INRAE

Optimisation des suivis de biodiversité en présence d'erreurs de détection.

Influence de l'effort d'échantillonnage et du niveau de réplication sur la qualité d'estimation de la tendance temporelle de l'occurrence.

Kreshnike MALOKU

Université de Paris

Remerciements

Tout d'abord, j'adresse mes remerciements à Frédéric ARCHAUX, Directeur de l'unité EFNO de m'avoir donné l'opportunité d'effectuer mon stage au sein de son unité. Durant mon stage, il a également été mon tuteur, je tiens ainsi à le remercier, pour ses conseils avisés, sa disponibilité, sa patience, son côté pédagogue ainsi que son soutien. Je le remercie énormément d'avoir cru en mes capacités et de m'être vu confié des tâches diverses et variées.

Je remercie également Frédéric GOSSELIN et Fabien LAROCHE qui ont pris sur leur temps afin de m'aider. Leur accueil et leur bienveillance m'ont fait progresser et monter en compétences dans une ambiance agréable et studieuse.

Je souhaite aussi remercier l'ensemble du personnel de l'unité de recherche, pour leur accueil et leur bonne humeur. Je remercie : Philippe Guillemard, Carl, Sylvie Le Roux, Marion Gosselin, Christophe Bouget, Richard, Sonia Launay, Laura Chevaux, Thierno, etc.

Préambule :

La mise en œuvre des simulations de ce stage a été facilité par l'accès, via une plateforme internet, à une station de calcul basée à INRAE de Nogent-sur-Vernisson, en effet ce cluster a permis de diminuer nettement le temps de calcul. Les simulations finales, réalisée avec 81 scénarios, ont nécessité près de un mois et demi de calcul.

Dans le contexte de compréhension de l'influence de facteurs écologiques sur la qualité de l'estimation de la tendance temporelle de l'abondance ou de l'occurrence, il étudie la modélisation de la tendance temporelle dans le cadre de données in silico, en s'intéressant spécifiquement à différents paramètres tels que l'effort d'échantillonnage (EE), le nombre de réplicats (J), la probabilité de détection (p) ou l'autocorrélation temporelle (acf). Enfin, le présent rapport a été entièrement rédigé sous Word.

Résumé :

Dans le cadre de ma dernière année de Master Mathématiques, Données et Apprentissage à l'Université de Paris, j'ai pu réaliser mon stage, en tant que stagiaire de recherche, de l'unité EFNO (« Écosystèmes Forestiers » de Nogent-sur-Vernisson) du centre INRAE Centre-Val de Loire.

Mon choix s'est porté sur ce stage, car je voulais approfondir mes connaissances en algorithme stochastique. Les algorithmes stochastiques nécessitent la connaissance et l'utilisation de chaînes de Markov permettant la résolution de problèmes d'optimisation et d'estimation complexes.

L'une des raisons pour laquelle j'ai choisi de rejoindre INRAE (Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement), est pour son excellence scientifique en matière de recherche en agriculture, environnement et alimentation. Pour devenir l'un des leaders mondiaux de la recherche, INRAE a mené 166 projets de recherche européens et possède près de 18 centres de recherche localisés dans toute la France. INRAE a pour but de répondre aux enjeux sociétaux comme la transition des agricultures, le changement climatique ou encore la gestion des ressources naturelles.

A cet effet, dans une première partie je présenterai INRAE, ainsi que l'équipe Biodiversité de l'unité EFNO. Puis en deuxième partie j'exposerai l'ensemble des missions et tâches que j'ai pu effectuer. Enfin, dans une troisième partie je m'attèlerai à vous exposer un bilan personnel de ce stage qui a été pour moi sur le plan professionnel très enrichissant.

Sommaire

PARTIE 05

Matériels et Méthodes

PARTIE 06

Résultats

PARTIE 07

Discussion

PARTIE 08

Conclusion

Liste des tableaux :

Liste des figures :

Liste des sigles et abréviations :

INRAE – Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement
EPST – Établissement Public à caractère Scientifique et Technologique)
MESRI – Ministère en charge de la Recherche, l'Enseignement Supérieur, la Recherche et l'Innovation
IRSTEA – Institut national de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture
INRA – Institut National de la Recherche Agronomique
ECODIV – Département d'ECOlogie et bioDIVersité
BIODIVERSITE – L'équipe Interactions gestion forestière et BIODIVERSITE spécifique
FONA – L'équipe Interaction Forêt ONgulés Activités humaines
FORHET – L'équipe FOrêts HETérogènes
GEEDAAF – L'équipe Groupe d'études et d'expertise sur la Diversité Adaptative des Arbres Forestiers
EFNO – L'Unité de recherche Écosystèmes Forestiers de Nogent-sur-Vernisson

i.i.d. – indépendantes et identiquement distribuées
MSE – Mean Squared Error
RMSE – Root Mean Squared Error
UICN – Union Internationale pour la Conservation de la Nature
EE – Effort d'échantillonnage (nombre total de passage dans une année t)
PR – Proportion de sites Répliquée J fois au cours d'une année t
p – Probabilité de detection
J – nombre de réplicats dans une année t pour les sites répliqués, c'est-à-dire échantillonnés plusieurs fois
T – nombre d'années de l'enquête
acf – autocorrelation temporelle

Présentation de l'institut de recherche :

INRAE est un EPST (Établissement public à caractère scientifique et technologique) français sous la tutelle conjointe du ministère en charge de la Recherche, l'Enseignement supérieur, la Recherche et l'Innovation (MESRI) et de celui en charge de l'Agriculture et de l'Alimentation (MAA). INRAE est issue de la fusion de deux instituts de recherche, IRSTEA (Institut national de Recherche en Sciences et Technologies pour l'environnement et l'Agriculture) et l'INRA (Institut National de la recherche agronomique).

Grâce à cette fusion INRAE incarne désormais le rôle de premier organisme de recherche spécialisé sur ses trois domaines scientifiques, agriculture, alimentation et environnement. Avec ses 12 000 collaborateurs, plus d'un milliard d'euros de budget et une richesse en compétences inégalée, INRAE veut représenter une nouvelle façon de lutter pour le développement durable.

INRAE est constitué de 14 départements scientifiques, regroupant 268 unités de recherche, service ou expérimentales dont le département ECOlogie et bioDIVersité (ECODIV) qui a pour mission d'étudier la structure, le fonctionnement et l'évolution des écosystèmes forestiers, prairiaux et aquatiques. Des recommandations de gestion et d'adaptation de ces écosystèmes aux changements globaux, de biodiversité et de bio économie seront proposées.

L'organigramme de la direction générale d'INRAE se présente de la manière suivante :

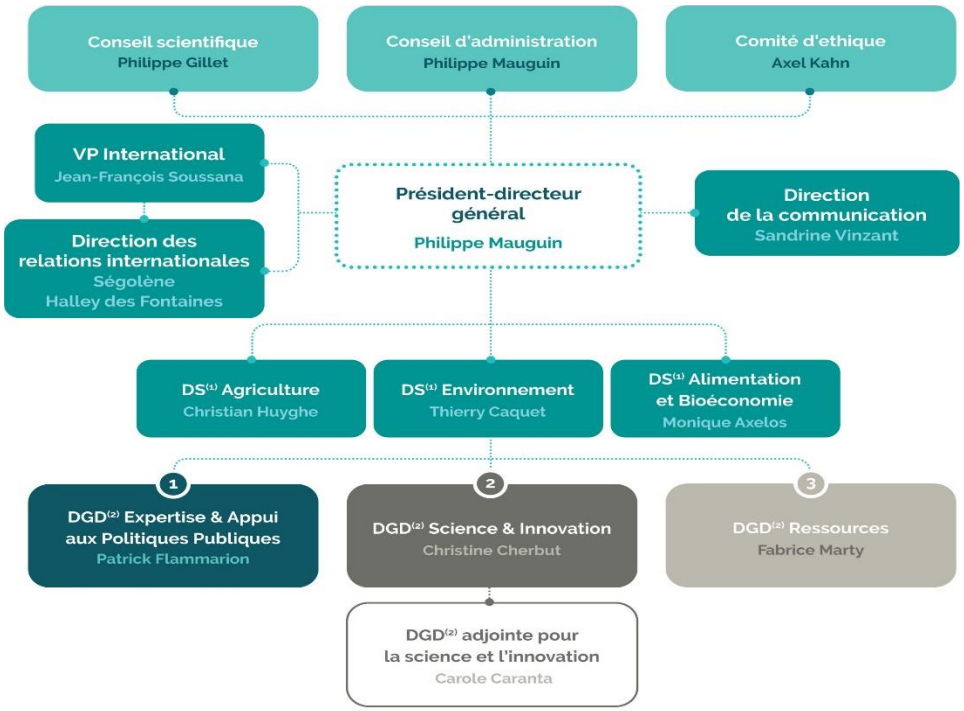


Figure 5 : Organigramme d'INRAE

Le département ECODIV est composé de 37 unités qui fonctionnent grâce à l'apport de 50 millions d'euros de crédits de subvention d'état et 15 millions d'euros de ressources extérieures. Rattachée au département ECODIV, l'unité de recherche « Ecosystèmes Forestiers » de Nogent-sur-Vernisson est constituée de 36 personnels permanents, dont 25 chercheurs et ingénieurs, 10 assistants-ingénieurs et techniciens et 1 secrétaire administrative. A cet effectif s'ajoutent 2 à 4 doctorants, 1 à 2 post-doctorants, 5 à 10 contractuels, 5 à 15 stagiaires.

L'unité est structurée en une équipe d'appui et quatre de recherche :

- BIODIVERSITE - Interactions gestion forestière et BIODIVERSITE spécifique

Maîtriser les pratiques de gestion forestière favorisant la préservation de la biodiversité forestière et animalière. Dans cette optique, l'équipe identifie les pressions qui pèsent sur la biodiversité afin de proposer des recommandations de gestion forestière et d'aménagement du territoire préservant la biodiversité.

- FONA - Interaction Forêt ONgulés Activités humaines

Comprendre les effets de l'abroutissement et éventuellement du piétinement des populations d'ongulés sauvages (cervidés, sangliers) sur la diversité de la flore. Des recommandations de gestion des habitats et des populations animales seront apportés afin d'assurer le bon fonctionnement des écosystèmes.

- FORHET - FOrêts HETérogènes

Etudier l'impact des pratiques sylvicoles sur le renouvellement et la croissance des forêts. L'équipe modélise la croissance des systèmes forestiers et l'évolution des systèmes forestiers afin d'apporter une aide à la bonne gestion des forêts.

- GEEDAAF - Groupe d'études et d'expertise sur la Diversité Adaptative des Arbres Forestiers

Apporter un appui aux décisions du ministère de l'Agriculture : sélection des graines forestières qui permettront le renouvellement et la croissance des forêts en dépit du changement climatique et élaboration de documents de conseils d'utilisation des Matériels Forestiers de Reproduction (MFR).

Présentation de l'unité EFNO :

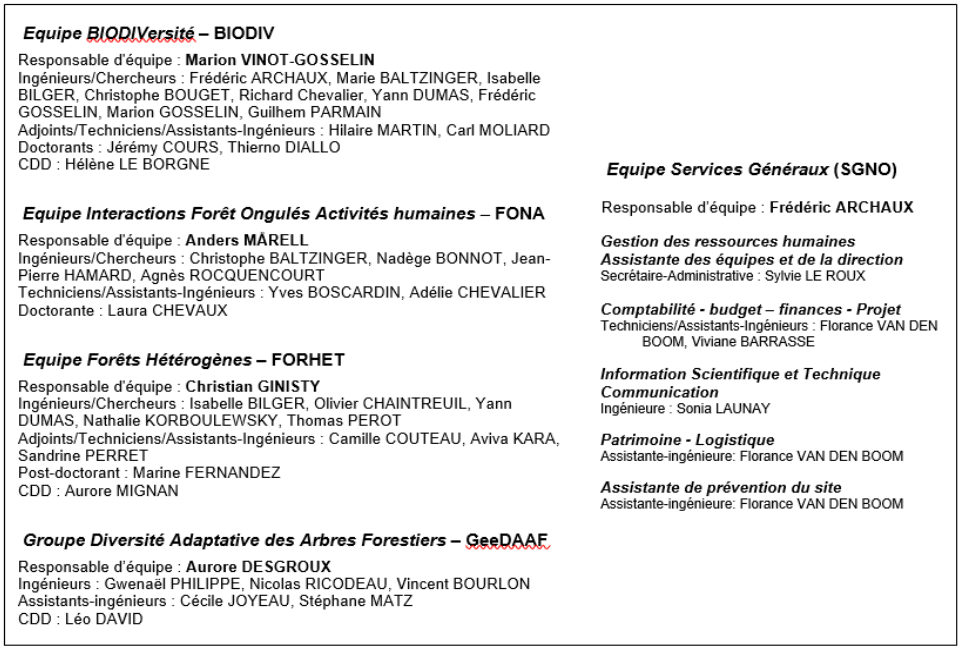


Figure 6 : Présentation de l'unité EFNO

Au sein de l'équipe BIODIVERSITE, mon stage s'inscrit dans le projet PASSIFOR2 (Proposition d'Amélioration du Système de Suivi de la bIodiversité FORestière), financé par le ministère de la Transition Ecologique et Solidaire. Il constitue la phase 2 d'un précédent projet PASSIFOR (2011-2015) soutenu par le ministère de l'agriculture.

Lexique et Abréviation :

Multi-saisons patch occupancy : Le modèle d'occupation multi-saisons (également « modèle d'occupation dynamique ») est utilisé dans la modélisation et la prédiction de la dynamique des espèces sur plusieurs années. Comme pour le modèle d'occupation pour une seule année, on retrouve les paramètres suivants : probabilité d'occupation (Ψ), probabilité de détection (p), probabilité de colonisation (γ) et probabilité de survie (ϕ). Un modèle d'occupation simple est le classique modèle de métapopulation.

Métapopulation : Ensemble de populations d'individus d'une même espèce séparées spatialement ou temporellement et étant interconnectées par la colonisation. Au sein d'une métapopulation, la population d'un site peut s'éteindre (extinction) ; par la suite, le site pourra être colonisé (colonisation) par une population d'un autre site, l'occupation du site sera déterminé par la probabilité d'occupation, ces trois paramètres font référence au processus biologique. La détection d'une population d'un site sera schématisée dans le processus d'observation. Cela a conduit à des modèles qui peuvent être utilisés pour prédire les schémas de déplacement des individus, la dynamique des espèces. Dans ce stage, on s'intéressera à l'évolution de la présence/absence des espèces dans un site.

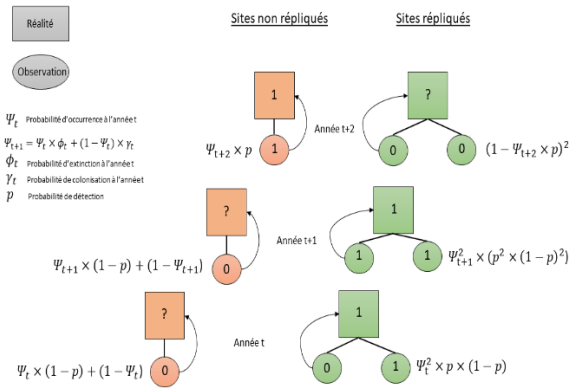


Figure 7 : Représentation d'une métapopulation

Processus biologique : Le processus biologique (ou « vrai statut d'occupation ») est noté $z(i,t)$, $\forall\ i\ =\ 1,2,\dots,\mathrm{J}$ les $\mathrm{z(i,t)}$ sont indépendants et identiquement distribués ($i.i.d$) pour chaque année t. Le vrai statut d'occupation d'un site permet de schématiser l'occupation d'un site par une espèce à travers la probabilité d'occupation (Ψ). Pour cela, on a recours à la probabilité de colonisation (γ) et de survie (ϕ) à l'année t.
Le statut d'occupation initial $z(i,1)$ suit une loi de bernoulli de paramètre Ψ_1 tel que $z(i,1) = Bernoulli(\Psi_1)$, ainsi d'une année t à une année $t+1$,

$$\mathrm{z(i,t)|z(i,t+1)} \sim Bernoulli(\mathrm{z(i,t)} \times \phi_\mathrm{t} + (1 - z(i,t) \times \gamma_t)$$

Probabilité d'occupation ou probabilité d'occurrence : Probabilité qu'une espèce soit présente d'un site i au cours d'une année t.

$$\Psi_t\ =\ \mathbb{P}(\mathrm{z(i,t)} = 1)$$

Un changement d'occupation au cours du temps peut s'expliquer par des processus d'extinction et de colonisation locales d'un site i au cours d'un année t à une année $t+1$.

$$\Psi_{t+1} =\ \Psi_t \times \phi_\mathrm{t} + (1 - \Psi_t) \times \gamma_t$$

Probabilité de survie : Probabilité qu'un site i occupé « survive » d'une année t à une année $t+1$.

$$\phi_\mathrm{t} =\ \mathbb{P}(\mathrm{z(i,t)} = 1|\mathrm{z(i,t+1)} = 1)$$

Probabilité de colonisation : Probabilité qu'un site i non occupé soit « colonisé » d'une année t à une année $t+1$. Cette probabilité de colonisation va dépendre dans l'absolu de la proximité des individus d'une population et/ou de l'importance numérique des populations sources. Dans ce stage, on va poser l'hypothèse d'indépendance à la fois de la probabilité d'occupation et de la distribution spatiale des sites occupés.

$$\gamma_\mathrm{t} =\ \mathbb{P}(\mathrm{z(i,t)} = 0|\mathrm{z(i,t+1)} = 1)$$

Processus d'observation : La probabilité de détecter une espèce durant l'étude notée (p). La probabilité de détection permet de conceptualiser le processus d'observation. Le processus d'observation permet de tenir compte des faux-négatifs, c'est-à-dire lorsque l'espèce, bien que présente mais n'a pas été détectée lors du recensement par l'observateur. Ainsi, il résulte des erreurs de détection par les observateurs des espèces dans les sites, la fréquence d'occupation observée sous-estime la vraie valeur. Pour un site i, au réplicat j de l'année t, le processus d'observation (ou « statut d'occupation observé ») est noté $\mathrm{y(i,j,t)}$, $\forall\ \mathrm{j} = 1,2,\dots,\mathrm{J}$ les $\mathrm{y(i,j,t)}$ sont indépendants et identiquement distribués $(i.i.d)$ pour chaque site i à une année t.

$$\mathrm{y(i,j,t)|z(i,t)} \sim Bernoulli(z(i,t) \times p)$$

Dans ce stage, nous ferons l'hypothèse qu'il n'y a pas de faux positifs (l'espèce est indiquée comme présente alors qu'elle ne l'était pas, par exemple en raison d'une confusion avec une espèce similaire).

Introduction :

Contexte de l'étude

Présentation du concept de « tendance temporelle »

La théorie d'équilibre des écosystèmes a suscité le doute chez bons nombres d'écologistes, DeAngelis et Waterhouse (1987) étaient convaincus que la répartition des espèces n'est pas due au hasard mais dépend de leurs besoins et des caractéristiques physique du milieu.
Les espèces sont présentes dans les milieux où les conditions de survie leur sont plus favorables et leur permettent de se reproduire. On parle alors du concept de « niche écologique », qui représente l'ensemble des conditions biotiques (désigne ce qui est en rapport avec la vie et les êtres vivants) et abiotiques (structure du sol, température, lumière, air) dans lesquelles une espèce est capable de persister et de maintenir la taille de sa population.
La répartition géographique d'une espèce est donc influencée par ses tolérances environnementales ainsi que par ses interactions avec d'autres espèces et de la dynamique spatiale, impliqué par les processus d'émigration/immigration à l'échelle des individus et d'extinction/colonisation locales à l'échelle des populations/patchs.

Un des plus grands défis de la conservation de la biodiversité est la prédiction de la tendance temporelle d'abondance ou d'occurrence des espèces des sites fragmentés et de concevoir des programmes de gestion de l'habitat pour assurer la persistance à long-terme des espèces. Dans la conception de programmes de conservation des espèces, la modélisation est un outil indispensable. Plusieurs résultats dérivés de modèles théoriques ont souligné les difficultés associées à la prédiction de la tendance temporelle.

Link et Sauer (2002) définissent la tendance comme le « taux moyen de changement sur un intervalle de temps spécifique »[1]. La tendance est souvent quantifiée comme la différence entre les valeurs ajustées des premières observations et des dernières observations.

L'application d'un modèle à des données empiriques pour estimer la tendance temporelle de l'abondance ou de l'occurrence d'espèces donne un résultat unique, difficile à généraliser (Saas et Gosselin 2014). Ce problème tient son origine de la non-stationnarité des jeux de données, l'autocorrélation temporelle entre deux relevés dépend du temps écoulé entre ces deux relevés, et ne permet pas de faire des recommandations générales. De plus, l'emploi de données empiriques ne permet pas d'estimer si les résultats sont bons ou mauvais puisque l'on ne connaît pas les vraies valeurs. Ces limites nous ont conduit à l'utilisation de données in silico, dont on pourra comparer les résultats aux propriétés simulées et on pourra explorer des gradients pour les différents paramètres du modèle.

La modification humaine des écosystèmes de la Terre a entraîné une altération de la biodiversité dans de nombreuses régions du monde, dans les écosystèmes marins, terrestres et d'eau douce (Sala et al., 2000). La recherche a documenté l'altération des variables essentielles de la biodiversité telles que la diminution des aires de répartition des espèces ou de l'abondance des groupes d'organismes clés. Ces changements dans la biodiversité ont conduit à ce que l'on appelle souvent une « crise de la biodiversité », avec des avertissements selon lesquels les taux actuels d'extinction sont exceptionnellement élevés, indiquant un phénomène d'extinction de masse à l'échelle mondiale.

Enjeux du projet dans le suivi de la biodiversité

Les politiques et les scientifiques ont répondu à la nécessité de suivre les changements de la biodiversité, les facteurs à l'origine de ces changements et leurs conséquences. Plus particulièrement, ces efforts ont conduit à la formulation d'objectifs pour la biodiversité, visant à stopper le déclin de la biodiversité.
Ainsi, face aux enjeux de préservation de la biodiversité, une meilleure connaissance de la réponse des espèces face aux changements climatiques et des répercussions sur leur distribution est importante. La modélisation se présente ainsi comme un procédé inéluctable pour estimer ces réponses. Elle permet de comprendre les facteurs qui influent sur la répartition des espèces.

L'objectif du projet PASSIFOR2 est d'élaborer à partir d'éléments existants et d'éléments à créer des « maquettes » de suivi de la biodiversité forestière. Il vise une aide aux décisions des politiques publiques dans le domaine du suivi continu de la biodiversité. En dépit d'acquis importants dans ce domaine, les indicateurs actuels de biodiversité forestière sont surtout des indicateurs indirects, ciblés sur les habitats d'espèces et mobilisant principalement des données dendrométriques (mesures des caractéristiques physiques quantifiables des arbres, hauteurs,…) ; il importe d'acquérir des informations qui permettent de (i) mieux cerner directement l'état et la dynamique de la biodiversité forestière et (ii) mieux évaluer le lien entre politiques publiques en forêt, pratiques de gestion et biodiversité.

Ce projet est constitué de cinq tâches et le stage fait partie de la tâche E « mesures, échantillonnage, analyses statistiques ». Cette dernière a pour objectif de fournir les éléments relatifs à l'échantillonnage (où et quand faire les relevés de biodiversité, combien), à la mesure (comment relever la biodiversité) et à l'analyse des données de biodiversité.

[1] William A. Link and John R Sauer (2002), USGS Patuxent Wildlife Research Center, A hierarchical analysis of population change with application to cerulean warblers

Ce projet, mené conjointement entre l'INRAE, l'UMS PATRINAT, le GIP ECOFOR, est piloté par Frédéric Gosselin, ingénieur chercheur en écologie forestière ainsi que Guy Landmann qui est directeur-adjoint du GIP ECOFOR.

Plusieurs études se sont concentrées sur l'estimation de la tendance temporelle de l'occurrence d'une espèce. Nous pouvons citer l'article de Jandt, Von Wehrden et Bruelheide (2011) ou celui de M.Banner, M.Irvine J.Rodhouse et R.Litte (2019) qui se sont penchés sur l'importance relative de différents facteurs tels que le nombre de sites, le nombre de réplicats, la probabilité de détection, l'autocorrélation temporelle sur la tendance temporelle. Relativement peu d'études ont apporté une correction à la détectabilité, dans cette étude la détection imparfaite sera prise en compte dans les analyses.

Problématique :

Les ONG internationales et les scientifiques ont abouti à la conclusion de l'érosion de la Biodiversité causée par la détérioration et fragmentation des sites, les activités humaines et les changements climatiques. Cette érosion se traduit par une augmentation du taux d'extinction des espèces et par la diminution des populations de certaines espèces.
Le taux d'extinction des espèces est tel qu'on parle aujourd'hui de sixième extinction de masse des espèces animales et végétales. Que ce soit sur des îles océaniques, dans des océans ou sur des continents la biodiversité est en danger sur toute la planète.

De nombreuses extinction auraient pu être suivies afin de déclencher des opérations à temps pour renverser la tendance (Yoan Paillet, 2017).
À cet égard, l'application d'un suivi efficace est essentielle pour informer la gestion et prévenir les extinctions d'espèces. Dans le contexte de la conservation des espèces menacées, le suivi est essentiel pour détecter les tendances d'abondance et d'occurrence, mesurer les impacts des facteurs menaçants (exemple : surpêche, déforestation, braconnage, …) et évaluer l'efficacité des mesures de conservation de la biodiversité mises en place.

Notre capacité à prédire la distribution réelle d'une espèce dépend d'un premier problème, la détectabilité. En effet, la présence ou l'absence d'une espèce peut être influencée par deux processus, à savoir le processus écologique et le processus d'observation. L'occurrence fait référence au véritable état de présence ou d'absence d'une espèce et dépend du processus écologique, tandis que la détection fait référence à la capacité de détecter l'espèce compte tenu des méthodes d'observation employées (Dorazio et al., 2006).
Ainsi, la détectabilité d'une espèce peut être définie comme la probabilité de détecter au moins un individu d'une espèce donnée sur un site particulier, étant donné que des individus de cette espèce sont présents dans la zone d'intérêts lors de l'enquête.
La réelle distribution d'une espèce est presque toujours affectée par une détection imparfaite. Par conséquent, la distribution réelle d'une espèce est sous-estimée lorsque la probabilité de détection est inférieure à 1.
Pour continuer à fournir des informations sur le suivi de la biodiversité et parer au problème de détectabilité, plusieurs méthodes sont mises en place pour estimer la détectabilité d'une espèce : augmentation du nombre de sites étudiés, c'est-à-dire augmenter l'effort d'échantillonnage (nombre total de passage dans une année), augmentation du nombre de réplicats.

Un deuxième problème qui va impacter la bonne estimation et prédiction de tendance temporelle de l'abondance ou de l'occurrence d'une espèce est l'autocorrélation temporelle des réplicats.

Les relevés très proches dans le temps seront plus similaires que ceux éloignés dans le temps. Ce manque d'indépendance des relevés conduit à des résidus aléatoires de la tendance temporelle de l'abondance ou de l'occurrence d'une espèce. L'autocorrélation temporelle peut résulter de la longévité des espèces, de taux de colonisation très proches de 0 conduisant à des relevés très corrélés, en effet une espèce recensée sur un site à une année ne va pas forcément coloniser un autre site l'année suivante.

Objectif de l'étude

L'estimation permet de comprendre le comportement d'une espèce à l'aide d'un petit échantillon. Ainsi, face aux enjeux de suivi de biodiversité, une meilleure compréhension des problèmes de détectabilité, des processus d'échantillonnage et des conséquences sur l'estimation de la tendance temporelle de l'abondance ou de l'occurrence des espèces est cruciale pour la préservation de la biodiversité.
Dans un contexte de déclin des espèces, appréhender et comprendre la tendance temporelle de l'abondance ou de l'occurrence d'espèces constitue un moyen efficace de soutenir les mesures de gestion et de conservation de la biodiversité.

Cette étude s'appuie sur des articles existants, l'originalité de ce travail par rapport aux études antérieures est la détectabilité.
Dans ce stage, on simule des données dont on contrôle parfaitement les paramètres écologiques qui les ont générées, qui permettent un test formel des propriétés sous-jacentes et d'évaluer l'importance relative de différents facteurs (nombre de sites, de réplicats, probabilité de détection, autocorrélation temporelle).

La première partie de cette étude établit les outils utilisés dans ce stage, nous présenterons l'Union Internationale pour la Conservation de la Nature, nous poursuivrons avec une représentation hiérarchique du modèle de métapopulation, avec le processus biologique et le processus d'observation. Nous étudierons par la suite l'influence de la probabilité de survie et de colonisation sur la probabilité d'occupation, ainsi que l'influence de l'autocorrélation temporelle sur la proportion réelle de sites occupés. Une modélisation des données sous R sera réalisée et permettra une étude graphique et numérique de notre problème, en conséquence de quoi le calcul d'indicateurs permettra de valider ou non le modèle.

Matériels et Méthodes :

Dans un premier temps, nous présenterons l'Union Internationale pour la Conservation de la Nature, ainsi que l'intérêt d'utiliser et d'adapter un seuil proposé par cette organisation dans ce stage.
Toutes les analyses seront faites avec le logiciel R 3.6.3 et en particulier pour la partie estimation de la tendance temporelle de la proportion de sites occupés nous utiliserons la fonction projected du package unmarked (MacKenzie et al., 2006).

Présentation de l'Union Internationale pour la Conservation de la Nature (UICN) :

Créée le 5 octobre 1948, l'UICN est devenue la source d'informations la plus complète au monde sur le statut de risque d'extinction des espèces animales et végétales.
L'UICN est la première union à l'échelle mondiale d'États, d'agences gouvernementales et d'organisations non gouvernementales (ONG) nationales et internationales, dont le siège est en Suisse.

Sa mission est d'influencer, d'encourager et d'assister les sociétés du monde entier, dans la conservation de la biodiversité, ainsi que de s'assurer que l'utilisation des ressources naturelles est faite de façon équitable et durable.

L'UICN a joué un rôle central dans l'élaboration de conventions internationales dans le but d'évaluer l'extinction des espèces nommé la Liste rouge des espèces menacées.
Dans la Liste rouge de l'UICN, chaque espèce et sous-espèce est classée suivant neuf catégories : En danger critique (CR), En danger (EN), Quasi menacée (NT), Préoccupation mineure (LC), Non évaluée (NE), Éteinte (EX), Éteinte à l'état sauvage (EW), Vulnérable (VU), Données insuffisantes (DD).

Dans cette étude, on utilisera une des catégories de l'UICN, notre choix se porte sur la catégorie des espèces vulnérables avec une baisse de 30% d'occurrence des espèces sur 10 ans.

Construction du modèle de métapopulation :

Les modèles d'occupation des sites ou « patch-occupancy models » sont des outils extrêmement importants et populaires pour comprendre la dynamique et prédire la persistance des populations. Un modèle d'occupation des sites simple est le classique modèle de « métapopulation », ou « population de populations », introduit par Levins en 1969.

La clé du concept de métapopulation résidant dans l'équilibre entre le taux d'extinction et de colonisation des sous-populations, les facteurs qui influencent ces paramètres sont donc déterminants pour maintenir la continuité et la persistance des espèces.
Un site vacant à l'année t peut être colonisé à l'année $t+1$ par une espèce prévenant d'un site voisin avec une probabilité γ_t. Inversement, un site occupé à l'année t peut ne pas survivre avec une probabilité ϕ_t, ce site deviendra alors vacant à l'année $t+1$.

La prédiction de la distribution d'une espèce dépend de la détectabilité, cependant, l'analyse de la probabilité de détection nécessite un nombre de réplicats et un effort d'échantillonnage adéquate Les paramètres contrôlant la tendance temporelle de l'abondance ou de l'occurrence, tels que la taille des sites, la probabilité de détection, la dimension des sous-populations sont, ainsi, des concepts importants pour la continuité et la persistance des espèces.

Royle et Kéry (2007) ont décrit une représentation hiérarchique, ou « espace d'états », du modèle de métapopulation composée de deux composants, un sous-modèle pour le processus biologique, il s'agit de l'occurrence vraie des espèces, et un autre pour les données conditionnelles au processus biologique, c'est-à-dire les observations faites sur le terrain (prise en compte de la détection imparfaite).

Description du processus biologique :

L'espace est divisé en unités appelées sites et sont indexés par $i \in [\![1\,;\mathrm{M}]\!]$. On suppose que chaque site est suivi $j = 1, \dots, J$ fois au cours d'une année $t = 1, \dots, T$, le nombre de réplicats par année peut dépendre des sites, c'est-à-dire $\mathrm{j}_{\max}$ dépend de i. On note z_{it} le vrai statut d'occupation d'un site i au cours de l'année t, ainsi le vrai statut d'occupation a deux valeurs possibles $z = 1$ pour le statut « occupé », l'espèce est présente et $z = 0$ pour le statut « non occupé », l'espèce est absente.
L'objectif est d'estimer la présence/absence pour tout site i et pour tout réplicat j d'une année t,
De plus, les sites sont considérés comme ouverts d'une année à une autre (de t à $t+1$) mais comme fermés d'un réplicat à un autre (de j à $j+1$), la présence/absence de l'espèce au site i pourra donc varier d'une année à une autre mais pas d'un réplicat à un autre. Certains sites peuvent être

observés chaque année 1 fois ou J fois et cela peut varier entre année, on va donc quantifier le pourcentage de site qui sont repliqués J fois, PR qui est noté $PR = \frac{\text{nombre de sites répliqués } J \text{ fois}}{\text{nombre total de sites}}$.

Un paramètre d'intérêt est la probabilité d'occupation ou d'occurrence au cours d'une année t et se note Ψ_t.
Dans une métapopulation, les variations d'occupations des sites au cours du temps sont dues aux processus de survie et de colonisation d'une année t à une année $t+1$. Un site occupé par une espèce au cours d'une année t peut survivre à l'année $t+1$, on notera ϕ_t (probabilité de survie) la probabilité qu'une espèce occupant un site i à l'année t survive à l'année $t+1$. La probabilité qu'une espèce s'éteigne d'une année t à une année $t+1$ est notée $\varepsilon_t = 1 - \phi_t$. Inversement, lorsqu'une population disparaît d'un site i, une population venant d'un autre site peut la coloniser, la probabilité de colonisation d'un site i d'une année t à une année $t+1$ est notée γ_t.

Déterminons maintenant le modèle d'état ou « 1ère couche », il s'agit du vrai statut d'occupation d'un site i à l'année t.

$$z(i,t) \sim Bernoulli(\Psi_t)$$

Le vrai statut d'occupation suit donc une loi de Bernoulli de paramètre Ψ_t. On suppose que le vrai statut d'occupation est identiquement distribué et indépendant ($i.i.d$) pour chaque site i.
Nous aurions donc à l'année $t = 1$,

$$z(i,1) \sim Bernoulli(\Psi_1) \quad \forall\, i = 1, \dots, \mathbb{R}$$

Ainsi le vrai statut d'occupation d'une année t à une année $t+1$ est noté :

$$z(i,t+1)|z(i,t) \sim Bernoulli(z(i,t) \times \phi_t + (1 - z(i,t)) \times \gamma_t)$$

Le vrai statut d'occupation z permet de déduire la proportion de sites réelles occupés psi_z, à l'aide de la formule R suivante :

```
psi_z=colMeans(z)
```

Description du processus d'observation :

Le processus d'observation permet de tenir compte de l'erreur d'observation.
La probabilité de détection est une espèce est notée $p \in [0,1]$. Par conséquent, le statut d'occupation observé d'un site i durant le réplicat j de l'année t sera noté y_{ijt} :

$$y(i,j,t)|z(i,t) \sim Bernouilli(z(i,t) \times\ p)$$

où p est la probabilité de détecter une espèce, $z(i,t)$ le vrai statut d'occupation d'un site i durant l'année t. On suppose que le statut d'occupation observé est identiquement distribué et indépendant ($i.i.d$) pour chaque site i.

Les modèles de métapopulation permettent d'étudier la dynamique des espèces. L'intérêt principale des modèles de métapopulation se situe dans la conceptualisation en terme d'équilibre ou pas entre extinction et colonisation.

Cependant, les modèles de métapopulations simplifient la réalité, ils font des hypothèses sur la probabilité de colonisation (γ) et de survie (ϕ) et supposent que toutes les espèces locales possède le même risque d'extinction, la recolonisation des sites peut avoir lieu même pour des sites éloignés et que la probabilité de dispersion des espèces est identique.

Paramètres du modèle à estimer :

Dans cette étude, on souhaiterais déterminer l'influence de différents facteurs écologiques tels que le nombre de sites, le nombre de réplicats (J), l'autocorrélation temporelle (acf), la probabilité de détection (p) et la proportion de sites répliqués (PR) sur la tendance temporelle de la proportion réelle de sites occupés. La proportion réelle de sites occupés est notée psi_z et est estimée à partir de la moyenne par année des valeurs du vrai statut d'occupation (z).
Dans cette étude à chaque fois que l'on fera mention de la « tendance temporelle », il s'agit en fait de la tendance temporelle de la proportion réelle de sites occupés, soit la tendance des psi_z.

Analyse de l'influence de ϕ et γ sur Ψ :

Une bonne estimation et modélisation des fluctuations de la tendance temporelle de l'abondance ou de l'occurrence (Ψ) passe par une bonne compréhension des paramètres qui influent sur la tendance.
La dynamique des modèles de métapopulation est principalement due aux fluctuations de probabilité de colonisation (γ) et de probabilité de survie (ϕ) d'une année t à une année $t+1$.
La probabilité d'occupation (Ψ) peut être relié à ces deux paramètres par la formule suivante :

$$\Psi_{t+1} = \Psi_t \times \phi_t + (1 - \Psi_t) \times \gamma_t$$

où ϕ_t désigne la probabilité de survie de l'espèce modélisée, et γ_t la probabilité de colonisation de l'espèce modélisée.

Ainsi, pour déterminer la probabilité de survie (ϕ), on doit adopter la formule suivante :

$$\phi_t = \frac{(1 - \Psi_t) \times \gamma_t - \Psi_{t+1}}{\Psi_t}$$

Dans cette formule, la probabilité de survie est dépendante de la probabilité de colonisation et de la probabilité d'occupation, cette double dépendance rend le calcul de la probabilité de survie très approximative. On souhaite alors paramétrer le modèle autrement, c'est-à-dire déterminer la probabilité de survie en terme d'autocorrélation temporelle et de probabilité d'occupation.

Une façon de calculer la probabilité de survie est d'utiliser l'acf à l'année t, la probabilité d'occupation à l'année t ainsi que la probabilité d'occupation à l'année $t+1$.
Avant de développer la formule de la probabilité de survie, définissons ce que l'acf de la proportion réelle de sites occupés psi_z. L'autocorrélation temporelle est définie par une absence d'indépendance entre observations temporelles. Dans le cas d'une autocorrélation temporelle, la répartition dans le temps des relevés est aléatoire ; en fonction du moment de l'échantillonnage il y a un phénomène de ressemblance des relevés. L'autocorrélation temporelle peut résulter de la longévité des espèces, de taux de colonisation très proches de 0 conduisant à des relevés très corrélés. Lorsque les observations proches dans le temps se ressemble, on parle alors d'autocorrélation positive.
Ainsi, pour la formule de l'acf de la proportion de sites occupés on aura :

$$acf_t = Corr(z_{i,t}, z_{i,t+1}) = \frac{Cov(z_{i,t}, z_{i,t+1})}{\sqrt{V(z_{i,t})} \times \sqrt{V(z_{i,t+1})}}$$

Or $V(z_{i,t})$ désigne la variance d'une Bernoulli de paramètre Ψ_t et $V(z_{i,t+1})$ la variance d'une Bernoulli de paramètre Ψ_{t+1}
Donc,

$$acf_t = \frac{\sqrt{\Psi_t \times (1-\Psi_t)} \times (\phi_t - \gamma_t)}{\sqrt{\Psi_{t+1} \times (1-\Psi_{t+1})}}$$

On peut maintenant utiliser cette formule afin d'exprimer la probabilité de survie indépendamment de la probabilité de colonisation.
Pour cela la résolution du système d'équation[2] suivant permet d'atteindre ce but :

$$\begin{cases} \Psi_{t+1} = \Psi_t \times \phi_t + (1-\Psi_t) \times \gamma_t \\ acf_t = \frac{\sqrt{\Psi_t \times (1-\Psi_t)} \times (\phi_t - \gamma_t)}{\sqrt{\Psi_{t+1} \times (1-\Psi_{t+1})}} \end{cases} \quad (1)$$

$$\phi_t = \frac{\sqrt{\Psi_{t+1} \times (1-\Psi_{t+1})} \times (1-\Psi_t) \times acf_t}{\sqrt{\Psi_t \times (1-\Psi_t)}} + \Psi_{t+1}$$

De manière semblable, la probabilité de colonisation dépend de la probabilité de colonisation et de la probabilité d'occupation, le calcul de la probabilité de colonisation sera approximatif. On décide de contourner ce problème et de paramétrer le modèle autrement, en exprimant la probabilité de colonisation en terme d'autocorrélation temporelle et de probabilité d'occupation.

De manière semblable, la probabilité de colonisation est calculée avec l'acf à l'année t et la probabilité d'occupation à l'année t et $t+1$ à partir du système d'équation (1).

$$\gamma_t = \Psi_{t+1} - \frac{\sqrt{\Psi_{t+1} \times (1-\Psi_{t+1})} \times \Psi_t \times acf_t}{\sqrt{\Psi_t \times (1-\Psi_t)}}$$

On remarque que dans le cas particulier où il n'y a pas d'autocorrélation temporelle dans les valeurs de z, $\phi_t = \gamma_t = \Psi_{t+1}$.

Analyse de l'influence de l'acf sur la proportion réelle de sites occupés :

Les jeux de données présentent la caractéristique d'être non-stationnaire temporellement : l'autocorrélation varie selon les relevés. Cette non stationnarité temporelle s'explique par la longévité des espèces, un taux de colonisation très proches de 0 conduisant à des relevés très corrélés. Lorsque les observations proches dans le temps se ressemble, on parle alors d'autocorrélation positive.
L'ambition de cette étude est d'appliquer par la suite notre modèle à des jeux de données, à cet effet il semble important de comprendre le rôle de l'autocorrélation temporelle, qui est un paramètre inconnu.

A l'échelle de la métapopulation, l'occurrence d'espèces au cours d'une année est le résultat des fluctuations du taux de colonisation (γ) et de survie (ϕ).

[2] Pour le calcul détaillé voir annexe n°3

De façon à quantifier le pourcentage de sites réellement occupés au cours d'une enquête, la proportion réelle de sites occupés est notée psi_z. Les relevés auto corrélés peuvent être statistiquement biaisés, les analyser sans tenir compte des biais pourrait conduire à des résultats erronés. L'autocorrélation temporelle de la proportion des sites occupés permet de quantifier ce biais.

$$Corr(z_{i,t}, z_{i,t+1}) = \frac{\sqrt{\Psi_t \times (1-\Psi_t)} \times (\phi_t - \gamma_t)}{\sqrt{\Psi_{t+1} \times (1-\Psi_{t+1})}}$$

On décrit ici le processus utilisé pour générer les données qui vont être ensuite soumises au modèle. Cela implique de considérer comme constante au cours du temps soit la valeur de probabilité de survie (ϕ), soit celle de la probabilité de colonisation (γ), soit de faire varier les deux probabilités. Un problème se pose lorsque les deux probabilités varient au cours du temps, il existe alors une infinité de solution possible. Dans le but de parer à ce problème on va fixer la valeur de l'autocorrélation temporelle qui lie la probabilité de colonisation (γ), la probabilité de survie (ϕ) et la probabilité d'occupation (Ψ). Des simulations ont été utilisées pour déterminer si cette stratégie était toujours valide quelle que soit l'occurrence initiale et la tendance temporelle de cette occurrence. Ces simulations ont montré que c'était le cas uniquement pour des valeurs strictement positives ou nulles d'acf mais pas pour des valeurs négatives.

On prétend ainsi observer que plus les réplicats sont proches dans le temps plus les phénomènes de colonisation (γ_t) et de survie (ϕ_t) vont faiblement influer sur la proportion de sites occupés. Autrement dit, plus les réplicats sont proches dans le temps plus l'acf sera proche de 1, la proportion de sites occupés restera inchangé. On s'attend à ce que l'autocorrélation temporelle de la proportion de sites occupés augmente lorsque la probabilité de survie augmente (ϕ_t).

<u>Modélisation des données et hypothèses du code R :</u>

Nous venons d'aborder le modèle statistique utilisé, nous allons désormais évoquer les hypothèses faites dans le code de simulation R des données utilisées dans la phase de simulation du modèle de métapopulation. En effet, certaines hypothèses peuvent être fortes d'un point de vue écologique.

Comme nous l'avons déjà dit l'estimation de la tendance temporelle de l'abondance ou de l'occurrence d'une espèce repose sur la simulation de données in silico, les données seront donc entièrement simulées. Les principaux paramètres du modèle de métapopulation introduit dès le début du code sont la probabilité de détection (p_t), l'autocorrélation temporelle (acf_t), le nombre d'années de l'étude (T), nombre de réplicats (J), l'effort d'échantillonnage (EE), la proportion de sites répliqués J fois et la probabilité d'occupation à l'année $T = 1$ (psi_0).
Dans le code nous posons $T = 30$, $acf = 0$ et $\text{psi}_0 = 0.2$. Auxquelles on ajoute la proposition de tendance temporelle proposée par l'UICN pour la catégorie des espèces vulnérable, qui est une décroissance de 30% sur 10 ans, $pente = -30\%$ (pente du temps), ces valeurs sont résumés dans le Tableau 1.
On a donc recouru à une enquête d'une durée de 30 années avec une autocorrélation temporelle de 0, une probabilité d'occupation des sites à l'année $t = 1$ de 0.2. Les espèces du jeu de données in silico sont considérés comme vulnérable, tous les ans la proportion d'espèces présentes sur les sites est multiplié par $\frac{\log(1 + (\text{pente}))}{30}$.

Paramètres contrôlés	Stables	T=30 (nombres d'années de l'étude)	acf=0 (autocorrélation temporelle de	$\text{psi}_0 = 0.2$ (probabilité d'occupation à l'année T=1)	$pente = -30\%$ sur 10 ans (pente du temps proposée par l'UICN)

			la proportion de sites occupés)		
	Variables	EE=500, 1000 ou 5000 (effort d'échantillonnage)	J=2, 3 ou 4 (nombres de réplicats)	PR=100%, 50% ou 10% (proportion de sites répliqués J fois)	p=0.2, 0.5 ou 0.8 (probabilité de détection)

Tableau 3 : Paramètres du modèle

Le modèle fait l'hypothèse que les observations répétées sur un site au cours d'une année sont indépendantes, on a donc une indépendance des réplicats.
Une autre hypothèse forte est que le vrai statut d'occupation z_{it} ne change pas d'un réplicat à un autre pendant une année t. Ce qui correspond à une situation écologique dans laquelle une espèce ne peut coloniser ou s'éteindre d'un réplicat à un autre. Les sites sont donc considérés comme « ouverts » à l'extinction et à la colonisation d'une année t à une année $t+1$ mais comme fermés d'un réplicat à un autre.
On peut également évoquer l'hypothèse faite que les fausses absences (faux positives) n'existent pas, c'est-à-dire l'espèce a été répertorié par l'observateur alors qu'elle n'était pas présente.

Une estimation de la tendance de la proportion réelle de sites occupés sera possible à partir de la proportion réelle de sites occupés psi_z, cette estimation sera plus tard comparée à la tendance de la vraie valeur de la proportion de sites occupés introduite avant.

Dans notre étude, nous avons fait le choix de fixer EE, PR et J plutôt que de fixer M. A partir de ces trois paramètres EE, PR et J, le nombre de sites M sera déduit.
Dans chaque scénario on va faire varier l'effort d'échantillonnage notés EE $(500, 1000$ et $5000\ relevés)$, la proportion de sites répliqués J fois notée PR $(100\%, 50\%$ et $10\%)$ et le nombre de réplicats par année pour les sites répliqués noté J $(2, 3$ et $4)$.
Le nombre de sites étudiés dans l'enquête noté M sera calculé avec la formule suivante de l'effort d'échantillonnage qu est fonction du nombre de sites, la proportion de d'entre eux qui sont répliqués et le nombre de réplicats pour ces sites :

$$EE = PR \times M \times J + (1 - PR) \times M$$

Le calcul de M n'aboutit pas systématiquement à une valeur entière pour M. Dans ces cas (par exemple $M = 1666.667$ pour un nombre de réplicats par année $J = 3$ et EE = 5000) on tirera une bernoulli suivant la loi de la différence entre la valeur de M c'est-à-dire $M_{non-tronquée} = 1666.667$ et la valeur tronquée de M càd $M_{tronquée} = 1666$, on aura donc une bernoulli de loi 0.667 (c'est-à-dire dans 2/3 des simulations M vaudra 1667 et dans le tiers restant 1666, le but étant d'avoir la bonne moyenne) pour s'approcher au mieux de la valeur attendue de M (non entière) sur l'ensemble des simulations du scénario.

	EE=500			EE=1000			EE=5000		
	J=2	J=3	J=4	J=2	J=3	J=4	J=2	J=3	J=4
PR=100%	250	166.6667	125	500	333.3333	250	2500	1666.667	1250
PR=50%	332	250	200	666	500	400	3332	2500	2000
PR=10%	454.5455	416.6667	384.6154	909.0909	833.3333	769.2308	4545.455	4166.667	3846.154

Tableau 4 : Nombres de sites (M) en fonction des réplicats (J), de la proportion de sites répliqués J fois (PR) et de l'effort d'échantillonnage (EE)

Il nous faudra dupliquer les différents scénarios par les différentes valeurs suivantes de probabilité de détection $p = 0.20, p = 0.50$ et $p = 0.80$. Ce qui va nous faire 81 scénarios en tout.

On rappelle que pour obtenir ces résultats nous avons analysé 100 simulations pour chacun des 81 scénarios et pour cela nous avons utilisé huit cœurs du processeur du Cluster de Nogent-sur-Vernisson. De plus par contrainte de temps nous avons limité les simulations en intégrant l'année seulement comme une variable continu pour le modèle de mélange (fonction colext) pour les cas où l'effort d'échantillonnage est supérieur à 500 (à l'exception des quelques tests réalisés présentés auparavant pour évaluer l'influence de ce choix sur la qualité des inférences). L'analyse des 81 scénarios sera faite en regroupant les scénarios par effort d'échantillonnages.

Simulation du modèle :

Après avoir introduit nos paramètres initiaux sur R, EE, PR, J, p, nombre de simulations, l'acf et après avoir définit la tendance linéaire de la probabilité d'occupation et calculé la probabilité de colonisation et la probabilité de survie.
On déterminer les deux sous-modèle du modèle de métapopulation, c'est-à-dire le vrai statut d'occupation et le statut d'occupation observé. La proportion réelle de sites occupés sera calculée à partir des valeurs du vrai statut d'occupation.

Comme précisé précédemment les simulations sont faites avec le package Unmarked. Pour organiser les données dans le format requis par colext, nous utilisons la fonction unmarkedMultFrame. Les arguments requis par cette fonction sont les données de présences/absences et le nombre d'années.

```
simUMF <- unmarkedMultFrame( y = yy, yearlySiteCovs = list(annee = annee),
numPrimary=(T+1))
```

Nous adopterons deux types de modèles un modèle avec la variables année en continu et un modèle avec la variable année en facteur.

Le modèle avec la variable année en continu est le suivant :

```
mod_annee_quanti<- colext(psiformula= ~1, gammaformula = ~ annee,
                          epsilonformula = ~ annee, pformula = ~ 1,
                          data = simUMF, method="BFGS")
```

Le modèle avec la variable année en facteur est le suivant :

```
mod_annee_facteur<- colext(psiformula= ~1, gammaformula = ~ annee-1,
                           epsilonformula = ~ annee-1, pformula = ~ 1,
                           data = simUMF, method="BFGS")
```

A partir de ces deux modèles, on procède à l'estimation de la tendance temporelle de la proportion réelle de sites occupés grâce à la fonction projected.

```
psi_annee_quanti=unmarked::projected(mod_annee_quanti)[2,]
```

Méthode d'ajustement du modèle :

Pour explorer l'importance de l'estimation de la proportion réelle de sites occupés, on déterminera le biais et l'erreur quadratique moyenne. Dans la mesure du possible on comparera la moyenne de la tendance temporelle estimée à la moyenne de la tendance temporelle simulée à l'aide d'un test de Student.

Résultats :

Dans cette partie, nous allons exposer les résultats obtenus à l'issue du protocole détaillé précédemment. Le temps de calcul des simulations varie d'un scénario à l'autre et sera indiqué dans la section du scénario correspondant.

Analyse de l'influence de l'acf sur la proportion réelle de sites occupés :

L'ambition de cette étude est d'appliquer par la suite notre modèle à des jeux de données, à cet effet il semble important de comprendre le rôle de l'autocorrélation temporelle, qui est un paramètre inconnu.

Rappelons que l'on souhaitait simuler toutes les valeurs de la probabilité d'occupation (Ψ), la figure disponible en annexe n°1, nous permet de répondre à ce problème.
Nous avons entrepris de travailler uniquement pour des valeurs d'acf positives, en effet, pour des valeurs d'acf négatives, l'espace des valeurs possibles de la probabilité d'occupation (Ψ) se réduit et est centrée autour de la valeurs $\Psi = 0.5$.
On voit sur la figure de l'annexe n°1 que pour des valeurs d'acf possitives, plus la probabilité d'occupation (Ψ) augmente, plus la probabilité de survie est forte (ϕ), on aura ainsi $\phi > 0.5$.
Inversement, toujours pour des valeurs d'acf positives, plus la probabilité d'occupation (Ψ) diminue, plus toutes les valeurs de la probabilité de survie (ϕ) sont atteignables.

Scénarios avec effort d'échantillonnage (EE) égal à 500 :

Pour ce premier lot de scénarios, un scénario avec un effort d'échantillonnage de 500 met 28 heures à tourner avec la variable année en continue et en factorielle. Avec uniquement la variable année en continue un scénario avec un effort d'échantillonnage de 500 met 32 minutes à tourner.
On voit grâce à la Figure 2, avec une proportion de sites répliqués de 100% (PR), une probabilité de détection de 0.20 (p) et un nombre de réplicats de 2 (J) (variable année en continue) que la tendance temporelle simulée qui est de -0.3 est dans l'intervalle de confiance (en pointillé bleu) de la tendance temporelle estimée. De même sur les histogrammes construit avec année en facteur, la tendance simulée est dans l'intervalle de confiance (en pointillé bleu) de la tendance temporelle estimée.
On peut faire la même conclusion avec les autres combinaisons (voir annexe n°5) de proportion de sites répliqués (PR), de probabilité de détection (p) et de nombre de réplicats (J).

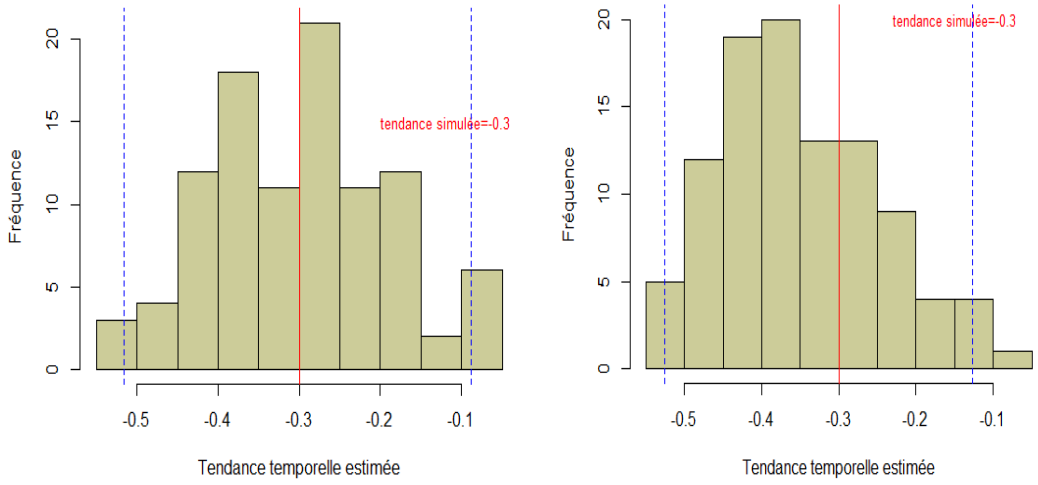


Figure 8 : Histogrammes de la tendance temporelle estimée avec année en continue (à gauche) et année en facteur (à droite), PR=100%, p=0.20 et J=2 et EE=500

Cela apparaît aussi à travers les courbes de densités en annexe n°6, toujours avec les mêmes valeurs de proportion de sites répliqués (PR), de probabilité de détection (p) et de nombre de réplicats (J). La tendance temporelle simulée de -0.3 est comprise dans les courbes de densités.

Pour une proportion de sites répliqués de 100% (PR), une probabilité de détection de 0.20 (p) et un nombre de réplicats de 2 (J), la variance est égale à 2.673857e-05 avec la variable année en continue et à 2.749297e-05 avec la variable année en facteur. L'écart-type est égale à 0.005170935 avec la variable année en continue et à 0.005243374 avec la variable année en facteur.
Cette analyse ne permet pas de dissocier les deux modèles, année en facteur et année en continue. Un moyen d'en déduire le meilleur modèle serait d'effectuer un test de Student, en effet il permettrait de déterminer le lien entre la tendance temporelle estimée avec la variable année en continue et la tendance temporelle simulée et réciproquement avec la variable année en facteur. Si on pose l'hypothèse nulle comme étant la moyenne de la tendance temporelle estimée et égale à la moyenne de la tendance temporelle simulée (-0.3).
Le test de comparaison avec la variable année en continue donne une p-valeur de 0.3648, la p-valeur étant supérieur au seuil de signification, il nous est alors impossible de rejeter l'hypothèse nulle. Avec la variable année en facteur, la p-valeur est de 7.426e-07, on ne rejette pas l'hypothèse nulle, la moyenne de la tendance temporelle estimée et égale à la moyenne de la tendance temporelle simulée avec la variable année en facteur.

Néanmoins, pour étudier l'influence du nombre de réplicats (J), de la probabilité de détection (p), de la proportion de sites répliqués (PR) et de l'effort d'échantillonnage (EE), il nous faut comparer les différents histogramme disponible en annexe n°5 et calculer l'erreur quadratique et le biais.

Scénarios avec effort d'échantillonnage (EE) égal à 1000 :

Pour ce deuxième lot de scénarios, un scénario avec un effort d'échantillonnage de 1000 met 58 heures soit près de 2 jours à tourner avec la variable année en continue et en factorielle. Avec uniquement la variable année en continue un scénario avec un effort d'échantillonnage de 1000 met 55 minutes à tourner.

Par contrainte de temps les simulations avec des efforts d'échantillonnages de 1000 et 5000 sont limitées à année en variable continue, les courbes de densités quant à elles ont étés réalisées avec la variable année en continue et en facteur uniquement pour le scénario suivant $PR = 100\%$, $p = 0.20$ et $\mathrm{J} = 2$.

Les histogrammes de la tendance temporelle estimée (Figure 4), avec une proportion de sites répliqués de 100% (PR), probabilité de détection de 0.20 (p) et un nombre de réplicat de 2 (J), montrent clairement que la tendance simulée qui est de -0.3 est dans l'intervalle de confiance (en pointillé bleu) de la tendance temporelle estimée.
Cela se retrouve avec les autres combinaisons de proportion de sites répliqués, de probabilité de détection et de nombre de réplicats.

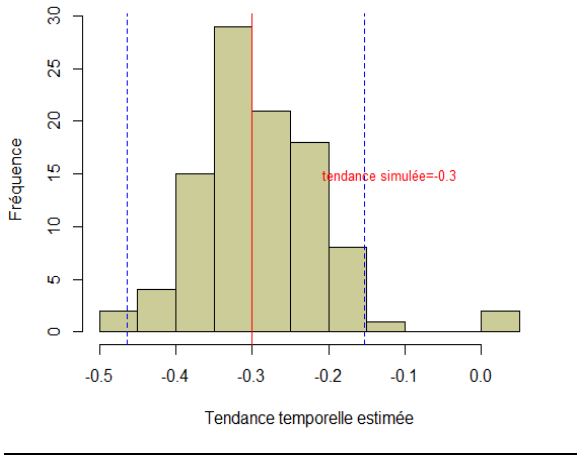


Figure 5 : Histogrammes de la tendance temporelle estimée avec année en continue, PR=100%, p=0.20 et J=2 et EE=1000

Toujours avec les mêmes valeurs de proportion de sites répliqués (PR), de probabilité de détection (p) et de nombre de réplicats (J) en annexe n°6. La tendance temporelle simulée de -0.3 est encore comprise dans les courbes de densités.

Le calcul de la variance donne 1.446512e-05 avec la variable année en continue, l'écart-type est égale à 0.003803304. La variance avec la variable année en continu et un effort d'échantillonnage de 1000 est inférieure à la variance avec un effort d'échantillonnage de 500. On fait la même conclusion avec l'écart-type, l'écart-type avec la variable année en continue et un effort d'échantillonnage de 1000 est inférieure à l'écart-type avec un effort d'échantillonnage de 500.
Le test de comparaison des moyennes année en continue donne une p-valeur de 0.4456, la p-valeur étant supérieur au seuil de signification, il nous est alors impossible de rejeter l'hypothèse nulle.

Scénarios avec effort d'échantillonnage (EE) égal à 5000 :

Pour ce troisième lot de scénarios, un scénario avec un effort d'échantillonnage de 5000 met 298 heures soit 12 jours à tourner avec la variable année en continue et en factorielle. Le temps d'exécution est de 8 heures avec uniquement la variable année en continue.
On s'attend à parvenir à la même conclusion que pour les efforts d'échantillonnages de 500, 1000, la tendance temporelle simulée est comprises dans l'intervalle de confiance de la tendance estimée.
A travers l'histogramme construit avec la variable année en continue, on peut faire cette observation, la tendance simulée est dans l'intervalle de confiance (en pointillé bleu) de la tendance temporelle estimée.

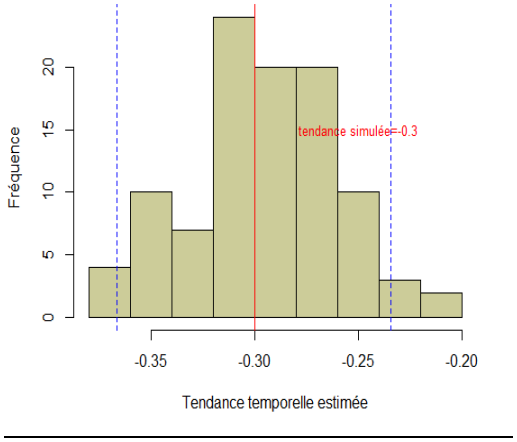


Figure 6 : Histogrammes de la tendance temporelle estimée avec année en continue, PR=100%, p=0.20 et J=2 et EE=5000

On remarque également dans la Figure 7 (annexe n°6) que la tendance temporelle simulée de -0.3 est comprise dans les courbes de densités. De plus, la courbe de densité de la tendance temporelle (Figure 7) avec la variable année en continue comporte plus de bruit que la courbe de densité avec année en variable factorielle.

On fait les mêmes calculs de variance et d'écart-type que précédemment, la variance est égale à 2.727803e-06 avec la variable année en continue, l'écart-type est égale à 0.001651606. La variance avec la variable année en continue et un effort d'échantillonnage de 5000 est inférieure à la variance avec un effort d'échantillonnage de 500 mais elle est également inférieure à la variance avec un effort d'échantillonnage de 1000. Il en est de même pour l'écart-type, l'écart-type avec la variable année en continue et un effort d'échantillonnage de 5000 est inférieure à l'écart-type avec un effort d'échantillonnage de 500 et est également inferieure à l'écart-type avec un effort d'échantillonnage de 1000.
Le test de Student donne une p-valeur de 0.4075, la p-valeur étant supérieur au seuil de signification, il nous est encore impossible de rejeter l'hypothèse nulle.

<u>Validation du modèle / Qualité prédictive :</u>

Un indicateur de la qualité de la prédiction pertinent est l'erreur quadratique moyenne ou RMSE.

			p=0.20				p=0.50				p=0.80			
EE	PR	J	RMSE cont.	Biais cont.	RMSE fact.	Biais fact.	RMSE cont.	Biais cont.	RMSE fact.	Biais fact.	RMSE cont.	Biais cont.	RMSE fact.	Biais fact.
500	100	2	0.00517	0.00047	0.00591	0.00277	0.00328	0.00025	0.00322	0.00037	0.00301	0.00013	0.00302	0.00373
		3	0.00601	0.00021	0.00821	0.00381	0.00377	-0.00036	0.00774	0.00411	0.00357	-0.00019	0.00367	1.98128e-05
		4	0.00609	0.00809	0.00628	0.00142	0.00450	0.00037	0.00715	0.00187	0.00438	0.00029	0.00447	0.0063
1000		2	0.00380	-0.00029			0.00207	-5.57767e-05			0.00207	8.55569e-05		
		3	0.00384	-0.00051			0.00280	-0.00020			0.00264	5.77990e-05		

		4	0.00505	-0.00056			0.00291 2	-1.15778			0.00289	1.79410 e-05		
5000		2	0.00165	-0.00014			0.00107	-8.26437 e-05			0.00096	-2.51414 e-05		
		3	0.00158	2.92979e -05			0.00111	-1.67665 e-05			0.00105	-1.67665 e-05		
		4	0.00105	-5.52388e -05			0.00216	-7.47837 e-07			0.00128	-5.06700 e-05		

Tableau 3 : Biais moyen et erreur quadratique moyenne de la tendance estimée pour une proportion de sites répliqués de 100%.

Scénarios avec effort d'échantillonnage (EE) égal à 500 :

Pour ce premier lot de scénarios constitués d'un effort d'échantillonnage 500, on va s'attacher à comparer le biais moyen et l'erreur quadratique moyenne de la tendance estimée, pour différentes valeurs de proportion de sites répliqués (PR), de probabilité de détection (p) et de nombre de réplicats (J).

On prédit que le RMSE diminue lorsque la probabilité de détection (p) augmente, qu'il diminue également lorsque le nombre de réplicats (J) augmente et qu'à contrario il augmente lorsque la proportion de sites répliqués J fois (PR) diminue. On s'attend également à avoir un biais qui diminue lorsque la probabilité de détection augmente et lorsque le nombre de réplicats augmente.

Contrairement aux résultats attendus, le RMSE augmente lorsque le nombre de réplicats augmente (J) augmente, ainsi pour $J = 2$ le RMSE est égal à 0.00517 et pour $J = 4$ le RMSE est de 0.00609. A l'inverse, pour $\mathrm{p} = 0.2$ le RMSE est de 0.00517 et quand $\mathrm{p} = 0.8$ le RMSE est de 0.00301. L'erreur quadratique moyenne augmente lorsque le nombre de réplicats (J) augmente. En revanche, l'erreur quadratique moyenne diminue lorsque la probabilité de détection augmente, lorsque la proportion de sites répliqués J fois diminue, le RMSE augmente.

L'observation faite précédemment avec le test de Student que l'estimation de la tendance temporelle est meilleure lorsque la variable année est factorielle n'est conforté par le calcul des RMSE. Ainsi, pour $J = 2$ le RMSE est égal à 0.00517 avec année en variable continue est à 0.00591 avec année en variable factorielle, ces différences de RMSE sont tellement petites que l'on ne peut rien conclure.

Etudions maintenant l'influence de ces paramètres à travers les productions de biais. Le biais se rapproche de la tendance temporelle simulée (-0.3) lorsque la probabilité de détection augmente et s'en éloigne lorsque le nombre de réplicats augmente.

Examinons maintenant les résultats du biais, le biais augmente lorsque le nombre de réplicats augmente. A l'inverse, il diminue lorsque la probabilité de détection augmente.

Scénarios avec effort d'échantillonnage (EE) égal à 1000 :

Pour ce deuxième lot de scénarios constitués d'un effort d'échantillonnage de 1000, on s'attend à faire les mêmes constatations qu'avec un effort d'échantillonnage de 500.

On fait les mêmes observations qu'avec un effort d'échantillonnage de 500.

En effet, pour $J = 2$ le RMSE est égal à 0.00380 et pour $J = 4$ le RMSE est de 0.00505. De même, pour $\mathrm{p} = 0.2$ le RMSE est de 0.00380 et quand $\mathrm{p} = 0.8$ le RMSE est de 0.00207. On peut donc dire que lorsque le nombre de réplicats (J) augmente, le RMSE augmente et lorsque la probabilité de détection augmente (p), le RMSE augmente également. On fait la même constatation qu'avec un effort d'échantillonnage de 500 lorsque la proportion de sites répliqués J fois (PR) diminue, le RMSE augmente.

Etudions maintenant l'influence de ces paramètres à travers les productions de biais. Le biais augmente lorsque le nombre de réplicats augmente et il diminue lorsque la probabilité de détection augmente.

Scénarios avec effort d'échantillonnage (EE) égal à 5000 :

Dans ce troisième lot de scénarios, on souhaite également comparer l'erreur quadratique moyenne et le biais, pour différentes valeurs de proportion de sites répliqués (PR), de probabilité de détection (p) et de nombre de réplicats (J). Et pressent que les calculs de RMSE donneront les mêmes résultats qu'avec un effort d'échantillonnage de 500 et de 1000.

Les calculs de RMSE donnent effectivement les mêmes résultats qu'avec un effort d'échantillonnage de 500.

Le calcul des RMSE et des biais nous conforte dans l'idée que plus le nombre de réplicats (J) et la proportion de sites répliqués J fois augmente plus le RMSE diminue, il diminue également lorsque la probabilité de détection augmente. Le biais diminue lorsque le nombre de réplicats augmente et il diminue lorsque la probabilité de détection augmente.

La bonne mise en œuvre des simulations peut être faite grâce à une méthode visuelle utilisant des boxplot.

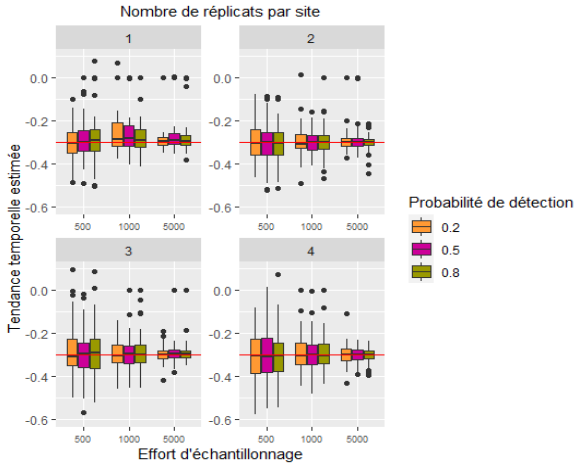


Figure 7 : Boxplots de la tendance temporelle estimée avec la probabilité de détection, l'effort d'échantillonnage, le nombre de réplicats par année et une proportion de sites répliqués égale à 100%

Les boxplots suivants (construit avec $PR = 100\%$) nous permettent de représenter de manière claire l'influence de l'effort d'échantillonnage (EE), le nombre de réplicats (J) et la probabilité de détection (p) sur l'estimation de la tendance temporelle estimée. Ils confirment les observations faites par les histogrammes et les calculs de RMSE, on remarque que plus l'effort d'échantillonnage augmente plus

l'estimation de la tendance temporelle se resserre autour de la tendance temporelle simulée (en rouge).

Discussion :

Nous avons développé des simulations pour tester l'influence des paramètres d'un suivi de biodiversité tels que le nombre de réplicats (J), la probabilité de détection (p), l'effort d'échantillonnage (EE), l'autocorrélation temporelle (acf) et la proportion de sites répliqués J fois (PR) sur la qualité d'estimation de la tendance temporelle.

Parmi ces facteurs, il ressort que l'effort d'échantillonnage est de loin celui qui influence le plus la qualité de l'inférence, quand tous les autres ont une influence relativement négligeable. Ce résultat n'est guère surprenant car il ressort que les estimations de la pente présentent de très faibles biais quelle que soit la probabilité de détection et le niveau de réplication : augmenter l'effort d'échantillonnage permet ainsi d'augmenter le nombre de site et donc d'améliorer la précision de l'estimateur.
Contrairement à nos attentes, la probabilité de détection n'a qu'un effet marginal sur la qualité d'inférence. Pour un nombre de réplicats de 1 et un effort d'échantillonnage de 1000, les boxplots sont moins importants que pour un effort d'échantillonnage de 500 et un nombre de réplicats de 2. Nous pensions qu'à effort d'échantillonnage constant, il y aurait un compromis entre le nombre de sites et l'effort de réplication. Cela se vérifie et il semble que dans tous les cas, il soit préférable d'augmenter le nombre de sites que le niveau de réplication.
Kéry et Royle (2016) ont montré l'importance relative du nombre de sites au nombre de réplicats, le nombre de sites d'échantillonnage a davantage influence sur l'estimation de la tendance temporelle que le nombre de réplicats. Il serait ainsi plus avantageux d'augmenter le nombre de sites d'échantillonnage que d'augmenter le nombre de réplicats.

Nous avons démontré l'influence de l'acf sur la proportion réelle de sites occupés.
Pour des valeurs d'autocorrélations positives, une faible probabilité d'occupation (Ψ) permet d'atteindre toutes les valeurs de probabilité de survie (ϕ).
Par conséquent, si la probabilité d'occupation d'un site est faible, la probabilité que l'espèce survive sur le site peut être faible avec une faible probabilité de colonisation de l'espèce, ou la probabilité de survie de l'espèce peut être forte avec une forte probabilité de colonisation de l'espèce. Toutes les valeurs de probabilité de survie seront ainsi possibles.
A contrario, lorsque la probabilité d'occupation est forte, la probabilité de survie est faible et la probabilité de colonisation est forte ou la probabilité de survie est forte et la probabilité de colonisation est faible, avec autocorrélation positives l'hypothèse d'une forte probabilité de survie sera privilégié.

Notre étude suggère également que comme prévu, l'estimation de la tendance temporelle avec le modèle dans lequel la variable année est continue et le modèle dans lequel la variable année est factorielle sont assez similaires. Les différences entre les deux modèles n'étaient évidentes que lors des simulations, le modèle avec la variable année en continue nécessite moins de temps à être généré que le modèle avec la variable année en facteur.
Dans le cas de l'utilisation d'un modèle avec la variable année en facteur, le RMSE était plus faible et les courbes de densités comportait moins de bruits. C'est probablement pourquoi ces modèles ont obtenu de meilleurs résultats d'estimation de la tendance temporelle. Dans ce cas, nous serions en mesure de tirer le meilleur parti du modèle avec la variable année en facteur pour augmenter la précision des estimations de la tendance temporelle.

Le RMSE ont permis de tester l'influence de facteurs tels que le nombre de réplicats (J), la probabilité de détection (p), l'effort d'échantillonnage (EE), autocorrélation temporelle (acf) et la proportion de sites répliqués J fois (PR) sur la tendance temporelle.
Les erreurs quadratiques moyennes ont montrer que lorsque le nombre de réplicats (J) augmente, le RMSE diminue, il en est de même pour la probabilité de détection (p) et pour l'effort d'échantillonnage (EE). D'un point de vue écologique, l'augmentation de l'effort d'échantillonnage dans un contexte de données réelles coûterait cher.

Toutefois les calculs de pour un effort d'échantillonnage de 1000, les calculs de RMSE n'étaient pas en adéquation avec les calculs de RMSE pour un effort d'échantillonnage de 500 ou 5000, bien au contraire lorsque le nombre de réplicats (J) augmente, le RMSE augmente.
Ce problème pourrait venir de l'estimation de la tendance temporelle qui possède une erreur-type inutilisable et dans le cas d'application de la méthode à un jeu de données réel, il faudrait ainsi réaliser une approche par bootstrap sur la base des paramètres estimés pour reconstituer un intervalle de confiance de la tendance temporelle.

Aussi les limites de notre étude sont que nous avons travaillé avec un lot de 81 scénarios et que nous avons posé l'hypothèse que p est constant en fonction de la probabilité d'occupation, qui ne couvre pas toutes les situations écologiques. En effet, nous avons posé trois niveaux de p pour couvrir la très grande majorité des cas d'étude, certaines enquêtes opportunistes ne permettent pas d'atteindre cette probabilité de détection. De plus, l'hypothèse faite d'avoir une probabilité de détection indépendante de l'abondance peut être discuté, effectivement, on peut supposer que plus une espèce est abondante plus elle sera visible par l'observateur. C'est-à-dire, une probabilité de détection constante quelle que soit l'année (pas de variabilité interannuelle, pas de tendance sur la probabilité de détection, pas de lien entre la probabilité d'occupation et la probabilité de détection) quand il est vraisemblable qu'il y ait des corrélations et que probabilité de détection varie certainement entre années dans la réalité écologique.
Par ailleurs, la question de la robustesse des comparaisons des RMSE se pose, en effet, faute de temps nous nous sommes limités à 100 simulations par scénario.

De plus, faute de temps, nous n'avons pas pu réaliser les histogrammes, les courbes de densités et les calculs de RMSE avec des efforts d'échantillonnages $EE = 1000$ ou $EE = 5000$, avec la variable année en factorielle. Or il semble que la prise en compte de l'effet année en variable continue pose parfois des problèmes de convergence des algorithmes d'optimisation, contrairement au cas où l'année est en facteur. Par ailleurs, cette vision correspond plus à la réalité écologique. Néanmoins, il ne semble pas qu'il y ait une grande différence de qualité d'estimation globale dans les deux cas, ne remettant ainsi pas en cause l'ensemble des résultats obtenu avec l'année intégrée comme variable continue.
Ceci aurait permis de mieux rentrer dans le cadre du projet PASSIFOR2 évoqué en début de ce rapport. On aurait alors pu mieux apercevoir les apports positifs et négatifs de l'estimation de la tendance temporelle avec un modèle où la variable année est factorielle. Il s'agit d'une piste à explorer par la suite, sachant que nos premiers essais sur la question semblaient indiquer que les simulations avec année en facteur donnaient une meilleure estimation de la tendance temporelle pour le premier lot de scénarios.

Conclusion :

Le rapport de stage touchant à sa fin, il est opportun de faire le bilan du travail réalisé d'une part mais aussi des apports de ce stage d'autre part.

Tout d'abord le but global était proposer une évaluation de l'importance relative de différents facteurs tels que le nombre de sites, le nombre de réplicats, la probabilité de détection et l'autocorrélation temporelle sur la tendance temporelle. Nous avons proposé des méthodes visuelles et quantitatives afin de démontrer les répercussions de l'augmentation ou de la diminution de ces facteurs écologiques sur l'estimation de la tendance temporelle.
On a voulu ensuite vérifier si les observations précédemment faites sur l'influence des paramètres écologiques sur la tendance temporelle étaient correctes en calculant le RMSE et la proportion de fois que l'intervalle de confiance de la tendance temporelle estimée comprend la tendance temporelle simulée.

Au cours de cette étude, on n'a pas trouvé les résultats escomptés, l'influence de ces facteurs écologiques sur la tendance temporelle n'a pas été démontré explicitement.

Ensuite, abordons les apports de ce stage d'un point de vue plus personnel. Tout d'abord ce stage m'a initié à la recherche scientifique et plus particulièrement à la recherche en statistiques appliquées à l'écologie. Lire des articles, les recouper entre eux, se poser des questions, travailler en autonomie voire en autodidaxie, faire une bibliographie sont autant de choses que j'ai apprises et qui rythment le quotidien d'un chercheur. En outre j'ai pu utiliser des clusters de calcul, en particulier celui de Nogent-sur-Vernisson. Bien entendu j'ai découvert et appris plein de nouvelles choses sur l'utilisation de Cluster, la modélisation de données et même les modèles statistiques. Par ailleurs plus je lisais et découvrais des choses plus je mesurais l'étendue des choses que j'ignorais.

Annexe n°1 : Tableau global avec le nom des pramètres et l'ensemble des scénarios

Pour appréhender l'influence du nombre de sites, le nombre de réplicats (J), l'autocorrélation temporelle (acf), la proportion de sites répliquée J fois (PR), la probabilité de détection (p) et le nombre d'années de l'enquête sur la tendance temporelle, nous avons considéré 3 valeurs pour chacun de ces facteurs, soit 729 scénarios s'ils étaient tous croisés.

Variables	EE=500, 1000 ou 5000	J=2, 3 ou 4	T=10,30 ou 50	acf=0, 0.3 ou 0.9	PR=100%, 50% ou 10%	p=0.2, 0.5 ou 0.8

Tableau 5 : Paramètres du modèle

On aura donc $3\ valeurs_{EE} \times 3\ valeurs_{J} \times 3\ valeurs_{T} \times 3 valeurs_{acf} \times 3\ valeurs_{PR} \times 3\ valeurs_{p}$, ce qui nous fait 729 scénarios.

Sachant qu'un scénario avec un effort d'échantillonnage de 500, $J = 2$, $p = 0.2$, $PR = 100\%$, $acf = 0$ et un nombre d'années de l'enquête $T = 30$ mets 28 heures et plus le nombre d'années de l'enquête augmente plus le temps de simulation augmente. Ainsi, faute de temps tous les scénarios n'ont pas pu être réalisés.

Voici la liste exhaustive des 81 scénarios réalisés au cours de ce stage : (acf et T en variables continues)

1) 100% des sites répliqués
   a) $EE = 500$
   - $J = 2$ et $M = 250$
   - $J = 3$ et $M = 166.6667$
   - $J = 4$ et $M = 125$
   b) $EE = 1000$
   - $J = 2$ et $M = 500$
   - $J = 3$ et $M = 333.3333$
   - $J = 4$ et $M = 250$
   c) $EE = 5000$
   - $J = 2$ et $M = 2500$
   - $J = 3$ et $M = 1666.667$
   - $J = 4$ et $M = 1250$
2) 50% des sites répliqués
   a) $EE = 500$
   - $J = 2$ et $M = 332$
   - $J = 3$ et $M = 250$
   - $J = 4$ et $M = 200$
   b) $EE = 1000$
   - $J = 2$ et $M = 666$
   - $J = 3$ et $M = 500$
   - $J = 4$ et $M = 400$
   c) $EE = 5000$
   - $J = 2$ et $M = 3332$
   - $J = 3$ et $M = 2500$
   - $J = 4$ et $M = 2000$
3) 10% des sites répliqués
   a) $EE = 500$
   - $J = 2$ et $M = 454.5455$
   - $J = 3$ et $M = 416.6667$
   - $J = 4$ et $M = 384.6154$
   b) $EE = 1000$

- $J = 2$ et $M = 909.0909$
- $J = 3$ et $M = 833.3333$
- $J = 4$ et $M = 769.2308$

c) $EE = 5000$

- $J = 2$ et $M = 4545.455$
- $J = 3$ et $M = 4166.667$
- $J = 4$ et $M = 3846.154$

Il nous faudra dupliquer les différents scénarios par les différentes valeurs suivantes de probabilité de détection $p = 0.20, p = 0.50$ et $p = 0.80$, mais en gardant des valeurs constantes pour acf et T ($acf = 0$ et $T = 30$), ce qui va nous faire 81 scénarii. Ce qui va nous faire 81 scénarios en tout.

Si maintenant le nombre d'années est définit comme une variable factorielle, on aura 243 scénarios possibles, les 81 scénarios cités précédemment dupliquer par le nombre d'années de l'enquête $T = 10, T = 30$ et $T = 50$.

<u>Annexe n°2 : Analyse de l'influence de l'acf sur la proportion réelle de sites occupés :</u>

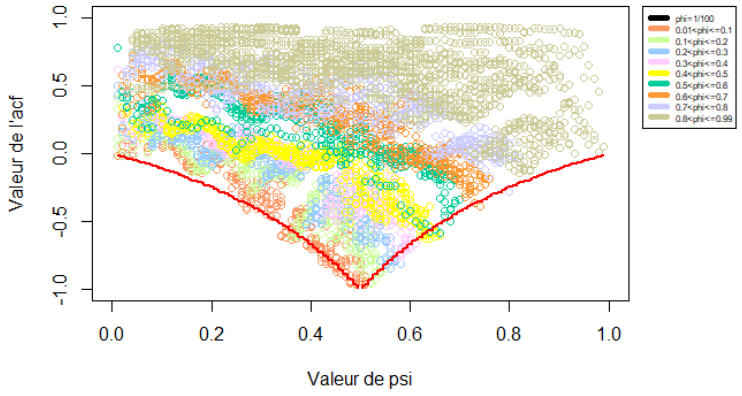


Figure 8 : Influence de l'acf sur la proportion réelle de sites occupés (psi) en fonction de la probabilité de survie (phi)

<u>Annexe n°3 : Calcul de la probabilité de colonisation (γ) et de la probabilité de survie (ϕ) indépendamment l'une de l'autre</u>

Démarrons du système suivant :

$$\begin{cases} \Psi_{t+1} = \Psi_t \times \phi_t + (1 - \Psi_t) \times \gamma_t & (1) \\ acf_t = \frac{\sqrt{\Psi_t \times (1-\Psi_t)} \times (\phi_t - \gamma_t)}{\sqrt{\Psi_{t+1} \times (1-\Psi_{t+1})}} & (2) \end{cases}$$

On a ainsi deux équations à deux inconnus (ϕ_t et γ_t). Pour l'équation (2), si on passe ϕ_t et γ_t d'un côté de l'égalité on obtient,

$$\begin{cases} \Psi_{t+1} = \Psi_t \times \phi_t + (1 - \Psi_t) \times \gamma_t & (1) \\ \frac{acf_t}{(\phi_t - \gamma_t)} = \frac{\sqrt{\Psi_t \times (1-\Psi_t)}}{\sqrt{\Psi_{t+1} \times (1-\Psi_{t+1})}} & (2) \end{cases}$$

On peut maintenant inverser l'égalité (2) :

$$\begin{cases} \Psi_{t+1} = \Psi_t \times \phi_t + (1 - \Psi_t) \times \gamma_t & (1) \\ \frac{(\phi_t - \gamma_t)}{acf_t} = \frac{\sqrt{\Psi_{t+1} \times (1 - \Psi_{t+1})}}{\sqrt{\Psi_t \times (1 - \Psi_t)}} & (2) \end{cases}$$

Ce qui nous donne,

$$\begin{cases} \Psi_{t+1} = \Psi_t \times \phi_t + (1 - \Psi_t) \times \gamma_t & (1) \\ \phi_t - \gamma_t = \frac{\sqrt{\Psi_{t+1} \times (1 - \Psi_{t+1})}}{\sqrt{\Psi_t \times (1 - \Psi_t)}} \times acf_t & (2) \end{cases}$$

Dans l'équation (1), on peut mettre ϕ_t à gauche de l'égalité, ce qui va nous donner,

$$\begin{cases} \phi_t = \frac{\Psi_{t+1} - (1 - \Psi_t) \times \gamma_t}{\Psi_t} & (1) \\ \phi_t - \gamma_t = \frac{\sqrt{\Psi_{t+1} \times (1 - \Psi_{t+1})}}{\sqrt{\Psi_t \times (1 - \Psi_t)}} \times acf_t & (2) \end{cases}$$

On peut maintenant dans l'équation (2) mettre ϕ_t à droite de l'égalité et le remplacer par le ϕ_t obtenu dans l'équation (1). L'équation (2) :

$$\phi_t - \gamma_t = \frac{\sqrt{\Psi_{t+1} \times (1 - \Psi_{t+1})}}{\sqrt{\Psi_t \times (1 - \Psi_t)}} \times acf_t$$

$$-\gamma_t = \frac{\sqrt{\Psi_{t+1} \times (1 - \Psi_{t+1})}}{\sqrt{\Psi_t \times (1 - \Psi_t)}} \times acf_t - \phi_t$$

$$\gamma_t = \phi_t - \frac{\sqrt{\Psi_{t+1} \times (1 - \Psi_{t+1})}}{\sqrt{\Psi_t \times (1 - \Psi_t)}} \times acf_t$$

On peut remplacer ce résultat dans l'équation (2) :

$$\begin{cases} \phi_t = \frac{\Psi_{t+1} - (1 - \Psi_t) \times \gamma_t}{\Psi_t} & (1) \\ \gamma_t = \phi_t - \frac{\sqrt{\Psi_{t+1} \times (1 - \Psi_{t+1})}}{\sqrt{\Psi_t \times (1 - \Psi_t)}} \times acf_t & (2) \end{cases}$$

Si on remplace ϕ_t dans l'équation (2) par la formule obtenu dans l'équation (1), on obtient :

$$\begin{cases} \phi_t = \frac{\Psi_{t+1} - (1 - \Psi_t) \times \gamma_t}{\Psi_t} & (1) \\ \gamma_t = \frac{\Psi_{t+1} - (1 - \Psi_t) \times \gamma_t}{\Psi_t} - \frac{\sqrt{\Psi_{t+1} \times (1 - \Psi_{t+1})} \times acf_t}{\sqrt{\Psi_t \times (1 - \Psi_t)}} & (2) \end{cases}$$

$$\begin{cases} \phi_t = \frac{\Psi_{t+1} - (1 - \Psi_t) \times \gamma_t}{\Psi_t} & (1) \\ \gamma_t = \frac{(\Psi_{t+1} - (1 - \Psi_t) \times \gamma_t) \times \sqrt{\Psi_t \times (1 - \Psi_t)} - \sqrt{\Psi_{t+1} \times (1 - \Psi_{t+1})} \times \Psi_t \times acf_t}{\Psi_t \times \sqrt{\Psi_t \times (1 - \Psi_t)}} & (2) \end{cases}$$

Si on développe le produit, on obtient :

$$\begin{cases} \phi_t = \dfrac{\Psi_{t+1} - (1-\Psi_t) \times \gamma_t}{\Psi_t} & (1) \\ \dfrac{\gamma_t \times \sqrt{\Psi_t \times (1-\Psi_t)}}{\phi_t \times \sqrt{\Psi_t \times (1-\Psi_t)}} = \dfrac{\Psi_{t+1} \times \sqrt{\Psi_t \times (1-\Psi_t)} - \sqrt{\Psi_{t+1} \times (1-\Psi_{t+1})} \times \Psi_t \times acf_t}{\Psi_t \times \sqrt{\Psi_t \times (1-\Psi_t)}} & (2) \end{cases}$$

$$\begin{cases} \phi_t = \dfrac{\sqrt{\Psi_{t+1} \times (1-\Psi_{t+1})} \times (1-\Psi_t) \times acf_t}{\sqrt{\Psi_t \times (1-\Psi_t)}} + \Psi_{t+1} & (1) \\ \gamma_t = \Psi_{t+1} - \dfrac{\sqrt{\Psi_{t+1} \times (1-\Psi_{t+1})} \times \Psi_t \times acf_t}{\sqrt{\Psi_t \times (1-\Psi_t)}} & (2) \end{cases}$$

1) Afin d'effectuer le contour analytique des nuages de points sous R, il s'agit du cas particulier où Ψ est constant, on peut donc utiliser $\Psi_t = \Psi_{t+1}$ pour terminer le calcul.

$$\phi = acf + \Psi \times (1 - acf)$$

2) Dans le cas non stationnaire, on aura :

$$\phi_t = \frac{\sqrt{\Psi_{t+1} \times (1-\Psi_{t+1})} \times (1-\Psi_t) \times acf_t}{\sqrt{\Psi_t \times (1-\Psi_t)}} + \Psi_{t+1}$$

1) De manière semblable, la probabilité de colonisation est calculée avec l'acf à l'année t et la probabilité d'occupation à l'année t et $t+1$ à partir du système d'équation (1). Dans le cas stationnaire :

$$\gamma = \Psi \times (1 - acf)$$

2) Dans le cas non stationnaire, on aura :

$$\gamma_t = \Psi_{t+1} - \frac{\sqrt{\Psi_{t+1} \times (1-\Psi_{t+1})} \times \Psi_t \times acf_t}{\sqrt{\Psi_t \times (1-\Psi_t)}}$$

Annexe n°4 : Calcul de l'autocorrélation temporelle en fonction de la probabilité d'occupation (Ψ), de la probabilité de colonisation (γ) et de la probabilité de survie (ϕ)

$$acf_t = Corr(z_{i,t}, z_{i,t+1}) = \frac{Cov(z_{i,t}, z_{i,t+1})}{\sqrt{V(z_{i,t})} \times \sqrt{V(z_{i,t+1})}}$$

On sait que $z_{i,t+1} = z_{i,t} \times C_{i,t} + (1 - z_{i,t}) \times S_{i,t}$ où $C_{i,t}$ est une variable aléatoire de Bernoulli de paramètre $\gamma_t$et $S_{i,t}$ une variable aléatoire de Bernoulli de paramètre ϕ_t. Les C et les S sont supposées indépendantes entre elles et avec le reste des variable aléatoire du système.

$$Cov(z_{i,t}, z_{i,t+1}) = Cov(z_{i,t}, z_{i,t} \times C_{i,t} + (1 - z_{i,t}) \times S_{i,t}$$

Par bilinéarité de la covariance, on obtient :

$$Cov(z_{i,t}, z_{i,t+1}) = Cov(C_{i,t}, z_{i,t}) - Cov(C_{i,t} \times z_{i,t}, z_{i,t}) + Cov(S_{i,t} \times z_{i,t}, z_{i,t})$$

Or, on sait que $C_{i,t}$ et $S_{i,t}$ sont indépendantes de $z_{i,t}$, donc $Cov(C_{i,t}, z_{i,t}) = 0$

$$Cov(\mathrm{S_{i,t}} \times z_{i,t}, z_{i,t}) = \mathbb{E}[\mathrm{S_{i,t}}]\left(\mathbb{E}[z_{i,t}^2] - \mathbb{E}[z_{i,t}]^2\right)$$
$$= \mathbb{E}[\mathrm{S_{i,t}}] \times \mathbb{V}(z_{i,t})$$
$$= \phi_t \times \Psi_t \times (1 - \Psi_t)$$

$$Cov(\mathrm{C_{i,t}} \times z_{i,t}, z_{i,t}) = -\gamma_t \times \Psi_t \times (1 - \Psi_t)$$

Donc,

$$Cov(z_{i,t}, z_{i,t+1}) = (\phi_t - \gamma_t) \times \Psi_t \times (1 - \Psi_t)$$

Calculons maintenant

$$\mathbb{V}(z_{i,t+1}) = \mathbb{E}[z_{i,t+1}^2] - \mathbb{E}[z_{i,t+1}]^2$$
$$= (\gamma_t - \gamma_t \times \Psi_t + \Psi_t \times \phi_t) - (\gamma_t \times (1 - \Psi_t) + \phi_t \times \Psi_t)^2$$
$$= \Psi_{t+1} - (\Psi_{t+1})^2$$

1) Afin d'effectuer le contour analytique des nuages de points sous $\mathbb{R}$, il s'agit du cas particulier où Ψ est constant, on peut donc utiliser $\Psi_t = \Psi_{t+1}$ pour terminer le calcul.
On va donc avoir,
$$Corr(z, z) = \frac{\sqrt{\Psi \times (1 - \Psi)} \times (\phi - \gamma)}{\sqrt{\Psi \times (1 - \Psi)}}$$
$$= \phi - \gamma$$

2) Dans le cas non stationnaire, on va avoir Ψ_t différent de Ψ_{t+1}. On va donc avoir,

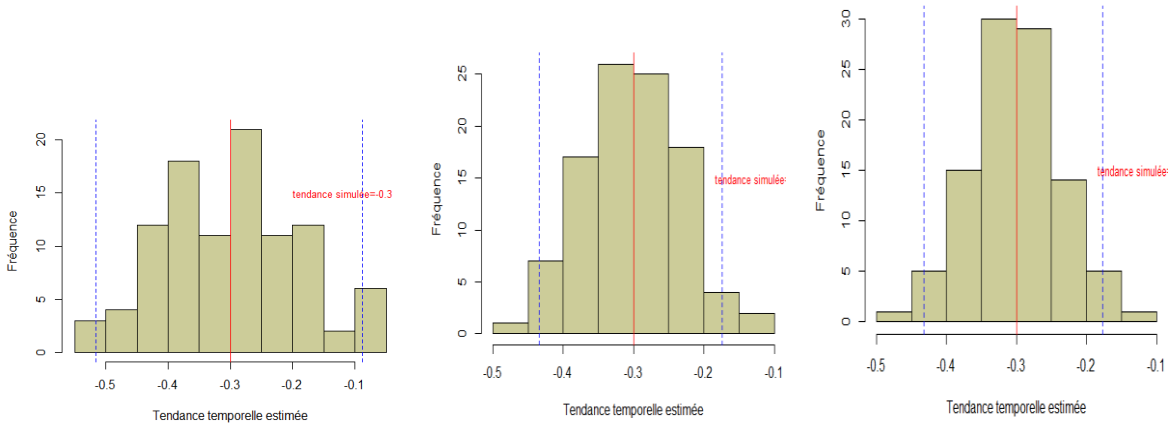
$$acf_t = Corr(z_{i,t}, z_{i,t+1}) = \frac{Cov(z_{i,t}, z_{i,t+1})}{\sqrt{V(z_{i,t})} \times \sqrt{V(z_{i,t+1})}}$$
$$= \frac{\sqrt{\Psi_t \times (1 - \Psi_t)} \times (\phi_t - \gamma_t)}{\sqrt{\Psi_{t+1} \times (1 - \Psi_{t+1})}}$$

Annexe n°5 : Histogrammes pour différentes valeurs de probabilité de détection, de nombre de réplicats, proportion de sites répliqués et d'effort d'échantillonnage

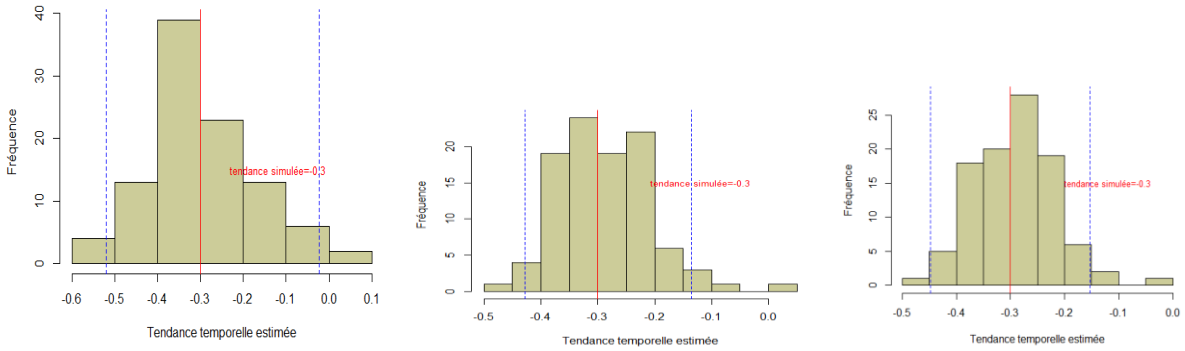
Les histogrammes sont construits avec une probabilité de détection $\mathrm{p} = 0.2$ (à gauche), $\mathrm{p} = 0.5$ (au centre) et $\mathrm{p} = 0.8$ (à droite).

EE=500 PR=100% (année en continue) :

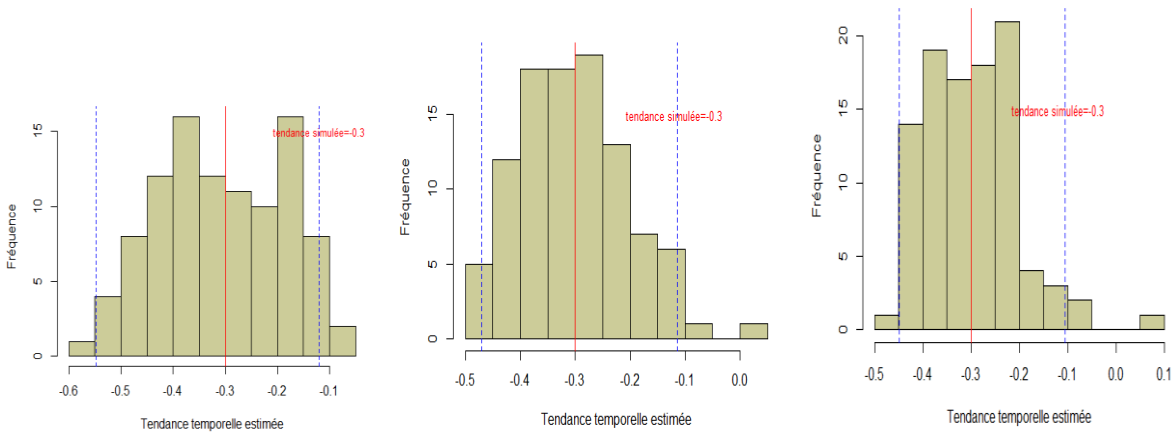
J=2



J=3

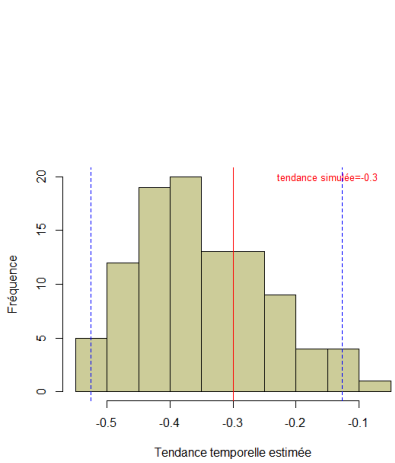


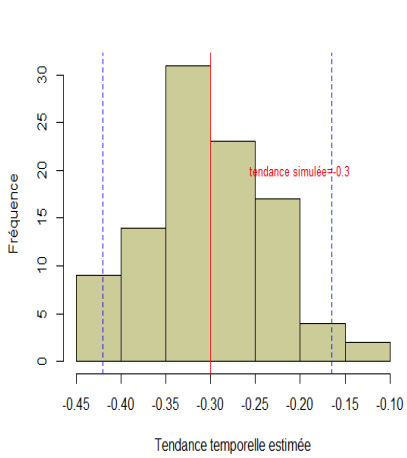
J=4

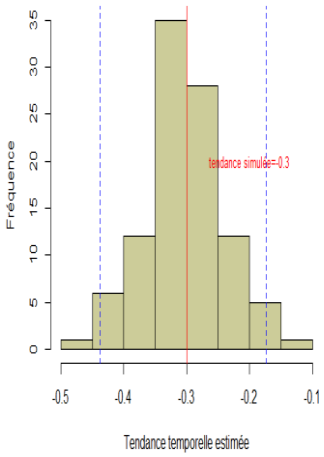


EE=500 PR=100% (année en facteur) :

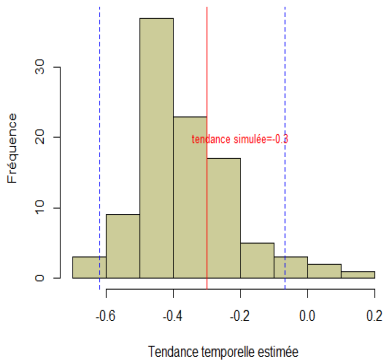
J=2 :

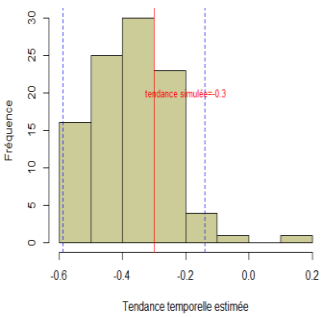


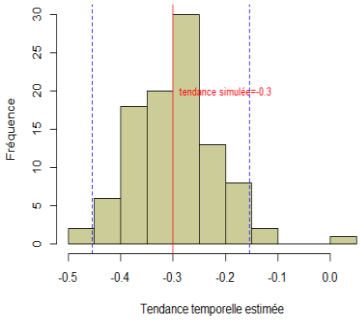




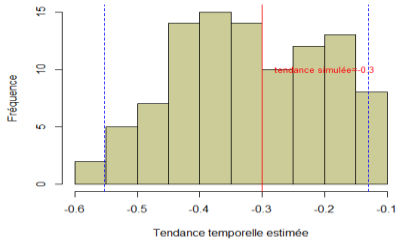
J=3 :

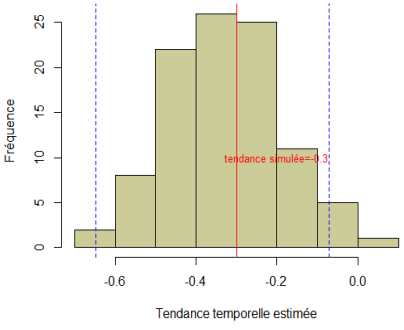


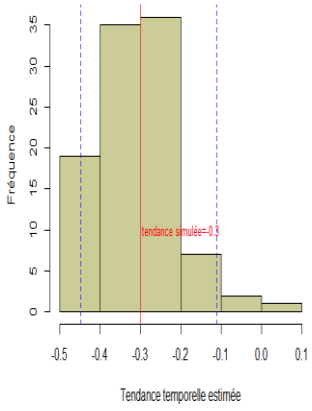




J=4 :

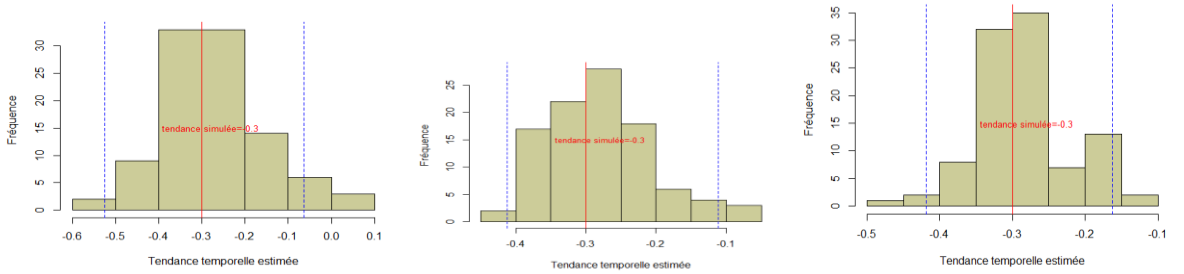




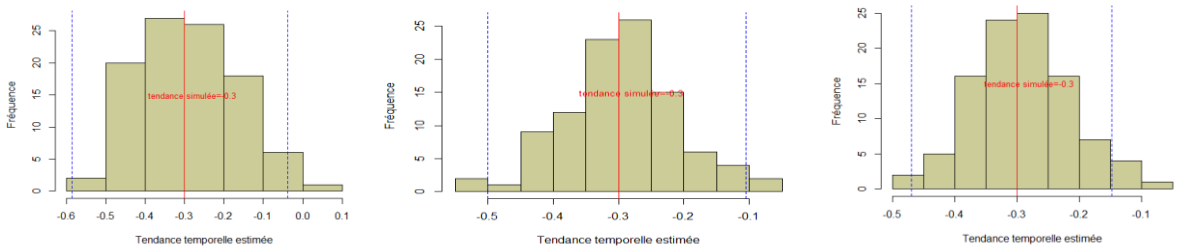


EE=500 PR=50% (année en continue) :

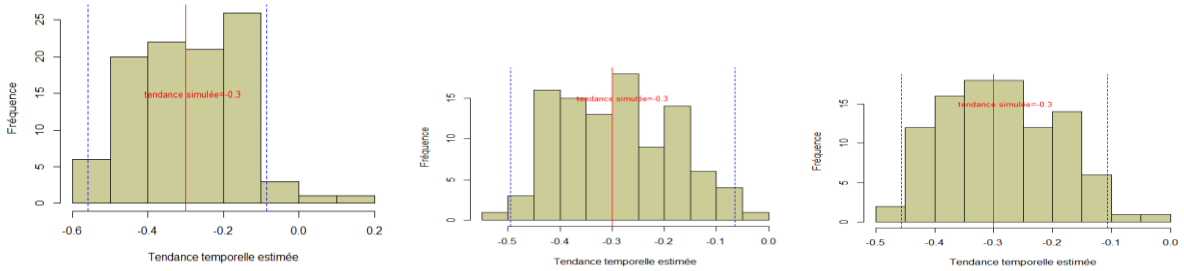
J=2 :



J=3 :

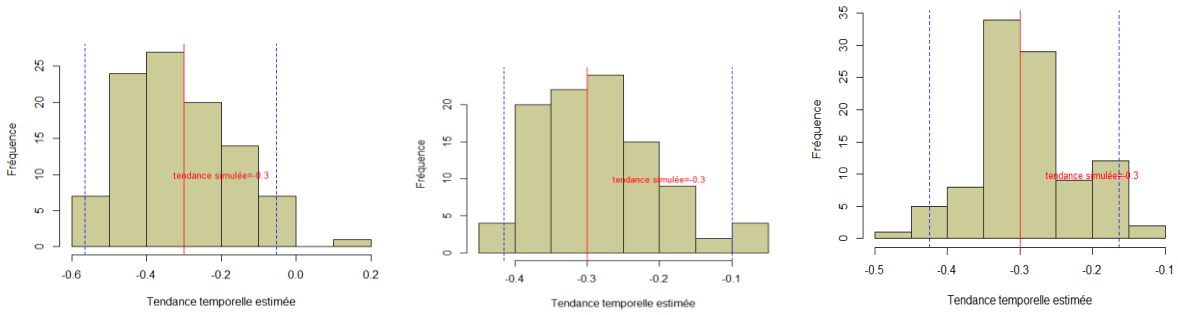


J=4 :

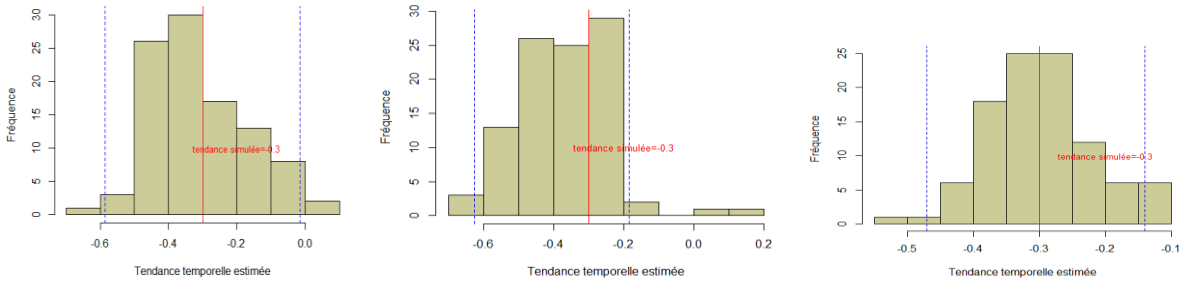


EE=500 PR=50% (année en facteur) :

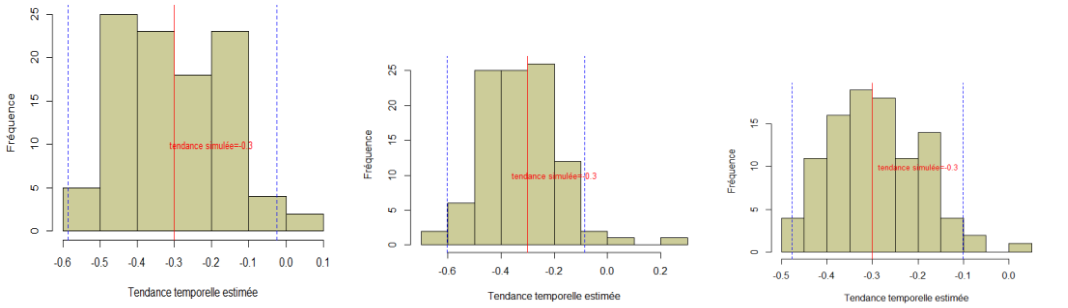
J=2 :



J=3 :

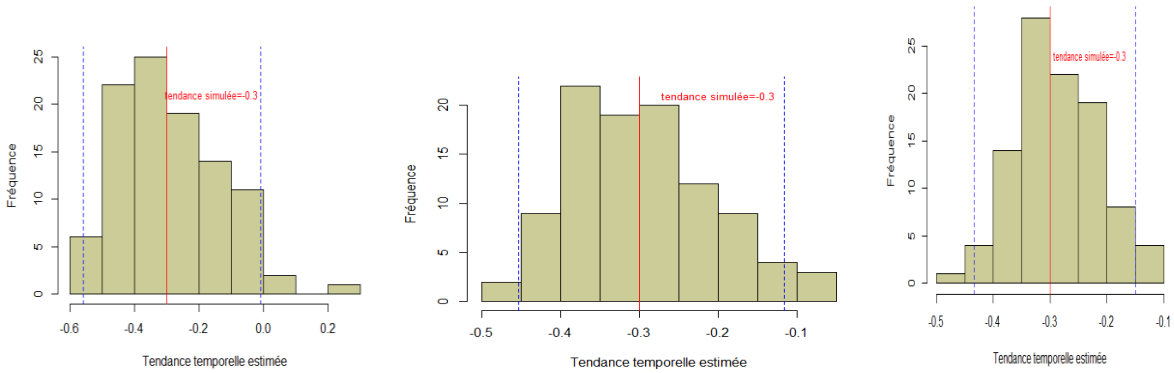


J=4 :

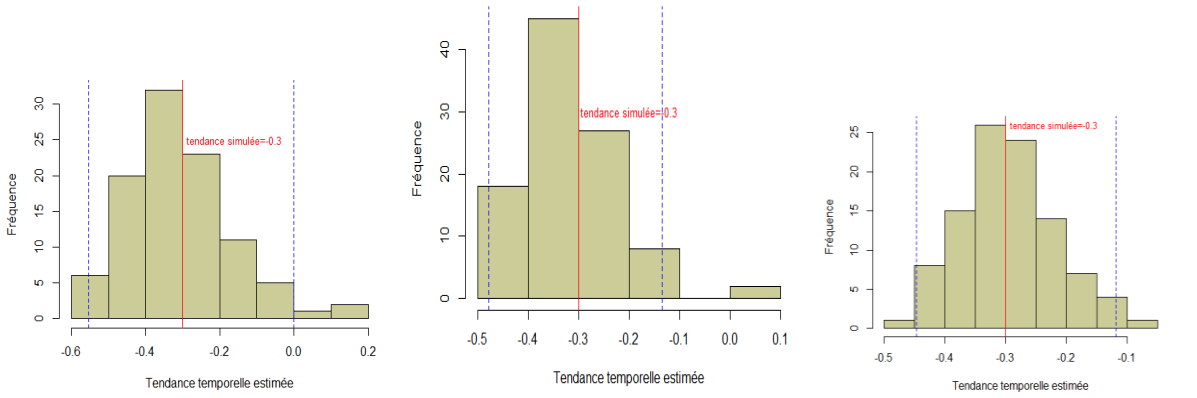


EE=500 PR=10% (année en continue) :

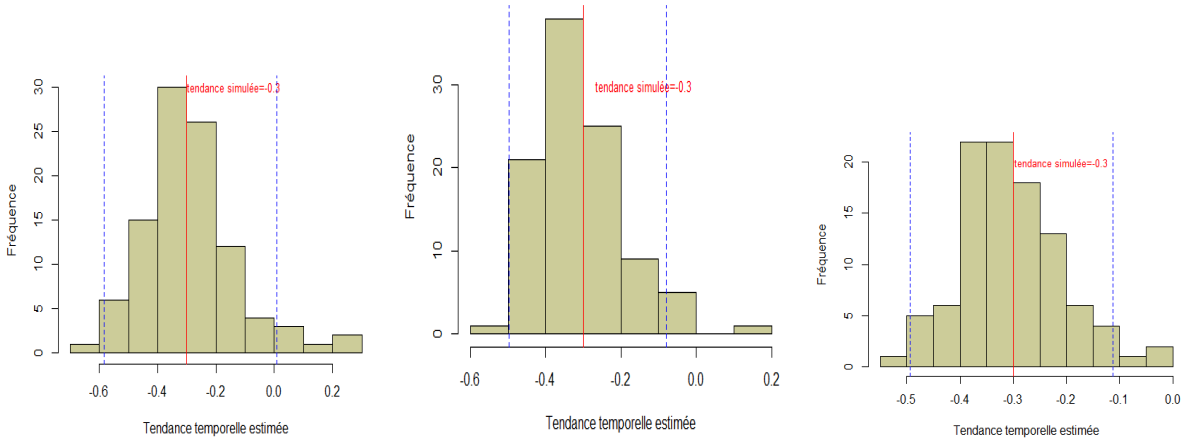
J=2 :



J=3 :

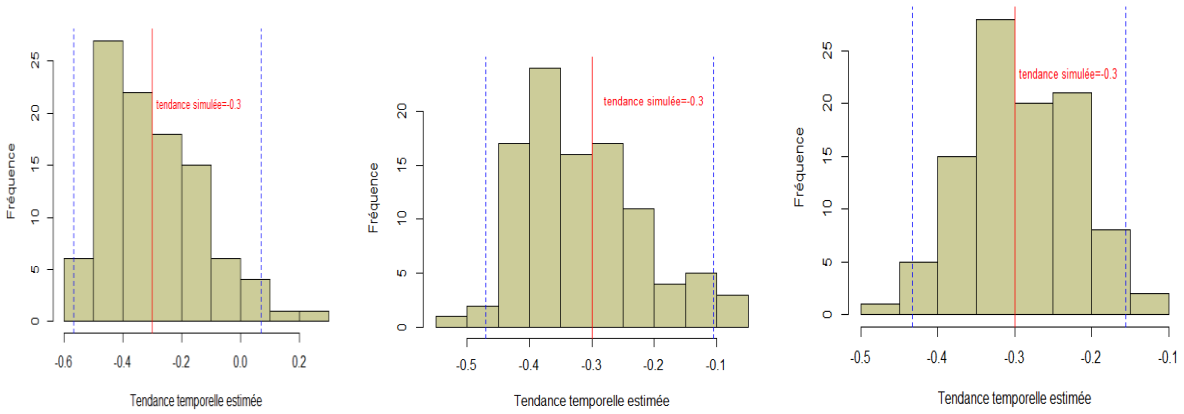


J=4 :

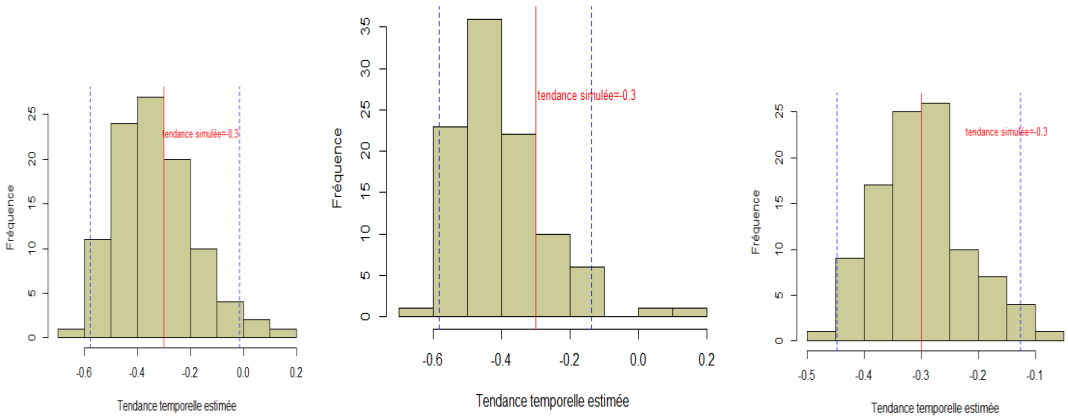


EE=500 PR=10% (année en facteur) :

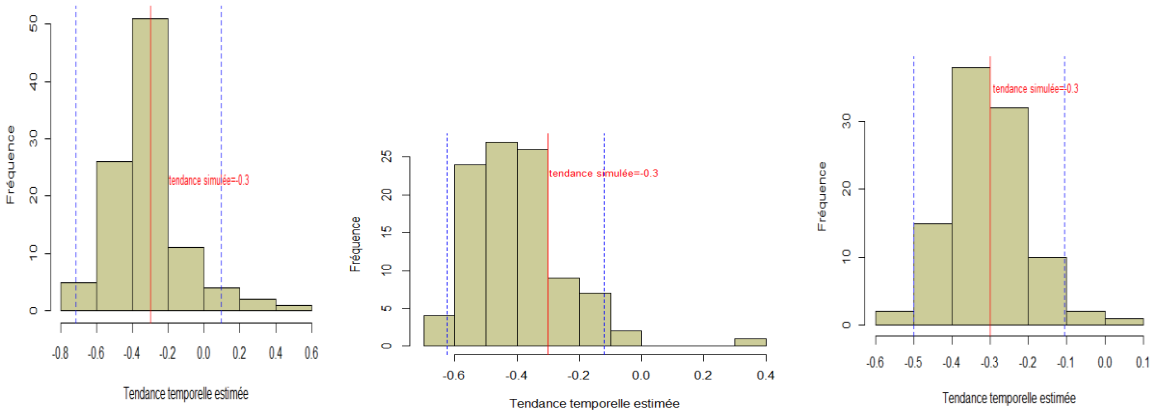
J=2 :



J=3 :

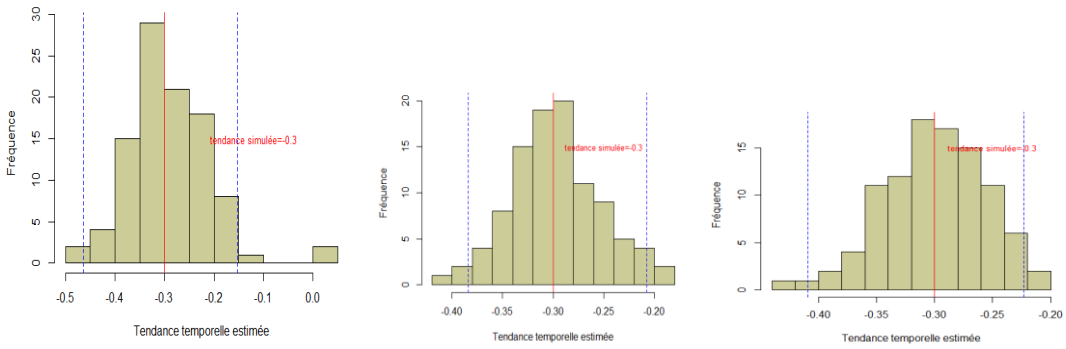


J=4 :

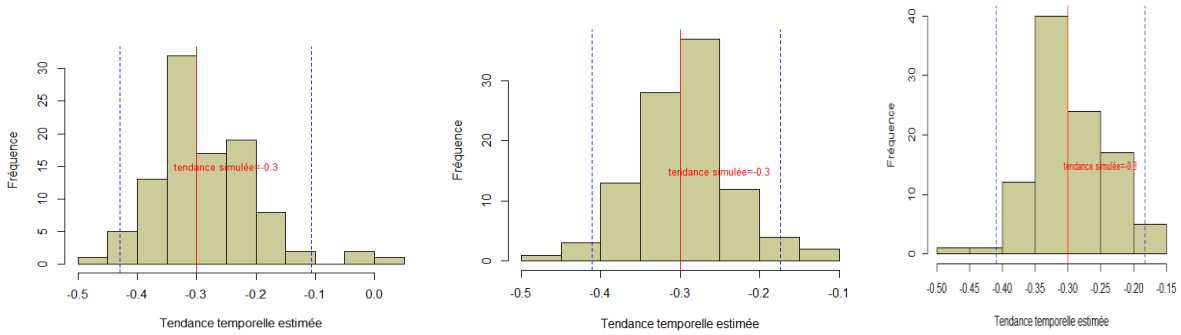


EE=1000 PR=100% (année en continue) :

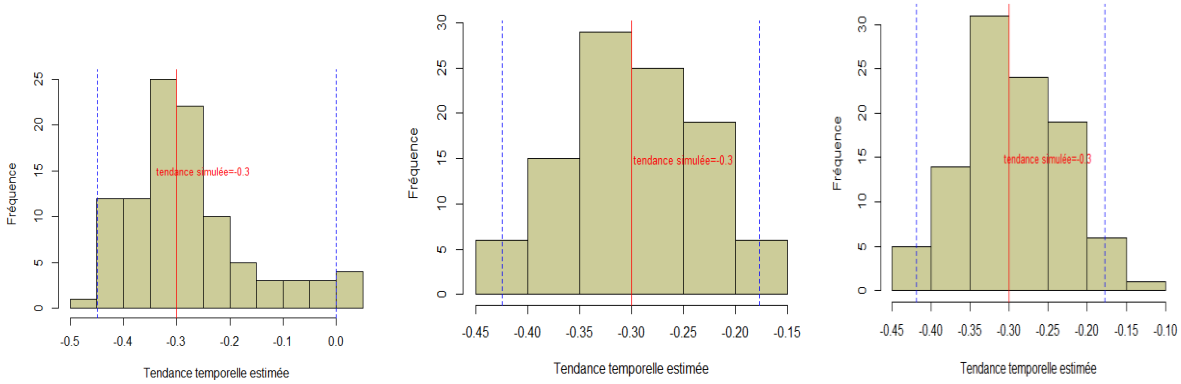
J=2 :



J=3 :

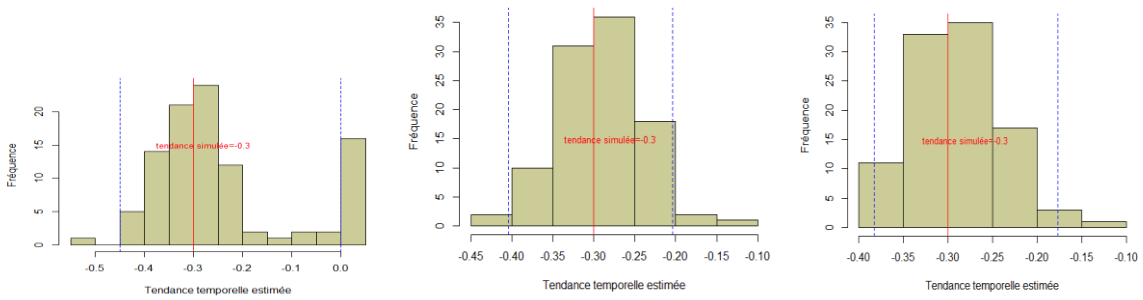


J=4 :

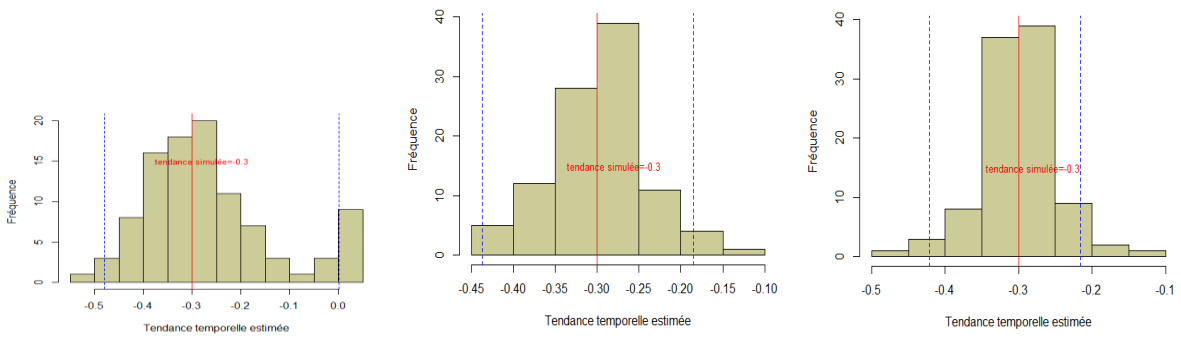


EE=1000 PR=50% (année en continue) :

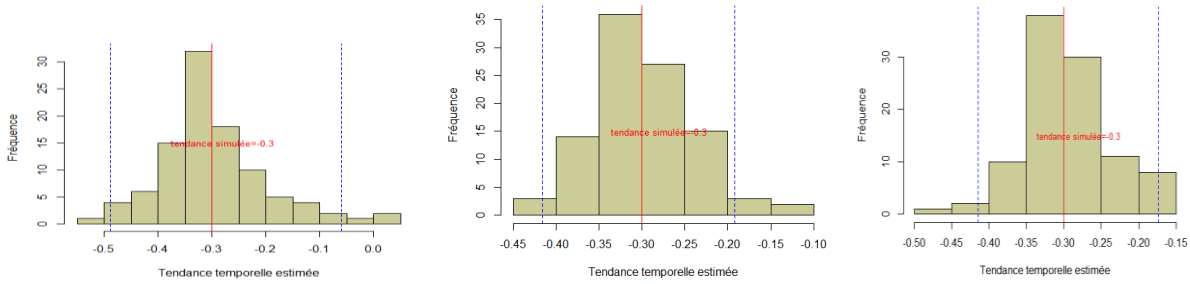
J=2 :



J=3 :

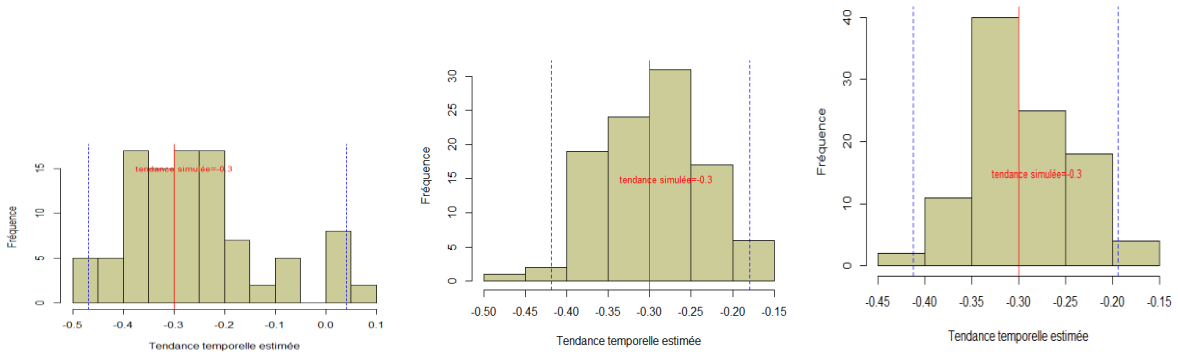


J=4 :

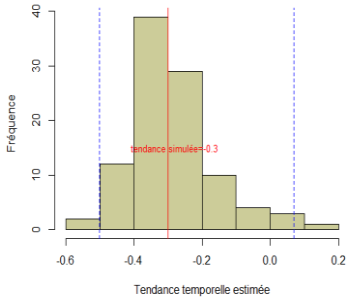


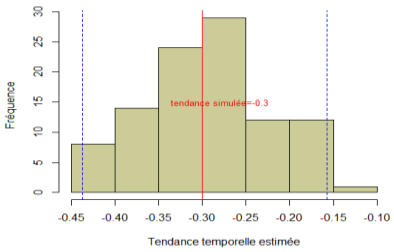
EE=1000 PR=10% (année en continue) :

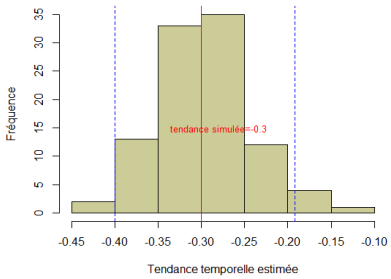
J=2 :



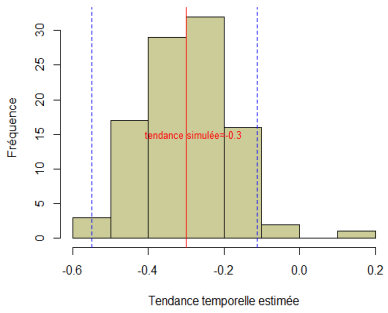
J=3 :

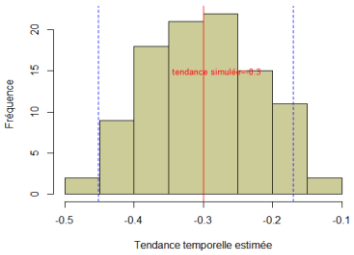


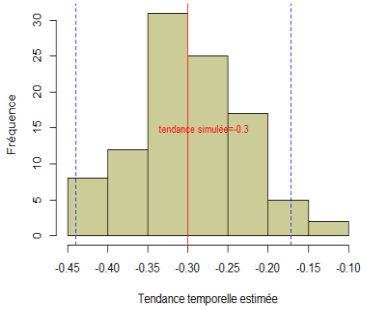




J=4 :

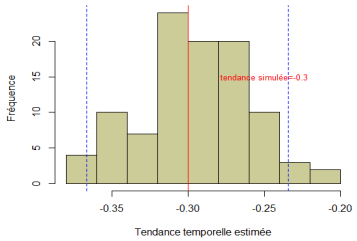


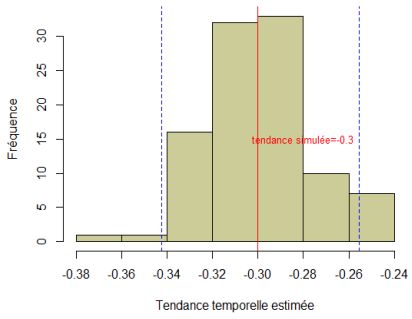


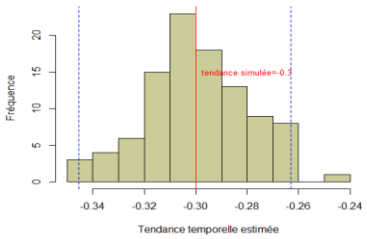


EE=5000 PR=100% (année en continue) :

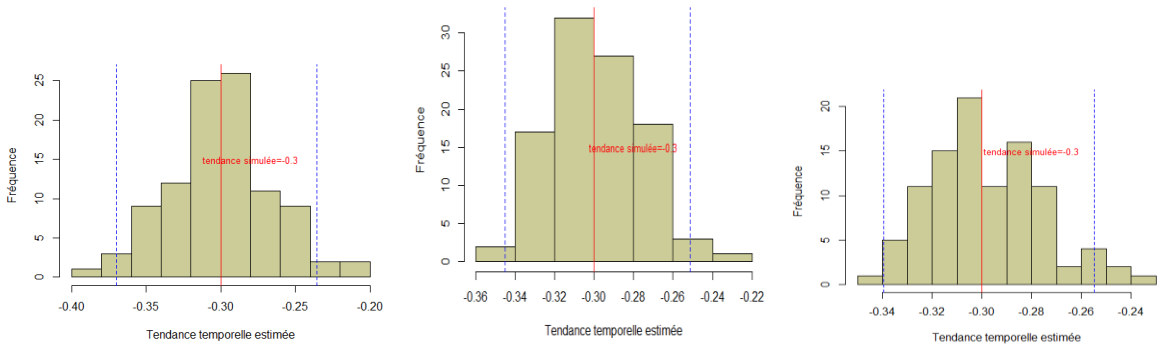
J=2 :



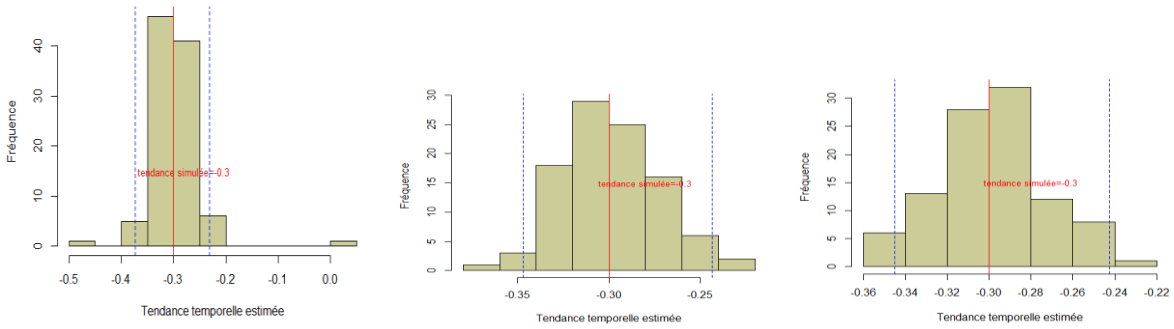




J=3 :

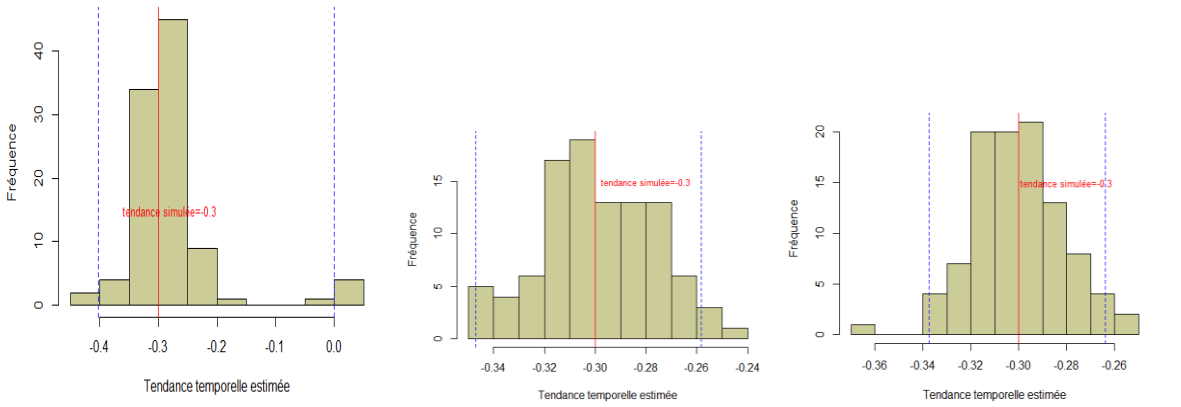


J=4 :

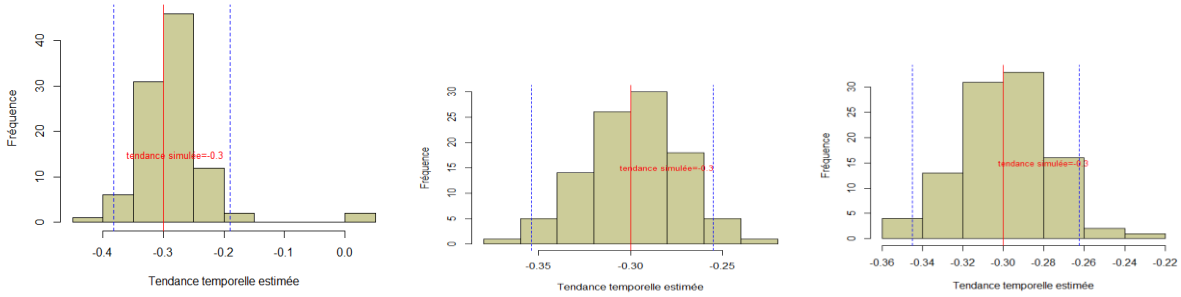


EE=5000 PR=50% (année en continue) :

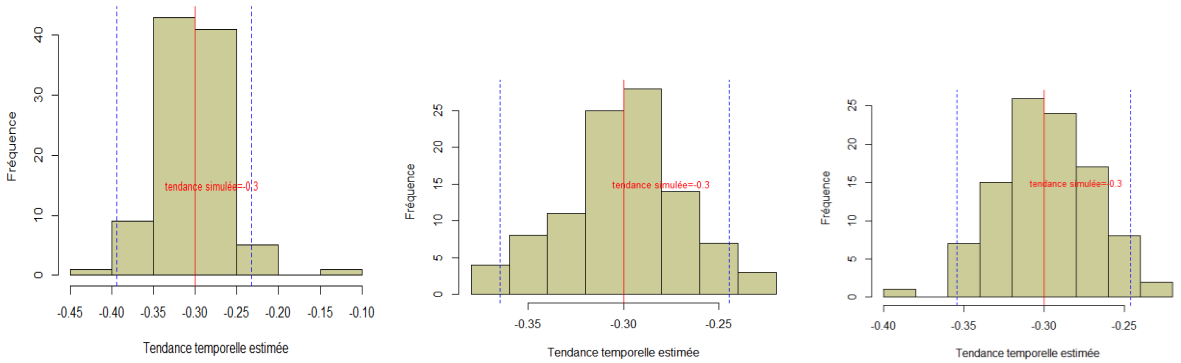
J=2 :



J=3 :

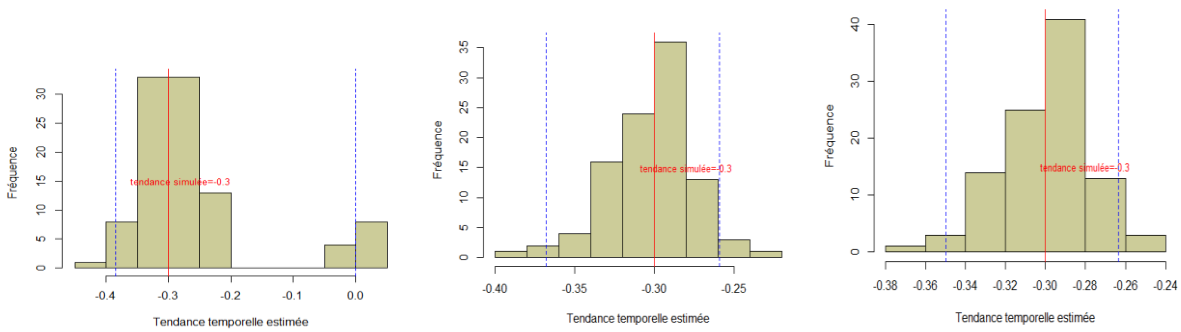


J=4 :

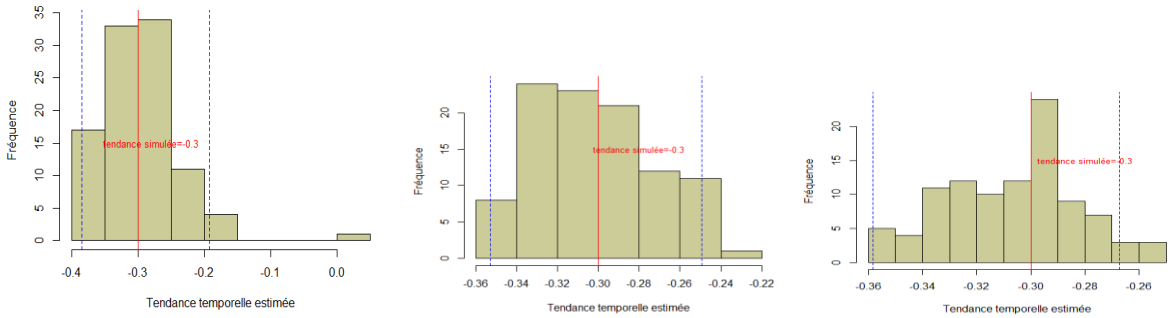


EE=5000 PR=10% (année en continue) :

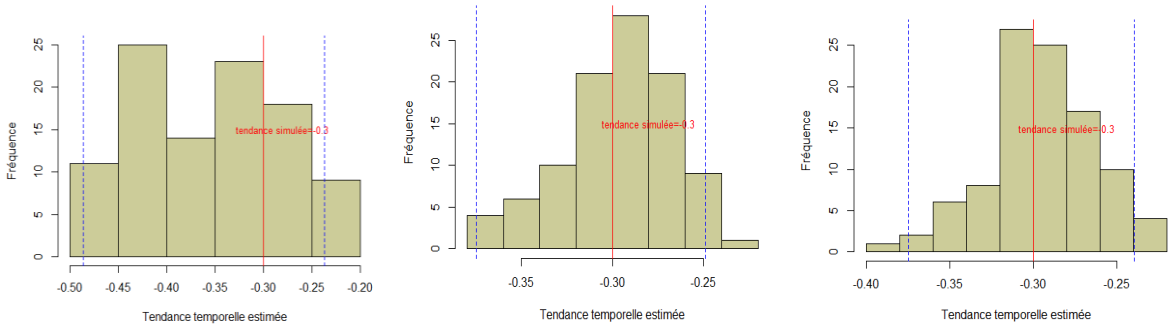
J=2 :



J=3 :



J=4 :



Annexe n°6 :

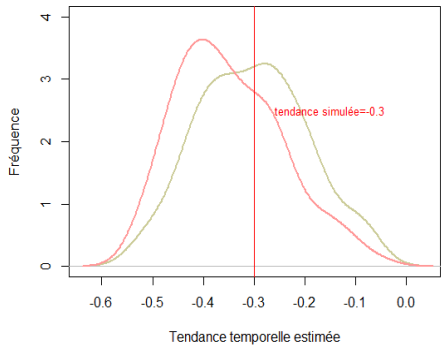


Figure 9 : Courbes de densités de la tendance temporelle avec année en variable continue (gris) et année en variable factorielle (rose), PR=100%, p=0.20 et J=2 et EE=500

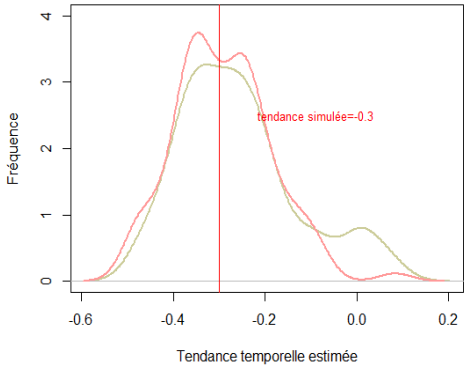


Figure 7 : Courbes de densités de la tendance temporelle avec année en variable continue (gris) et année en variable factorielle (rose), PR=100%, p=0.20 et J=2 et EE=1000

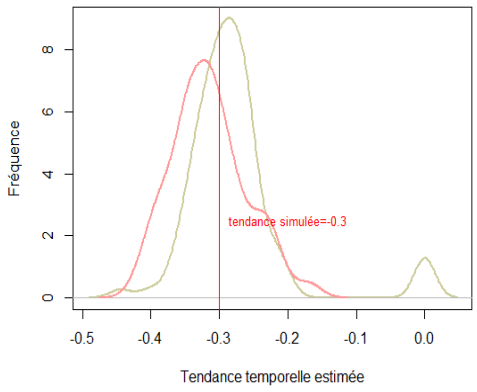


Figure 7 : Courbes de densités de la tendance temporelle avec année en variable continue (gris) et année en variable factorielle (rose), PR=100%, p=0.20 et J=2 et EE=5000

Annexe n°7 : Tableaux

PR=10 % EE=5000 :

			p=0.20				p=0.50				p=0.80			
EE	PR	J	RMSE cont.	Biais cont.	RMSE fact.	Biais fact.	RMSE cont.	Biais cont.	RMSE fact.	Biais fact.	RMSE cont.	Biais cont.	RMSE fact.	Biais fact.
500	50	2	0.00715	0.00035	0.00793	0.00085	0.00419	0.00018	0.00466	0.00120	0.00340	-0.00048	0.00337	-0.00020
		3	0.00681	0.00120	0.00769	0.00233	0.00479	0.00117	0.00916	0.00627	0.00479	0.00117	0.00916	0.00627
		4	0.00736	0.00031	0.01021	0.00221	0.00512	0.00072	0.01004	0.00626	0.00458	0.00047	0.00506	0.00092
1000		2	0.00583	-0.00147			0.00289	7.70130e-06			0.00253	5.79690e-06		
		3	0.00563	-9.12003e-05			0.00334	-0.00013			0.00250	-0.00013		
		4	0.00556	0.00047			0.00368	0.00019			0.00320	9.32205e-05		
5000		2	0.00452	-0.00137			0.00128	0.00011			0.00106	-3.36255e-05		
		3	0.00258	-0.00021			0.00140	0.00013			0.00120	0.00034		
		4	0.00496	0.00307			0.00149	9.66368e-05			0.00153	-0.00020		

Tableau 6 : Tableau RMSE et biais pour une proportion de sites répliqués J fois de 10%

PR=50 % :

			p=0.20				p=0.50				p=0.80			
EE	PR	J	RMSE cont.	Biais cont.	RMSE fact.	Biais fact.	RMSE cont.	Biais cont.	RMSE fact.	Biais fact.	RMSE cont.	Biais cont.	RMSE fact.	Biais fact.
500	10	2	0.00550	-0.00060	0.00689	0.00149	0.00359	-0.00070	0.00373	-0.00042	0.00309	-0.00054	0.00318	-0.00036
		3	0.00589	-0.00012	0.00739	0.00157	0.00430	5.78929e-05	0.00806	0.004134	0.00373	-8.92713e-05	0.00382	0.00017
		4	0.00668	0.00039	0.00706	0.00077	0.00513	-0.00011	0.00700	0.00193	0.00452	-0.00035	0.00462	-0.00013
1000		2	0.00603	-0.00197			0.00245	-0.00030			0.00236	-0.00032		
		3	0.00579	-0.00090			0.00272	2.65479e-05			0.00242	0.00011		
		4	0.00457	0.00025			0.00283	4.83193e-05			0.00266	-8.05855e-05		
5000		2	0.00330	-0.00083			0.00106	-2.86155e-05			0.00088	3.08862e-05		
		3	0.00248	-0.00054			0.00119	-0.00011			0.00108	-1.99562e-05		
		4	0.00208	9.91472e-05			0.00147	1.65600e-05			0.00142	-1.26304e-05		

Tableau 7 : RMSE et biais pour une proportion de sites répliqués J fois de 50%

PR=10 %

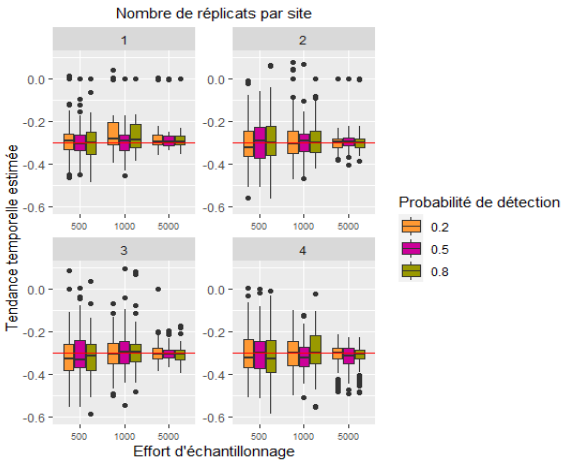


Figure 10 : Boxplots de la tendance temporelle estimée avec la probabilité de détection, l'effort d'échantillonnage, le nombre de réplicats par année et une proportion de sites répliqués égale à 10%

PR=50 %

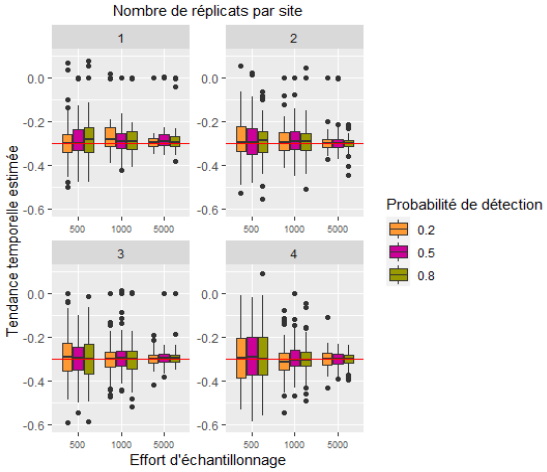


Figure 2 : Boxplots de la tendance temporelle estimée avec la probabilité de détection, l'effort d'échantillonnage, le nombre de réplicats par année et une proportion de sites répliqués égale à 50%

Bayarri, M. J., & Berger, J. O. (2000). P-values for composite null models. Journal of the American Statistical Association, 95(452), 1127-1142. doi) Dec
Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. Statistical Science, 19(1), 58-80.
Box, G. E. P. (1976). Science and statistics. Journal of the American Statistical Association, 71(356), 791-799.
Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society, Series A, 143(4), 383-430.
Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. The American Statistician, 46(3), 167-174.
Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. The American Statistician, 49(4), 327-335.
Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. The American Statistician, 59(2), 121-126.
Coron, C., Calenge, C., Giraud, C., & Julliard, R. (2018). Bayesian estimation of species relative abundances and habitat preferences using opportunistic data. Environmental and Ecological Statistics, 25(1), 71-93.
Cowles, M. K., Roberts, G. O., & Rosenthal, J. S. (1999). Possible biases induced by MCMC convergence diagnostics. Journal of Statistical Computation and Simulation, 64(1), 87-104.
Dennis, R. L. H., & Thomas, C. D. (2000). Bias in Butterfly Distribution Maps: The Influence of Hot Spots and Recorder's Home Range. Journal of Insect Conservation, 4(2), 73-77 %! Bias in Butter.
Evans, M. (2007). Comment: Bayesian checking of the second levels of hierarchical models. Statistical Science, 22(3), 344-348. doi:10.1214/07-sts235c
Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). Bayesian Data Analysis (3rd ed.). Boca Raton: Chapman & Hall.
Geweke, J. (1991). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (Vol. 196): Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN.
Giraud, C., Calenge, C., Coron, C., & Julliard, R. (2015). Capitalizing on opportunistic data for monitoring relative abundances of species. Biometrics, 72(2), 649-658. doi:10.1111/biom.12431
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(3), e14770. doi:10.1371/journal.pone.0014770
Hjort, N. L., Dahl, F. A., & Hognadottir, G. (2006). Post-processing posterior predictive p values. Journal of the American Statistical Association, 101(475), 1157-1174. doi) Sep
Johnson, V. E. (2004). A Bayesian chi(2) test for goodness-of-fit. Annals of Statistics, 32(6), 2361-2384.
Johnson, V. E. (2007). Bayesian Model Assessment Using Pivotal Quantities. Bayesian Analysis, 2(4), 719-734.
Kass, R. E., Carlin, B. P., Gelman, A., & Neal, R. M. (1998). Markov chain Monte Carlo in practice: a roundtable discussion. The American Statistician, 52(2), 93-100.
Kéry, M., Royle, J. A., Schmid, H., Schaub, M., Volet, B., Häfliger, G., & Zbinden, N. (2010). SiteOccupancy Distribution Modeling to Correct Population-Trend Estimates Derived from Opportunistic Observations. Conservation Biology, 24(5), 1388-1397.
Kuussaari, M., Heliölä, J., Pöyry, J., & Saarinen, K. (2007). Contrasting trends of butterfly species preferring semi-natural grasslands, field margins and forest edges in northern Europe. Journal of Insect Conservation, 11(4), 351-366 %! Contrasting t.
Link, W. A., & Sauer, J. R. (1998). Estimating population change from count data: application to the north american breeding bird survey. Ecological Applications, 8(2), 258-268. doi:1998
O'Hagan, A. (2003). HSSS model criticism. In P. J. Green, N. L. Hjort, & S. T. Richardson (Eds.),

Highly Structured Stochastic Systems (pp. 423-444): Oxford University Press.
Piccinato, L. (2000). Comments on Asymptotic distribution of P values in composite null models by J.
van Strien, A. J., van Swaay, C. A. M., & Termaat, T. (2013). Opportunistic citizen science data of animal species produce reliable estimates of distribution trends if analysed with occupancy models. Journal of Applied Ecology, 50(6), 1450-1458.
Zhang, J. L. (2014). Comparative investigation of three Bayesian p values. Computational Statistics and Data Analysis, 79, 277-291