Académie de Montpellier

Université Montpellier II

- Sciences et Techniques du Languedoc -

Habilitation à Diriger des Recherches

Des modèles de capture-recapture pour l'écologie évolutive

Olivier Gimenez

Soutenue le 29 Mars 2010, devant la commission d'examen :

Dominique Pontier	Professeur, Université Claude Bernard, Lyon	Rapporteur
Denis Couvet	Professeur, MNHN, Paris	Rapporteur
James D. Nichols	Directeur de recherche (équiv.), USGS, Etats-Unis	Rapporteur
Isabelle Olivieri	Professeur, UM2, Montpellier	Examinateur
Thierry Boulinier	Directeur de recherche, CNRS, Montpellier	Examinateur

Table des matières

Des modèles de capture-recapture pour l'écologie évolutive

En écologie évolutive, la disponibilité croissante de données récoltées dans le temps permet de comprendre les processus évolutifs en conditions naturelles mieux que par le passé. Les biologistes montrent ainsi un intérêt de plus en plus marqué dans les tests empiriques des forces évolutives à l'oeuvre dans des populations sauvages. Bien que ces études soient nécessaires pour mettre à jour des processus en environnement naturel qui ne peuvent être aisément imités en laboratoire, elles posent aussi des problèmes méthodologiques récurrents. En particulier, mesurer la valeur sélective sur le terrain s'avère être une tâche difficile. Idéalement, estimer la fitness [1] requiert un suivi continu de tous les individus d'une population de la naissance à la mort. En pratique toutefois, il est peu réaliste de pouvoir suivre tous les éléments d'une population. Les individus peuvent être vus (ou capturés) ou pas à plusieurs occasions au cours de leur vie, ce qui pose le problème d'une probabilité de détection < 1.

Alors que ce problème de détectabilité imparfaite a depuis longtemps été intégré en écologie et en biologie de la conservation, il est surprenant de constater qu'il a été négligé en écologie évolutive. Le plus souvent, on suppose que la probabilité de détection vaut 1 pour se ramener commodément à des méthodes statistiques classiques comme des régressions ou des analyses de survie. Or, parce que cette hypothèse forte revient à supposer qu'un individu vu pour la dernière fois avant la fin de l'étude est mort, on sur-estime la mortalité (entre autres) si la probabilité de détection est effectivement < 1.

Les modèles de Capture-Marquage-Recapture (CMR ; voir Encadré 1) incorporent de manière explicite le processus de détection dans l'estimation des paramètres démographiques (e.g. probabilités de survie, de reproduction ou de dispersion). Toutefois, le potentiel des méthodes de CMR

1. Dans ce document, j'utiliserai valeur sélective ou fitness indifféremment.

Encadré 1. Modèles et données de CMR uniétats

Un protocole de CMR consiste en plusieurs occasions d'échantillonnage au cours desquelles des individus sont capturés ou observés. A chaque occasion, les individus non marqués reçoivent des marques uniques puis sont relâchés. L'identité des individus précédemment marqués est relevée avant leur relâché. Les données se ramènent à un ensemble d'histoires de détection individuelles faites de « 1 » et de « 0 » selon que l'individu a été capturé ou pas (Lebreton et al. 1992). Par exemple, pour 3 occasions de capture, l'individu i avec l'histoire $h_i = (1, 0, 1)$ a été capturé pour la première fois à la première occasion, marqué puis relâché, non capturé à la deuxième occasion, puis capturé à la troisième et dernière occasion.

L'estimation des paramètres démographiques s'appuie sur le modèle probabiliste suivant. Notons $\phi_{i,t}$ la probabilité qu'un individu i vivant à l'occasion t survive jusqu'a l'occasion $t + 1$ et p_t la probabilité de détection à l'occasion t d'un individu vivant. Par simplicité, on suppose que les probabilités de détection ne diffèrent pas entre individus (pas d'indice i) mais varient au cours du temps. La probabilité de l'histoire h_i ci-dessus est alors $\phi_{i,1}(1 - p_2)\phi_{i,2}p_3$.

En pratique, les données de CMR ne permettent pas d'estimer une survie différente pour chaque individu et l'on s'appuiera sur diverses hypothèses d'homogénéité. Le modèle de base (CJS pour Cormack-Jolly-Seber du nom des 3 chercheurs qui ont contribué au développement de ce modèle, voir Lebreton et al. 1992) suppose que la probabilité de survie est identique pour tous les individus et varie au cours du temps. Diverses généralisations et particularisations ont été proposées : effets groupes (mâle vs. femelle par exemple, e.g. P23), âge (jeune vs. adulte, e.g. S31), contraintes sur le temps (survie constante ou variant linéairement avec des covariables environnementales, e.g. P14, P25).
Cette flexibilité conduit à une stratégie de modélisation basée sur un modèle « parapluie » contenant tous les effets biologiquement plausibles, à partir duquel on obtient d'autres modèles plus restrictifs par des contraintes linéaires, de façon analogue aux modèles linéaires généralisés. La qualité d'ajustement du modèle CJS peut être évaluée (voir la revue P9 ; logiciel P45), et la sélection d'un meilleur modèle s'effectue ensuite en général par minimisation d'un critère comme l'AIC (Burnham et Anderson 2002). L'estimation des paramètres, qu'elle soit menée dans un contexte bayésien ou fréquentiste (voir Encadré 2), repose sur la fonction de vraisemblance. Sous l'hypothèse d'indépendance des individus, la vraisemblance d'un jeu de données CMR est proportionnelle au produit $\prod_i h_i$.
Divers logiciels facilitent considérablement la mise en œuvre de ces méthodes pour les biologistes : MARK (White et Burnham 1999), M-SURGE (P47), E-SURGE (Choquet et al. 2009 ; P43) et WinBUGS (CL46).

n'a été que peu exploré en écologie évolutive, et des discussions récurrentes avec des collègues m'ont convaincu qu'un effort pédagogique était d'abord nécessaire pour convaincre de l'intérêt de tenir compte de la probabilité de détection. Quoi de mieux alors que des exemples revisitant avec l'outil CMR des questions bien connues en écologie évolutive (voir P40 et Figure 1) ?

Voilà la boîte de Pandore ouverte ! Alors que les paramètres démographiques peuvent être influencés par plusieurs caractéristiques individuelles comme l'âge, le sexe, le rang social, le phénotype ou le génotype [2], il faut avouer que ces diverses sources de variabilité ne sont incorporées que de manière imparfaite dans les modèles de CMR, voire complètement ignorées. Plus ennuyeux, l'hétérogénéité entre individus est souvent considérée simplement comme une nuisance susceptible de biaiser les estimateurs ; les efforts pour évaluer et caractériser cette hétérogénéité sont donc essentiellement motivés par la réduction de ce biais et ainsi le fait de statisticiens. Pourtant, l'un des objectifs majeurs en écologie évolutive n'est-il pas justement la détection et la caractérisation des différences entre individus ou groupes d'individus, notamment les différences touchant les composantes de la valeur sélective (et donc les paramètres démographiques) ?

Dans ce document, je plaide pour que la variabilité individuelle soit mise au centre de l'effort d'estimation et d'inférence dans les modèles de CMR, et qu'elle ne soit plus seulement considérée comme une nuisance. Pour illustrer cette idée, je présente comment quatre questions centrales en écologie évolutive (et autant de chapitres) peuvent être abordées grâce aux modèles de CMR.

Cette motivation explique pourquoi j'ai demandé à passer mon HDR dans l'école doctorale SIBAGHE (Systèmes Intégrés en Biologie, Agronomie, Géosciences, Hydrosciences, Environnement) et non dans ISS (Information, Structures, Systèmes) dans laquelle j'ai obtenu ma thèse en Biostatistique. Aujourd'hui, mon travail se situe clairement à l'interface de la modélisation, des statistiques appliquées et de la biologie des populations : des cas d'étude générent des questions biologiques qui motivent le développement de modèles et méthodes statistiques ; je tiens aussi à assurer le transfert de ces connaissances et techniques par ma participation à l'encadrement d'étudiants, l'enseignement (je n'en fais pas assez à mon goût), à des ateliers de travail et au développement et la popularisation de logiciels. Je souhaitais donc que la pertinence de mes travaux soit évaluée par des biologistes.

2. Les paramètres démographiques peuvent aussi être influencés par des facteurs environnementaux abiotiques (climat, température, précipitations) et biotiques (compétition, prédation, maladies). Etudier l'impact des facteurs abiotiques sur les paramètres démographiques occupe une part importante de mes recherches. Plutôt que d'être exhaustif, j'ai fait le choix d'une présentation orientée sur mes travaux de recherche qui semblent les plus prometteurs.

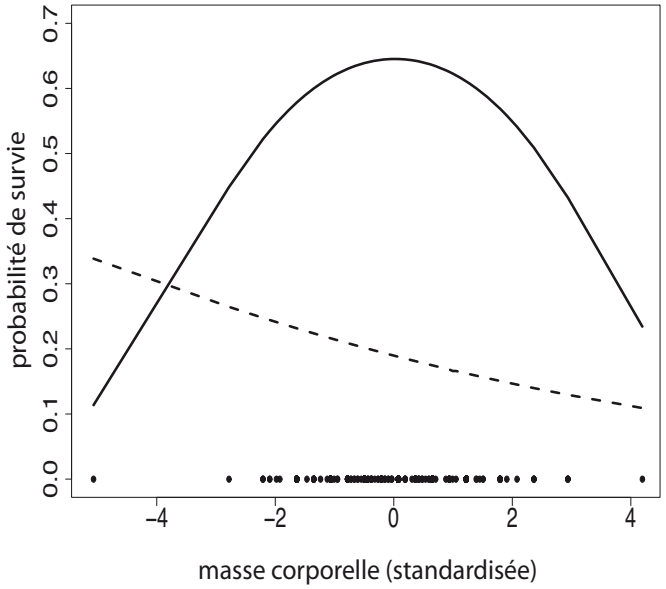


FIGURE 1 – Relation entre survie et masse corporelle chez le tisserin social *Philetairus socius* en Afrique du Sud estimée sur la période 1993-2000 à partir du marquage de 435 jeunes oiseaux (P41). La relation est obtenue par deux approches : une analyse de CMR (trait continu) et une analyse naïve supposant une probabilité de détection des individus $= 1$ (tirets). Les valeurs observées de la masse sont représentées par des cercles pleins. Comme les probabilités de détection sont < 1 (et varient au cours du temps : entre 0.124 ± 0.045 et 0.829 ± 0.085), l'analyse naïve sous-estime les probabilités de survie. De plus, en comparant les modèles reliant la probabilité de survie de façon respectivement linéaire et quadratique avec la masse corporelle (voir plus bas), on peut discriminer entre sélection directionnelle (relation linéaire) et sélection stabilisante (pic de survie pour des valeurs intermédiaires de la masse : relation quadratique). Alors que l'approche naïve favorise le modèle linéaire et indiquerait donc une sélection directionnelle en faveur des individus les plus légers, l'approche par CMR sélectionne le modèle quadratique et plaide donc pour une sélection stabilisante, avec une survie optimale autour de la masse moyenne (voir chapitre 3). Cet exemple illustre bien les risques d'inférence erronée sur la forme de la sélection quand la détectabilité imparfaite n'est pas prise en compte. Techniquement, en partant de l'Encadré 1, la probabilité de survie peut être exprimée comme fonction d'une ou plusieurs variables explicatives : $\text{logit}(\phi_{i,t}) = f(x_{i,t})$ où $x_{i,t}$ est la valeur de la variable explicative pour l'individu i à l'occasion t. Une fonction, $\text{logit}(\phi(x)) = \log(x/(1\breve{}x))$, permet de contenir l'estimateur de la probabilité de survie entre 0 et 1. Le prédicteur $f(x_{i,t})$ peut être linéaire ($f(x) = \beta_0 + \beta_1 x$) ou bien quadratique ($f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$), ou basé sur une forme plus souple sans a priori (voir chapitre 3).

Sans rentrer dans l'analyse, je dois avouer qu'il m'a fallu plusieurs années pour comprendre et assumer cette place, ma place. Cela est probablement dû à une tradition française élitiste autour des études en mathématiques dont je suis le produit. Mes séjours post-doctoraux en Grande-Bretagne m'ont aidé à faire « exploser » ce cloisonnement et évacuer un certain complexe. Je dois aussi énormément à mon laboratoire, le Centre d'Ecologie Fonctionnelle et Evolutive, et aux ingénieurs, techniciens, administratifs, étudiants, chercheurs et enseignants-chercheurs qui le font vivre.

Bonne lecture !

1

Décrire des patrons de sénescence

1.1 Sénescence en conditions naturelles

La sénescence (un déclin des paramètres démographiques avec l'âge) est un paradoxe à première vue difficile à expliquer. C'est en effet un cas où la sélection naturelle favoriserait un déclin de la valeur sélective avec l'âge. On sait néanmoins que la sénescence évolue du fait de la mortalité extrinsèque qui entraîne une diminution de la force de la sélection naturelle avec l'âge. Si les prédictions de la théorie sur l'évolution de la sénescence ont été vérifiées en laboratoire, des études empiriques ont remis en cause leur validité en conditions naturelles (e.g. Reznick et al. 2004, Williams et al. 2006).

Jusqu'à récemment, les preuves empiriques d'une sénescence de survie restaient ambigües. Si chez les populations sauvages de mammifères, on a produit rapidement des preuves de sénescence de survie, on prédisait que la sénescence était rare voire inexistante chez les oiseaux. Pourquoi ? Une partie de l'explication réside dans certains problèmes d'ordre méthodologique. La sénescence peut être masquée par des changements dans les facteurs extrinsèques qui affectent la survie, par un biais dans l'estimation de la survie généré par exemple par un individu qui disperse pour aller se reproduire ailleurs, ou bien parce que la probabilité de détection n'est pas prise en compte. Si les modèles de CMR ont permis de démontrer des phénomènes de sénescence sur la

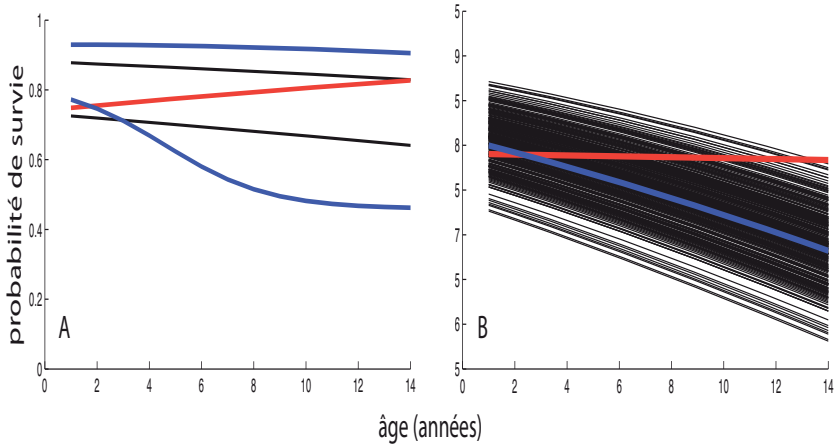


FIGURE 2 – Patrons de sénescence de survie en présence d'hétérogénéité A. *discrète* et B. *continue*. Cas A. On simule des données selon deux classes d'individus : des individus robustes qui ont une forte survie initiale et des individus fragiles qui ont une survie initiale plus faible, les 2 groupes sénescent au même rythme (2 lignes en noir). Cas B. On simule des données avec un total de 300 individus dont la survie en fonction de l'âge pour l'individu i est donnée par $\text{logit}(\phi_{i,t}) = 1.5 - 0.05a_{i,t} + e_i$ où $a_{i,t}$ est l'âge de l'individu i à t et la variable latente $e_i \sim N(0, \sigma = 0.5)$. Certains individus ont une forte survie initiale et d'autres plus faible, et tous sénescent au même rythme (nombreuses lignes en noir). Les individus de survie faible (fragiles) sortent de la population plus rapidement que les individus robustes comme leur chances de survie sont plus faibles. Si la variabilité individuelle ne peut être mesurée (i.e. le caractère robuste ou fragile des individus), la sélection intra-génération masque le vrai patron de sénescence (connu puisque les données ont été simulées) et produit des estimations de la survie âge-spécifique au niveau de la population qui induisent une inférence erronée puisque dans les 2 cas on conclut pas à de la sénescence (la ligne rouge dans les 2 figures). Pour corriger le problème d'hétérogénéité individuelle dans la probabilité de survie, il faut ajuster aux données de CMR des modèles spécifiques (cas A : modèle de mélange à deux classes d'individus, voir Encadré 3 ou cas B : modèle avec un effet aléatoire continu sur la survie, voir Encadré 2) qui permettent d'obtenir des patrons de survie conforme au patron de sénescence attendu (cas A : deux courbes bleues ou cas B : ligne bleue). Dans les 2 cas, la probabilité de détection est fixée à 0.7. Ces 2 exemples miment les résultats obtenus par Vaupel and Yashin (1985) dans le cas d'une probabilité de détection égale à 1.

survie en conditions naturelles en tenant compte de ces différents problèmes, il restait néanmoins une difficulté à lever.

1.2 Le problème de l'hétérogénéité individuelle

Un des problèmes pour mettre en évidence la sénescence est l'existence d'hétérogénéité individuelle. Par hétérogénéité individuelle, j'entends une variation systématique entre individus dans les paramètres démographiques (par exemple des individus avec une meilleure probabilité de survie que les autres).

Si l'on néglige l'hétérogénéité sur le processus de détection générée par exemple par des différences comportementales entre individus (P13) ou bien par des spécificités du protocole d'échantillonnage, alors un biais dans l'estimation des variations de la survie avec l'âge peut apparaître. Je ne m'étendrai pas sur l'hétérogénéité de détection ici. Certainement plus connue (mais pas forcément plus répandue [1] que l'hétérogénéité dans la détection), l'hétérogénéité individuelle dans la probabilité de survie peut être générée par des différences génétiques (voir chapitre 4) ou non-génétiques comme des différences dans l'allocation des ressources (voir chapitre 2). Ce phénomène bien géré dans l'analyse de données de survie chez l'humain (Vaupel et Yashin 1985) est à l'origine des mêmes biais dans l'analyse des données de CMR. L'hétérogénéité individuelle dans la probabilité de survie peut mener à des patrons à l'échelle de la population (courbes en rouge dans les Figure 2A et 2B) qui ne sont pas toujours représentatifs de la vraie relation survie vs. âge à l'échelle individuelle (courbes en noir dans les Figure 2A et 2B).

L'explication est assez simple et tient à un changement dans la composition de la population avec l'âge. Comme la proportion d'individus avec une forte probabilité de survie a tendance à augmenter avec l'âge (sélection phénotypique intra-génération), la probabilité moyenne de survie de la population, utilisée dans la plupart des études sur la sénescence, peut ne pas décroître et même augmenter avec l'âge. Une telle hétérogénéité dans le risque de mortalité pourrait expliquer l'apparente contradiction entre les prédictions de la théorie de l'évolution de la sénescence et les observations. Les analyses de CMR ne font pas exception, et l'impossibilité jusqu'à récemment de pouvoir intégrer l'hétérogénéité individuelle dans la probabilité de survie pourrait expliquer certains échecs pour détecter la sénescence (e.g. Nichols et al. 1997).

1. L'hétérogénéité individuelle est aussi prise en compte depuis récemment dans l'étude de la sénescence de reproduction (van de Pol et Verhulst 2006).

Encadré 2. Hétérogénéité sous forme d'un effet aléatoire

Pour prendre en compte l'hétérogénéité individuelle, nous utilisons un effet aléatoire individuel sur les paramètres qui permet d'estimer une moyenne pour ce paramètre ainsi qu'une variabilité individuelle autour de cette valeur.

Dans un papier récent, Royle (2008) développe une approche bayésienne pour ajuster ce modèle. En bref, l'approche bayésienne suppose qu'aux paramètres sont associées des distributions plutôt que des valeurs ponctuelles comme dans l'approche fréquentiste classique (voir plus bas). On part alors d'une distribution a priori non-informative pour les paramètres (une loi uniforme entre 0 et 1 pour la probabilité de détection par exemple) que l'on met à jour via l'information contenue dans les données de CMR pour obtenir la distribution a posteriori des paramètres. L'a posteriori est proportionnel au produit de la vraisemblance par l'a priori : c'est une application directe de la formule de Bayes $\Pr(A \text{ sachant } B) \propto \Pr(B \text{ sachant } A)\Pr(B)$ où A est le vecteur des paramètres et B les données. En pratique, on génère des observations de la distribution des paramètres que l'on obtient via des algorithmes de Monte Carlo par chaîne de Markov (MCMC) : on accepte ou rejette selon certaines règles des valeurs générées via une chaîne de Markov dont la distribution stationnaire est précisément la distribution a posteriori des paramètres. La distribution d'un ou plusieurs paramètres peut être résumée par la moyenne a posteriori (la moyenne empirique des valeurs générées par une approximation de Monte Carlo) et un intervalle de confiance bayésien dit de crédibilité (voir McCarthy 2007 pour une introduction, ainsi que L0 et CL46).

Pour des modèles complexes et / ou des jeux de données conséquents, les algorithmes sont gourmands en temps de calcul et leur utilisation peut donc limiter l'ajustement et la comparaison de plusieurs modèles (P5).

Dans le cas particulier des modèles CMR avec un effet aléatoire individuel, le problème se gère dans le cadre fréquentiste classique. On cherche les valeurs des paramètres qui maximisent la probabilité d'occurence des données, i.e. la vraisemblance : c'est l'estimation par maximum de vraisemblance. Le problème de l'estimation des paramètres dans un modèle avec effets aléatoires est qu'il faut intégrer la vraisemblance par rapport à la distribution de ces effets aléatoires pour obtenir la fonction à maximiser. Dans un travail récent (P1), nous avons proposé une méthode pour approcher cette intégrale, comme il est fait dans les logiciels statistiques de référence comme SAS ou R. En temps de calculs, notre approche est toujours plus rapide que les algorithmes MCMC et repose sur un cadre bien connu des biologistes. L'approche est disponible dans le logiciel E-SURGE (P43).

1.3 Intégrer l'hétérogénéité individuelle dans les modèles de CMR

Pour étudier la sénescence dans les populations sauvages, il faut donc utiliser des modèles statistiques (CMR ou pas) prenant l'hétérogénéité individuelle des paramètres démographiques en compte.

Dans le cas où des caractéristiques individuelles peuvent être mesurées sur le terrain, elles peuvent être utilisées pour rendre compte de l'hétérogénéité individuelle dans les modèles via des covariables individuelles (des mesures morphologiques par exemple, voir Figure 1 et chapitre 3). Cependant, les covariables mesurables ne sont pas toujours suffisantes pour rendre compte de l'hétérogénéité individuelle, précisément puisque la cause exacte de cette hétérogénéité est souvent non identifiée ou non mesurée. Parfois même, on ne peut mesurer ces caractéristiques individuelles (P13).

Si les modèles d'analyse de la sénescence chez les humains ont très tôt intégré une variation individuelle non-expliquée (Vaupel et Yashin 1985), ce n'est que récemment qu'elle a pu être incorporée dans les modèles de CMR. Deux approches existent. Une approche consiste à utiliser un effet aléatoire individuel (P1, P41, Royle 2008 ; voir Encadré 2), une variable aléatoire qui n'est pas directement observée (on parle de variable latente) mais dont la variance permet de quantifier l'écart entre la réponse d'un individu et la réponse moyenne de la population. Dans le contexte de la sénescence, cette réponse est la survie, et chaque individu a son propre risque de mortalité, qu'on appelle souvent « frailty » pour fragilité, une réalisation de la variable latente (Cam et al. 2002). L'autre approche est basée sur des modèles dits de mélange qui considèrent des classes discrètes d'hétérogénéité (Pledger et al. 2003). Comme les modèles avec effets aléatoires, ce sont aussi des modèles à structure cachée (Pradel 2005 ; voir Encadré 3) où l'indicateur d'appartenance d'un individu à l'une des classes est une variable latente (Pradel 2009) [2]. Ces deux approches permettent de détecter de la sénescence de survie en présence d'hétérogénéité, qu'elle soit discrète (courbes bleues dans la Figure 2A) ou continue (courbe bleue dans la Figure 2B).

Dans la section suivante, je m'attarde sur deux cas d'étude qui montrent comment les modèles de CMR récemment proposés pour prendre en compte l'hétérogénéité individuelle non-observée permettent de décrire de manière fiable des patrons de sénescence de survie.

2. Nous avons aussi développé de tels modèles pour une utilisation en biologie de la conservation. Dans une étude sur la recolonisation par le loup des Alpes, nous nous intéressons à l'estimation des effectifs et du taux de croissance de la population. Une difficulté qui surgit dans l'utilisation des modèles de CMR est la présence d'hétérogénéité individuelle dans le processus de détection des loups. On pense qu'elle vient de la structure sociale de l'espèce (les individus dominants seraient plus détectables que les individus dominés). Ce travail a commencé avec le M2 Biométrie et Biologie Evolutive de Sarah Cubaynes (voir P13), et se continue actuellement au travers de la thèse de Lucile Marescot.

Encadré 3. Hétérogénéité sous forme d'un mélange discret d'individus

Pour prendre en compte l'hétérogénéité individuelle, nous utilisons des modèles dits de mélange (Pledger 2000, Pledger et al. 2003) qui incorporent des groupes cachés d'individus. Autrement dit, on suppose qu'il existe plusieurs classes d'individus, mais on ne sait pas à quelle classe un individu appartient.

Ces modèles de mélange pour données de CMR sont un cas particulier (Pradel 2009) de modèles CMR à structure cachée qui permettent de prendre en compte l'incertitude sur l'assignation d'un état à un individu (HMMs pour hidden Markov models ; Pradel 2005). Ces modèles incluent à la fois des états et des observations générées à partir de l'état sous-jacent des individus (voir Encadré 4 pour plus de détails).

Par exemple, supposons que l'hétérogénéité individuelle puisse s'expliquer par deux groupes d'individus aux caractéristiques spécifiques, et que cette hétérogénéité n'affecte que la probabilité de survie. Alors, si l'on distingue les individus de la classe 1 ($C1$) en proportion π des individus de la classe 2 ($C2$) en proportion $1-\pi$, la probabilité de l'histoire $h_i = (1,0,1)$ de l'individu i s'écrit :

$$\pi\phi_1^{C1}(1-p_2)\phi_2^{C1}p_3 + (1-\pi)\phi_1^{C2}(1-p_2)\phi_2^{C2}p_3$$

où ϕ_t^{C1} (respectivement ϕ_t^{C2}) est la probabilité qu'un individu de la classe $C1$ (respectivement $C2$) vivant à l'occasion t survive jusqu'à l'occasion $t+1$.

Si l'on suppose que tous les individus ont la même probabilité de survie (modèle homogène), on retombe sur le modèle CJS (voir Encadré 1). A l'inverse, on peut construire des modèles dans lesquels l'hétérogénéité agit sur d'autres paramètres en plus de la survie (la probabilité de détection par exemple ; voir P13 et P33).

1.4 Des cas d'étude

1.4.1 Modèle de CMR avec frailty sur la survie

Quelques études seulement ont été menées sur la sénescence chez des oiseaux à courte durée de vie montrant que la variation avec l'âge de la survie des adultes ne peut être négligée. Toutefois, ces résultats ont été obtenus en supposant une probabilité de détection égale à 1. Nous avons étudié les patrons de sénescence de survie chez le Cincle plongeur (*Cinclus cinclus*) (S38). A ma connaissance, c'est la première fois que des modèles de CMR incorporant une frailty sur la survie sont proposés pour étudier la sénescence de survie.

Le jeu de données s'étale sur 29 années et la longévité maximale observée chez cette espèce est de 10 ans. Les individus sont marqués à la naissance et on ne considère que les femelles. Pour tenir compte du passage des individus de l'état non-reproducteur à l'état reproducteur, nous utilisons une extension des modèles CMR uniétats à plusieurs états (voir chapitre 2). Concernant le processus de détection, les individus non-reproducteurs ne sont pas observables comme ils errent en dehors de l'aire d'étude et n'ont pas un territoire établi alors que les reproducteurs peuvent être revus, mais avec probabilité < 1.

On se concentre maintenant sur la probabilité de survie annuelle des individus reproducteurs ($\phi(R)$). Nous avons testé plusieurs modèles dont certains sans l'effet de l'âge et d'autres stipulant différentes formes pour la relation survie vs. âge, dont le modèle quadratique :

$$\text{logit}(\phi_{i,t}(R)) = \beta_1 + \beta_2 a_{i,t} + \beta_3 a_{i,t}^2 + b_t + e_i$$

où $a_{i,t}$ est l'âge de l'individu i l'année t. Le terme $b_t \sim N(0, \sigma_b^2)$ est un effet aléatoire temporel pour prendre en compte la variabilité environnementale interannuelle et $e_i \sim N(0, \sigma_e^2)$ est un effet aléatoire individuel pour prendre en compte l'hétérogénéité (voir Encadré 2). Les coefficients de régression β ainsi que les variances des effets aléatoires sont des paramètres à estimer.

Ce modèle quadratique domine les autres modèles, suggérant une sénescence de survie chez le cincle. Ignorer l'hétérogénéité individuelle ne masque pas la sénescence qui est forte ici (Figure 3), mais conduit à une sur-estimation de l'âge de début de sénescence (2.45 [1.07, 2.70] sans contre 2.11 [0.51, 2.83] avec). Le pic dans la probabilité de survie annuelle est atteint pour un âge approximativement de 2 ans, ce qui coïncide avec l'âge de première reproduction dans cette population et correspond à la prédiction que la sénescence devrait commencer peu de temps après le déclenchement de la reproduction, et donc affecter un nombre non-négligeable d'individus, même en conditions naturelles.

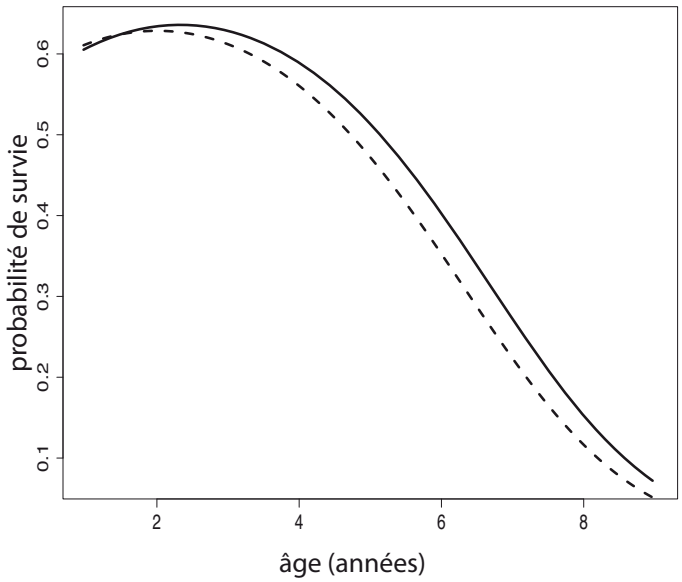


FIGURE 3 – Probabilité de survie en fonction de l'âge dans une population de cincles (femelles reproductrices). Un modèle quadratique est considéré, incluant une hétérogénéité individuelle (tirets) ou pas (trait continu).

1.4.2 Quand l'hétérogénéité de survie n'explique pas tout

La thèse de Guillaume Péron (co-dirigée par P.-A. Crochet et R. Pradel), dans laquelle je me suis investi, nous a permis d'explorer des patrons de sénescence de survie chez la Mouette rieuse (*Chroicocephalus ridibundus*) (P33). Deux différences importantes par rapport à l'étude précédente sur les cincles méritent d'être soulignées. D'abord, nous montrons ici qu'ignorer l'hétérogénéité individuelle sur un autre compartiment que la survie peut aussi masquer la sénescence de survie. Ensuite, nous nous inspirons de notre connaissance du système pour utiliser, plutôt qu'un effet aléatoire continu, un mélange de deux classes d'individus pour modéliser cette hétérogénéité. Les données viennent d'un suivi à long terme sur l'étang de La Ronze, une colonie de reproduction suivie par notre équipe et située dans le centre de la France. Le jeu de données s'étale sur 28 années et la longévité maximale observée chez cette espèce est de 30 ans.

Dans cette population, la probabilité de détection varie entre individus car ils construisent leur nid à l'intérieur de la végétation ou à la lisière. Comme une grande proportion des réobservations est faite sur les nids, on s'attend à une hétérogénéité dans la probabilité de détection. D'autre part, il y a de l'émigration temporaire des individus vers d'autres sites qui ne sont pas suivis, de telle sorte que des individus ratés à plusieurs reprises sur La Ronze peuvent en fait être vivants ailleurs. Ignorer ce phénomène a un impact sur l'estimation de la « vraie » survie par opposition à la survie « locale ». Cette probabilité est modélisée dans un cadre multiétat qui généralise les modèles CMR uniétats (voir chapitre 2) en considérant un état vivant sur La Ronze et un supplémentaire « vivant ailleurs » de détection nulle sous lequel on regroupe les colonies non-suivies (P28). Des analyses préliminaires montrent que cette dispersion varie individuellement, en particulier entre mâles et femelles. On s'attend donc à une hétérogénéité individuelle dans la probabilité d'émigration temporaire. Dans les deux cas, il est difficile d'avoir des indicateurs fiables de ces sources d'hétérogénéité sous la forme de covariables : la détectabilité est difficilement estimable pour des nids qui ne sont pas visibles et l'espèce présente un faible dimorphisme sexuel. Cette hétérogénéité est prise en compte grâce à un mélange de 2 classes d'individus (voir Encadré 3).

Le meilleur modèle montre une sénescence de survie chez la mouette (Figure 4). La comparaison de plusieurs modèles (voir Tableau 2 dans P33) montre une forte probabilité d'hétérogénéité à la fois sur la probabilité de détection et la probabilité d'émigration temporaire (mais pas sur la probabilité de survie). Le deuxième meilleur modèle (« très loin » du premier en AIC) n'incorpore pas d'hétérogénéité individuelle sur la probabilité d'émigration temporaire, et surtout, n'inclut pas de variation de la probabilité de survie avec l'âge [3]. Ce résultat suggère que, dans

3. Dans l'analyse des données de CMR, l'âge d'un individu correspond au temps écoulé depuis sa capture initiale (TFC pour Time Since First Capture). Les mouettes sont marquées poussins et sont donc d'âge connu, mais elles rentrent dans l'étude comme adultes à un âge variable. Le TFC ne correspond donc pas à l'âge véritable. Pour modéliser cet âge réel, il faut en pratique contraindre la survie à varier avec le temps via autant de groupes qu'il y

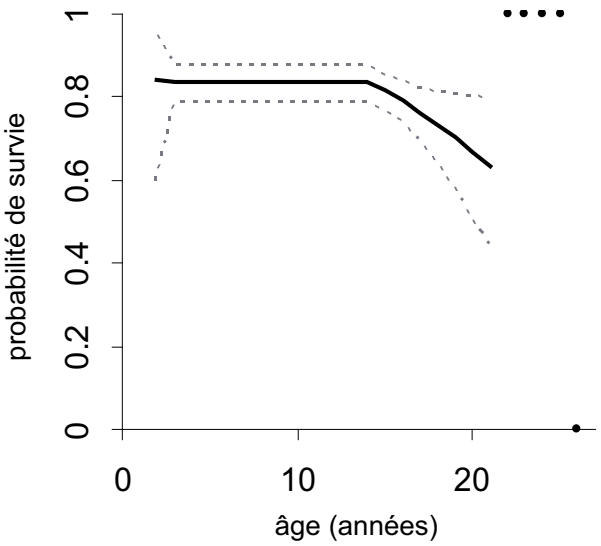


FIGURE 4 – Probabilité de survie chez la Mouette rieuse en fonction de l'âge. Les paramètres estimés sont ceux d'un modèle incluant une hétérogénéité individuelle (mélange de 2 classes d'individus) sur les probabilités de détection et d'émigration temporaire, et une relation contrainte par morceaux de la probabilité de survie en fonction de l'âge. Les lignes en pointillés correspondent aux limites de l'intervalle de confiance à 95%, et les points noirs à des estimations aux borne qui viennent sans erreur standard (voir chapitre 2). La fonction par morceaux reliant la probabilité de survie et l'âge est construite comme suit : on considère une probabilité de survie à 1 an distincte, puis un plateau qui s'étend jusqu'à 14 ans (cette valeur est espèce spécifique et peut s'obtenir grâce à un modèle développé par Guillaume Péron qui prédit le déclenchement de la sénescence en fonction de traits de début d'histoire de vie, S36) puis une décroissance linéaire au-delà.

cette étude, la sénescence de survie est masquée par l'hétérogénéité individuelle dans l'émigration temporaire.

On pouvait s'attendre à un tel résultat puisqu'un événement d'émigration temporaire se produisant à la fin de la vie d'un individu, s'il n'est pas suivi d'une détection, peut passer pour une mort précoce. Ainsi, l'hétérogénéité dans la probabilité d'émigration temporaire peut créer des patrons similaires à ceux générés par l'hétérogénéité dans la probabilité de survie (voir Figures 2A et 2B) et masquer la sénescence. Néanmoins, il semble que ce résultat soit en fait plus du à un défaut d'ajustement du modèle affectant le processus de sélection de modèles (P33).

1.5 Conclusions et perspectives

Grâce à l'incorporation de l'hétérogénéité individuelle, les modèles de CMR rattrapent un certain retard sur les modèles de sénescence utilisés en démographie humaine. Intégrer un effet aléatoire continu revient à considérer qu'il y autant de classes de mélange que d'individus. Si l'on n'a aucune hypothèse a priori sur le nombre de classes, un exercice de sélection de modèles peut permettre de déterminer la structure d'hétérogénéité la mieux adaptée (P1). Si la présence d'hétérogénéité non expliquée dans les données fourni du travail intéressant au statisticien, il me semble que le recours à de tels modèles fait toutefois perdre en clarté du point de vue du biologiste. A quoi peut-on attribuer cette hétérogénéité ? Que représente-t-elle ? Au risque d'enfoncer une porte ouverte, on ne saurait trop encourager à utiliser si possible des covariables individuelles pour réduire cette hétérogénéité, voire à réfléchir en amont aux indicateurs à mesurer sur le terrain pour la prendre en compte.

Influence de l'environnement sur la sénescence de survie. Si les développements de méthodes statistiques permettent de documenter de plus en plus la sénescence de survie dans des populations naturelles, rares sont les études qui ont identifié les facteurs abiotiques et biotiques impliqués dans la sénescence. En particulier, une forte densité a un impact négatif sur la survie et la croissance des jeunes d'une population. Pourtant, chez les espèces à longue durée de vie, beaucoup d'individus survivent à cette période stressante de leur développement, et les effets de ces

a d'âges d'entrée dans l'étude. Cette procédure est très gourmande en temps de calcul (pour un même modèle, 24h avec l'âge réel vs. 1h avec le TFC !) et peu réaliste si l'on veut ajuster plusieurs modèles. Nous utilisons donc le TFC comme un proxi de l'âge réel. Cet argument pratique est appuyé par un argument statistique. Une analyse de puissance a montré que c'est la taille d'échantillon plutôt que le fait d'utiliser le TFC vs. l'âge réel qui est le facteur critique dans la détection de la sénescence de survie (Crespin et al. 2006). Le problème est qu'en utilisant le TFC, on mélange des individus d'âge réel différent, ce qui crée du bruit et augmente l'erreur standard associé à l'effet de l'âge sur la survie. Ici, nous considérons que la sélection de modèles est conservative vis-à-vis de la détection de la sénescence de survie. Par acquis de conscience, nous avons vérifié que, pour le meilleur modèle, utiliser l'âge réel ou le TFC produisait des résultats similaires.

conditions sur leur survie plus tard au cours de leur vie restent peu explorés. Grâce à un suivi sur 40 ans d'une population d'Oies des neiges (*Chen caerulescens*) en croissance rapide, nous proposons d'étudier comment, dans une population dont la densité croît dans le temps et l'espace, la sénescence de survie est affectée. Ce projet fait l'objet d'une demande de financement NSF pilotée par Dave Koons (Université de l'Utah) dont je suis co-responsable.

Tester la pléiotropie antagoniste en conditions naturelles. La sénescence évoluerait comme résultat de deux mécanismes génétiques non mutuellement exclusifs [4]. En particulier, la théorie de la pléiotropie antagoniste prédit un compromis génétique entre investissement précoce dans la reproduction et performance tard dans la vie. Les expériences menées en laboratoire ont montré l'existence de tels compromis, mais on sait peu de choses sur l'importance de ces processus dans les populations naturelles. Pour tester cette théorie, il faudrait par exemple pouvoir étudier une covariation entre l'âge de première reproduction, l'âge de début de sénescence, et le rythme du déclin de la survie avec l'âge (e.g. Nussey et al. 2008). En pratique, il s'agit de modéliser conjointement plusieurs paramètres démographiques et d'estimer la covariation entre ces paramètres au niveau individuel. Cette idée est un volet du projet de délégation CNRS que nous avons construit avec Emmanuelle Cam (Université de Toulouse) et sur lequel nous travaillerons lors de son séjour dans notre équipe (à partir de la rentrée 2010 si la délégation est, je l'espère, acceptée). Ce projet est directement relié au chapitre suivant.

4. La sénescence évolue comme la conséquence a) de la pléiotropie antagoniste selon laquelle des gènes pléiotropiques avantageux en début de vie ont des effets délétères à des âges avancés (Williams 1957), ou b) de l'accumulation de mutations qui exercent des effets négatifs aux âges avancés (Medawar 1952, Hamilton 1966).

2

Mettre en évidence des compromis évolutifs

2.1 Compromis évolutifs chez des populations sauvages

L'évolution des traits d'histoire de vie est une question fondamentale de la théorie de l'évolution. Si ces traits étaient indépendants, les individus optimiseraient chaque trait de manière à maximiser leur valeur sélective. Or, comme les ressources sont limitées, les individus doivent décider de comment allouer ces ressources aux différentes fonctions qui assurent leur survie et leur reproduction (Van Noordwijk et De Jong 1986). Si deux traits dépendent positivement d'une même ressource, alors ils sont reliés négativement l'un à l'autre : on parle d'un compromis évolutif (Roff 1992 ; Stearns 1992). Des compromis souvent étudiés sont ceux liant reproduction et survie ainsi que reproduction et reproduction. L'on peut ainsi se demander : la survie dépend-elle ou non du statut reproducteur et, sachant qu'un individu est vivant, son statut reproducteur à la date t a-t-il une influence sur son statut reproducteur à la date $t + 1$?

L'étude et la mise en évidence de compromis en conditions naturelles sont difficiles. D'abord, le suivi exhaustif des individus est impossible en conditions naturelles, survie et reproduction ne sont donc que partiellement observées : si un individu n'est pas détecté, est-il vivant ou mort ? S'il est vivant, est-il reproducteur ou pas ? De plus, prendre en compte l'hétérogénéité individuelle dans les analyses de compromis est importante du fait d'inégalités possibles entre les individus

dans l'acquisition des ressources. Cette variabilité dans la qualité individuelle peut interférer dans la détection de compromis, les individus de meilleure qualité ayant par exemple des fortes probabilités de survie et de reproduction contrairement à la prédiction d'un coût de la reproduction sur la survie.

Dans ce chapitre, j'introduis deux approches qui permettent d'étudier des compromis évolutifs en conditions naturelles via l'utilisation de données de CMR. D'abord, je présente des modèles dits multiétats qui ont fait l'objet de ma thèse et qui généralisent les modèles uniétats (Encadré 1). Mon travail de l'époque a plutôt porté sur le développement de méthodes statistiques pour ces modèles et présente un intérêt limité pour le biologiste (traduction : le lecteur pourra passer rapidement sur cette partie sans conséquence sur la lecture du reste du document). Je présente dans une seconde partie un cadre de modélisation dit à espace d'états que j'ai proposé récemment et dont les applications occupent une part de mes recherches actuelles. Dans la thèse de Mathieu Buoro en particulier, nous cherchons à mettre en évidence des compromis chez le Saumon Atlantique en tenant compte de la qualité individuelle.

2.2 Modèles de CMR multiétats

Nichols et al. (1994) proposent les modèles de CMR multiétats comme un outil pour l'étude des compromis évolutifs entre paramètres démographiques. Dans les modèles multiétats, les individus peuvent non seulement survivre au cours du temps comme dans les modèles uniétats (Encadré 1), mais aussi se déplacer entre différents sites ou états (Hestbeck et al. 1991). Si l'on considère 2 états A et B, l'histoire d'un individu est constituée de 0 (non-détection), A (détection dans l'état A) et B (détection dans l'état B). En termes de paramètres, les probabilités de survie ou de détection peuvent dépendre de l'état et du temps. On introduit en plus les probabilités de transition, $\psi_t(\mathrm{p} \rightarrow \mathrm{q})$, qu'un individu vivant à l'occasion t dans l'état $\mathrm{p} = \mathrm{A}$ ou $\mathrm{p} = \mathrm{B}$ se « déplace » vers l'état $\mathrm{q} = \mathrm{A}$ ou $\mathrm{q} = \mathrm{B}$ entre les occasions t et $t + 1$. Les états peuvent être des sites géographiques mais aussi des états définis au niveau individuel comme un état reproducteur ou non-reproducteur. Ces modèles permettent d'aborder une grande variété de questions en écologie évolutive (Nichols et Kendall 1995 ; Lebreton et al. 2009) comme l'évolution de la dispersion, la sélection de l'habitat ou l'accession à la reproduction. Malgré leur potentiel, ces modèles sont gourmands en paramètres et sont à l'origine de problèmes statistiques qui n'existent pas ou plus pour les modèles uniétats. Comment évaluer la qualité de l'ajustement d'un modèle multiétat aux données ? Quels sont les paramètres estimables ? Comment estimer ces paramètres et avoir une idée de la précision de ces estimations ? Dans cette section, je prends le prétexte de l'étude des compromis évolutifs pour illustrer ces problèmes. Si l'état reproducteur est noté R et l'état

non-reproducteur NR, un coût de la reproduction l'année t sur la survie jusqu'à l'année $t+1$ est mis en évidence par $\phi_t(\text{R}) < \phi_t(\text{NR})$, alors qu'un coût de la reproduction l'année t sur une reproduction future est démontrée si $\psi_t(\text{R} \rightarrow \text{R}) < \psi_t(\text{NR} \rightarrow \text{R})$ (e.g. Townsend et Anderson 2009 pour une application récente).

2.2.1 Tests d'ajustement

Dans toute analyse de données, la construction d'un modèle qui ajuste les données [1] est une étape cruciale puisque c'est sur lui que reposent l'estimation et l'inférence. Le choix d'un modèle erroné est la cause d'un biais dans l'estimation des paramètres ou la source d'erreurs dans les procédures de sélection de modèles [2].

Pour les modèles de CMR uniétat, l'approche classique d'un test du χ^2 entre observés et attendus est impossible du fait d'un grand nombre d'histoires dont la probabilité est faible, voire nulle. L'alternative repose sur des caractéristiques propres des modèles de CMR qui ont une signification biologique. Un partitionnement astucieux de la statistique de test en composantes spécialisées permet de tester ces hypothèses (un mélange d'individus de passage et d'individus résidents qui peut biaiser la survie ou bien une hétérogénéité dans la détection générée par un effet du piégeage par exemple) en gagnant du même coup en puissance [3]. S'il y un défaut d'ajustement, on peut alors proposer une modélisation plus fine (voir P9 pour une revue).

Nous avons proposé une généralisation de cette approche au cas des modèles multiétats (P12). Une particularité de cette procédure est qu'elle met en lumière le rôle du phénomène dit de mémoire comme une alternative plus plausible au modèle multiétat standard. Je pense ici à un modèle qui spécifie que les transitions à partir du site occupé à une date t dépendent aussi du site occupé à la date $t-1$. Ces modèles permettent d'explorer l'hypothèse d'une fidélité d'oiseaux à leur site d'hivernage (Hestbeck et al. 1991). Si l'on remplace les sites par des états reproducteur et non-reproducteur, cette approche permettrait de tester formellement l'existence d'une qualité individuelle.

1. Par qualité de l'ajustement d'un modèle aux données, j'entends que la distance entre les valeurs prédites par ce modèle et les données observées est faible ; cette notion est à distinguer de l'étape de sélection d'un meilleur modèle par AIC par exemple, qui, si l'étape du test d'ajustement est ignorée, peut conduire à sélectionner un moins mauvais modèle parmi un ensemble de modèles tous mauvais.

2. La problématique des tests d'ajustement fait aussi l'objet d'un chapitre de la thèse de François Guilhaumon que je co-encadre avec David Mouillot. Il s'agit de proposer une démarche pour tester la qualité de l'ajustement des modèles aires-espèces très utilisés en écologie (voir S44, P49 et P51).

3. On parle de test directionnel, par opposition à un test d'hypothèses omnibus dont l'hypothèse alternative n'est pas spécifiée dans une direction particulière. On peut penser à un test de moyenne unilatéral vs. un test bilatéral par exemple.

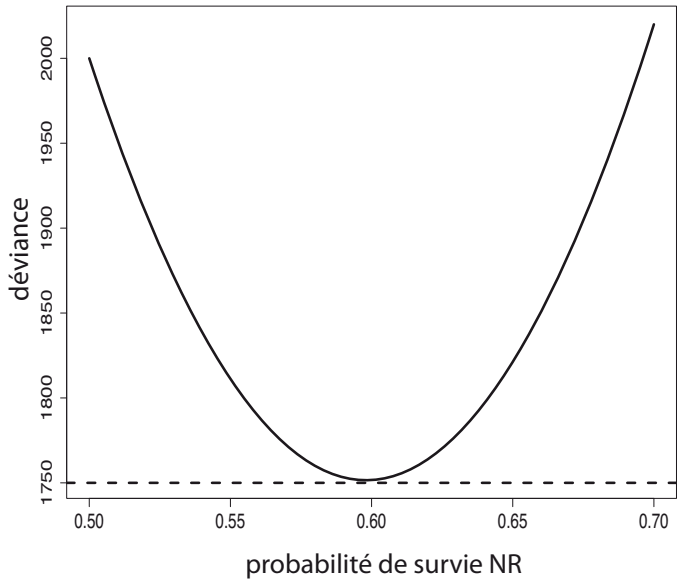


FIGURE 5 – Redondance en paramètres pour un modèle à 2 états reproducteur (R) vs. non-reproducteur (NR) sur des données simulées. L'état non-reproducteur n'est pas observable (la probabilité de détection dans l'état NR est nulle), et les probabilités de survie et de transition entre les états dépendent de l'âge. Le graphique montre la déviance de profil comme une fonction de la probabilité de survie des individus adultes non-reproducteurs. Modèle non-identifiable (tirets) : les survies des adultes reproducteurs sont différentes des adultes non-reproducteurs. La déviance ne change pas avec la survie des adultes non-reproducteurs. Il y a une infinité de minima ce qui signifie que la survie des adultes non-reproducteurs est non-identifiable. Modèle identifiable (trait continu) : les survies des adultes reproducteurs et non-reproducteurs sont contraintes à être égales. La survie des adultes devient estimable. Ce diagnostic peut être fait de manière automatique en utilisant une approche formelle (P11, voir revue dans P10). En appliquant cette démarche, on trouve que pour le modèle identifiable (trait continu), tous les paramètres sont estimables. En revanche, dans le modèle non-identifiable (tirets), seuls les paramètres survie adulte des reproducteurs et probabilité de détection sont estimables. Les autres paramètres (survie des jeunes, survie adulte des non-reproducteurs et paramètres de transition) ne sont pas séparément estimables mais participent à l'estimation dans les fonctions identifiables non montrées ici. Ce diagnostic peut se faire pour un petit nombre d'occasions de détection à l'aide de logiciels de calcul formel type Maple ou Mathematica. Pour généraliser à un nombre quelconque d'occasions, on a recours à un raisonnement par récurrence (P11).

2.2.2 Redondance en paramètres

On dit qu'un modèle est redondant en paramètres si sa vraisemblance peut s'exprimer comme une fonction de moins de paramètres que le nombre original (Catchpole et al. 1996), sinon on dit qu'il est de plein rang (Figure 5). Il est fréquent que des modèles biologiquement pertinents soient surparamétrés et par conséquent, que certains paramètres ne puissent être estimés séparément [4]. Or, identifier quels paramètres (ou fonctions de paramètres) et combien sont séparément estimables est crucial (1) pour une sélection de modèles fiable basée sur l'AIC par exemple (puisque le nombre de paramètres intervient dans le calcul de ce critère) et (2) tout simplement pour l'interprétation des estimations obtenues.

Il existe des méthodes numériques pour diagnostiquer la redondance en paramètres, mais elles sont peu fiables et ne permettent pas de déterminer quels les paramètres ou fonctions de paramètres sont estimables dans les modèles redondants (voir revue dans P10).

Nous avons montré comment vérifier si tous les paramètres d'un modèle multiétat sont estimables (P11 ; voir aussi P28 pour une application). Le cas échéant, nous avons proposé une procédure permettant de réécrire le modèle comme une fonction de paramètres possédant un unique estimateur.

2.2.3 Intervalles de confiance

Généralement, on associe aux paramètres estimés un intervalle de confiance qui reflète l'incertitude d'échantillonnage. Les bornes de l'intervalle de confiance à 95% du paramètre θ sont obtenues comme $\hat{\theta} \pm 1.96 \times \text{SE}(\hat{\theta})$. Construire un tel intervalle (1) repose sur le comportement asymptotique selon une loi normale (le terme ± 1.96 où 1.96 est le quantile à 5% d'une loi normale) de l'estimateur du maximum de vraisemblance $\hat{\theta}$, et (2) nécessite de pouvoir estimer la variance (le terme $\text{SE}(\hat{\theta})$ qui est l'écart-type de l'estimateur).

Bien utile pour le statisticien, la notion d'asymptotique est toute relative pour le biologiste. Pour des données récoltées sur le terrain, la taille d'échantillon est limitée. Le biais dans l'estimation est alors important et l'utilisation de la normalité asymptotique douteuse. De plus, quand l'estimateur d'un paramètre est près de ou sur la frontière de son domaine de définition, sa va-

4. Dans le modèle de CJS par exemple (Encadré 1), la survie au dernier pas de temps et la détection à la dernière occasion ne sont pas identifiables autrement que sous la forme de leur produit. Il faudrait une session de recapture supplémentaire pour pouvoir les estimer séparément.

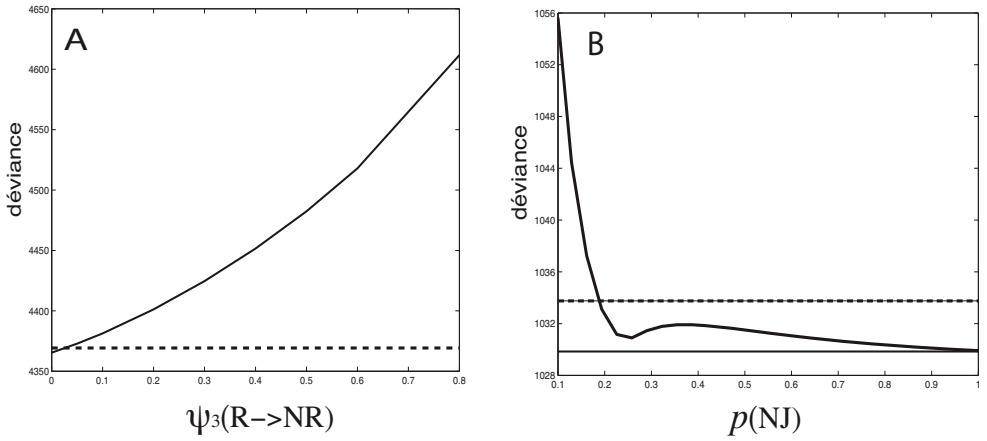


FIGURE 6 – Construction d'un intervalle de confiance basé sur la déviance de profil (P8) pour un paramètre A/ estimé aux bornes et B/ en présence d'un minimum local. Cas A. On considère un modèle multiétat où les probabilités de survie et de recapture ont été contraintes à ne dépendre que de l'état (reproducteur R ou non-reproducteur NR), alors que les probabilités de transition dépendent de l'état et du temps. Ce modèle est ajusté à des données sur la reproduction des Puffins fuligineux (Paul Scofield, comm. pers.). Le paramètre $\psi_3(\text{R} \rightarrow \text{NR})$ est estimé à 0 par maximum de vraisemblance, sa variance est nulle. L'intervalle de confiance à 95% basé sur la déviance de profil est $[0; 0.0266]$; les bornes de celui-ci sont calculées comme les valeurs de $\psi_3(\text{R} \rightarrow \text{NR})$ telle que la déviance (courbe en trait plein) soit égale à son minimum augmenté de 3.84 (ligne en trait pointillé). Cas B. La déviance de profil (courbe en trait plein) est représentée pour un modèle multiétat où les probabilités de survie, de recapture et de transition ont été contraintes à ne dépendre que de l'état et pas du temps (en trait plein). Ce modèle est ajusté à des données de femelles chevreuils (Jean-Michel Gaillard, comm. pers.), avec 3 états considérés sont vivant sans jeunes (SJ), avec un jeune et avec plus d'un jeune. La ligne en tirets est la déviance évaluée en l'estimateur du maximum de vraisemblance (le minimum) qui vaut 1029.9, celle en trait continu est ce minimum augmenté de 3.84. La probabilité $p(\text{SJ})$ de détection dans l'état vivant sans jeunes est estimée à 1 (le minimum global). Il y a un minimum local dans la déviance pour $p(\text{SJ}) \approx 0.25$. L'intervalle de confiance à 95% basé sur la déviance de profil est $[0.1870; 1]$.

riance est quasi voire nulle [5]. Enfin, la présence avérée de maxima locaux [6] dans la vraisemblance de certains modèles multiétats posent problème dans le calcul de l'estimateur du maximum de vraisemblance, affectant l'estimation par intervalle de confiance.

Nous avons considéré la construction d'intervalle de confiance par la vraisemblance (ou déviance) de profil (P8). Les extrémités de l'intervalle de confiance sont calculées en se basant sur la distribution du χ^2 asymptotique de la déviance. On cherche les valeurs du paramètres pour lesquelles, après optimisation par rapport à tous les autres paramètres, la déviance a augmenté de 3.84 (le seuil à 5% de la distribution du χ^2 à 1 degré de liberté). Pour déterminer ces valeurs, on imagine bien combien une procédure manuelle du type essai–erreur pour approcher la valeur 3.84 peut être fastidieuse. En s'inspirant d'un algorithme existant (Venzon and Moolgavkar 1988), nous avons proposé une approche automatisée dont, comme nous l'avons montré dans des simulations, les performances se sont révélées au moins comparables voire meilleures à celles de l'approche classique. De plus, nous avons montré que cette approche fournit un intervalle de confiance dans le cas d'un paramètre estimé aux bornes (Figure 6A) ou de la présence de maxima locaux (Figure 6B).

2.3 Modèles à espace d'états

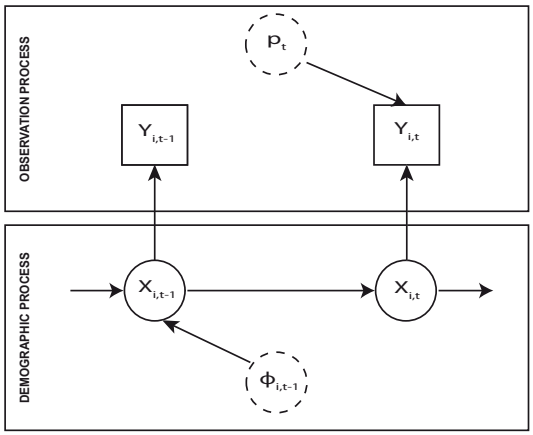
Dans la section précédente, nous avons vu qu'il est possible d'étudier des compromis malgré une probabilité de détection < 1 et plusieurs problèmes statistiques liés aux données CMR et à la structure des modèles multiétats. Récemment, nous avons développé un cadre de travail général pour mettre en évidence des compromis en conditions naturelles en prenant en compte à la fois le problème de détectabilité mais aussi celui d'hétérogénéité individuelle évoqué plus haut. Ce cadre repose sur une nouvelle formulation des modèles CMR (voir Encadré 4) qui permet, en bref, de séparer les processus démographiques qui nous intéressent (l'évolution des traits d'histoire de vie ainsi que des corrélations entre ces traits) et les processus d'observation (la détection des individus). C'est au cours de mon post-doc à StAndrews en Ecosse que j'ai commencé à

5. On imagine un lancé de pièces à 10 reprises qui se solde par l'obtention de pile à chaque coup. Il n'y a pas de variabilité sur l'issue de l'expérience, même si l'on sait que la variance de l'estimateur de la probabilité d'obtenir pile n'est pas nulle. Ce cas se produit dans des données CMR quand par exemple tous les animaux à risque à une occasion survivent jusqu'à l'occasion d'après. La probabilité de survie sur cet intervalle est estimée à 1 et on ne peut y associer une variance.

6. Pour obtenir les estimations des paramètres par la méthode du maximum de vraisemblance (voir Encadré 2), il faut maximiser la vraisemblance (ou minimiser la déviance qui vaut $-2\log(\text{vraisemblance})$). Cette optimisation peut se faire de manière explicite dans quelques cas rares. Généralement, on fait appel à des algorithmes itératifs d'optimisation (type Quasi-Newton). Si la déviance présente plusieurs minima, il arrive que l'algorithme reste piégé dans un de ces minima locaux, alors que l'estimateur du maximum de vraisemblance est atteint pour le minimum global.

Encadré 4. Modélisation à espace d'état des CMR

Tout comme les HMMs (Encadré 3), les modèles à espace d'états (SSMs pour state-space models) stipulent deux séries temporelles qui évoluent en parallèle : l'une pour les états qui évoluent selon une chaîne de Markov, et l'autre pour les observations générées par ces états. Les SSMs sont une généralisation des HMMs à un espace infini d'états, même si en pratique cela fait peu de différence comme le choix se fait sur la base de l'approche pour estimer les paramètres, fréquentiste pour HMM et bayésien pour SSM (voir Encadré 2).
Une représentation graphique des modèles à structure cachée (SSMs ou HMMs) est donnée. .
Le *modèle d'état* est un processus démographique caractérisé par une succession d'états cachés (cercles en trait continu) dans lesquels l'individu évolue selon des probabilités de transition (cercles en tirets). L'état de l'individu i au temps t est vivant ($X_{i,t} = 1$) ou mort ($X_{i,t} = 0$) avec probabilité $\phi_{i,t-1}$ ou $1 - \phi_{i,t-1}$. Le *modèle d'observation* précise que les observations (carrés en trait continu) sont obtenues conditionnellement aux états cachés et aux paramètres associés (cercles en tirets). Si un individu est vivant au temps t, il est observé ($Y_{i,t} = 1$) ou pas ($Y_{i,t} = 0$) avec probabilité p_t ou $1-p_t$. Le processus démographique (la survie ici) est bien caché (au moins partiellement) puisque quand un individu n'est pas vu, impossible de dire s'il est vivant ou pas.



Ces modèles (Pradel 2005, P6) font une distinction entre les paramètres d'observation souvent considérés comme une nuisance (la détection) et les paramètres d'intérêt (la survie) impliqués exclusivement dans le modèle d'état. Cela permet une description aisée de la dynamique et permet de construire des modèles complexes. En particulier, nous utilisons les SSMs à plusieurs reprises dans ce document (compromis évolutifs P34 et ce chapitre ; héritabilité S37 et chapitre 4 ; sénescence S38 et chapitre 1, surface de fitness P39, P41 et chapitre 3).

m'intéresser à cette approche. Dans la thèse de Mathieu Buoro, nous proposons une application de ce cadre pour mettre en évidence des compromis évolutifs chez le Saumon Atlantique (*Salmo salar*) sur la base des données récoltées sur le Scorff en Bretagne (P34).

2.3.1 Quels coûts pour les juvéniles saumons ?

Le saumon Atlantique effectue son cycle de vie en eau douce et dans l'océan. La phase juvénile se passe en eau douce et dure 1 à 2 ans. Il y a ensuite migration vers l'océan puis retour après 1 ou 2 ans vers le cours d'eau natal pour se reproduire. On se concentre sur la phase juvénile durant laquelle les saumons peuvent adopter différentes tactiques. D'abord, ils doivent décider s'ils vont migrer vers l'océan après leur première année de vie ou bien s'ils vont résider en eau douce une année supplémentaire. La migration vers l'océan s'accompagne d'un processus dit de smoltification qui prépare les individus à la vie en eau salée. Ensuite, pour les mâles qui sont en eau douce depuis 2 ans, ils doivent décider de maturer ou pas avant de migrer vers l'océan. Smoltification (migration vers l'océan), maturation (reproduction) et survie sont trois traits qui dépendent des mêmes ressources, les réserves énergétiques accumulées, d'où la prédiction de l'existence de compromis.

2.3.2 Comment mettre en évidence ces coûts ?

Grâce à la distinction que l'on peut faire entre processus démographiques et processus d'observation dans les modèles à espace d'états (voir Encadré 4), il devient aisé de modéliser et tester des coûts. Pour un coût de la migration, on relie la probabilité de survivre au premier hiver $\phi_{1,i}$ pour un individu i à la décision de smoltification $smolt_i$:

$$\text{logit}(\phi_{1,i}) = \alpha_1 + \alpha_2 smolt_i + e_i.$$

Pour un coût de la reproduction, on relie la probabilité de survivre au second hiver $\phi_{2,i}$ pour un mâle i à la décision de maturation mat_i :

$$\text{logit}(\phi_{2,i}) = \delta_1 + \delta_2 mat_i + e_i.$$

Le terme e_i est un effet aléatoire individuel, distribué selon une $N(0, \sigma_e^2)$ qui capture une variabilité individuelle de survie [7] non expliquée par la décision de smoltifier ou de maturer. Les

7. On suppose qu'avoir une survie forte au premier hiver est signe d'un individu en bonnes conditions (fort potentiel de survie) qui se transmet aux autres événements de survie (en particulier la survie au second hiver) via l'effet aléatoire individuel qui a une variance commune entre ces probabilités de survie.

variables binaires $smolt_i$ et mat_i prennent la valeur 1 s'il y a smoltification ou maturation, et 0 sinon. Les paramètres α_2 et δ_2 reflètent donc l'influence des décisions de smoltifier et maturer : si l'un d'eux est différent de 0, alors il existe un compromis.

Comment une possible hétérogénéité individuelle est-elle prise en compte dans la détection des compromis ? De deux manières, selon que cette hétérogénéité est observable ou pas. Pour la migration, on sait que l'âge à la smoltification dépend positivement de la croissance au cours des premiers mois de vie en eau douce. On peut donc représenter cette hétérogénéité via une norme de réaction en écrivant que la probabilité de smoltifier κ_i pour l'individu i est une fonction de sa longueur L_i :

$$\text{logit}(\Pr(smolt_i = 1)) = \text{logit}(\kappa_i) = \beta_1 + \beta_2 L_i.$$

Quant à la reproduction, nous n'avons aucun indicateur mesurable sur le terrain. On écrit donc que la probabilité de maturer ψ_i pour un individu mâle i a une certaine moyenne γ et une variabilité individuelle σ^2 capturée par l'effet aléatoire e_i :

$$\text{logit}(\Pr(mat_i = 1)) = \text{logit}(\psi_i) = \gamma + e_i.$$

L'effet aléatoire dans cette relation entraîne une corrélation positive entre les probabilités de maturer ψ_i et de survie au second hiver $\phi_{2,i}$: un mâle qui a une survie forte est de bonne qualité (il stocke beaucoup d'énergie) et aura donc une forte probabilité de maturer [8]. En incorporant une variabilité individuelle observée ou non-observée, on peut ainsi estimer les compromis au travers d'un différentiel de survie (α_2 et δ_2) en corrigeant pour une éventuelle hétérogénéité.

2.3.3 Des résultats compromettants ?

Le paramètre α_2 est estimé positif avec une probabilité > 0.99 (1.44 [0.66 ; 3.53]), ce qui suggère une meilleure survie en faveur des individus qui décide de smoltifier : cette différence de survie hivernale entre futurs migrants et futurs résidents est estimée positive (0.35 [0.09 ; 0.68]) (Figure 7A) [9]. Ce résultat est contraire à notre prédiction d'un coût de la migration sur la survie, mais pourrait s'expliquer par la variabilité dans l'acquisition des ressources au cours du premier hiver. Les futurs migrants continuent d'acquérir des ressources et maintiennent une croissance ce qui leur permet d'assurer la smoltification et la survie. L'énergie ainsi accumulée en période hivernale permet d'assurer à la fois smoltification et survie. A contrario, les résidents

8. Dans un volet du post-doctorat de Sabrina Servanty, nous avons proposé une méthode qui permet de modéliser explicitement la corrélation entre deux paramètres démographiques, voir S31.

9. La survie hivernale des futurs migrants est plus élevée (0.53 [0.32 ; 0.83]) que celles des futurs résidents (0.19 [0.07 ; 0.33]).

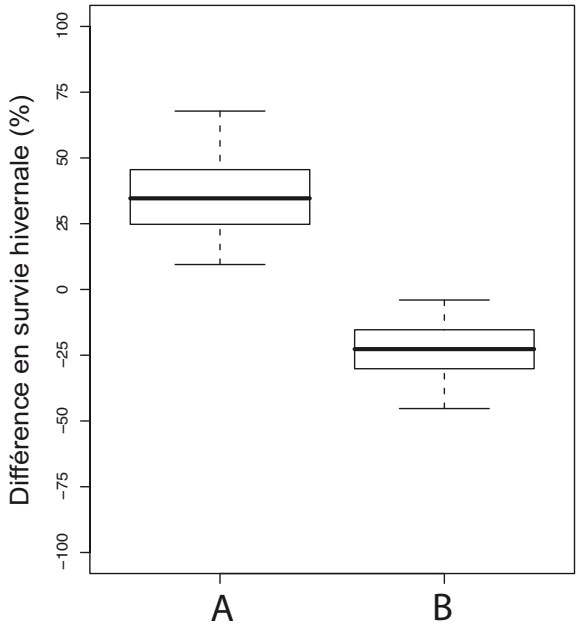


FIGURE 7 – Différentiels de probabilités de survie pour (A) futurs migrants vs. futurs résidents durant le premier hiver (avantage de la migration sur la survie ?) et (B) individus matures vs. immatures durant le second hiver (coût de la reproduction sur la survie ?). Le différentiel de survie est obtenu de la manière suivante : pour la relation entre probabilités de smoltifier et de survivre au premier hiver, on a logit$(\phi_{1,i}) = \alpha_1 + \alpha_2 smolt_i + \epsilon_i$, d'où la survie des individus futurs résidents ($smolt_i = 0$) donnée par logit$^{-1}(\alpha_1)$ et celle des futurs migrants ($smolt_i = 1$) donnée par logit$^{-1}(\alpha_1 + \alpha_2)$. Le différentiel s'obtient alors comme la différence entre survie des futurs migrants et survie des futurs résidents. Idem pour le différentiel de survie au second hiver entre individus matures et immatures. Les diagrammes à moustache sont délimités par les percentiles 2.5 (extrémité inférieure de l'intervalle de crédibilité à 95%), 25, 50 (médiane), 75, 97.5 (extrémité supérieure de l'intervalle de crédibilité à 95%).

réduisent leur activité, deviennent anorexiques et doivent puiser dans leurs réserves énergétiques pour assurer leur survie.

Le paramètre δ_2 est estimé négatif avec une probabilité > 0.97 (-1.41 [-2.92 ; -0.20]), ce qui suggère une survie sélective en fonction du statut de maturation sexuelle, i.e. un coût de la reproduction sur la survie au second hiver : cette différence de survie est estimée négative (-0.23 [-0.04 ; -0.45]) (Figure 7B) [10].

2.4 Conclusions et perspectives

Les modèles multiétats permettent d'étudier les compromis évolutifs entre traits d'histoire de vie en conditions naturelles, en tenant compte d'une détectabilité imparfaite. La reformulation des modèles de CMR comme modèles à espace d'états fiabilise l'étude de ces compromis en facilitant la prise en compte d'une éventuelle hétérogénéité individuelle non-observée [11].

Cette qualité des modèles à espace d'états est toutefois plus le produit de la méthode adoptée pour estimer les paramètres que des modèles. Les modèles multiétats sont traditionnellement ajustés dans un cadre fréquentiste via des approches d'optimisation, les modèles à espace d'états dans un cadre bayésien via des approches MCMC (voir Encadrés 2 et 4). Ces dernières permettent d'incorporer facilement des effets aléatoires, d'où une préférence pour ces modèles lorsqu'on souhaite intégrer une possible hétérogénéité individuelle. Néanmoins, cet avantage n'est plus l'apanage de l'approche bayésienne puisqu'on sait gérer de mieux en mieux des effets aléatoires individuels dans les modèles de CMR dans un cadre fréquentiste (P1, voir aussi Encadré 2).

Une question pratique se pose alors : pourrions-nous modéliser les coûts sur la survie des saumons par une approche multiétat via la méthode du maximum de vraisemblance ? Probablement, mais au prix de gros efforts de programmation vu la complexité du cycle de vie de l'espèce. En comparaison, on peut construire un modèle à espace d'états dans un contexte bayésien assez aisément grâce aux logiciels disponibles [12] sans trop se focaliser sur l'implémentation des méthodes d'estimation des paramètres. Quel intérêt cela aurait-il alors ? Peut-être celui d'avoir des outils disponibles pour faire de la sélection de modèles où chaque modèle correspond à un scénario biologique ? Dans le travail de thèse de Mathieu, nous avons en effet adopté une approche basée sur

10. La survie hivernale d'un mâle mature (0.14 [0.02 ; 0.35]) est plus faible que celle d'un immature (0.39 [0.12 ; 0.62]).

11. L'hétérogénéité observée capturable par des covariables individuelles mesurées sur le terrain est relativement facile à prendre en compte dans les modèles à espace d'états (la longueur dans l'exemple saumon par exemple) et dans les modèles multiétats. Relativement car se pose le problème des valeurs manquantes dans le cas de covariables individuelles dont les valeurs changent au cours du temps : quelle est la valeur de cette covariable quand un individu n'est pas détecté ?

12. Voir par exemple le code écrit par Mathieu pour les saumons www.cefe.cnrs.fr/biom/salmonOpenBUGS.txt ou bien les codes associés à CL46.

un modèle intégrant toutes nos connaissances biologiques sur le système plutôt que l'approche sélection de modèles. Les critères d'information comme l'AIC ne font malgré tout pas consensus quand il s'agit de modèles avec effets aléatoires [13]. L'approche pragmatique s'impose alors.

Certains problèmes statistiques des modèles multiétats que j'ai attaqués pendant ma thèse sont à ma connaissance peu abordés pour les modèles à espace d'états. Comment tester l'ajustement de tels modèles ? Comment déterminer quels sont les paramètres estimables quand la tentation est forte d'ajuster des modèles de plus en plus complexes avec les mêmes données ? Qu'on adopte une démarche fréquentiste ou bayésienne, peu importe puisque ces problèmes ont souvent à voir avec la structure du modèle. Comme les modèles multiétats sont des cas particuliers de modèles à espace d'états (P6 et S31), on peut s'aider des méthodes existantes que j'ai exposées pour les modèles multiétats pour fournir des éléments de réponse pour les modèles à espace d'états (e.g. P13, S31).

Hétérogénéité fixe vs. dynamique. L'articulation entre modèles multiétats et à espace d'états permet d'envisager des applications excitantes. Dans un papier récent, Tuljapurkar et al. (2008) introduisent l'hétérogénéité dynamique définie comme des différences d'histoires de vie entre individus générées par le processus stochastique qui décrit les changements d'états observés dans un modèle multiétat (e.g. chez les saumons : migrant, résident, immature, mature non-reproducteur, mature reproducteur, ...). Les auteurs distinguent hétérogénéité dynamique et hétérogénéité fixe définie comme des différences d'histoires de vie entre individus générées par des différences non-observées entre individus et fixées à la naissance (e.g. le génotype ou des effets maternels). On reconnait par hétérogénéité fixe la notion de frailty introduite au chapitre 1. Sur plusieurs espèces, les auteurs montrent que l'hétérogénéité dynamique suffit à générer la variabilité observée dans les trajectoires de vie. Ils proposent ce modèle pour explorer le rôle des différences de qualité individuelle non-observée. Les populations naturelles sont probablement soumises aux deux types de variabilité individuelle, et un modèle qui incluerait hétérogénéités fixe (via des effets aléatoires) et dynamique (via des transitions entre états) permettrait de quantifier la part relative de chacune. C'est un projet en cours avec Shripad Tuljapurkar, Uli Steiner et Emmanuelle Cam.

13. Si l'on reprend l'exemple chez les juvéniles de saumon d'un possible coût de la reproduction sur la survie au second hiver, $\text{logit}(\phi_{2,i}) = \delta_1 + \delta_2 mat_i + e_i$, quel est le nombre de paramètres à compter dans la formule de l'AIC en plus des 2 paramètres de régression δ_1 et δ_2 ? Doit-on considérer juste 1 paramètre en plus pour la variance de l'effet aléatoire e_i ou bien autant de paramètres que d'individus pour refléter les réalisations i de cet effet aléatoire ? Il n'existe pas de réponse définitive à cette question à ma connaissance.

3

Explorer des surfaces de valeur sélective

3.1 Sélection naturelle en conditions naturelles

Comprendre comment la sélection naturelle agit sur un ensemble de traits phénotypiques est central en écologie évolutive. Quantifier l'action de la sélection sur des traits phénotypiques dans des populations naturelles permet, si ces traits sont héritables, de prédire leur réponse à l'évolution (Kingsolver et al. 2001). Il existe deux approches utilisées en routine pour estimer et visualiser la relation entre la valeur sélective (caractérisée par la survie, le succès reproducteur ou d'autres combinaisons de traits) et un ensemble de traits phénotypiques (surface de fitness).

Premièrement, la méthode proposée par Lande and Arnold (1983) permet d'estimer des gradients de sélection. Techniquement, il s'agit d'une régression multiple dont les termes linéaires, quadratiques et croisés (interactions) - on parle de régression polynomiale du second ordre - correspondent respectivement aux gradients de sélection directionnels, stabilisants ou disruptifs et corrélationnels.

Deuxièmement, même si la méthode de Lande et Arnold fournit une approximation (quadratique) de la surface de fitness, des méthodes non-paramétriques [1] sont aussi utilisés car elles permettent plus de flexibilité dans la visualisation des surfaces. En effet, elles ne requièrent pas

1. Par non-paramétrique, j'entends ici une approche qui n'implique pas une fonction paramétrique connue des paramètres. Rien à voir avec les tests du même nom.

de modèle a priori (linéaire ou quadratique par exemple) pour la relation entre valeur sélective et traits. Schluter (1998) puis Schluter et Nychka (1994) ont introduit l'utilisation de telles méthodes pour visualiser en 2 ou 3 dimensions l'action de la sélection sur des traits phénotypiques.

Pour mettre en oeuvre ces deux approches, il faut évidemment pouvoir estimer la valeur sélective et donc utiliser les modèles de CMR. Par ailleurs, quand les modèles de CMR sont utilisés pour étudier la sélection, on focalise généralement sur des gradients linéaires ou quadratiques. On ne considère que très rarement plusieurs traits, et jamais à ma connaissance on n'a envisagé l'analyse de termes croisés (ou sélection corrélationnelle). Enfin, les outils pour visualiser la sélection n'existent pas. C'est au cours de mon post-doctorat à l'Université du Kent à Canterbury que j'ai commencé à m'intéresser au sujet.

3.2 Méthode analytique d'exploration de la fitness

L'idée qui consiste à étudier l'effet de la sélection sur des traits via une approche CMR a d'abord été proposée par Kingsolver et Smith (1995). Malgré tout, elle n'a été que peu utilisée et presqu'uniquement appliquée à un seul trait phénotypique. Nous avons proposé récemment l'intégration de l'approche de Lande et Arnold à des modèles CMR (P39). On écrit la survie comme la variable réponse d'une régression multiple sur les traits (Brodie et al. 1995) :

$$\text{logit}(\phi_{i,t}) = \beta_0 + \sum_{p=1}^{P} \beta_p x_i^p + 1/2 \sum_{p=1}^{P} \sum_{q=1}^{P} \gamma_{pq} x_i^p x_i^q + e_i + \delta_t$$

où P est le nombre de traits, x_i^p est la valeur du p-ème trait pour l'individu i, les δ_t sont des effets temps fixes et e_i un effet aléatoire individuel. On utilise les effets fixes b_t pour prendre en compte les variations temporelles dans la survie, et l'effet aléatoire e_i pour incorporer une possible variation individuelle (résiduelle) en plus de celle captée par les traits. La signification des paramètres de régression est illustrée dans ce qui suit.

Pour déterminer les combinaisons pertinentes des gradients de sélection, nous voyons le problème comme un exercice de sélection de variables. Nous utilisons une extension des algorithmes MCMC standards (voir Encadré 2 et L0) - MCMC à sauts réversibles (RJMCMC ; Green 1995, CL46, L0) - pour chercher parmi le grand nombre de combinaisons possibles et ainsi trouver le meilleur modèle.

paramètre	modèle linéaire	modèle quadratique	modèle cubique
β_0	0.333 (0.109)	0.473 (0.115)	0.509 (0.117)
β_1	0.180 (0.094)	0.189 (0.096)	0.016 (0.131)
β_2	-	-0.200 (0.072)	-0.268 (0.088)
β_3	-	-	0.081 (0.044)
σ_ϕ	0.861 (0.161)	0.831 (0.163)	0.823 (0.163)
p	0.428 (0.023)	0.428 (0.023)	0.428 (0.023)
AIC (np)	2399.6 (4)	2392.9 (5)	2390.8 (6)

TABLE 1 – Paramètres estimés de 3 modèles décrivant la relation survie vs. masse corporelle chez le Tisserin social. Les β sont les paramètres de régression, σ_ϕ est l'écart-type de l'effet aléatoire individuel sur la survie, p est la probabilité de détection et np le nombre de paramètres.

3.2.1 Un trait

Covas et al. (2002) étudient l'effet de la sélection sur la masse corporelle du Tisserin social (*Philetairus socius*) en Afrique du Sud de 1993-2000 à partir du marquage de 435 jeunes oiseaux. Les auteurs relient la probabilité de survie annuelle de façon respectivement linéaire et quadratique avec la masse (Encadré 1) pour discriminer entre sélection directionnelle (relation linéaire ; $\text{logit}(\phi_{i,t}) = \beta_0 + \beta_1 x_i$) et sélection stabilisante (pic de survie pour des valeurs intermédiaires de la masse : relation quadratique ; $\text{logit}(\phi_{i,t}) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$). On se ramène à la méthode de Lande et Arnold (1983) si l'on suppose que la détectabilité est parfaite. L'approche par CMR sélectionne le modèle quadratique et plaide donc pour une sélection stabilisante, avec une survie optimale autour de la masse moyenne (Tableau 1). Il s'agit du modèle décrit plus haut avec 1 trait ($P = 1$). On verra néanmoins dans la section suivante que l'analyse est incomplète et qu'une approche non-paramétrique suggère un autre scénario.

3.2.2 Plusieurs traits

Plus récemment, nous avons étudié l'action de la sélection sur un ensemble de traits phénotypiques du merle commun (*Turdus merula*) à Dijon de 1998-2002 à partir du marquage de 199 oiseaux (P39 ; voir Grégoire et al. 2004 pour plus de détails sur l'étude). Nous avons considéré les traits suivants : longueur du tarse, longueur du phalanx, hauteur du bec, longueur de l'aile et longueur de la rectrice. Plusieurs de ces traits étant fortement corrélés, nous avons d'abord effectué une analyse en composantes principales (Tableau 2).

L'interprétation des composantes principales ne pose pas de problème particulier. Par exemple, la composante PC3 est expliquée par une contribution forte et positive de la hauteur du bec et sera interprétée comme telle. Ou encore, la composante principale PC4 oppose la longueur de l'aile et la longueur de la rectrice, deux traits fortement reliés aux performances de vol ; on l'interprète comme la composante « agilité » (plus la rectrice est longue et les ailes courtes, plus grande est l'agilité).

On applique le modèle décrit plus haut avec $P = 5$. Les résultats montrent qu'il n'y a pas de sélection directionnelle ni quadratique, mais une interaction entre PC3 et PC4 (Tableau 3). En d'autres termes, nous détectons un gradient de corrélation entre PC3 et PC4. Cette corrélation entre la hauteur du bec et l'agilité est négative, ce qui suggère que les merles partagent une combinaison particulière de ces deux composantes, telles qu'une augmentation de l'agilité (correspondant à une augmentation de la longueur de rectrice et une diminution dans la longueur de l'aile, i.e. une diminution des valeurs de PC4) est toujours associée à une augmentation de la taille du bec (i.e. augmentation des valeurs de PC3).

Principal component	PC1	PC2	PC3	PC4	PC5
Bec	0.37	0.50	0.78	0.04	0.08
Tarsus	0.75	-0.55	0.06	-0.15	0.34
Phalanx	0.77	-0.47	0.17	0.27	-0.30
Aile	0.66	0.48	-0.45	0.34	0.15
Rectrice	0.80	0.34	-0.20	-0.41	-0.19

TABLE 2 – Analyse en composantes principales des traits phénotypiques du Merle noir. Les contributions des différents traits aux axes principaux sont données. En particulier, PC3 représente la taille du bec et PC4 l'agilité.

				γ		
	β	PC1	PC2	PC3	PC4	PC5
PC1	-0.00 (0.07)	0.05 (0.16)				
PC2	0.002 (0.11)	-0.00 (0.05)	0.12 (0.39)			
PC3	0.02 (0.15)	0.01 (0.07)	0.04 (0.22)	0.00 (0.08)		
PC4	-0.19 (0.65)	-0.01 (0.16)	0.07 (0.46)	-1.67* (0.48)	-0.22 (0.65)	
PC5	0.01 (0.29)	-0.01 (0.18)	0.01 (0.23)	0.12 (0.58)	0.19 (0.90)	-0.01 (0.42)

TABLE 3 – Tableau regroupant le vecteur des gradients de sélection directionnels (β) et la matrice des gradients de sélection quadratiques et corrélationnels (γ) - sur l'échelle logit. On reporte les médianes a posteriori ainsi que l'écart-type entre parenthèses. Une covariable sélectionnée par l'algorithme RJMCMC est affublée d'un signe *.

3.3 Méthode visuelle d'exploration de la fitness

Plutôt que de passer par une approche paramétrique comme dans la section précédente, l'exploration des surfaces de fitness peut se faire via une approche non-paramétrique.

3.3.1 Un trait

Idéalement, on aimerait pouvoir écrire que

$$\text{logit}(\phi_{i,t}) = f(x_i) + e_i + \delta_t$$

où x_i est toujours la valeur du trait x pour l'individu i, e_i est un effet aléatoire individuel et δ_t sont des effets fixes temporels. La nouveauté ici est la fonction f qui n'est pas forcément linéaire ou quadratique, mais dont la forme est guidée par les données. Dans nos travaux (P39, P41 ; voir aussi P7 et P24 pour des applications à des covariables environnementales), nous avons opté pour une famille particulière de fonctions appelées fonctions splines. Une spline est une fonction construite à partir de morceaux de fonctions linéaires ou quadratiques qui sont mises bout à bout. Les points de jonction sont appelés des noeuds, et le coeur du problème de l'ajustement non-paramétrique revient à déterminer leur nombre et leur position (voir Ruppert et al. 2003 pour une introduction).

Reprenons l'exemple des tisserins. Covas et al. (2002) ont trouvé qu'une relation quadratique expliquait le mieux la relation entre survie et masse corporelle (points, Figure 8). Si l'on ajuste le modèle spline proposé (trait continu, Figure 8), on trouve certaines ressemblances mais aussi des différences marquantes entre les deux modèles (P41). D'abord, en accord avec les résultats de Covas et al. (2002), les individus plus légers présentent une forte mortalité. Ensuite, alors que le modèle quadratique est symétrique, les individus aux extrêmes ont des survies faibles similaires, le modèle spline est très asymétrique, les individus plus lourds survivant mieux que les individus plus légers [2].

Cette réanalyse remet en question la conclusion de sélection stabilisante et montre qu'il est difficile de la distinguer d'une sélection directionnelle en faveur des individus les plus lourds. Malgré tout, l'utilisation des fonctions splines peut permettre d'aller un cran plus loin dans la modélisation paramétrique de la section précédente en suggérant un modèle alternatif. Dans le cas des tisserins, l'ajustement de la spline suggère qu'un modèle cubique pourrait être plus approprié qu'un modèle quadratique (tirets, Figure 8), ce qui est confirmé par la sélection de modèles (Table 1).

2. Ces conclusions sont basées sur un petit nombre d'observations aux extrêmes du phénotype et sont donc à prendre avec précaution.

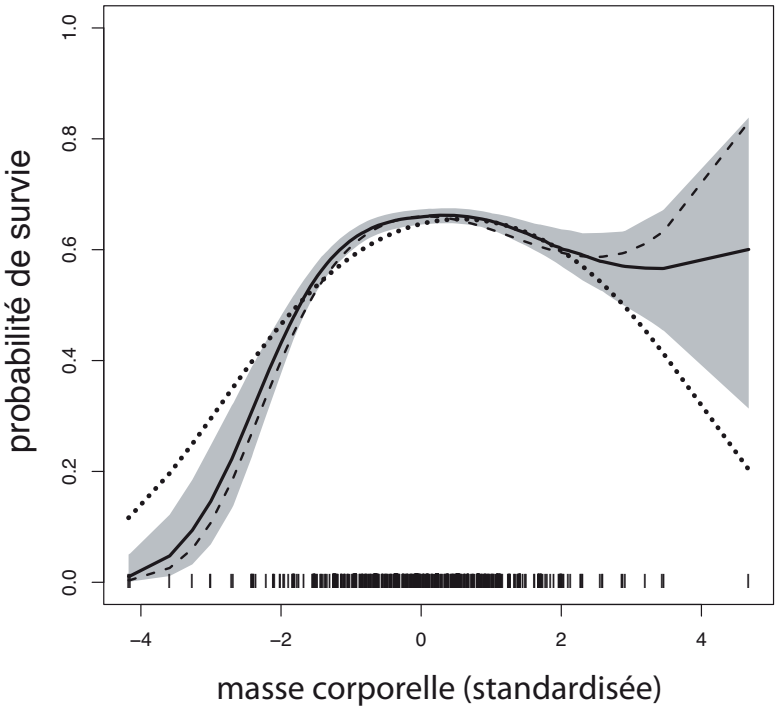


FIGURE 8 – Relation entre survie et masse corporelle chez le Tisserin social. On reporte les médianes a posteriori (trait continu) et l'intervalle de crédibilité à 95% (aire grisée) pour le modèle spline, ainsi que les médianes pour les modèles quadratique (points) et cubique (tirets). Les valeurs de la masse sont aussi reportées sur l'axe des abscisses.

3.3.2 Plusieurs traits

Pour étendre le modèle spline à deux traits, nous adaptons des outils développés en géostatistique pour la construction de carte d'abondance d'espèces par exemple (P41 ; voir aussi CL29 pour une application à des covariables environnementales). Notre objectif ici est de détecter des patrons intéressants qui ne sont pas juste dus à de la variabilité échantillonnale. On considère un modèle dit semi-paramétrique (voir P7) dans lequel deux traits p et q entrent dans le modèle sous forme d'une interaction non-paramétrique, et les traits restants entrent dans le modèle dans une composante linéaire :

$$\text{logit}(\phi_{i,t}) = f(x_i^p, x_i^q) + \sum_{s=1, s\neq p,q}^{P} \beta_s x_i^s + e_i + \delta_t$$

où x_i^p et x_i^q sont les valeurs des deux traits considérés pour l'individu i, et f est une fonction spline à 2 dimensions.

Appliquons ce modèle à l'étude du Merle noir, en particulier sur les deux composantes principales PC3 (taille du bec) et PC4 (agilité). La représentation graphique de la surface de survie (Figure 9) confirme la sélection corrélationnelle détectée dans la section précédente. La surface présente une crête de survie forte (d'en haut à gauche à en bas à droite), ce qui est cohérent avec la corrélation négative trouvée dans l'analyse paramétrique (voir aussi le contour en tirets rouges illustrant cette corrélation estimées par le modèle paramétrique).

3.4 Conclusions et perspectives

Dans ce chapitre, nous proposons un cadre pour l'étude de la sélection sur des traits phénotypiques via l'utilisation de données CMR. Il s'agit d'outils généraux, flexibles et complémentaires qui permettent de quantifier la force de la sélection sur une combinaison de traits (extension de la méthode de Lande et Arnold pour les données CMR) et de visualiser la forme de la sélection (extension de la méthode de Schluter pour les données CMR).

La méthode paramétrique permet d'étudier les gradients de sélection de manière formelle, comme cela se fait avec des données classiques pour lesquelles on n'a pas le problème de détectabilité, et elle permet ainsi de mettre en évidence des patrons de sélection non-linéaires, comme dans l'exemple merles. La méthode d'ajustement par splines permet de capturer des patrons dans la surface de fitness qui ne sont pas capturés par l'approche paramétrique, comme dans l'exemple des tisserins.

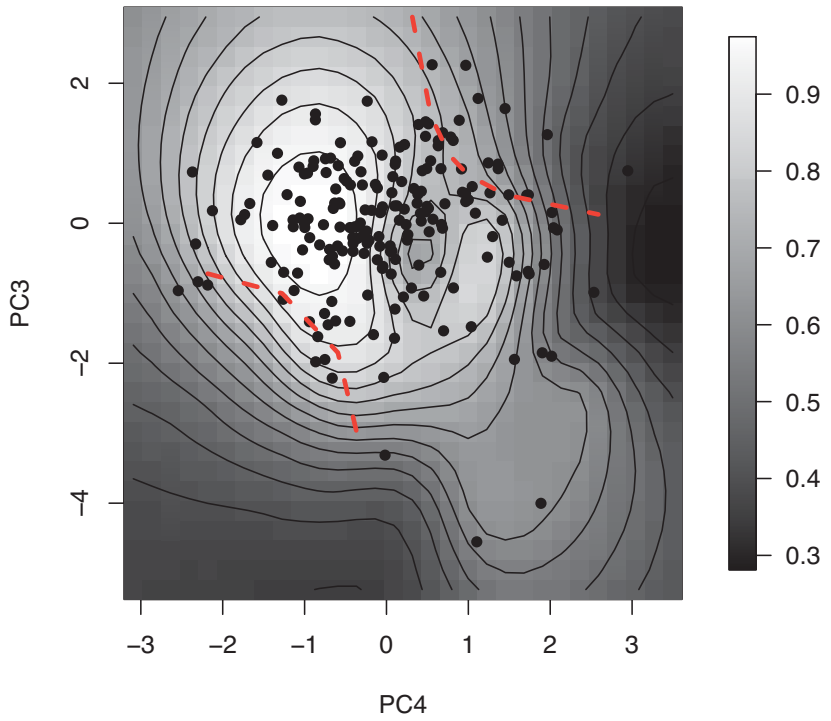


FIGURE 9 – Visualisation de la surface de survie comme une fonction de PC4 (axe des abscisses ; agilité) et PC3 (axe des ordonnées ; taille du bec) pour le Merle noir. Un contour bi-dimensionnel des médianes a posteriori est utilisé. Pour comparaison, le modèle paramétrique (basé sur notre méthode analytique) est illustrée par le contour correspondant à une survie égale à 0.8 (tirets rouges).

Une approche analytique-visuelle unifiée ? J'insiste sur le fait que l'approche analytique fait l'hypothèse, comme l'approche de Lande et Arnold (1983), que la forme sous-jacente de la surface de fitness est effectivement quadratique, alors qu'elle pourrait avoir une toute autre forme (avec un ou plusieurs pics par exemple). Pour relâcher cette hypothèse, Schluter et Nychka (1994) proposent une méthode - la régression par poursuite de projection - qui réduit le nombre de dimensions à la manière de l'analyse en composantes principales, mais utilise aussi des splines pour relier ces nouvelles variables synthétiques à la valeur sélective. Avec une stagiaire M2 Biostatistique (Soad Bouhazama), nous avons obtenu des résultats encourageants pour l'adaptation d'une méthode équivalente pour les CMR.

Comment comparer des gradients de sélection ? La méthode visuelle ne fournit pas des gradients de sélection facilement interprétables et comparables à d'autres études. Elle permet néanmoins de suggérer des modèles analytiques alternatifs aux formes classiques. Dans ce cas, les gradients sont obtenus sur l'échelle logit et sont donc difficilement comparables avec d'autres valeurs obtenues dans d'autres études. Une linéarisation de la fonction logit par développement de Taylor résout le problème pour des valeurs de survie autour de 0.5, mais pour les espèces à longue durée de vie par exemple, il faut aller plus loin.

Le problème des valeurs manquantes. Dans les approches évoquées dans ce chapitre, on a considéré des covariables individuelles (les traits phénotypiques) qui ne varient pas au cours du temps. Même si nos modèles permettent une dépendance vis-à-vis du temps, le problème de données manquantes se pose alors puisque lorsqu'un individu n'est pas capturé, quelle valeur du trait lui attribuer ? A l'heure actuelle, il existe deux approches pour pallier à ce problème. On peut discrétiser la covariable (par exemple, pour le tisserin, trois classes de poids : léger, intermédiaire et lourd) pour se ramener à des modèles multiétats (chapitre 2) dans lesquels les transitions entre états gèrent le problème des valeurs manquantes. Une autre approche consiste à « remplir les trous » en modélisant la distribution des valeurs manquantes. Cela demande de faire des hypothèses sur les changements observés au cours du temps dans la covariable sur les individus capturés (e.g. CL46). A ce jour, les avantages et inconvénients de ces deux approches n'ont pas été évalués pour l'étude de la sélection sur des traits phénotypiques.

4

Quantifier l'héritabilité de paramètres démographiques

4.1 Héritabilité des traits de populations sauvages

La base génétique des traits est au cœur de la sélection naturelle. Comprendre l'évolution par sélection naturelle, prédire son rythme et sa direction sont autant d'étapes qui passent par la compréhension de l'influence génétique et environnementale sur ces traits.

Les modèles de génétique quantitative (Lynch et Walsh 1998) ont pour but de séparer les sources de variation phénotypiques en analysant des données sur les traits quantitatifs (morphologiques, démographiques, comportementaux ou physiologiques) ainsi que sur les relations de parenté entre les individus (pedigree). En particulier, le « modèle animal » permet, via l'utilisation d'effets aléatoires, d'estimer simultanément la part de variance phénotypique qui peut être attribuée à des facteurs génétiques, environnementaux ou d'autres facteurs inconnus (Kruuk 2004). L'héritabilité d'un trait est calculée comme la part de variance expliquée par les effets génétiques additifs.

Alors que l'estimation de l'héritabilité est bien développée en agronomie, ce n'est que récemment que le modèle animal a été proposé pour les populations naturelles de plantes et d'animaux.

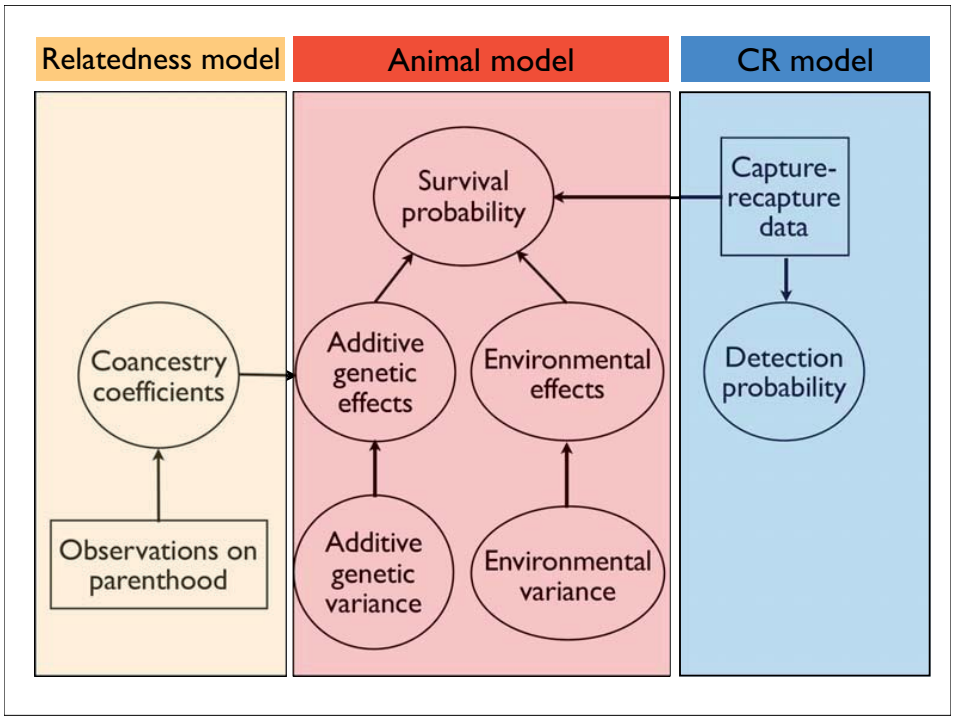


FIGURE 10 – Structure d'un modèle combinant information sur le pedigree et données de CMR. Adapté de O'Hara et al. (2008). Les données de CMR sont obtenues sur plusieurs générations et l'on souhaite estimer les composantes de la variance dans la survie. Le modèle a 3 composantes, un modèle de parenté (en jaune), un modèle animal (en rouge) et un modèle de CMR (en bleu). Le modèle de parenté est utilisé pour déterminer la structure du pedigree à partir d'observations directes sur les individus (via la structure sociale ou des marqueurs génétiques). Dans l'application de notre approche aux mésanges bleues, cette partie est supposée connue. Le modèle animal se nourrit du pedigree au travers des coefficients de parenté nécessaires pour estimer les effets additifs. Enfin, la probabilité de survie dépend des effets additifs et environnementaux. Chaque partie du modèle général peut être modifiée ou étendue séparément. Les rectangles représentent des constantes et les ellipses des variables aléatoires.

Avec la disponibilité croissante de suivis à long terme, cette approche est de plus en plus utilisée.

Toutefois, quantifier l'héritabilité dans des populations sauvages reste une tâche difficile. D'abord, comme les méthodes n'existent (ou n'existaient) pas, son estimation se fait dans les cas rares où la probabilité de détection est proche de 1, ou bien en supposant que cette probabilité vaut 1 (Cam 2009). En outre, la description des patrons évolutifs en milieu naturel peut être masquée par des variations environnementales (Téplitsky et al. 2009) qui dominent les autres composantes de la variance phénotypique. Néanmoins, d'après Kruuk (2004), ces « pièges potentiels (...) peuvent être évités, au moins en partie, par l'utilisation de techniques statistiques plus sophistiquées que celles ayant été traditionnellement utilisées dans la majorité des études de populations sauvages. »

C'est l'esprit de ce chapitre dans lequel nous proposons un cadre pour l'estimation de l'héritabilité des paramètres démographiques à partir de données CMR. Notre approche combine modèles animaux et CMR. Cette idée est dans l'air puisque O'Hara et al. (2008) publient une figure illustrant la structure d'un tel modèle (Figure 10).

Toutefois, les conditions requises pour son développement exigent des efforts transdisciplinaires (Cam 2009). Qu'à cela ne tienne ! Avec deux étudiants (Julien Papaïx qui a fait son stage de M2 Biostatistique sur le sujet, et Sarah Cubaynes au travers d'un chapitre de sa thèse), nous avons proposé des modèles animaux pour données de CMR (S37) [1] afin d'estimer l'héritabilité de la survie.

4.2 Brancher le modèle animal sur les modèles de CMR

Le coeur de notre approche repose sur la formulation à espace d'états des modèles de capture-recapture (P6 ; voir Encadré 4) dans laquelle nous considérons deux couches, l'une pour le processus dynamique (le modèle d'état) et l'autre connectant ce processus démographique à son observation via la détection (ou pas) d'individus (le modèle d'observation). Pour simplifier, on se concentre ici sur le processus démographique de survie dont on veut calculer l'héritabilité. Si l'on parcourt la littérature en génétique quantitative, on s'aperçoit que l'héritabilité d'un trait discret (la survie est bien binaire, vivant ou mort) est calculée grâce à des modèles dits à seuil (e.g. Gianola 1982). Cette approche suppose qu'il existe une variable aléatoire continue latente (« liability ») à partir de laquelle des valeurs discrètes du trait sont générées. La clé de la combi-

1. La combinaison de différentes sources d'information, ici le pedigree avec des données de CMR, participe d'un effort plus large de modélisation intégrée qui me tient à coeur comme en témoigne l'encadrement de deux étudiants du M2 Biostatistique (Marie Cheminat et Blaise Doris), ma participation à l'encadrement des thèses de Fitsum Gebreselassie et Rachel McCrea ainsi que plusieurs publications (avec Fitsum : P2, P19 ; avec Rachel : CL27 et S3 ; voir aussi P16).

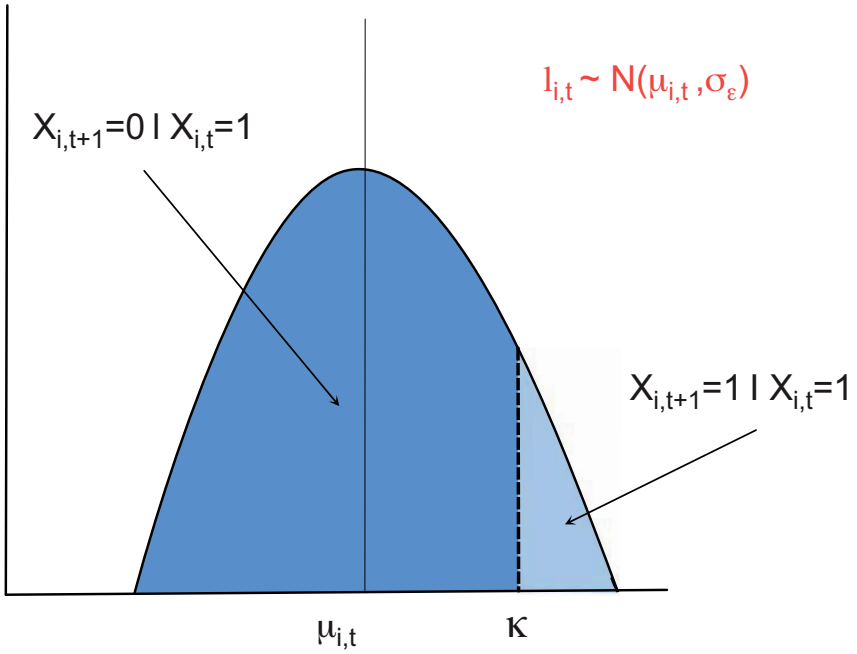


FIGURE 11 – Schéma du modèle à seuil pour la survie. La variable $X_{i,t}$ est l'état de l'individu i à l'occasion t (1 si vivant, 0 si mort) et $l_{i,t}$ sa liability distribuée selon une $N(\mu_{i,t}, \sigma^2_\epsilon)$. On raisonne conditionnellement à $X_{i,t} = 1$. Alors, si la liability de l'individu i à t est plus grande que le seuil κ, cet individu survit jusqu'à l'occasion suivante $X_{i,t+1} = 1$, sinon il meurt sur l'intervalle $X_{i,t+1} = 0$.

naison entre données CMR et information du pedigree réside dans la définition de cette liability. On suppose que $X_{i,t+1}$ l'état (vivant ou mort) d'un individu i au temps $t+1$ est généré par une variable latente continue $l_{i,t}$, qui, sachant que cet individu est vivant à l'occasion précédente (i.e. $X_{i,t} = 1$), satisfait :

$$X_{i,t+1} = \begin{cases} 1 & \text{if } l_{i,t} > \kappa, \\ 0 & \text{if } l_{i,t} \leq \kappa. \end{cases}$$

où κ est la valeur seuil (Figure 11). En mots, si la liability de l'individu i à t est plus grande que le seuil κ, alors cet individu survit jusqu'à l'occasion suivante $t+1$, sinon il meurt sur l'intervalle. Nous supposons que la liability est distribuée selon une loi normale d'espérance $\mu_{i,t}$ et de variance σ^2. Pour des raisons d'identifiabilité des paramètres (voir Chapitre 2), on fixe $\sigma^2 = 1$ et $\kappa = 0$ sans perte de généralité [2].

Cette construction permet d'exprimer la survie d'un individu i entre les occasions t et $t+1$ comme $\phi_{i,t} = F(\mu_{i,t})$ où F est la fonction de répartition [3] d'une loi $N(0,1)$ [4]. Autrement dit, la survie est directement reliée à la liability. Ce lien se fait via la fonction F, ou sa réciproque F^{-1} souvent utilisée dans l'analyse de données discrètes et qu'on appelle fonction probit. On peut alors spécifier le modèle animal sur cette échelle probit :

$$\text{probit}(\phi_{i,t}) = F^{-1}(\phi_{i,t}) = \mu_{i,t} = \eta + b_t + e_i + a_i$$

où :

- η est un terme constant pour la survie moyenne sur l'échelle probit ; ce terme peut être modifié en incluant des covariables en effets fixes qui affectent la survie (climat, voir P25 et CL20 ; exploitation par l'homme, voir P15),
- b_t est un effet aléatoire temps avec $b_t \sim N(0, \sigma_b^2)$,
- e_i est un effet aléatoire individuel non-génétique avec $e_i \sim N(0, \sigma_e^2)$,
- a_i est un effet aléatoire pour la valeur génétique de l'individu i, où le vecteur des éléments a_i est distribué comme une loi normale multivariée $MN(0, \sigma_a^2 \boldsymbol{A})$, avec σ_a^2 la variance additive génétique et $\boldsymbol{A}$ la matrice de relation additive génétique. La matrice $\mathbf{A}$ est connue et se construit à partir du pedigree (e.g. entre un individu i et lui-même, on a $A_{i,i} = 1$, et entre un descendant i et son parent j, on a $A_{i,j} = 0.5$) [5].

2. Cette dernière condition entraîne que la valeur du seuil est absorbée dans le terme constant du modèle animal défini plus bas.

3. La fonction de répartition d'une variable aléatoire réelle X est la fonction F telle que $F(x) = \Pr(X \leq x)$. Si cette fonction est connue, alors la loi de probabilité de X est entièrement caractérisée.

4. $\phi_{i,t} = \Pr(X_{i,t+1} = 1 | X_{i,t} = 1) = \Pr(l_{i,t} > \kappa) = F(\mu_{i,t})$

5. La matrice $\mathbf{A}$ a autant de lignes et de colonnes que le nombre d'individus, et nécessite un traitement particulier (Damgaard 2007).

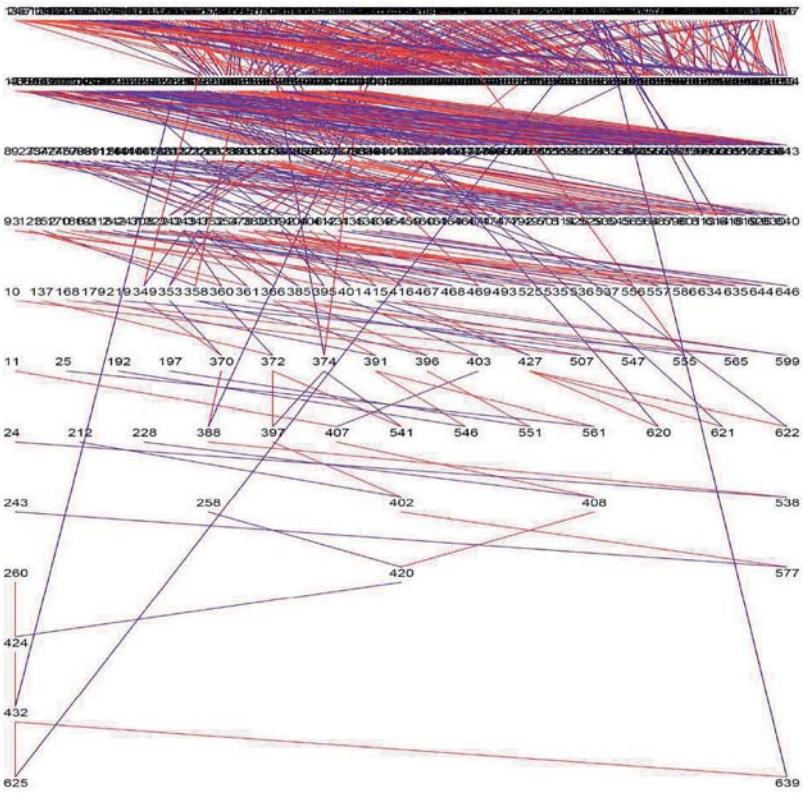


FIGURE 12 – Représentation graphique du pedigree sur les mésanges bleues à partir de 654 individus (218 pères en rouge, 215 mères en bleu, 12 générations en ligne). Ce graphe illustre une dépendance complexe entre les individus (numérotés de 1 à 654) du jeu de données CMR.

L'héritabilité se calcule comme la contribution de la variance additive génétique à la variance totale :

$$h^2 = \frac{\sigma_a^2}{\sigma_b^2 + \sigma_e^2 + \sigma_a^2 + 1}.$$

4.3 Application aux données CMR sur la mésange bleue

Pour illustrer notre approche, nous utilisons un jeu de données sur des individus marqués de mésanges bleues (*Cyanistes caeruleus*) dans une population naturelle en Corse. Nous avons à disposition 327 individus reproducteurs suivis de 1979 à 2007. Un pedigree est construit à partir des observations sur le nid (Figure 12).

La probabilité de détection p est forte, et la survie (probit$^{-1}(\eta) = F(\eta)$) conforme à ce que l'on attend pour un petit passereau (Figure 13). La variance additive génétique σ_a^2 est faible, d'où une héritabilité h^2 faible aussi. La variabilité environnementale σ_b^2 est non-négligeable. On trouve que le meilleur modèle ne contient pas l'effet aléatoire individuel a_i [6], ce qui confirme que l'héritabilité de survie n'est pas significative chez cette population de mésanges.

4.4 Conclusions et perspectives

On développe dans ce chapitre un modèle pour estimer et faire de l'inférence sur la base génétique des composantes de la valeur sélective en conditions naturelles. Si l'idée est simple et consiste en la combinaison des modèles animaux et de CMR, sa formalisation passe par une modélisation avancée et sa mise en oeuvre requiert des méthodes MCMC (voir Encadré 2) pour gérer plusieurs effets aléatoires de nature différente (individuel et temporel).

L'analyse des données mésanges montre une héritabilité de la survie faible, ce qui va dans le sens de l'interprétation classique du théorème fondamental de Fisher sur la sélection naturelle qui prédit une héritabilité faible des traits fortement associés à la valeur sélective. Toutefois, le peu d'estimations de l'héritabilité de la longévité en conditions naturelles montrent des résultats contrastés à ce propos (Kruuk et al. 2000, Coltman et al. 2005) et l'héritabilité de la survie adulte n'a jamais été estimée à ma connaissance. Qui plus est, c'est la première valeur d'héritabilité de survie entre saisons de reproduction chez un vertébré sauvage fournie grâce à un modèle tenant compte de la probabilité de détection < 1.

Héritabilité d'autres paramètres démographiques que la survie. Si nous nous sommes concentrés sur la survie, la généralisation de notre approche à d'autres paramètres démographiques es-

6. La procédure de sélection de modèles qui permet d'arriver à ce résultat n'est pas détaillée ici, voir S39.

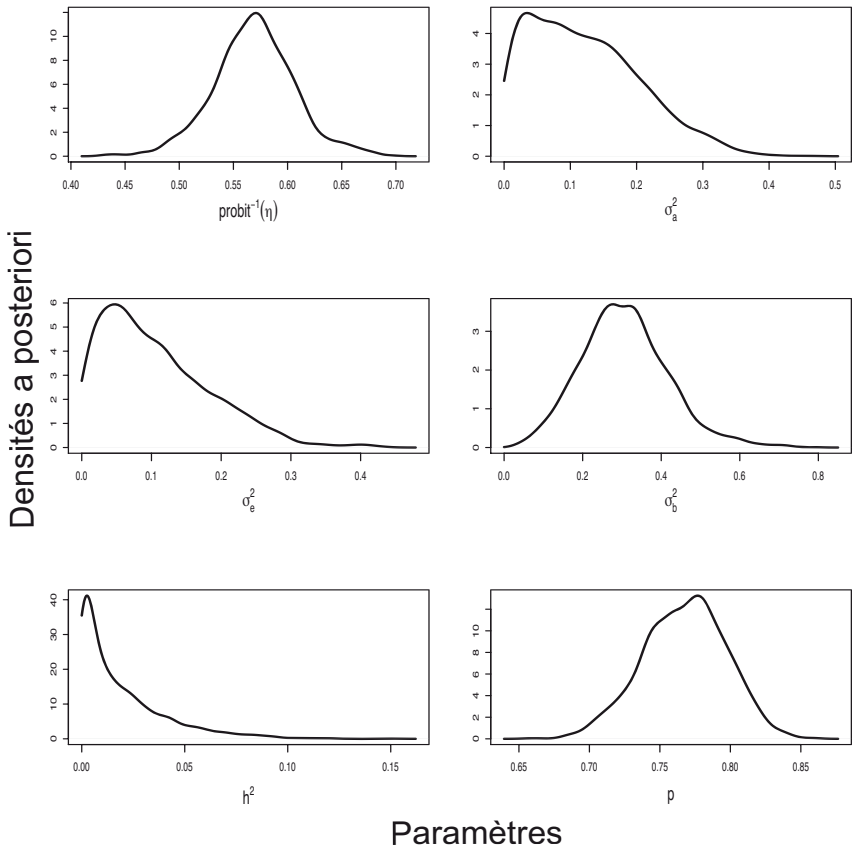


FIGURE 13 – Densités de probabilité a posteriori des paramètres du modèle animal pour données de CMR appliqué aux mésanges bleues. Les valeurs possibles des paramètres sont en abscisses, et en ordonnées on retrouve la fréquence à laquelle on obtient ces valeurs. La signification, ainsi que les médianes et intervalles de crédibilité à 95% des différents paramètres sont (de gauche à droite et de haut en bas) : probit$^{-1}(\eta)$ (0.57 ; [0.49, 0.65]) est la survie moyenne, σ_a^2 (0.11 ; [0.01, 0.31]) est la variance additive génétique, σ_e^2 (0.09 ; [0.01, 0.28]) est la variance de l'effet aléatoire individuel non-génétique, σ_b^2 (0.30 ; [0.11, 0.56]) est la variance de l'effet aléatoire année, h^2 (0.01 ; [0.00, 0.08]) est l'héritabilité et p (0.77 ; [0.71, 0.82]) est la probabilité de détection.

timables via les modèles de CMR ne devrait pas poser de problème. Dans un chapitre de sa thèse, Sarah Cubaynes étudie l'héritabilité de l'âge de première reproduction. Parallèlement, nous proposons de quantifier l'héritabilité de la dispersion avec Blandine Doligez au travers d'une demande de financement pour un post-doctorant via le CNRS et d'une ANR blanche. Techniquement, il faut utiliser une formulation à espace d'états des modèles multiétats de CMR (voir Encadré 4) et étendre la définition de la liability à plusieurs seuils (Sorensen et al. 1995).

Héritabilité de la sénescence de survie. Comme discuté dans les perspectives du Chapitre 1, l'environment peut être un déterminant majeur de la sénescence dans les populations naturelles. Par exemple, une étude à long terme dans une population de Gobemouches à collier (*Ficedula albicollis*) a montré que la variance additive génétique pour la survie (apparente) à des âges avancés était nulle bien qu'elle soit non-nulle à des âges plus précoces (Brommer et al. 2007). Dans un travail en collaboration avec Anne Charmantier et Ben Sheldon, nous souhaitons développer des modèles CMR incluant cette relation entre la variation additive génétique et l'âge en s'inspirant des modèles « random regression » (Meyer 1998).

Conclusions et réflexions

Quels modèles de CMR pour l'écologie évolutive ?

Au risque de passer pour un « intégriste » des CMR, j'ai beaucoup insisté sur l'importance de prendre en compte la probabilité de détection. Toutefois, si dans certains cas ignorer ce paramètre peut entraîner une inférence erronée (voir introduction), dans d'autres la différence peut être négligeable (par exemple si la détectabilité est forte et constante dans le temps, comme dans le chapitre 4). Il est néanmoins difficile d'énoncer des règles fiables pour prédire si le problème peut être ignoré et ainsi pouvoir se ramener à des analyses plus classiques que les CMR. Par conséquent, on ne saurait trop recommander de faire l'analyse CMR. Concernant les études déjà menées et ignorant le problème de détectabilité, je ne prétends pas qu'elles sont fausses, mais plutôt qu'il est difficile de savoir si on peut se fier aux résultats obtenus tant qu'une réanalyse CMR n'est pas conduite (voir Nichols et al. 1997 et un exemple avec Gaillard et al. 1994).

La motivation principale de ce mémoire (outre le diplôme...) était de mettre la variabilité individuelle au centre des problèmes d'estimation et d'inférence dans les modèles CMR. Y suis-je parvenu ? Je l'espère, au moins dans ce document, grâce à l'examen de quatre questions importantes en écologie évolutive. Ma contribution est principalement d'ordre méthodologique via le traitement de l'hétérogénéité dans les modèles de CMR. En bref, si l'on dispose de critères observables supposés refléter l'essentiel de l'hétérogénéité intra-population, on peut les intégrer dans les analyses (âge, chapitre 1 ; traits morphologiques, chapitres 2 et 3). Si l'on est en présence d'une hétérogénéité dont l'origine n'est pas forcément identifiée et qui ne peut être correctement quantifiée en utilisant des critères observables (le concept de frailty introduit au chapitre 1), alors il faut recourir à des modèles de mélange (chapitre 1) ou à effets aléatoires (tous les chapitres). Les deux approches sont complémentaires (chapitres 1 et 3).

Le traitement d'une frailty nécessite donc le recours à des modèles à structure cachée (modèles

de Markov caché ou à espace d'état, Encadrés 3 et 4). S'ils sont souvent complexes du point de vue numérique ou statistique, il me semble que ces modèles correspondent bien à l'intuition du biologiste quant à son appréhension du terrain et du fonctionnement de son système d'étude : on focalise sur la façon dont on pense que ça marche (le processus caché) plutôt que sur ce qu'on voit (une observation bruitée du signal biologique).

Je suis convaincu qu'une des pistes prometteuses pour incorporer cette frailty sont les modèles mixtes de CMR (P1). Ces modèles permettent la prise en compte de la variabilité individuelle par l'incorporation d'effets aléatoires [7], en plus de la tendance ou du patron général pris en compte par les effets fixes typiques des modèles linéaires généralisés (e.g. Bolker et al. 2009). Ces modèles mixtes de CMR occupent une grande place dans mes recherches actuelles, en particulier au travers de la thèse de Sarah Cubaynes qui s'attaque à plusieurs cas d'études et des développements méthodologiques et du post-doctorat d'Eleni Papadatou sur des modèles multi-espèces et multi-populations (S21 ; voir aussi P24) [8].

Vers une éco-évo-statistique

La rédaction de mon HDR est l'occasion de faire le point sur la manière dont je vois mon travail. Est-ce celui d'un statisticien ? Non, je ne développe pas de nouvelles méthodes statistiques en tant que telles. D'un biométricien alors ? J'avoue que ce terme me fait de plus en plus penser (comme au grand public j'imagine) au passeport du même nom... D'un biostatisticien ? Même si c'est l'intitulé de mon diplôme de thèse, on observe que ce terme est malheureusement monopolisé par les sciences médicales et la bioinformatique (il suffit de prendre la table des matières du volume de janvier 2010 de la revue Biostatistics pour s'en convaincre). Je suis tenté d'utiliser le terme statisticien écologique proposé par mes collègues britanniques (http ://www.ncse.org.uk/), mais on en oublierait presque l'évolution. Pour être au plus près de mon activité, j'introduirais le terme d'éco-évo statisticien.

Plus sérieusement, j'aimerais pour conclure insister sur mon attachement aux collaborations

7. On pense à des effets génétiques (chapitre 4) différent selon les individus mais également à l'influence des conditions durant le développement (e.g. effets familles ou effets parentaux) ou des conditions rencontrées plus tard dans la vie (e.g. l'exposition à un pathogène).

8. Ce thème autour des modèles mixtes concentre aussi une partie de mes efforts de transfert de méthodes statistiques vers d'autres domaines que la biologie des populations comme l'écologie fonctionnelle (co-encadrement avec Eric Garnier de Baptiste Testi M2 Biostatistique et divers projets en cours avec Eleni Kazakou ; voir aussi P48 et P53) et la primatologie (projet d'un article de revue avec Marie Charpentier et Jo Stechell ; voir aussi P17, P18, P52, P55 et P57).

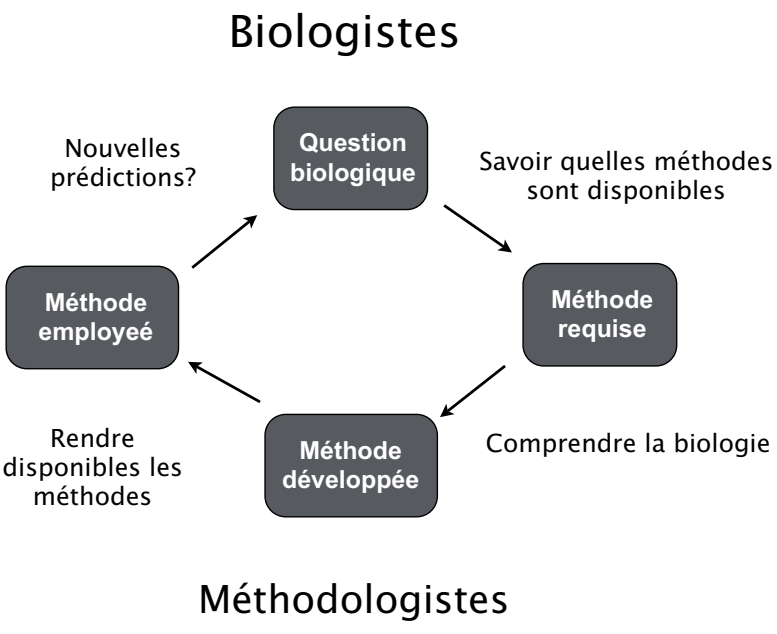


FIGURE 14 – Cercle vertueux de collaborations (d'après une figure imaginée par Paul Doherty et que j'ai librement adaptée). Les biologistes formulent des questions biologiques qu'ils souhaitent tester sur la base de données (expérimentales ou observations). Ces études sont basées sur des analyses combinant des protocoles (« design-based) » et des modèles (« model-based »). Dans les protocoles, on se pose des questions sur le nombre d'individus et d'échantillons à collecter, on envisage des études pilotes basées sur des études de puissance. En gros, c'est tout ce que le biologiste fait sur le terrain. Le présent document fournit un exemple de modèles, avec en particulier la prise en compte de l'hétérogénéité individuelle. En gros, c'est tout ce que le méthodologiste fait devant son ordinateur. Pour mener à bien ces analyses, il est nécessaire de connaître les méthodes disponibles. Si l'existant ne suffit pas, le biologiste, souvent en collaboration avec le méthodologiste, met à jour le besoin de nouvelles méthodes. Les méthodologistes développent alors de nouvelles méthodes, ce qui suppose qu'ils comprennent la biologie. Pour que ces méthodes puissent être utilisées par les biologistes, le méthologiste doit les mettre à disposition. Puis en générant des résultats, le biologiste génère de nouvelles questions biologiques et ainsi de suite. Par ce processus, biologistes et méthodologistes s'enrichissent.

entre biologistes et méthodologistes auxquelles j'accorde une grande importance (Figure 14) ; ces collaborations s'articulent autour de plusieurs points.

- Le développement et la popularisation (articles, ateliers de travail, enseignement, ...) de logiciels pour analyser ces données et pour assurer le transfert des méthodes. Cet aspect est une partie importante du travail effectué dans notre équipe et portée par nos « informathématiciens » (Rémi Choquet assisté d'Erika Nogué).
- Les données et en particulier les suivis à long terme si précieux pour biologistes et méthodologistes. On ne soulignera jamais assez l'importance du travail de terrain dont j'ai eu la chance de rencontrer des représentants dévoués (je pense à toutes les personnes gravitant autour des programmes à long terme du Museum National d'Histoire Naturelle, de l'Office National de la Chasse et de la Faune Sauvage, des Terres Australes et Antarctiques Françaises et du CEFE bien sûr). Qu'il me soit permis de leur rendre hommage ici.
- La formation d'étudiants sous la forme d'encadrements, d'enseignements et d'ateliers de travail. Ce que je peux enseigner consiste essentiellement à utiliser et à adapter si besoin des méthodes existantes et concepts pour répondre à des questions en biologie des populations. J'espère aussi pouvoir saupoudrer cet enseignement du pragmatisme d'un non-puriste [9]. En retour, qu'ils soient assurés que le bout de vie passé ensemble est pour moi une source d'apprentissage intarissable (entre autres...).
- Enfin, et avant tout, les collaborations passent par des rencontres. Chacun des quatre chapitres et plus généralement mon travail au jour le jour en a été, en est, et en sera le fruit. Ils se reconnaîtront.

9. La polémique bayésien vs. fréquentiste en est une bonne illustration. Plutôt qu'une opposition de l'approche bayésienne vs. l'approche fréquentiste, on peut choisir l'une ou l'autre approche selon ses besoins (P5), voire les utiliser en combinaison (voir CL29 et S31 pour des exemples).

Bibliographie

Bolker, B. M. et al. 2009. Generalized linear mixed models : a practical guide for ecology and evolution. Trends Ecol. Evol. 24 : 127-135.

Brodie, E. D. et al. 1995. Visualizing and quantifying natural selection. Trends Ecol. Evol. 10 : 313–318.

Brommer J.E. et al. 2007. Exploring the genetics of ageing in a wild passerine. Am. Nat. 170 : 643-650.

Burnham, K. P. et D. R. Anderson. 2002. Model selection and multimodel inference : a practical information-theoretic approach. 2nd Edition. Springer-Verlag, New York, USA.

Cam, E. 2009. Contribution of capture-mark-recapture modeling to studies of evolution by natural selection. Pp. 83-129 in D. L. Thomson, E. G. Cooch, and M. J. Conroy, eds. Modeling demographic processes in marked populations, Vol. 3. Springer Series : Environmental and Ecological Statistics, New York.

Cam, E. et al. 2002. Individual covariation between life-history traits : seeing the trees despite the forest. Am. Nat. 159 : 96-105.

Catchpole, E. A. et al. 1996. Steps to parameter redundancy in age-dependent recovery models. J. R. Statistic. Soc., B, 58 : 763-774.

Choquet, R., L. et al. 2009. Program E-SURGE : a software application for fitting multievent models. Pp. 845-865 in D. L. Thomson, E. G. Cooch, and M. J. Conroy, eds. Modeling demographic processes in marked populations, Vol. 3. Springer Series : Environmental and Ecological Statistics, New York.

Coltman, D. et al. 2005. Selection and genetic (co)variance in bighorn sheep. Evolution 59 : 1372-1382.

Covas, R. et al. 2002. Stabilizing selection on body mass in the sociable weaver Philetairus socius. Proc. R. Soc. Lond. B. 269 : 1905-1909.

Crespin, L. et al. 2006. Increased adult mortality and reduced breeding success with age in a population of common guillemot Uria aalge using marked birds of unknown age. J. Avian Biol. 37 : 273-282.

Damgaard, L.H. 2007. Technical note : How to use WinBUGS to draw inferences in animal models. J. Anim. Sci. 85 : 1363-1368.

Gaillard, J.M. et al. (1994). Senescence in natural-populations of mammals - a reanalysis. Evolution 48 : 509-516.

Gianola, D. 1982. Theory and analysis of threshold characters. J. Anim. Sci., 54 :1079-1096.

Green, P. J. 1995. Reversible jump MCMC computation and Bayesian model determination. Biometrika 82 : 711-732.

Grégoire, A. et al. 2004. Stabilizing natural selection on the early expression of a secondary sexual trait in a passerine bird. J. Evol. Biol. 17 : 1152-1156.

Hamilton, W. D. 1966 The moulding of senescence by natural selection. J. Theor. Biol. 12 : 12-45.

Hestbeck, J. B. et al. 1991. Estimates of movement and site fidelity using mark-resight data of wintering canada geese. Ecology 72 : 523-533.

Kingsolver, J. G. et S. G. Smith. 1995. Estimating selection on quantitative traits using capture-recapture data. Evolution 49 : 384–388.

Kingsolver, J. G. et al. 2001. The strength of phenotypic selection in natural populations. Am. Nat. 157 : 245-261.

Kruuk, L.E.B. et al. 2000. Heritability of fitness in a wild mammal population. Proc. Natl. Acad. Sci. U. S. A. 97 : 698-703.

Kruuk, L.E.B. 2004. Estimating genetic parameters in natural populations using the “animal model”. Phil. Trans. R Soc. B Biol. Sci. 359 : 873-890.

Lande, R. et S. J. Arnold. 1983. The measurement of selection on correlated characters. Evolution 37 : 1210-1226.

Lebreton, J.-D. et al. 1992. Modelling survival and testing biological hypotheses using marked animals : a unified approach with case studies. Ecol. Monogr. 62 :

Lebreton J.-D. et al. 2009. Modeling individual animal histories with multistate capture-recapture models. pp 88-159 In Caswell, H. Advances in Ecological Research 41. Academic Press.

Lynch M. et Walsh B. 1998. Genetics and analysis of quantitative traits. Sinauer Associates Inc. Publishers. Sunderland, Massachusetts, U. S. A.

McCarthy, M. A. 2007. Bayesian methods for ecology. Cambridge Univ. Press, Cambridge.

Medawar, P. B. 1946. An unsolved problem of biology. London, UK : Lewis.

Meyer, K. 1998. Estimating covariance functions for longitudinal data using a random regression model. Genet. Sel. Evol. 30 : 221-240.

Nichols, J. D. et al. 1994. Estimating breeding proportions and testing hypotheses about costs of reproduction with capture-recapture data. Ecology 75 : 2052-2065.

Nichols, J. D. et Kendall, W. L. 1995. The use of multi-state capture-recapture models to address questions in evolutionary ecology. J. of Appl. Stat. 22 : 835-846.

Nichols, J. D. et al. 1997. Test for senescent decline in annual survival probabilities of common pochard, Aythia ferina. Ecology 78 : 1009-1018.

Nussey D. H. et al. 2008. Testing for genetic trade-offs between early- and late-life reproduction in a wild red deer population. Proc. R. Soc. B. 275 : 745-750.

O'Hara, R. B. et al. 2008. Bayesian approaches in evolutionary quantitative genetics. J. of Evol. Biol. 21 : 949-957.

Pledger, S. 2000. Unified maximum likelihood estimates for closed capture-recapture models using mixtures. Biometrics 56 : 434-442.

Pledger, S. et al. 2003. Open capture-recapture models with heterogeneity : I. Cormack-Jolly-Seber model. Biometrics 59 :786-794.

Pradel, R. 2005. Multievent : an extension of multistate capture-recapture models to uncertain states. Biometrics 61 : 442-447.

Pradel, R. 2009. The stakes of capture-recapture models with state uncertainty. Pp. 781-795 in D. L. Thomson, E. G. Cooch, and M. J. Conroy, eds. Modeling demographic processes in marked populations, Vol. 3. Springer Series : Environmental and Ecological Statistics, New York.

Reznick, D.N. et al. 2004 Effect of extrinsic mortality on the evolution of senescence in guppies. Nature 431 : 1095-1099.

Roff, D. A. 1992. The evolution of life histories, theory and analysis. Chapman and Hall, N.Y., U.S.A.

Royle, J. A. 2008. Modeling individual effects in the Cormack-Jolly-Seber model : a state-space formulation. Biometrics 64 : 364-370.

Ruppert, D. et al. 2003. Semiparametric regression. Cambridge Univ. Press, Cambridge.

Schluter, D. 1998. Estimating the form of natural selection on a quantitative trait. Evolution 42 : 849-861.

Schluter, D. et D. Nychka. 1994. Exploring fitness surfaces. Am. Nat. 143 : 597-616.

Sorensen, S. et al. 1995. Bayesian inference in threshold models using gibbs sampling. Genet. Sel. Evol. 27 : 229-249.

Stearns, S. C. 1992. The evolution of life histories. Oxford University Press, New York, USA.

Téplitsky, C. et al. 2009. Heritability of fitness components in a wild bird population. Evolution 63 : 716-726.

Townsend, H. M. et D.J. Anderson 2009. Assessment of costs of reproduction in a pelagic seabird using multistate mark-recapture models. Evolution 61 : 1956-1968.

Tuljapurkar, S. et al. 2008. Dynamic heterogeneity in life histories. Ecol. Let. 12 : 93-106.

van de Pol, M. et S. Verhulst. 2006. Age-dependent traits : a new statistical model to separate within- and between-individual effects. Am. Nat. 167 : 766-73.

Van Noordwijk, A. J. et G. De Jong. 1986. Acquisition and allocation of resources - Their influence on variation in life-history tactics. Am. Nat. 128 : 137-142.

Vaupel, J. W. et Yashin, A. I. 1985. Heterogeneity's ruses : some surprising effects of selection on population dynamics. Am. Stat. 39 : 176-185.

Venzon, D. J. et Moolgavkar, S. H. 1988. A method for computing profile-likelihood-based confidence intervals. Appl. Stat. 37 : 87-94.

White, G. C. et K. P. Burnham. 1999. Program MARK : survival estimation from populations of marked animals. Bird Study 46 : 120-138.

Williams, G.C. 1957. Pleiotropy, natural selection and the evolution of senescence. Evolution 11 : 398-411.

Williams, P. D. et al. 2006. The shaping of senescence in the wild. Trends Ecol. Evol. 21 : 458-463.