Chapitre 13

MODÈLES DE MÉLANGE EN CAPTURE-RECAPTURE

Lauriane Rouan, Sarah Cubaynes, Christophe Duchamp, Christian Miquel, Anne-Marie Reboulet, Olivier Gimenez, Jean-Dominique Lebreton, Rémi Choquet et Roger Pradel

13.1 Introduction

La méthode dite de capture-recapture consiste à suivre des individus marqués, mais laissés libres, dans une population animale naturelle. C'est le moyen privilégié pour étudier la démographie et plus généralement la dynamique d'une telle population. Les applications de cette méthode ne cessent d'ailleurs de s'élargir avec la prise en compte des mouvements mais aussi du comportement. Le cadre d'analyse actuel est en effet extrêmement souple puisqu'il permet de considérer des états à définir, reflétés indirectement par des événements également à définir ; seuls les événements sont directement perceptibles. On reconnait là, au vocabulaire près, « événement » au lieu d' « observation », le contexte d'une chaîne de Markov cachée.

De fait, longtemps développées à l'écart des courants statistiques généraux, les méthodes d'analyse de données de capture-recapture tendent de plus en plus à les rejoindre. Une première étape importante (Lebreton *et al.* [1992]) avait été franchie avec l'introduction des outils de sélection de modèles, notamment l'AIC, et d'un cadre conceptuel directement inspiré des modèles linéaires généralisés. L'étape actuelle correspond à une fusion plus profonde puisque les modèles d'analyse sont désormais reconnus comme des cas particuliers d'un outil statistique très général, les modèles de Markov cachés (Pradel [2005]). Ce rapprochement a permis d'importer des techniques très utiles puisées dans la boîte à outils des modèles de Markov cachés. On peut citer l'application de

l'algorithme de Viterbi à l'estimation du nombre de jeunes produits par un individu au cours de sa vie (le très important « lifetime reproductive success » des biologistes) ; ou encore les tentatives d'utilisation d'algorithmes de type EM au lieu des traditionnels algorithmes de type quasi-Newton (Rouan [2007] ; Rouan *et al.* [2008]).

Les modèles de capture-recapture gardent cependant des particularités. Si les observations sont faites en temps discret comme dans les modèles de Markov cachés standards, les variations des conditions environnementales imposent de considérer des chaines de Markov cachées hétérogènes. Autre particularité : ces chaînes sont courtes (quelques dizaines d'éléments au plus) mais nombreuses (une par individu et une centaine d'individus au minimum). Toutes ses particularités font que l'analyse des données de capture-recapture garde encore aujourd'hui des spécificités telles que le développement de logiciels spécialisés continue à être nécessaire (Choquet *et al.* [2008]).

Le cadre conceptuel nouveau permet de traiter aussi bien des états cachés réels (par exemple, statut malade ou sain lors d'une étude d'épidémiologie) que des états purement théoriques dont l'intérêt principal n'est pas l'existence réelle mais leur utilité dans la description des données. C'est en particulier le cas lorsqu'ils portent sur le processus de détection. Ce processus intervient en effet nécessairement dans la description des données mais n'est généralement pas un processus d'intérêt. Cependant, il doit être pris en compte de manière satisfaisante pour éviter des biais sur l'estimation de paramètres principaux comme les taux de survie. La difficulté le plus souvent rencontrée est une hétérogénéité individuelle de détection.

Cette détectabilité inégale des individus n'a pas une influence considérable sur l'estimation de la survie annuelle, ou de tout autre paramètre démographique assimilable à un rapport d'effectifs ; cependant, elle introduit un biais négatif qui peut être très sévère sur l'estimation de l'effectif lui-même. L'effectif de la population est en effet essentiellement obtenu comme le ratio du nombre d'individus détectés à une date donnée par la probabilité de détection estimée à cette date. Une approche naturelle pour traiter cette hétérogénéité de détection est la considération de modèles de mélange à un nombre fini, généralement petit, de classes. Une estimation plus juste de l'effectif, critique pour des problèmes de gestion, peut alors être obtenue sur cette base. Nous allons voir la mise en oeuvre de cette idée dans l'estimation de l'effectif de loups dans le massif du Mercantour en Provence.

13.2 Les enjeux

La recolonisation spontanée par le loup (*Canis lupus*) du massif alpin depuis l'Italie pose le problème d'un « développement contrôlé » de la population, étant donné les contraintes induites par la présence de cette espèce sur la filière de l'élevage.

l'algorithme de Viterbi à l'estimation du nombre de jeunes produits par un individu au cours de sa vie (le très important « lifetime reproductive success » des biologistes) ; ou encore les tentatives d'utilisation d'algorithmes de type EM au lieu des traditionnels algorithmes de type quasi-Newton (Rouan [2007] ; Rouan *et al.* [2008]).

Les modèles de capture-recapture gardent cependant des particularités. Si les observations sont faites en temps discret comme dans les modèles de Markov cachés standards, les variations des conditions environnementales imposent de considérer des chaines de Markov cachées hétérogènes. Autre particularité : ces chaînes sont courtes (quelques dizaines d'éléments au plus) mais nombreuses (une par individu et une centaine d'individus au minimum). Toutes ses particularités font que l'analyse des données de capture-recapture garde encore aujourd'hui des spécificités telles que le développement de logiciels spécialisés continue à être nécessaire (Choquet *et al.* [2008]).

Le cadre conceptuel nouveau permet de traiter aussi bien des états cachés réels (par exemple, statut malade ou sain lors d'une étude d'épidémiologie) que des états purement théoriques dont l'intérêt principal n'est pas l'existence réelle mais leur utilité dans la description des données. C'est en particulier le cas lorsqu'ils portent sur le processus de détection. Ce processus intervient en effet nécessairement dans la description des données mais n'est généralement pas un processus d'intérêt. Cependant, il doit être pris en compte de manière satisfaisante pour éviter des biais sur l'estimation de paramètres principaux comme les taux de survie. La difficulté le plus souvent rencontrée est une hétérogénéité individuelle de détection.

Cette détectabilité inégale des individus n'a pas une influence considérable sur l'estimation de la survie annuelle, ou de tout autre paramètre démographique assimilable à un rapport d'effectifs ; cependant, elle introduit un biais négatif qui peut être très sévère sur l'estimation de l'effectif lui-même. L'effectif de la population est en effet essentiellement obtenu comme le ratio du nombre d'individus détectés à une date donnée par la probabilité de détection estimée à cette date. Une approche naturelle pour traiter cette hétérogénéité de détection est la considération de modèles de mélange à un nombre fini, généralement petit, de classes. Une estimation plus juste de l'effectif, critique pour des problèmes de gestion, peut alors être obtenue sur cette base. Nous allons voir la mise en oeuvre de cette idée dans l'estimation de l'effectif de loups dans le massif du Mercantour en Provence.

13.2 Les enjeux

La recolonisation spontanée par le loup (*Canis lupus*) du massif alpin depuis l'Italie pose le problème d'un « développement contrôlé » de la population, étant donné les contraintes induites par la présence de cette espèce sur la filière de l'élevage.

Pour ce faire, un diagnostic sur le statut de l'espèce en France est vital et se base, entre autres, sur la collecte d'informations concernant la dynamique de la population. Toutefois, l'utilisation des protocoles classiques de capture-recapture est particulièrement difficile pour des espèces dites discrètes (difficilement observables) telles que le loup puisqu'elle requiert la capture physique des individus lors de leur marquage initial.

L'utilisation de l'outil moléculaire permet de caractériser les signatures génétiques individuelles à partir d'échantillons d'excréments, de poils ou d'urine (Taberlet *et al.* [1999]). Cette méthode dite non-invasive ou non-intrusive (pas de capture effective des individus) permet d'appliquer les modèles de CR pour l'estimation des effectifs et de la survie (Lukacs et Burnham [2005]).

Le modèle de base en Capture-recapture, sur lequel nous allons nous appuyer, est le modèle dit de Jolly-Seber. Il fait intervenir deux types de paramètres : les probabilités de détection et les probabilités de survie annuelle, toutes deux susceptibles de varier au cours du temps.

Ce modèle repose sur plusieurs hypothèses fortes dont l'hypothèse clef pour l'estimation des effectifs d'une population : l'homogénéité des paramètres et notamment l'homogénéité du taux de détection. Or, plusieurs éléments montrent que cette hypothèse n'est clairement pas remplie, tous convergent vers une forte hétérogénéité de détection dans le jeu de données.

D'abord un examen du nombre de capture par individu révèle une bimodalité dans la distribution qui signifie que notre échantillon provient d'un mélange d'individus peu capturables et d'individus fortement capturables.

Cette indication est renforcée par la mise en œuvre des tests d'ajustement qui sont hautement significatifs (Burnham *et al.* [1987]). Un premier test signale un excès d'individus jamais revus, l'équivalent d'animaux en transit ; le deuxième signale des problèmes de ce qu'il est convenu d'appeler « trap happiness » (Pradel [1993]). La présence de ces 2 tests simultanément significatifs indique une forte hétérogénéité de détection entre individus.

Enfin, en l'absence de tests statistiques permettant de mettre formellement en évidence une hétérogénéité de détection, nous sommes néanmoins parvenus à simuler un scénario reproduisant les données loups, et nous avons pu montrer qu'elles correspondent assez bien au degré de non-ajustement induit par une hétérogénéité de capture permanente. Or, il est connu qu'une hétérogénéité de détection provoque un biais conséquent dans l'estimation de l'effectif (Carothers [1973]). Ce résultat a été confirmé dans la configuration particulière de notre jeu de données par simulation : nous montrons que, si le problème d'hétérogénéité n'est pas pris en compte, le modèle de Jolly-Seber sous-estime de plus de 40% l'effectif réel de la population.

Le problème d'hétérogénéité de détection est donc le principal problème méthodologique auquel nous avons été confrontés.

13.3 La méthode

13.3.1 Une extension du modèle de Jolly-Seber prenant en compte l'hétérogénéité de capture

Afin de traiter l'hétérogénéité de capture dans la population de loups, nous sommes partis du modèle de Jolly-Seber dont la formulation comme modèle de Markov caché est représentée schématiquement dans la figure 13.1.

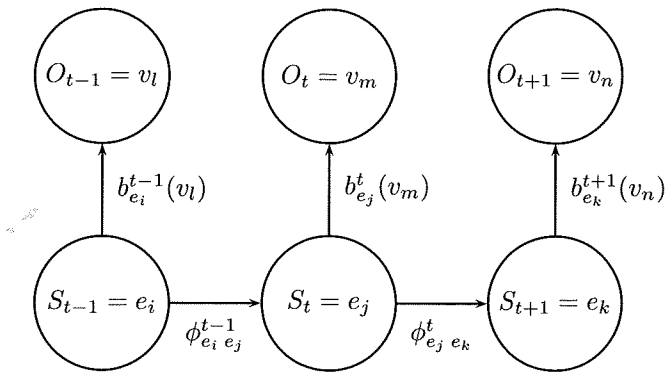

Figure 13.1 – *Les paramètres du modèle multiévénement (Pradel [2005])* [1].

Les probabilités de survie correspondent aux transitions entre états qui sont au nombre de 2 : « vivant » et « mort » ; et les probabilités de rencontre décrivent le processus de génération des événements également au nombre de 2 : « détecté » ou « pas détecté ». Le modèle peu être résumé par les matrices stochastiques ci-dessous.

Les ensembles regroupant les observations d'une part et les états de l'autre sont :

- $\Omega = \{$pas vu (codé par 0), vu (codé par 1)$\}$,
- $E = \{$vivant (V), mort (†)$\}$.

Les paramètres du modèle sont :

1. les probabilités de survie-transition :
$$\phi^t,\ t \in \{1, \cdots, T-1\}.$$
2. les probabilités de capture : $p^t,\ t \in \{2, \cdots, T\}$.

1. Les paramètres se répartissent en deux groupes distincts : ceux régissant l'évolution du processus d'état i.e. la succession des états sous-jacents par lesquels passe l'individu avec en particulier les probabilités de survie-transition ϕ et ceux reliant le processus d'état au processus d'observation à savoir les paramètres b. Ces derniers correspondent à la probabilité d'observer un événement lors de la session de capture sachant que l'individu est dans un état sous-jacent donné.

La représentation matricielle de ces paramètres permet aussi de mieux cerner les liens possibles entre états d'une part :

$$\Phi_t = \begin{array}{c} \\ V \\ \dagger \end{array} \begin{array}{c} \begin{array}{cc} V & \dagger \end{array} \\ \begin{pmatrix} \phi & 1-\phi \\ 0 & 1 \end{pmatrix}_t \end{array}$$

et entre états et événements d'autre part :

$$B_t = \begin{array}{c} \\ V \\ \dagger \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1-p & p \\ 1 & 0 \end{pmatrix}_t \end{array}.$$

Les modèles de capture-recapture conditionnant à la première observation de chaque individu, la matrice d'observation est triviale à cette date et s'obtient en fixant les taux de capture p^t à 1 :

$$B_t^0 = \begin{array}{c} \\ V \\ \dagger \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}_t \end{array}.$$

Nous avons modifié le modèle de Jolly-Seber en considérant deux classes d'individus qui diffèrent par leur paramètre de détection : les fortement et les faiblement capturables. Cette dichotomie peut correspondre à une différence de comportement entre individus dominants et dominés, mais comme indiqué dans l'introduction, ce n'est pas nécessaire. Il s'agit simplement de prendre en compte l'hétérogénéité de capture afin d'estimer plus exactement l'effectif de la population. L'introduction de ces 2 classes de détectabilité amène à considérer 2 états vivants et conséquemment à introduire de nouveaux paramètres. En particulier, lors de leur première détection, les individus peuvent appartenir à l'une ou l'autre des 2 classes de détectabilité selon une probabilité à estimer. Dans ce modèle simple, nous considérons que les individus ne changent pas de classe.

Nous avons donc deux classes de capturabilité : « facilement capturable » (F) et « difficilement capturable » (D). Les ensembles regroupant les observations d'une part et les états de l'autre sont alors :

- $\Omega = \{$pas vu (codé par 0), vu (codé par 1)$\}$,
- $E = \{F, D, \dagger\}$.

Les paramètres du modèle sont :

1. les probabilités des états initiaux : π_F^t et $\pi_D^t = 1 - \pi_F^t$, $t \in \{1, \cdots, T\}$;
2. les probabilités de survie-transition :
$$\phi_{FF}^t,\ \phi_{FD}^t,\ \phi_{DF}^t \text{ et } \phi_{DD}^t,\ t \in \{1, \cdots, T-1\} ;$$
3. les probabilités de capture : p_F^t, des individus appartenant à la classe « facilement capturable », et p_D^t, de ceux appartenant à la classe « difficilement capturable », $t \in \{2, \cdots, T\}$.

La représentation matricielle de ces paramètres devient :

$$\Pi_t = (\pi_F, \pi_D, 0)_t,$$

$$\Phi_t = \begin{array}{c} \\ F \\ D \\ \dagger \end{array} \begin{array}{c} \begin{array}{ccc} F & D & \dagger \end{array} \\ \begin{pmatrix} \phi_{F\,F} & \phi_{F\,D} & 1-\phi_{F\,F}-\phi_{F\,D} \\ \phi_{D\,F} & \phi_{D\,D} & 1-\phi_{D\,F}-\phi_{D\,D} \\ 0 & 0 & 1 \end{pmatrix}_t \end{array},$$

$$B_t = \begin{array}{c} \\ F \\ D \\ \dagger \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1-p_F & p_F \\ 1-p_D & p_D \\ 1 & 0 \end{pmatrix}_t \end{array}.$$

Les paramètres d'observation relatifs à la première rencontre des individus regroupés dans la matrice B_t^0 s'obtiennent à nouveau en fixant les taux de capture p_F^t et p_D^t à 1 :

$$B_t^0 = \begin{array}{c} \\ F \\ D \\ \dagger \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}_t \end{array}.$$

13.3.2 Une formule corrigée pour le calcul de l'effectif

Dans le modèle de Jolly-Seber, pour une probabilité de détection estimée p_t au temps t et un nombre total capturé n_t, l'estimation de l'effectif N_t est n_t/p_t. Cette formule générique valable pour tous les modèles de capture-recapture à une seule classe de capturabilité permet donc de passer des probabilités de détectabilité estimées aux effectifs estimés. C'est un cas particulier trivial de l'estimateur de Horvitz-Thompson (voir par exemple Seber [1982]) qui estime la taille de la population par la somme $\sum_i 1/p_i$ où p_i est la probabilité de détection de l'individu i. Si maintenant on définit m_t le nombre d'individus déjà marqués pris au temps t, u_t le nombre d'individus nouvellement marqués capturés au temps t, $p_t(L)$ la probabilité de capture des individus faiblement capturables et $p_t(H)$ la probabilité de capture des individus fortement capturables, ces individus étant en proportion π et $1-\pi$ dans les nouvellement marqués, l'estimateur de Horvitz-Thompson montre que l'estimation de l'effectif de la population tenant compte de l'hétérogénéité de capture est approximativement égale à :

$$m_t/p_t(H) + \pi u_t/p_t(L) + (1-\pi)u_t/p_t(H).$$

L'approximation porte sur le fait que les individus déjà marqués sont traités comme s'ils étaient tous fortement capturables bien qu'une proportion d'entre

eux, inconnue mais généralement faible, soit faiblement capturable. Avec cette approximation, on tend donc à sous-estimer la taille de la population.

13.4 Les résultats

Le fichier de données mis à notre disposition était formé de listes d'événements de « capture » par individu, jour, mois, année et lieu. Les données retenues pour l'analyse, 96 individus, s'étalent d'avril 1995 à décembre 2001 et sont regroupées en 27 trimestres. Elles concernent le seul Mercantour. La figure 13.2 représente les estimations de l'effectif de loups en Mercantour selon les différentes saisons.

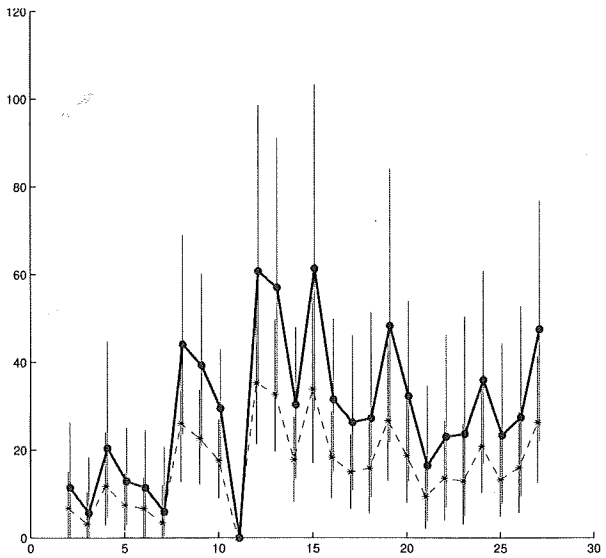

Figure 13.2 – *Effectifs de la population de loups en Mercantour*[2].

En utilisant le modèle corrigeant pour l'hétérogénéité (en traits continus), le dernier effectif estimé à l'automne 2001 est de 48 individus avec un intervalle de confiance assez large de 22 à 77 individus. Les résultats du modèle de Jolly-Seber en pointillé sous-estiment systématiquement le nombre de loups. On voit apparaître une oscillation des effectifs suivant les saisons avec des pics qui sont souvent hivernaux et ce malgré la plus forte valeur de taux de détection en

2. La période de suivi s'étale d'avril 95 à décembre 2001 ; le pas de temps utilisé est le trimestre. Les estimations obtenues par le modèle de Jolly-Seber sont en pointillé et celles obtenues par le modèle corrigeant pour l'hétérogénéité de capture sont en continu. Les intervalles de confiance à 95% sont donnés sous forme de barres verticales. Note : les résultats obtenus selon le modèle avec hétérogénéité ont légèrement été décalés sur l'axe des abcisses pour une meilleure lisibilité.

hiver. Ces oscillations de saison ne semblent pas induites par le choix d'un modèle saison sur la probabilité de détection [3].

13.5 Conclusion

Il est impératif de corriger l'estimation des effectifs pour tenir compte de la détectabilité mais aussi dans le cas particulier des loups pour l'hétérogénéité de détection, sous peine d'une estimation biaisée des effectifs de la population. Bien que des paramètres supplémentaires soient à considérer (2 probabilités de capture au lieu d'une), le coût quant à la précision sur les paramètres est acceptable. Globalement, le comportement des modèles que nous avons développés présente une bonne stabilité numérique malgré le faible nombre de données. En outre, ces modèles peuvent être facilement généralisés pour prendre en compte le changement de statut des individus de faiblement à fortement capturables et inversement.

D'un point de vue biologique, il conviendrait d'étudier les variations saisonnières en relation avec le cycle de vie du loup. Il faudrait aussi prendre en compte dans les données le mélange de jeunes et d'adultes, ainsi que de mâles et de femelles. Enfin, il serait bon de relier les informations spatiales des individus à leur statut d'animal en transit et pour ce, de faire une évaluation et un bilan avec une équipe italienne qui assure le même suivi de l'autre côté de la frontière.

Les modélisations par capture-recapture proposées ici peuvent être adaptées pour prendre en compte ce type d'éléments. Elles débouchent donc sur d'excellentes perspectives à condition que les difficultés bien compréhensibles pour un jeu de données aussi original puissent être explorées et levées.

3. Les auteurs tiennent à remercier E. Marboutin pour son investissement dans l'analyse des données et l'interprétation des résultats.