variable aléatoire multivariée normale de moyenne nulle et de matrice de variance-covariance  $\Sigma$  définie par :

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \gamma^{d_{1,2}} & \dots & \gamma^{d_{1,N}} \\ \gamma^{d_{2,1}} & 1 & \dots & \gamma^{d_{2,N}} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma^{d_{N,1}} & \gamma^{d_{N,2}} & \dots & 1 \end{pmatrix}$$

au site  $i$  au temps  $t$  avec  $var(E_i^t) = \sigma^2$  où  $\gamma$  est la corrélation spatiale entre sites.  $E_i^t$  représente la variation environnementale au site  $i$  au temps  $t$ .  $d_{i,j}$  est la distance géographique entre les sites  $i$  et  $j$ ,  $N$  le nombre de sites dans  $D$ . La quantité  $-0.5\sigma^2$  assure que la tendance attendue soit la même que celle du modèle déterministe, c'est à dire, elle assure que le taux de croissance attendu soit le même que celui pour la version déterministe équivalente.

$K_i^t$  est la carrying capacity au site  $i$  au temps  $t$ , elle représente la dynamique en population d'équilibre et est déterministe au temps  $t \geq 1$  :

$$K_i^t = K_i^{t-1} \exp(r_i) \quad (2)$$

où  $r_i$  est la tendance temporelle de la dynamique temporelle au site  $i$ . Au départ  $t = 0$ ,  $\ln K_i^0$  suit une loi normale de moyenne  $\alpha + \beta X$  et de variance  $\sigma_0^2$  où  $X$  est une variable écologique influente (ex. le climat ...).

### 3.2.2 Le modèle de Thorson

L'article de [Thorson et al. \[2015\]](#) relate la nécessité d'inclure de la dépendance spatiale dans les questions traitant de la densité-dépendance des populations d'espèces. Ils étudient l'estimation de la densité-dépendance dans le modèle de Gompertz en non-spatial et spatial. Le résultat dans le cas du modèle de Gompertz non-spatial est que les estimations des paramètres sont biaisées alors que celles du modèle spatial sont précises et sans biais et cela pour différents scénarios de simulations.

Le modèle spatial de Gompertz est donné par cette formulation [[Thorson et al., 2015](#)] :

$$\log(Z_i^t) = \Omega_i + \rho \log(Z_i^{t-1}) + E_i^t, \text{ si } t > 0 \quad (3)$$

avec  $\log(Z_i^0) = \frac{\Omega_i}{1-\rho} + \phi + E_i^0$ , si  $t = 0$  où  $\rho$  est la force de densité-dépendance,  $\Omega_i$  un champ aléatoire gaussien qui modélise la moyenne en abondance au site  $i$ ,  $E_i^t$ , la variation environnementale (variation du processus) au site  $i$  au temps  $t$ , est un champs aléatoire gaussien de moyenne nulle.  $\phi$  est un vecteur où  $\phi$  est le log-ratio entre l'abondance à l'équilibre et la médiane de l'abondance en  $t = 0$ ,  $\frac{\Omega_i}{1-\rho}$  est la carrying capacity.

Pour des besoins d'adéquation de la formulation de [Thorson et al. \[2015\]](#) avec celle de [Rhodes and Jonzén \[2011\]](#), nous avons remplacé l'équation (3) par l'équation (4) en faisant varier la moyenne en abondance dans le temps (cf 7.2).

$$\log(Z_i^t) = \Omega_i^t + \rho \log(Z_i^{t-1}) + E_i^t, \text{ si } t > 0 \quad (4)$$

avec  $\log(Z_i^0) = \frac{\Omega_i^0}{1-\rho} + \phi 1 + E_i^0$ , si  $t = 0$  et  $\Omega_i^t = \Omega_{0i} + t\Omega_{1i}$ .

$\Omega_0 \sim MVN(\alpha 1, \Sigma_{\Omega_0})$  et  $\Omega_1 \sim MVN(r, \Sigma_{\Omega_1})$  suivent des lois multivariées normales (MVN) de matrice de variance covariance  $\Sigma_{\Omega_0} = \sigma_{\Omega_0} R_h$  et  $\Sigma_{\Omega_1} = \sigma_{\Omega_1} R_h$ . Elles sont approximées par des champs gaussiens aléatoires de Markov (GMRF) (cf. 3.2.2.3),  $h$  est la distance entre sites,  $E \sim MVN(0, \Sigma_E)$ .

Nous allons introduire dans ce qui suit (parties 3.2.2.1, 3.2.2.2 et 3.2.2.3) le passage de la matrice  $\Sigma$  à la matrice de précision  $Q$  telle que  $Q^{-1}$  creuse en plus, représente  $\Sigma$  dans une certaine norme.

### 3.2.2.1 Matrice de covariance Matérn

Soit  $\|\cdot\|$  la distance euclidienne dans  $R^d$ . La matrice de covariance Matérn entre les sites  $i$  et  $j$  est donnée par :

$$C(i, j) = \frac{\sigma^2}{2^{\nu-1} \Gamma(\nu)} (\kappa \|i - j\|)^{\nu} K_{\nu}(\kappa \|i - j\|) \quad (5)$$

où  $K_{\nu}$  est la fonction de Bessel modifiée de deuxième espèce,  $\nu > 0$  contrôle la régularité,  $\kappa > 0$  est un paramètre d'échelle,  $\sigma^2$  la variance marginale. La fonction de corrélation Matérn  $C(h)$  avec  $h = \|i - j\|$  sur la figure (4) illustre une corrélation forte au voisinage et décroît au fur et à mesure qu'on s'éloigne.

- Si  $\nu = 1/2$ , on retrouve une corrélation exponentielle :  $C(i, j) = \sigma^2 e^{-\kappa \|i - j\|}$
- Si  $\nu = +\infty$ , on a une corrélation gaussienne :  $C(i, j) = \sigma^2 e^{-\kappa^2 \|i - j\|^2 / 2}$

La portée pratique  $Range = \sqrt{8\nu}/\kappa$  [Thorson et al., 2015, Appendix B, gambas.cirad.fr] est la distance à partir de laquelle il n'y a plus de corrélation spatiale.  $Range$  est un paramètre de plage qui correspond à une corrélation proche de 0.1 à une distance égale à  $Range$  pour tout  $\nu$ .

Les travaux du projet Gambas (Generating Advances in Modeling Biodiversity And Ecosystem Services), dont l'unité EFNO est partenaire, ont montré empiriquement qu'il faudrait un maillage inférieur ou égal à 1/11 de la portée pratique pour des données réelles et 1/14 en simulation. Donc plus la grille latente est fine et plus on reproduit de structures locales.

### 3.2.2.2 Approximation par une équation aux dérivées partielles (SPDE)

Un vieux résultats de WHITTLE [1954] dit que si  $Z_i$  est un champ gaussien de matrice de covariance Matérn  $C$  définie dans (5) alors  $Z_i$  est solution de l'équation aux dérivées partielles (SPDE) suivante :

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} Z_i = \tau W_i, i \in R^d, \alpha = \nu + d/2, \kappa > 0, \nu > 0 \quad (6)$$

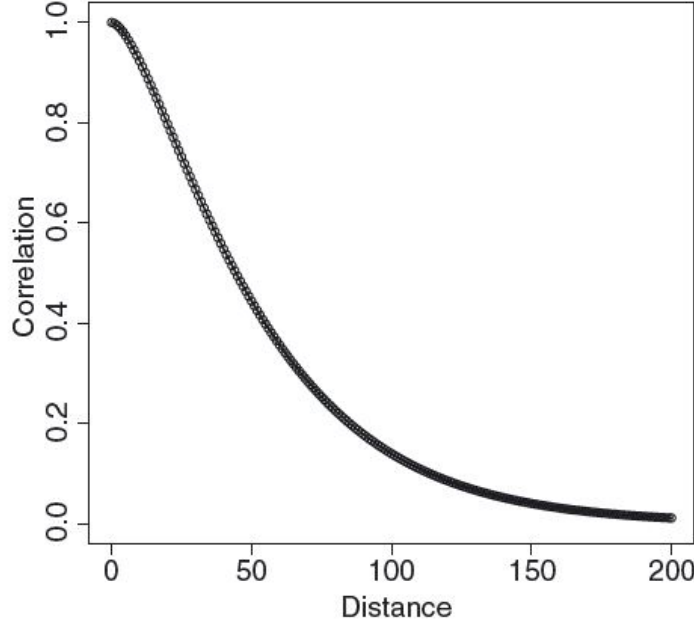


FIGURE 4 – Corrélacion spatiale en fonction de la distance  $h$  entre sites issue de la fonction de corrélacion Matérn Lindgren (2011).

où  $\tau$  est un paramètre de dispersion,  $\Delta$  est le Laplacien

$$\Delta = \sum_{i=1}^d \frac{\partial^2}{\partial z_i^2},$$

et  $W_i$  est un bruit blanc.

La variance marginale  $\sigma^2$ , proportionnelle à  $\frac{1}{\tau^2}$ , est donnée par

$$\sigma^2 = \frac{1}{\tau^2} \left( \frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^{2\nu}} \right),$$

Soit  $(u, v)$  les indices (longitude-latitude) du site  $i$ . Pour  $\alpha = 2$ ,  $d = 2$  on a à partir de (6) :

— Grâce à la méthode des différences finies

$$(\kappa^2 - \Delta)Z_{u,v} = \tau W_{u,v}$$

$$\kappa^2 Z_{u,v} - \frac{1}{dx} \left( \frac{Z_{u+1,v} - Z_{u,v}}{dx} - \frac{Z_{u,v} - Z_{u-1,v}}{dx} \right) - \frac{1}{dy} \left( \frac{Z_{u,v+1} - Z_{u,v}}{dy} - \frac{Z_{u,v} - Z_{u,v-1}}{dy} \right) = \tau W_{u,v}$$

En termes de matrice et en rangeant les coefficients des  $Z_{u,v}$  dans une matrice  $B$  creuse, on arrive à la forme suivante

$$BZ = \tau W,$$

dont la matrice de variance covariance de  $Z$  est  $\Sigma = cov(Z)$  définie par

$$B\Sigma B^T = \tau^2 I \Rightarrow \Sigma^{-1} = \frac{1}{\tau^2} B^T B$$

D'où la matrice de précision  $Q = \frac{1}{\tau^2} B^T B$

— Grâce à la méthode des éléments finis ([gambas.cirad.fr](http://gambas.cirad.fr))

On note par  $\langle \cdot, \cdot \rangle$  le produit scalaire dans  $R^d$ . Soit  $\psi$  une fonction qui vérifie

$$\langle (\kappa^2 - \Delta)Z, \psi \rangle = \tau \langle W, \psi \rangle,$$

avec

$$\langle \phi, \psi \rangle = \int \phi_s \psi_s ds$$

En appliquant les conditions aux limites de Neumann, nous avons :

$$\begin{aligned} \langle \Delta Z, \psi \rangle &= -\langle \nabla Z, \nabla \psi \rangle \\ \Rightarrow \kappa^2 \langle Z, \psi \rangle + \langle \nabla Z, \nabla \psi \rangle &= \tau \langle W, \psi \rangle \end{aligned}$$

où  $Z_i = \sum_{j=1}^n Z_j \psi_{ij}$  et  $\psi_{ij} = 1$  si  $i = j$  sinon  $\psi_{ij}$  décroît linéairement vers 0 en dehors du voisinage du site  $i$ .

$$\Rightarrow \sum_{j=1}^n \kappa^2 Z_j \langle \psi_j, \psi_i \rangle + \langle \nabla \psi_j, \nabla \psi_i \rangle = \tau \langle W, \psi_i \rangle$$

En forme matricielle on a :

$$(\kappa^2 M + G)Z = \tau F$$

- $Z = (Z_j)_j$
- $M = (\langle \psi_j, \psi_i \rangle)_{i,j}$  la matrice masse de masse creuse,
- $G = (\langle \nabla \psi_j, \nabla \psi_i \rangle)_{i,j}$ , est la matrice de rigidité creuse,
- $F = \langle W, \psi_i \rangle$

$$\Rightarrow Q = (\kappa^2 M + G)M^{-1}(\kappa^2 M + G)$$

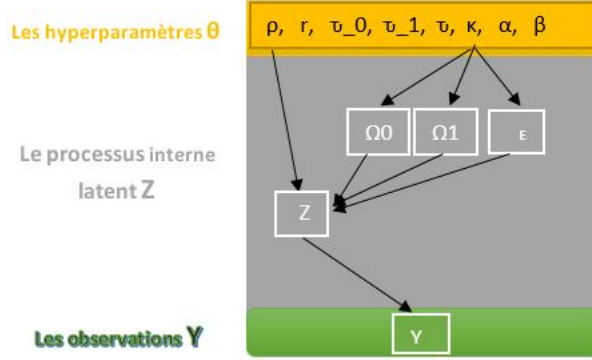


FIGURE 5 – Graphe acyclique hiérarchique du modèle hiérarchique de Thorson modifié.

### 3.2.2.3 Approximation des champs gaussiens dans INLA

Les champs gaussiens  $\Omega_0, \Omega_1$  et  $E$  sont approximés par des GMRFs en utilisant *R-INLA* qui calcule  $G_0, G_1$  et  $G_2$  des matrices creuses, pour  $\nu = 1$  et  $d = 2$  la dimension.

L'utilisation des GMRF permet de gagner du temps de calcul à l'ordre de  $O(n^{3/2})$  au lieu d'inverser une matrice  $n \times n$  dont le coût en terme de temps est à l'ordre de  $O(n^3)$ .

$$\begin{aligned}\Omega_0 &\sim MVN(\alpha 1, \Sigma_{\Omega_0}), \Sigma_{\Omega_0} = (\tau_{\Omega_0}^2 (\kappa^4 G_0 + 2\kappa^2 G_1 + G_2))^{-1} \\ \Omega_1 &\sim MVN(0, \Sigma_{\Omega_1}), \Sigma_{\Omega_1} = (\tau_{\Omega_1}^2 (\kappa^4 G_0 + 2\kappa^2 G_1 + G_2))^{-1} \\ E &\sim MVN(0, \Sigma_E \otimes \Sigma), \Sigma_E = (\tau_E^2 (\kappa^4 G_0 + 2\kappa^2 G_1 + G_2))^{-1},\end{aligned}\tag{7}$$

[Thorson et al., 2015]

avec [Lindgren et al., 2011]

$$\Sigma = \frac{\sigma_E^2}{1 - \rho^2} \begin{pmatrix} 1 & \rho & \dots & \rho^{T-1} \\ \rho & 1 & \dots & \rho^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \dots & 1 \end{pmatrix},$$

La fonction de corrélation  $Corr$  et  $G_0, G_1, G_2$  sont obtenues grâce aux commandes suivantes :

```
Corr <- inla.matern.cov(nu,kappa,dist,d=2,corr=T) # dist=matrice de distance
mesh <- inla.mesh.create(locations) # locations=coordonnées long-lat
spde <- inla.spde2.matern(mesh)
G0 <- spde$param.inla$M0
G1 <- spde$param.inla$M1
G2 <- spde$param.inla$M2
```



### 3.3 Modèle d'observation

Pour le modèle d'observation, nous avons opté d'interpoler sur les quatre sites latents qui entourent une placette comme suit :

- Chercher sur la grille les quatre sites virtuels qui entourent chaque placette observée ;
- Changer de base en prenant comme origine le nœud situé en bas et à gauche de la placette ;
- Calculer les nouvelles coordonnées de cette placette et des quatre sites latents dans la nouvelle base de sorte que ses coordonnées soient comprises entre 0 et 1.
- calculer la moyenne pondérée des simulations à l'endroit des quatre sites latents qui l'entourent.

Soit  $m$  la placette (relevé d'inventaire), de coordonnées dans la nouvelle base long-lat  $(x_m, y_m)$ , et si  $A, B, C, D$  sont les nœuds qui l'entourent, et tenant compte que l'origine du repère  $A$  se situe en bas et à gauche de  $M$  et que  $A, B, C, D$  sont placés dans le sens des aiguilles d'une montre,  $\log Z_A^t, \log Z_B^t, \log Z_C^t, \log Z_D^t$  sont les simulations à l'endroit des quatre sites latents au temps  $t$ , on a :

$$\hat{Y}_m^t = (1 - x_m)(1 - y_m)\log Z_A^t + (1 - x_m)y_m\log Z_B^t + x_my_m\log Z_C^t + x_m(1 - y_m)\log Z_D^t$$

Cette conception de modélisation nous permet de statuer que le phénomène qui se déroule au point de relevé IFN a une structure semblable à ceux non observés se trouvant au niveau des quatre nœuds de la grille latente qui l'entourent, de sorte que si ce relevé est proche d'un nœud quelconque, les autres vont être négligés et la dynamique en ce relevé tend linéairement vers celle du nœud le plus proche.

Ainsi, pour modéliser la présence-absence, on calcule la probabilité  $\hat{p}$  que l'espèce soit présente au site  $m$  :

- $\hat{p} = 1 / (1 + \exp(-\hat{Y}_m^t))$  ;
- Spécifier que  $Y_m^t$  suit une loi de Bernoulli de paramètre  $\hat{p}$  :  $Y_m^t \sim \text{Bernoulli}(\hat{p})$ .

### 3.4 Inférence statistique

Nous avons utilisé *TMB* (*Template Model Builder* en anglais) pour estimer les paramètres du modèle  $\theta$ . TMB utilise l'approximation de Laplace en minimisant la négative log-vraisemblance (marginale) pour estimer et calculer les intervalles de confiance. Il s'appuie sur des techniques de dérivation (Automatic Differentiation) très puissantes comme *CppAD* dans R codé en C++ en faisant appel aux packages *Matrix* et *Eigen* de R. Dans ce qui suit nous allons introduire cette méthode d'approximation.

On note  $u \in R^p$  le vecteur des variables latentes et  $\theta \in R^q$  les hyper-paramètres du modèle. Soit  $f(u, \theta)$ , deux fois différentiables, la négative jointe vraisemblance des données et des variables latentes (exemple cf eq.15). C'est cette fonction qui est codé dans TMB sous forme de code source C++. Le maximum de vraisemblance pour estimer  $\theta$  est donné par :

$$L(\theta) = \int_{R^p} \exp(-f(u, \theta)) du \quad (8)$$

Dans l'équation (8), les variables latentes sont intégrées, ainsi elle ne dépend que des effets fixes  $\theta$ .

On pose  $\hat{u}(\theta) = \operatorname{argmin}_u f(u, \theta)$  et  $H(\theta)$  la matrice hessienne de  $f(u, \theta)$  en  $u$  évaluée en  $\hat{u}(\theta)$ , i.e.  $H(\theta) = \frac{\partial^2 f}{\partial u^2}(\hat{u}(\theta), \theta)$ . Le théorème de Taylor-Young au point  $\hat{u}(\theta)$  nous donne :

$$f(u, \theta) = f(\hat{u}(\theta), \theta) + \nabla f(\hat{u}(\theta), \theta)(u - \hat{u}(\theta), \theta) + \frac{1}{2}H(\theta)(u - \hat{u}(\theta), \theta)^2 + O(u - \hat{u}(\theta); \theta)^3 \quad (9)$$

Comme on a  $\nabla f(\hat{u}(\theta), \theta) = 0$ , il en résulte l'équation (10) :

$$f(u, \theta) = f(\hat{u}(\theta), \theta) - \frac{1}{2}H(\theta)(u - \hat{u}(\theta), \theta)^2 \quad (10)$$

De (8) on a :

$$L(\theta) = \exp(-f(\hat{u}(\theta), \theta)) \int_{R^p} \exp\left(-\frac{1}{2}H(\theta)(u - \hat{u}(\theta), \theta)^2\right) du \quad (11)$$

Le terme qui accompagne l'intégrale dans l'équation (11) s'identifie naturellement comme étant la densité de probabilité d'une loi normale multivariée centrée en  $\hat{u}(\theta)$  et de matrice de variance-covariance  $H$ .

$$\begin{aligned} \Rightarrow \int_{R^p} \exp\left(-\frac{1}{2}H(\theta)(u - \hat{u}(\theta), \theta)^2\right) du &= \int_{R^p} \exp\left(-\frac{1}{2}(u - \hat{u}(\theta), \theta)^T H(\theta)(u - \hat{u}(\theta), \theta)\right) du \\ &= \sqrt{2\pi^p} \det(H(\theta))^{-\frac{1}{2}} \end{aligned}$$

Ce qui donne finalement le maximum de vraisemblance  $L^*(\theta)$

$$L^*(\theta) = \exp(-f(\hat{u}(\theta), \theta)) \times \sqrt{2\pi^p} |\det(H(\theta))|^{-\frac{1}{2}} \quad (12)$$

L'estimateur du maximum de vraisemblance  $\hat{\theta}$  des paramètres  $\theta$  de dimension  $(1 \times q)$  va alors minimiser la log-négative de l'approximation de Laplace appelée *fonction objective* :

$$-\log L^*(\theta) = -p \log \sqrt{2\pi} + \frac{1}{2} \log(\det(H(\theta))) + f(\hat{u}(\theta), \theta) \quad (13)$$

La précision de l'incertitude sur  $\theta$  est donnée par  $VAR(\hat{\theta}) = -\hat{\theta} (\nabla^2 \log L^*(\theta))^{-1} \hat{\theta}^T$  [Kristensen et al., 2015].

Nous considérons l'équation (4) avec laquelle nous travaillerons par la suite en simulation et qui servira à faire l'inférence statistique.

La densité jointe au temps  $t$ , que l'on note ici par  $[\cdot, \cdot]$ , est donnée par (14) :

$$\begin{aligned} [Y^t, Z^t, \Omega^t, E^t, \theta] &= [Y^t | Z^t, \theta] \times [Z^t | \Omega^t, E^t, \theta] \times [E^t | \theta] \times [\Omega^t | \theta] \times [\theta] \\ &= \prod_{k=1}^{n_{obs}} [Y_k^t | Z^t, \theta] \prod_{i=1}^n [Z_i^t | \Omega^t, E^t, \theta] \prod_{i=1}^n [E_i^t | \theta] \prod_{i=1}^n [\Omega_i^t | \theta] \cdot [\theta] \end{aligned} \quad (14)$$

$$\begin{aligned} \Rightarrow -\log[Y^t, Z^t, \Omega^t, E^t, \theta] &= -\sum_{k=1}^{n_{obs}} \log[Y_k^t | Z^t, \theta] - \sum_{i=1}^n \log[Z_i^t | \Omega^t, E^t, \theta] \\ &\quad - \sum_{i=1}^n \log[E_i^t | \theta] - \sum_{i=1}^n \log[\Omega_i^t | \theta] - \log[\theta] \end{aligned} \quad (15)$$

où  $n_{obs}$  est le nombre d'observations (de placettes),  $n$  le nombre de sites latents. Les  $n$  sites sont construits sur une grille latente sur la France métropolitaine  $D$ . La figure (2) montre une grille de  $n=2548$  sites latents où l'on simule les  $\log(Z_i^t)$  suivant l'équation 4.

La densité jointe a posteriori est donnée par  $[Z, E, \Omega, \theta | Y] \propto [Z, E, \Omega, \theta] [Y | Z]$ ,

avec  $[Z, E, \Omega, \theta] = [Z | E, \Omega, \theta] [E | \theta] [\Omega | \theta] [\theta]$

La loi marginale a posteriori peut être calculée en faisant cette intégration grâce à la formule des probabilités totales :

$$[\theta | Y] = \int [Z, E, \Omega, \theta | Y] dz d\epsilon d\omega \quad (16)$$

### 3.5 Présentation des données de l'inventaire forestier

Nous disposons des données de l'Inventaire forestier national collectées depuis octobre 2005 (campagne de 2006) jusqu'à fin décembre 2020. Le jeu de données est composé de la date de campagne, CAMPAGNE, le jour du relevé DATEECO, les longitudes et latitudes en Lambert 93 respectivement dans les colonnes XL et YL. Les noms scientifiques des espèces

TABLE 1 – Extrait des 3 premiers lignes du jeu de données IFN

CAMPAGNE	IDP	DATEECO	XL	YL	LB_NOM	ABOND
2006	118450	2005-10-03	1019571	6318249	Aphyllanthes monspeliensis	1
2006	118450	2005-10-03	1019571	6318249	Brachypodium pinnatum	3
2006	118450	2005-10-03	1019571	6318249	Buxus sempervirens	3
...	...	...	...	...	...	...

sont stockés dans la variable LB\_NOM et ABOND représente l’abondance correspondante de l’espèce. La colonne IDP est un identifiant unique pour chaque relevé (cf Table (1)).

Les données comptent 87187 relevés et 3296 espèces de plantes après avoir enlevé la campagne de 2006 . Cette suppression s’opère suite à un changement dans le protocole de récoltes de données IFN après la campagne de 2006, ce qui fait que nous n’avons pas pu exploiter cette partie du jeu de données. La figure (6) montre le niveau de relevés dans les données de l’Inventaire forestier national pour chaque année sur laquelle on voit bien que la campagne de 2006 présente un niveau significativement faible de placettes par rapport aux relevés des autres campagnes.

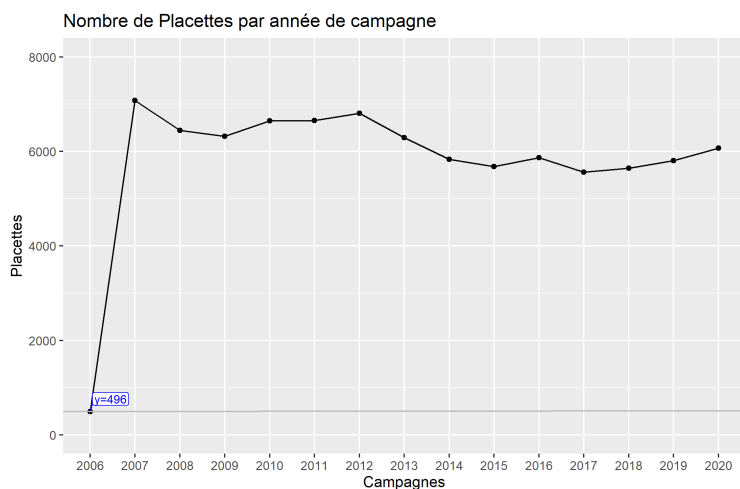


FIGURE 6 – Nombre de relevés par campagne

Nous allons par la suite nous intéresser à trois espèces de plantes avec lesquelles nous travaillerons en présence-absence, le sapin blanc (*Abies alba*), le houx (*Ilex aquifolium*) et l’anémone (*Anemone nemorosa*).

La figure (7) montre une diminution assez importante de la fréquence de l’anémone tandis que la fréquence du houx augmente entre 2009 et 2020. Le sapin blanc diminue progressivement à partir de 2010.

Dans la partie qui suit, nous allons introduire quelques modèles pour étudier la phénologie.

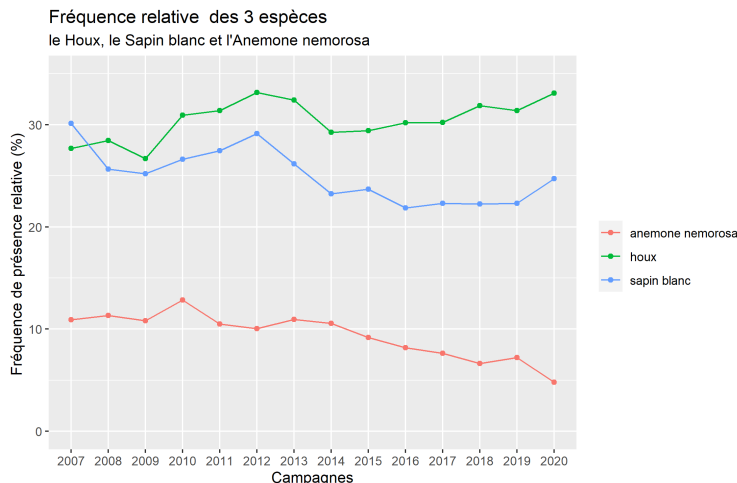


FIGURE 7 – Évolution de la fréquence de présence relative annuelle de l’anémone, du houx et du sapin blanc entre 2007 et 2020 dans les relevés de l’Inventaire forestier national

### 3.6 Phénologie

La phénologie est l’étude des phénomènes périodiques intra-annuels de certaines espèces animales ou végétales. Dans le cas de la flore vasculaire, il s’agit de déterminer les phases de développement des plantes, du débourrement à la fanaison en passant par la floraison et le début de l’automne.

Des mesures précises de l’index de végétation doivent être effectuées pour améliorer les modèles et pour appréhender cette variabilité inter-annuelle dans l’écosystème [Zhang et al., 2003].

Différentes méthodes ont été développées ces 2 dernières décennies pour déterminer le temps de la verdure et de la senescence de la végétation, utilisant des index de végétation à différence normalisée (NDVI) grâce à la télédétection satellitaire.

#### 3.6.1 Fonction sigmoïde à double asymétrie

$$ph(t) = (\omega_1 + \omega_2) + \frac{1}{2}(\omega_1 - \omega_2) \{ \tanh(\omega_3(t - u)) - \tanh(\omega_4(t - v)) \} \quad (17)$$

avec  $ph(t)$  l’index de végétation,  $t$  est le jour de l’année,  $u$  et  $v$  les dates correspondant respectivement aux pics de verdure au printemps et de feuilles fanées en automne.  $\omega_1 + \omega_2$  la moyenne de  $ph(t)$  et  $\omega_1 - \omega_2$  l’amplitude de variation,  $\tanh$  est la tangente hyperbolique.  $(\omega_1 - \omega_2) > 0$ . [Soudani et al., 2020]

#### 3.6.2 Fonction à double logistique

— Forme additive :

Fisher et al. [2006] modélisent la phénologie par une fonction à double logistique donnée par :

$$ph(t) = \beta + \alpha \left( \frac{1}{1 + e^{a+bt}} - \frac{1}{1 + e^{c+dt}} \right) \quad (18)$$

Comme (17),  $\beta$  est la moyenne de  $ph(t)$ ,  $a, b$  les paramètres qui décrivent le déroulement au printemps (débourrement et pic de floraison) et  $c, d$  ceux d'automne.

Beck et al. [2006] modélisent la phénologie par une double logistique additive donnée par :

$$ph(t) = mean_{amp} + (max_{amp} - min_{amp}) \left( \frac{1}{1 + e^{a(t-\mu)}} + \frac{1}{1 + e^{b(t-\nu)}} - 1 \right) \quad (19)$$

$\mu$  et  $\nu$  sont les jours de pics au printemps et en automne respectivement.

— Forme multiplicative :

Huisman et al. [1993] (équation 4 page 38) utilisent une forme multiplicative

$$ph(t) = M \frac{1}{1 + e^{a+bt}} \frac{1}{1 + e^{c-bt}}$$

dont  $a, b, c, M$  sont les paramètres à estimer.

Il existe d'autres modèles pour étudier la phénologie notamment celui de Zhang et al. [2003] par une approche gaussienne ou transformée en logit-gaussienne, Rose and Nagle [2021] par une régression harmonique ...

L'estimation de ces paramètres, dans notre cas, ne dépend que de l'espèce et des données. La figure (8) montre les courbes observées pour les 3 espèces choisies.

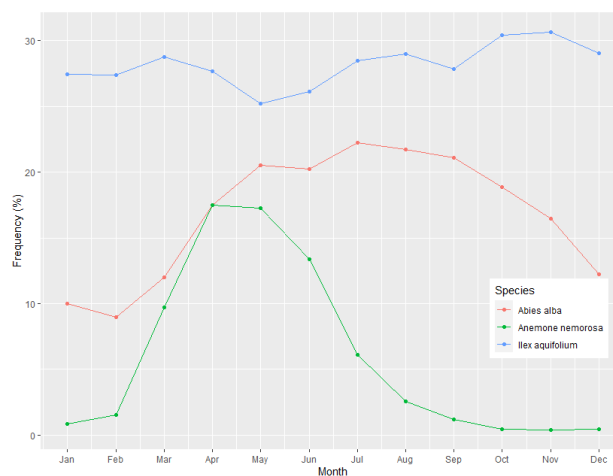


FIGURE 8 – Courbes de tendances saisonnières du houx, de l'anémone et du sapin blanc en fonction des mois.

## 4 Simulation et résultats

Le travail de simulation a été effectué suivant le modèle de Thorson modifié (4) avec  $n = 400$  sites latents sur toutes les placettes de l'inventaire sur une machine Intel core i5 8 cœurs 2.40Ghz.

Soit  $\theta$  l'ensemble des paramètres.  $\theta = \{r, \alpha, \phi, \log(\sigma_E), \log(\sigma_{\Omega_0}), \log(\sigma_{\Omega_1}), \log(\kappa), \text{logitrho}\}$ . la variable écologique  $X$  n'a pas été prise en compte sur cette partie, ainsi que la phénologie et la détectabilité.

Dans cette partie de simulation, on fait varier les effets aléatoires **random** =  $\{\Omega_0, \Omega_1, E\}$ , les autres paramètres sont supposés fixes.

Les variances de  $\Omega_0$  et  $\Omega_1$  sont données par  $\sigma_{\Omega_0} = 0.01$  et  $\sigma_{\Omega_1} = 0.001$ , nous choisissons d'utiliser une fonction de covariance Matérn avec les paramètres  $\nu = 1$  et  $\kappa = 1$ , la variance de  $E$ , donnée par  $\sigma_E^2 = 0.6^2$ ,  $\phi = 1$ . La tendance et la corrélation temporelle sont respectivement fixées à 0.133 et 0.5 .

Les 400 sites latents correspondent à un maillage régulier de 45 km dont nous considérons égal à 1 *unit* le pas. Les résultats de la simulation sont donnés dans le tableau (2) avec les vraies valeurs simulées (*trues\_values*), leurs estimations (*estimates*), leurs erreurs-types (*sd*) et les intervalles de confiance (*CI*).

La procédure d'estimation se fait comme suit dans R :

- écrire tout d'abord le code du modèle sous R en C++ (nommé fichier DDL.cpp) puis construire la liste des données DATA et des paramètres  $\theta$ ,
- compiler le fichier DDL.cpp avec la commande `TMB::compile` et faire un lien avec le DDL sans l'extension,
- calculer le gradient, l'optimiser grâce à la commande de R `nlminb` ou `optim` et reporter les estimateurs et leurs erreurs types (*sd*)

```
TMB::compile("DDL.cpp")
dyn.load(dynlib("DDL"))
obj <- MakeADFun(DATA,theta,random=random, DDL="DDL")
opt <- nlminb(start = obj$par,objective = obj$fn, gradient=obj$gr)
sdreport <- sdreport(opt)
```

Le résultat obtenu dans le tableau (2) montre que les vraies valeurs simulées dans le modèle sont toutes retrouvées avec plus ou moins un écart et qui tombent toutes dans les intervalles de confiance correspondants.

TABLE 2 – Tableau des vraies valeurs, des valeurs estimées, leurs erreurs-types (sd) et leurs intervalles de confiance à 95% (Conf.Int(95%))

	trues_values	estimates	sd	Conf.Int(95%)
alpha	1.0000	1.240000	0.15812	[ .93 ; 1.55 ]
r	0.0133	0.017600	0.01360	[ -.009 ; 0.04426 ]
phi	1.0000	0.700000	0.16000	[ .38 ; 1.0136 ]
rho	0.5000	0.420000	0.06015	[ .3 ; .54 ]
kappa	1.0000	1.008635	0.10050	[ .811 ; 1.205 ]
Range	2.8300	2.804210	0.28183	[ 2.252 ; 3.3566 ]
sigma_E	0.6000	0.558440	0.02527	[ .5089 ; .6079 ]
sigma_O0	0.0100	0.061700	0.06155	[ 0 ; .1823 ]
sigma_O1	0.0010	0.008830	0.00813	[ 0 ; .02477 ]

## 5 Ajustement sur les données IFN

Nous avons effectué l'inférence sur les trois espèces : l'anémone (*Anemone anemorosa*), le houx (*Ilex aquifolium*) et le sapin blanc (*Abies alba*) avec une grille régulière maillée à 16 km.

La phénologie, la détectabilité et la variable  $X$  n'ont pas été incluses. La tendance temporelle, fixée, est supposée la même pour chaque site et nous prenons comme effets aléatoires  $\Omega_0, \Omega_1$  et  $E$ .

Les calculs numériques ont été réalisés sur les machines du Centre de Calcul Scientifique en région Centre-Val de Loire et ont duré 13 heures.

Les résultats obtenus sont dans le tableau (3) pour les valeurs estimées des paramètres (estimates) et leurs écarts-types (sd) et dans le tableau (4) où sont stockés les intervalles de confiances de chaque paramètres estimés et pour chaque espèce.

Les écarts-types de  $E$ ,  $\Omega_0$  et  $\Omega_1$  sont données respectivement par sig\_E, sig\_O0 et sig\_O1.



TABLE 3 – Tableau des estimations pour les 3 espèces.

Espèces	values	alpha	phi	r	rho	kappa	Range	sig_E	sig_O0	sig_O1
Houx	estimates	-1.58	-.001	.01	.2052	.5037	5.62	.0875	1.5	.019
	sd	1.0258	.0470	.0071	.5014	.0767	.431	.0305	.954	.0126
Anémone	estimates	-3.68	-.453	-.082	.10	.5456	5.2	1.386	1.182	.0315
	sd	.2669	.2506	.018	.05	.049	.2535	.0461	.091	.0227
Sapin blanc	estimates	-1.91	-.0175	.0029	.508	.3667	7.713	.056	1.3	.0123
	sd	1.4505	.066	.0045	.3676	.1	.7706	.0432	.98	.0087

TABLE 4 – Intervalle de confiance des paramètres estimés pour les 3 espèces.

Espèces	Houx	Anémone	Sapin blanc
alpha	[-3.5 ; .42]	[-4.2 ; -3.2]	[-4.7 ; 1]
phi	[-.1 ; 1]	[-1 ; .04]	[-.15 ; .11]
r	[-.005 ; .02]	[-.12 ; -.05]	[-.01 ; .012]
rho	[-.1 ; 1.2]	[.01 ; .2]	[-.21 ; 1.23]
Range	[4.8 ; 6.5]	[4.7 ; 5.7]	[6.2 ; 9.22]
sig_E	[.03 ; .15]	[1.3 ; 1.5]	[0 ; .14]
sig_O0	[0 ; 3.4]	[1 ; 1.4]	[0 ; 3.22]
sig_O1	[0 ; 0.4]	[0 ; 3]	[0 ; .03]

## 6 Discussion

- Pour le houx, sa tendance temporelle est estimée à 0.01 par an (à l'échelle logit). Elle varie avec un écart-type de 0.019 et une corrélation temporelle de 0.2052. Sa carrying capacity est égale à 0.26 et le log-ratio  $\phi = -.001$  montre que sa population d'équilibre est sensiblement égale à sa médiane de départ. Sur 14 ans, on prévoit une augmentation de sa probabilité de présence qui passe de 0.206 à 0.23, soit en moyenne environ 2.4% de sa population. Le houx a une structuration locale portée environ sur 90 km dans notre modèle.
- On assiste à une diminution de la fréquence de l'anémone avec une tendance temporelle négative à l'ordre de 0.082 par année mais qui varie environ à  $C_v = 38\%$  autour. Il a une faible auto-corrélation temporelle de 0.10 et une portée proche à celle du houx environ égale à 83 km. Son log-ratio  $\phi = -.453$  indique que sa fréquence au départ est loin de sa fréquence à l'équilibre avec une carrying capacity de 0.028 et avec une faible probabilité de présence en moyenne égale à 0.025. Sur les 14 années de campagne, sa fréquence prédite est 0.0081, ce qui implique une baisse de sa fréquence en moyenne de 1.71%
- Le sapin blanc présente une tendance temporelle d'environ 0.03, faible mais très corrélée dans le temps (0.508). Il a une probabilité de présence moyenne d'environ 0.1481 et sa carrying capacity est d'ordre 0.1485. Cela révèle une faible variation de sa fréquence qui fluctue autour de sa fréquence médiane à l'équilibre ( $\phi = -.0175$ ). Nous

prédissons une augmentation de sa fréquence d'environ 0.54% sur les 14 années de campagne qui passe de 0.1481 à 0.1535. Il a une portée pratique structurée sur, à peu près, 123 km en France métropolitaine.

Ce résultat sur les trois espèces confirme l'intuition que nous avons en analysant la figure (7). Cependant, le fait que nous n'avons pas tenu compte de la détectabilité des espèces peut avoir une influence sur l'estimation de ces paramètres. Elle est intrinsèquement liée à l'effet équipe, qui est malencontreusement absent dans notre modèle du fait des données IFN non exhaustives.

Aussi nous n'avons pas pu intégrer la phénologie dans le modèle par manque de temps, néanmoins les modèles décrits dans (17) ou dans (18) pourraient satisfaire notre curiosité en l'incluant dans le modèle d'observation par le biais d'une probabilité  $ph$  obtenue grâce à l'index de végétation :  $Y_m^t \sim \text{Bernoulli}(\hat{p} \times ph)$ .

D'autres perspectives sur notre travail consistait à faire varier la densité-dépendance, l'auto-corrélation spatiale (dans notre cas le paramètre  $\kappa$  dans (4)), ce que nous essayerons prochainement. Il faudrait aussi prendre en compte du fait que la tendance temporelle est supposée constante pour chaque site. Ainsi, cette assertion pourrait être relaxée par la suite en incluant un nugget pour les observations qui fait de sorte que deux relevés de l'Inventaire forestier national n'aient pas une corrélation proche de 1 mais inférieure à 1. Autrement dit, que la variabilité n'est pas que structurée dans l'espace. Cette variabilité, noté  $\eta$ , serait incluse dans la partie observation comme suit :

$$\hat{p} = \frac{1}{1 + \exp((-\hat{Y}_m^t + \eta_m))}$$

Nous avons modélisé les espèces séparément comme si elles n'interagissaient pas entre elles. La loi fondamentale de l'écologie préconise que les espèces entretiennent des relations de compétitions [Ancelet, 2008]. Ainsi, étudier leurs fréquences dans leurs milieux de vies en tenant compte cette inter-dépendance peut s'avérer adéquat pour le suivi de ce travail. Une idée pertinente serait aussi d'étudier leur abondance-dominance.

## 7 Appendices

### 7.1 Rhodes et Jonsén

$$\log Z_i^t = (1 - \rho_0) \log Z_i^{t-1} - 0.5\sigma^2 + \rho \log K_i^{t-1} + \mu_i^t \quad (20)$$

Population d'équilibre au site  $i$  au temps  $t$  :  $K_i^t = K_i^{t-1} \exp(r_i)$

$$\Rightarrow \log K_i^t = \log K_i^{t-1} + r_i \quad (21)$$

$$\log K_i^{t-1} = \log K_i^0 + (t-1)r_i; t > 0 \quad (22)$$

### 7.2 Thorson modifié

$$\log D_i^0 = \frac{\Omega_i^0}{1-\rho} + \Phi 1 + E_i^0, \text{ si } t = 0 \quad (23)$$

$$\log D_i^t = \Omega_i^t + \rho \log D_i^{t-1} + E_i^t, \text{ si } t > 0 \quad (24)$$

avec  $\Omega_i^t = \Omega_{0i} + t\Omega_{1i}$  et  $\Omega_0 \sim MVN(\alpha 1, \Sigma_\Omega)$ ,  $\Omega_1 \sim MVN(r, \Sigma_\Omega)$  et  $E \sim MVN(0, \Sigma_E)$

où  $\Sigma_E = \sigma_E^2 R_h$  et  $\Sigma_\Omega = \sigma_\Omega^2 R_h$ ,  $\alpha 1$  est un vecteur.

$R_h$  est une matrice de corrélation entre les sites.

Posons  $\log D_i^t = \log Z_i^t$ , il en découle par identification :

$$\begin{cases} \Omega_i^t = \rho \log K_i^{t-1} - 0.5\sigma^2 \\ E_i^t = \mu_i^t \Rightarrow \sigma_E = \sigma \\ \rho = 1 - \rho_0 \end{cases}$$

$$\log K_i^{t-1} = \frac{\Omega_i^t + 0.5\sigma^2}{1-\rho_0} = \frac{\Omega_{0i} + t\Omega_{1i} + 0.5\sigma^2}{1-\rho_0} \quad (25)$$

$$\Rightarrow \log K_i^{t-1} = \frac{\Omega_{0i} + \Omega_{1i} + 0.5\sigma^2}{1-\rho_0} + (t-1) \frac{\Omega_{1i}}{1-\rho_0}$$

## Références

- Sophie Ancelet. *Exploring the bayesian hierarchical approach for the statistical modeling of spatial structures : application in population ecology*. Theses, AgroParisTech, July 2008. URL <https://pastel.archives-ouvertes.fr/pastel-00004396>.
- Pieter SA Beck, Clement Atzberger, Kjell Arild Høgda, Bernt Johansen, and Andrew K Skidmore. Improved monitoring of vegetation dynamics at very high latitudes : A new method using modis ndvi. *Remote sensing of Environment*, 100((3)) :321–334, 2006.
- Chave, J. Dynamique spatio-temporelle de la forêt tropicale. *Ann. Phys. Fr.*, 25(6) :1–184, 2000. doi : 10.1051/anphys:200006001. URL <https://doi.org/10.1051/anphys:200006001>.
- Elsa E Cleland, Isabelle Chuine, Annette Menzel, Harold A Mooney, and Mark D Schwartz. Shifting plant phenology in response to global change. *Trends in ecology & evolution*, 22(7) :357–365, 2007.
- Noel Cressie, Catherine A. Calder, James S. Clark, Jay M. Ver Hoef, and Christopher K. Wikle. Accounting for uncertainty in ecological analysis : the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, 19(3) :553–570, 2009. doi : <https://doi.org/10.1890/07-0744.1>. URL <https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/07-0744.1>.
- Collin Edwards and Elizabeth E. Crone. Fitting phenological curves with generalized linear mixed models (glmms). *bioRxiv*, 2020. doi : 10.1101/2020.06.01.127910. URL <https://www.biorxiv.org/content/early/2020/06/04/2020.06.01.127910>.
- Jeremy Isaac Fisher, John F. Mustard, and Matthew A. Vadeboncoeur. Green leaf phenology at landsat resolution : Scaling from the field to the satellite. *Remote Sensing of Environment*, 100(2) :265–279, 2006. ISSN 0034-4257. doi : <https://doi.org/10.1016/j.rse.2005.10.022>. URL <https://www.sciencedirect.com/science/article/pii/S0034425705003597>.
- Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2nd edition, 2004.
- David Higdon. A process-convolution approach to modelling temperatures in the north atlantic ocean. *Environmental and Ecological Statistics*, 5(2) :173–190, 1998. doi : <https://doi.org/10.1023/A:1009666805688>.
- J. Huisman, H. Olf, and L. F. M. Fresco. A hierarchical set of models for species response analysis. *Journal of Vegetation Science*, 4(1) :37–46, 1993. ISSN 11009233, 16541103. URL <http://www.jstor.org/stable/3235732>.
- Brian D. Inouye, Johan Ehrlén, and Nora Underwood. Phenology as a process rather than an event : from individual reaction norms to community metrics. *Ecological Monographs*, 89((2)) :e01352, 2019. doi : <https://doi.org/10.1002/ecm.1352>. URL <https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecm.1352>.

- Ian D. Jonsen, Ransom A. Myers, and Joanna Mills Flemming. Meta-analysis of animal movement using state-space models. *Ecology*, 84(11) :3055–3063, 2003. doi : <https://doi.org/10.1890/02-0670>.
- Kasper Kristensen, Anders Nielsen, Casper W Berg, Hans Skaug, and Brad Bell. Tmb : automatic differentiation and laplace approximation. *arXiv preprint arXiv:1509.00660*, 2015. doi : <https://doi.org/10.18637/jss.v070.i05>.
- Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields : the stochastic partial differential equation approach. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 73(4) :423–498, 2011.
- Peter P Marra, Charles M Francis, Robert S Mulvihill, and Frank R Moore. The influence of climate on the timing and rate of spring bird migration. *Oecologia*, 142(2) :307–315, 2005.
- Cécile Nivet, Marion Gosselin, and H el ene Chevalier. Evaluation des indicateurs nationaux de biodiversit e foresti ere. In *Les indicateurs de biodiversit e foresti ere. Synth ese des r eflexions issues du programme de recherche "Biodiversit e, gestion foresti ere et politiques publiques"*, pages p. 41 – p. 55. GIP Ecofor, 2012. URL <https://hal.archives-ouvertes.fr/hal-00794558>.
- J orn Pagel and Frank M Schurr. Forecasting species ranges by statistical estimation of ecological niches and spatial population dynamics. *Global Ecology and Biogeography*, 21(2) :293–304, 2012.
- M.W. Pedersen, C.W. Berg, U.H. Thygesen, A. Nielsen, and H. Madsen. Estimation methods for nonlinear state-space models in ecology. *Ecological Modelling*, 222(8) :1394–1400, 2011. ISSN 0304-3800. doi : <https://doi.org/10.1016/j.ecolmodel.2011.01.007>. URL <https://www.sciencedirect.com/science/article/pii/S0304380011000299>.
- Jonathan R. Rhodes and Niclas Jonz en. Monitoring temporal trends in spatially structured populations : how should sampling effort be allocated between space and time ? *Ecography*, 34(6) :1040–1048, 2011. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0587.2011.06370.x>.
- M. Brooke Rose and Nicholas N. Nagle. Characterizing forest dynamics with landsat-derived phenology curves. *Remote Sensing*, 13(2), 2021. ISSN 2072-4292. doi : 10.3390/rs13020267. URL <https://www.mdpi.com/2072-4292/13/2/267>.
- DB Roy and TH Sparks. Phenology of british butterflies and climate change. *Global change biology*, 6(4) :407–416, 2000.
- Yannick Saas and Fr ed eric Gosselin. Comparison of regression methods for spatially-autocorrelated count data on regularly-and irregularly-spaced locations. *Ecography*, 37(5) :476–489, 2014.

- Kamel Soudani, Nicolas Delpierre, Daniel Berveiller, Gabriel Hmimina, Jean-Yves Pontailler, Lou Seureau, Ga lle Vincent, and Éric Dufrêne. A survey of proximal methods for monitoring leaf phenology in temperate deciduous forests. *bioRxiv*, 2020. doi : 10.1101/2020.09.24.311191. URL <https://www.biorxiv.org/content/early/2020/09/25/2020.09.24.311191>.
- James T Thorson, Hans J Skaug, Kasper Kristensen, Andrew O Shelton, Eric J Ward, John H Harms, and James A Benante. The importance of spatial models for estimating the strength of density dependence. *Ecology*, 96(5) :1202–1212, 2015. doi : <https://doi.org/10.1890/14-0739.1>.
- P. WHITTLE. ON STATIONARY PROCESSES IN THE PLANE. *Biometrika*, 41(3-4) : 434–449, 12 1954. ISSN 0006-3444. doi : 10.1093/biomet/41.3-4.434. URL <https://doi.org/10.1093/biomet/41.3-4.434>.
- Xiaoyang Zhang, Mark A Friedl, Crystal B Schaaf, Alan H Strahler, John CF Hodges, Feng Gao, Bradley C Reed, and Alfredo Huete. Monitoring vegetation phenology using modis. *Remote sensing of environment*, 84(3) :471–475, 2003.