Association
mathématique
du Québec

ISM
Institut des sciences mathématiques

Théorie et applications des modèles de capture-recapture

Louis-Paul Rivest, Département de mathématiques et de statistique,
Université Laval
lpr@mat.ulaval.ca
http ://www.mat.ulaval.ca/pages/lpr

Résumé

Ce travail donne les bases théoriques sous-jacentes aux modèles de capture-recapture à deux et à trois occasions de capture. Ces modèles sont utilisés pour estimer la taille de populations difficiles à rejoindre, telles le nombre d'animaux dans un certain territoire ou bien le nombre de personnes souffrant d'une maladie. Il traite aussi de modèles pour populations ouvertes avec lesquels il est un peu plus difficile d'obtenir de bons estimateurs des caractéristiques démographiques de la population étudiée. Le sujet présenté pourrait faire l'objet d'un travail long dans le cadre d'un projet d'activité d'intégration avec des étudiants ayant fait un cours de calcul à plusieurs variables. L'auteur a déjà publié un article dans *Accromath* [3] qui montre une utilisation particulière des modèles de capture-recapture.

1 Les données

En capture-recapture, le problème classique est l'estimation de la taille d'une population de poissons. Il est représenté schématiquement à la figure 1. Au jour 1, six poissons sont capturés et marqués ; une marque noire les identifie à la figure 1. On répète la procédure au jour 2 : les poissons capturés sont marqués, s'ils ne le sont pas déjà, et relâchés ; une marque rouge les identifie à la figure 1. Après deux jours, il y a quatre historiques de capture possibles pour un poisson, soit $(0,0)$, $(1,0)$, $(0,1)$ et $(1,1)$, où 1 représente une capture et 0 une non-capture. Les poissons avec un historique $(0,0)$ ne sont pas vus ; ainsi leur nombre n_0 est inconnu. Les fréquences des autres historiques sont connues ; pour l'expérience de la figure 1, elles sont données au tableau 1.

© Association mathématique du Québec

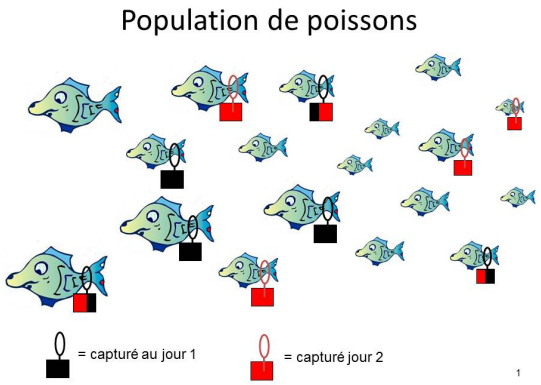

FIGURE 1 – Représentation schématique d'une population de poissons ayant subi une expérience de capture-recapture. Une marque noire (resp. rouge) indique une capture au jour 1 (resp. 2)

	Manqué jour 2	Capturé jour 2
Manqué jour 1	$n_0 = ?$	$n_{01} = 4$
Capturé jour 1	$n_{10} = 3$	$n_{11} = 3$

Tableau 1 – Fréquences des historiques de capture pour l'expérience de la figure 1.

En tout, 10 poissons différents ont été capturés dans l'expérience rapportée à la figure 1. La taille de la population est forcément supérieure à 10. Pour estimer cette taille il faut un modèle mathématique qui pose des hypothèses quant au mécanisme de capture et qui permettra d'estimer n_0.

À la figure 1, les poissons forment une population difficile à rejoindre et les deux jours de capture permettent d'obtenir des listes incomplètes, ou partielles, des unités de la population. Ce cadre conceptuel a de nombreuses applications. Par exemple, Tauvo [8] cherche à estimer le nombre d'enfants mendiants à Dakar au Sénégal dans le cadre d'une étude réalisée pour la Banque mondiale. La capture est effectuée par des préposés qui essaient d'interviewer des enfants dans des lieux publics à Dakar. La recapture utilise *le sondage déterminé par les répondants* (SDR) qui est défini ainsi : des primes sont données aux enfants qui acceptent de participer. L'enfant reçoit une récompense pour répondre au questionnaire d'enquête et pour recruter de nouveaux enfants pour participer à l'enquête. L'estimation du nombre total d'enfants mendiants est un des objectifs de l'étude. Elle utilise les données du tableau 2. Les expériences considérées jusqu'ici, sur les poissons dans un lac et sur les mendiants à Dakar, sont très différentes. Elles ont cependant un objectif commun, celui d'estimer la taille d'une population à l'aide de deux listes partielles de ses membres.

	Manqué SDR	Capturé SDR
Manqué préposé	$n_0 = ?$	$n_{01} = 486$
Capturé préposé	$n_{10} = 425$	$n_{11} = 92$

Tableau 2 – Fréquence des historiques de capture pour l'estimation du nombre d'enfants mendiants à Dakar

Souvent plus de deux listes incomplètes sont disponibles pour estimer la taille d'une population. Le tableau 3 présente trois exemples tirés de la littérature scientifique qui font intervenir trois listes. Dans chaque cas, il donne les fréquences des sept historiques de captures observables. Le premier, "11 Sept", est tiré de Murphy [5] qui veut estimer le nombre de personnes qui, bien qu'elles aient été présentes dans les tours jumelles, ont survécu aux attentats du 11 septembre 2001. La première liste contient les noms des personnes qui se sont portées volontaires pour participer à une étude sur les effets du 11 septembre ; la deuxième a été fournie par des entreprises hébergées dans les deux tours et la troisième, soumise par le *Port Authority of New York*, contient le nom des personnes qui avaient un droit d'accès aux tours jumelles.

Historique de capture			"11 Sept"	Arapiraca	Recensement
1	1	1	174	3	72
1	1	0	88	8	155
1	0	1	1658	13	7
1	0	0	1702	133	71
0	1	1	750	12	13
0	1	0	270	83	53
0	0	1	4323	64	43
Total			8965	316	414

Tableau 3 – Données pour trois expériences de capture-recapture faisant intervenir des listes incomplètes : "11 Sept" réfère aux données de Murphy [5], Arapiraca à Bezerra et coll. [2] et Recensement à Rivest [7]

Le deuxième jeu de données est tiré de Bezerra et coll. [2]. Il cherche à estimer le nombre d'enfants itinérants à Arapiraca au Brésil. Les trois listes ont été compilées à partir des arrestations d'enfants répertoriées au bureau de procureur de la ville et de deux enquêtes sur la rue. Le troisième jeu de données, traité dans Rivest [7], cherche à estimer le nombre d'Afro-Américains dans un district des États-unis en 1988. Les trois listes incomplètes sont : le recensement du district, un dénombrement post-censitaire des résidents et une liste administrative de résidents compilée par des agences gouvernementales.

2 Modèles et estimation pour deux occasions de capture

Dans la suite du texte, N représente la taille de la population, le paramètre que l'on cherche à estimer. Le modèle pour deux occasions de capture s'appuie sur l'hypothèse d'indépendance entre les deux occasions de capture. Pour la population de la figure 1, cette hypothèse signifie qu'au jour 2 un poisson marqué a la même probabilité d'être capturé qu'un poisson non marqué. Soient p_1 et p_2 les probabilités de capture aux jours 1 et 2. Les valeurs prédites μ pour les fréquences des quatre historiques de capture sont données au tableau 4 pour le modèle d'indépendance.

	Manqué jour 2	Capturé jour 2
Manqué jour 1	$\mu_{00} = N(1-p_1)(1-p_2)$	$\mu_{01} = N(1-p_1)p_2$
Capturé jour 1	$\mu_{10} = Np_1(1-p_2)$	$\mu_{11} = Np_1p_2$

Tableau 4 – Valeurs prédites des fréquences des historiques de capture sous le modèle d'indépendance

Le rapport de cotes θ de ce tableau correspond au produit des termes sur la diagonale du tableau divisé par le produit des termes hors diagonale. On vérifie facilement que $\theta = 1$, c'est-à-dire

$$\theta = \frac{\mu_{00} \times \mu_{11}}{\mu_{10} \times \mu_{01}} = 1 \quad \text{ou bien} \quad \mu_{00} = \frac{\mu_{10} \times \mu_{01}}{\mu_{11}}\,.$$

Sous l'hypothèse d'indépendance, une bonne estimation du nombre de poissons manqués est donc $\hat{n}_0 = n_{10} \times n_{01}/n_{11}$. Ceci conduit à *l'estimateur de Petersen*, du nom de l'écologiste qui a été l'un des premiers utilisateurs de l'estimateur du total, soit

$$\hat{N} = n_{10} + n_{01} + n_{11} + \frac{n_{10} \times n_{01}}{n_{11}} = (n_{10} + n_{11}) \Big/ \frac{n_{11}}{n_{01} + n_{11}}\,.$$

C'est le nombre de poissons capturés à l'occasion 1 divisé par la proportion des poissons capturés à l'occasion 2 qui ont aussi été capturés à l'occasion 1.

En général $\hat{N} \neq N$ et on mesure la qualité de $\hat{N}$ comme estimation de N à l'aide de la variance échantillonnale, notée $\mathrm{var}(\hat{N})$. En effet, la statistique $\hat{N}$ est, au moins approximativement, une variable aléatoire de loi normale, centrée à la vraie valeur N et de variance $\mathrm{var}(\hat{N})$. Ainsi si $v(\hat{N})$ est une estimation de variance, un intervalle de confiance approximatif à 95% pour la taille de population inconnue N est donné par $\hat{N} \pm 2\sqrt{v(\hat{N})}$. Pour la plupart des estimateurs considérés ici, l'estimateur de variance s'écrit

$$v(\hat{N}) = \hat{n}_0 + \hat{n}_0^2 \times \left(\sum_{\omega \in \Omega} \frac{1}{n_\omega} \right),$$

où Ω est l'ensemble des historiques de capture ω observables et n_ω représente la fréquence de ω. Ainsi pour l'estimateur de Petersen, on a

$$v(\hat{N}) = \frac{n_{01} \times n_{10}}{n_{11}} + \left(\frac{n_{01} \times n_{10}}{n_{11}}\right)^2 \times \left(\frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{11}}\right) = \hat{N} \times \frac{n_{01} \times n_{10}}{n_{11}^2}\,.$$

Pour les données du tableau 1,

$$\hat{N} = 3 + 4 + 3 + \frac{3 \times 4}{3} = 14 \quad \text{et} \quad v(\hat{N}) = 14 \times \frac{3 \times 4}{3^2} = 4.3^2.$$

L'intervalle de confiance à 95% pour N est donc $[5, 23]$. Puisque 10 poissons ont été vus, on peut toutefois écrire $[10, 23]$. Dans la figure 1, il y a exactement $N = 18$ poissons ; cette valeur est bien dans l'intervalle de confiance à 95%.

L'hypothèse d'indépendance entre occasions de capture est cruciale pour la validité de l'estimateur de Petersen. En termes mathématiques, elle s'écrit $\theta = 1$, où θ est le rapport de cotes du tableau 4. Elle est violée si la population est hétérogène, c'est-à-dire si les probabilités de capture varient d'une unité à l'autre. Si tel est le cas, il est vraisemblable qu'une unité capturée à l'occasion 1 aura une probabilité de capture élevée ; elle aura donc plus de chance d'être recapturée à l'occasion 2 qu'une unité non capturée en 1. Dans ce cas, on parle d'une association positive entre les deux captures et le rapport de cote satisfait $\theta > 1$. L'estimateur de Petersen, qui suppose que $\theta = 1$, sous-estime alors la taille de la population. Un exemple numérique du biais associé à un estimateur de Petersen calculé avec une population hétérogène est donné dans Genest et Rivest [3].

3 Modèles et estimation pour trois occasions de capture

Dans cette section, le fait de disposer de trois occasions de capture va permettre de s'affranchir de l'hypothèse d'indépendance. Dans le tableau 5, les trois occasions de capture sont appelées J1, J2 et J3. Les données sont divisées en deux sous-tableaux selon que le sujet ait ou non été capturé à J1.

Manqué J1	Manqué J3	Capturé J3	Capturé J1	Manqué J3	Capturé J3
Manqué J2	$n_0 = ?$	n_{001}	Manqué J2	n_{100}	n_{101}
Capturé J2	n_{010}	n_{011}	Capturé J2	n_{110}	n_{111}

Tableau 5 – Fréquences observées pour une expérience avec trois occasions de capture

Supposons que les captures en J2 et J3 soient dépendantes. Cette dépendance peut se mesurer à l'aide du rapport de cotes du tableau des capturés en J1,

$$\hat{\theta}_1 = \frac{n_{111} \times n_{100}}{n_{101} \times n_{110}} .$$

On peut ensuite reporter ce rapport de cotes au tableau des manqués en J1 pour obtenir une estimation de n_0 qui tient compte de la dépendance entre J2 et J3, à savoir :

$$\hat{n}_0 = \hat{\theta}_1 \times \frac{n_{001} \times n_{010}}{n_{011}} .$$

L'estimation de la taille de la population est alors donnée par

$$\hat{N} = n + \frac{n_{111} \times n_{100} \times n_{001} \times n_{010}}{n_{101} \times n_{110} \times n_{011}} ,$$

où n est le total des sept fréquences n_ω données au tableau 5. C'est l'estimateur le plus général lorsque l'on dispose de trois occasions de capture ; sa variance peut être estimée par

$$v(\hat{N}) = \hat{n}_0 + \hat{n}_0^2 \sum_{\omega \in \Omega} \frac{1}{n_\omega} .$$

Cette variance est élevée lorsque certaines fréquences n_ω sont faibles. Plutôt que d'utiliser à tout coup l'estimateur le plus général, on va chercher à déterminer si certains rapports de cotes θ_i sont égaux à 1. Si tel est le cas, on pourra utiliser un estimateur de N plus simple, avec une plus petite variance. Supposant par exemple que $\theta_2 = 1$, ce qui veut dire que les captures à J1 et à J3 sont indépendantes étant donné que l'unité est capturée en J2. On peut alors estimer n_0 en appliquant l'estimateur pour deux occasions de capture au tableau des manqués en J2. L'estimateur de N est alors

$$\hat{N} = n + \frac{n_{100} \times n_{001}}{n_{101}} .$$

Sa variance estimée a une forme semblable à celle de l'estimateur de Petersen, soit

$$v(\hat{N}) = \frac{n_{100} \times n_{001}}{n_{101}} + \left(\frac{n_{100} \times n_{001}}{n_{101}} \right)^2 \times \left(\frac{1}{n_{100}} + \frac{1}{n_{001}} + \frac{1}{n_{101}} \right) ;$$

elle est plus petite que celle de l'estimateur général qui ne fait aucune hypothèse sur les rapports de cotes θ_i.

Si deux des trois rapports de cotes sont égaux à 1, disons $\theta_2 = \theta_3 = 1$, une dépendance entre les captures aux jours 2 et 3 est possible. Dans ce cas, on peut appliquer l'estimateur de Petersen en combinant les jours 2 et 3 en une seule occasion de capture. Il y alors deux occasions de capture, J1 et "J2–3" ; les fréquences des nouveaux historiques sont alors données au tableau 6. L'estimateur de Petersen pour les données du tableau 6 s'écrit

	Manqué jour 2–3	Capturé jour 2–3
Manqué jour 1	$n_0 = ?$	$n_{010} + n_{001} + n_{011}$
Capturé jour 1	n_{100}	$n_{110} + n_{101} + n_{111}$

Tableau 6 – Historique de capture obtenu en combinant J2 et J3

$$\hat{N} = n + \frac{n_{100} \times (n_{010} + n_{001} + n_{011})}{n_{110} + n_{101} + n_{111}}\,.$$

Si les trois rapports de cotes sont égaux à 1, le modèle d'indépendance entre les trois occasions de capture est vraisemblable. Dans ce cas, $\hat{N}$ n'a pas une forme explicite et nous ne traitons pas de cet estimateur par la suite.

Lorsqu'il y a trois occasions de capture, il faut donc faire une analyse de données pour choisir le modèle à partir duquel N est estimé. Cette analyse vise à déterminer si 1 est une valeur plausible pour chacun des trois rapports de cotes. Puisqu'ils sont inconnus, il faut les estimer et faire des tests statistiques pour déterminer si 1 est une valeur plausible. Intéressons-nous à θ_3. Pour tester $H_0 : \theta_3 = 1$, on utilise un estimateur de variance pour le logarithme du rapport de cotes. On a

$$\hat{\theta}_3 = \frac{n_{001} \times n_{111}}{n_{011} \times n_{101}} \quad \text{et} \quad v\{\ln(\hat{\theta}_3)\} = \frac{1}{n_{001}} + \frac{1}{n_{111}} + \frac{1}{n_{101}} + \frac{1}{n_{011}}\,.$$

Le test d'hypothèse rejette H_0 au seuil 5% si la statistique z_{obs} suivante est, en valeur absolue, supérieure à 2 :

$$z_{\text{obs}} = \ln(\hat{\theta}_3) \Big/ \sqrt{\frac{1}{n_{001}} + \frac{1}{n_{111}} + \frac{1}{n_{101}} + \frac{1}{n_{011}}}\,.$$

Ici, "ln" représente le logarithme naturel ou népérien. Les tests pour θ_1 et θ_2 se font de la même façon. Considérons par exemple les données du 11 septembre et estimons θ_1. On trouve

$$\hat{\theta}_1 = \frac{174 \times 1702}{88 \times 1658} = 2.03 = e^{.71} \quad \text{et} \quad v(\hat{\theta}_1) = \frac{1}{174} + \frac{1}{1702} + \frac{1}{88} + \frac{1}{1658} = .14^2.$$

Ainsi un simple calcul donne $z_{\text{obs}} = .71/.14 \approx 5.1$. Puisque cette valeur est plus grande que 2, l'hypothèse nulle $H_0 : \theta_1 = 1$ est rejetée. Il y a donc une association positive entre les deux dernières listes. On trouve de même $\hat{\theta}_2 = .71$, $\ln(\hat{\theta}_2) = -.34$, $v(\hat{\theta}_2) = .15^2$, $z_{\text{obs}} = -2.3$ et $\hat{\theta}_3 = .60$, $\ln(\hat{\theta}_3) = -.50$, $v(\hat{\theta}_3) = .09^2$, $z_{\text{obs}} = -5.5$. Ainsi les trois statistiques z sont plus grandes que 2 en valeur absolue et, pour les données du 11 septembre, on doit utiliser le modèle le plus général. On obtient

$$\hat{N} = n + \hat{n}_0 = 8965 + \frac{174 \times 1702 \times 270 \times 4323}{88 \times 1658 \times 750} = 12\ 124,$$

$$v(\hat{N}) = 3159 + 3159^2 \times \left(\frac{1}{174} + \cdots + \frac{1}{4323}\right) = 488^2.$$

et un intervalle de confiance à 95% donné par $[11\ 147, 13\ 100]$.

4 Un peu de théorie

Les fréquences n_ω dans une expérience de capture-recapture suivent une loi multinomiale de paramètres N et p_ω, où p_ω est la probabilité qu'une unité ait l'historique de capture ω. Pour l'expérience à trois occasions de capture, il y a huit p_ω et leur somme donne 1, peu importe le modèle retenu pour les données. Il est souvent utile de considérer que les fréquences n_ω sont indépendantes et suivent des lois de Poisson d'espérance $\mu_\omega = Np_\omega$. La masse de probabilité de ces variables aléatoires est donnée par

$$\Pr(n_\omega = k) = \frac{\mu_\omega^k}{k!}\, e^{-\mu_\omega}, \quad k = 0, 1, 2, \ldots$$

Le modèle le plus général pour l'expérience à trois occasions de capture fait intervenir sept paramètres log-linéaires. Il y a l'ordonnée à l'origine γ, les effets simples β_1, β_2, β_3, et les interactions deux à deux, λ_{12}, λ_{13}, λ_{23}. Les fréquences prédites μ_ω s'expriment en fonction des sept paramètres de la manière suivante :

$$\mu_\omega = \exp(\gamma + \omega_1\beta_1 + \omega_2\beta_2 + \omega_3\beta_3 + \lambda_{12}\omega_1\omega_2 + \lambda_{13}\omega_1\omega_3 + \lambda_{23}\omega_2\omega_3),$$

où ω_1, ω_2 et ω_3 sont les composantes du vecteur ω ; elles valent 1 pour une capture et 0 si l'unité est manquée. On note que $\mu_0 = e^\gamma$ représente la fréquence prédite des unités manquées ; son estimation permet d'estimer N. Le tableau 7 reprend le tableau des capturés en J1 du tableau 5 et donne les fréquences théoriques des quatre cellules en fonction des paramètres du modèle log-linéaire.

Capturé en J1	Manqué J3	Capturé J3
Manqué J2	$\mu_{100} = \exp(\gamma + \beta_1)$	$\mu_{101} = \exp(\gamma + \beta_1 + \beta_3 + \lambda_{13})$
Capturé J2	$\mu_{110} = \exp(\gamma + \beta_1 + \beta_2 + \lambda_{12})$	$\mu_{111} = \exp(\gamma + \beta_1 + \beta_2 + \beta_3 + \lambda_{12} + \lambda_{13} + \lambda_{23})$

Tableau 7 – Fréquences théoriques du tableau des capturés en J1

Le rapport de cotes théorique du tableau des unités capturées en J1 s'écrit

$$\theta_1 = \frac{\mu_{111} \times \mu_{100}}{\mu_{101} \times \mu_{110}} = e^{\lambda_{23}}.$$

Ainsi les logarithmes des rapports de cotes sont les paramètres d'interaction dans le modèle log-linéaire et les tests présentés à la section précédente cherchent à déterminer un modèle log-linéaire simple qui s'ajuste bien aux données, basé sur l'hypothèse de la loi de Poisson évoquée précédemment.

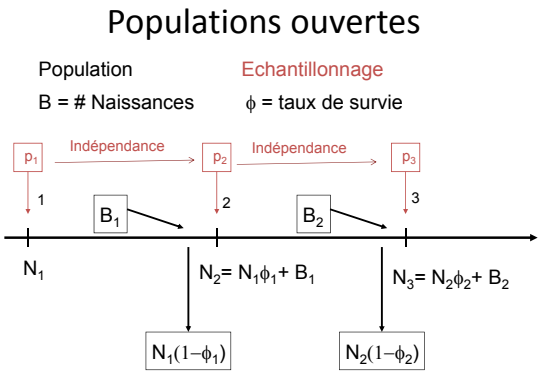

FIGURE 2 – Représentation schématique de l'évolution d'une population ouverte. Les paramètres ϕ (survie) et B (naissance) sont des caractéristiques de la population. Les probabilités de capture p dépendent du mécanisme d'observation.

Il faut finalement souligner que les estimateurs de N présentés aux sections 2 et 3 sont des estimateurs du maximum de vraisemblance. La vraisemblance, vue comme fonction des paramètres, est la probabilité d'observer les résultats obtenus. Pour le modèle à sept paramètres, la vraisemblance de Poisson s'écrit

$$L(\lambda, \beta_1, \beta_2, \beta_3, \lambda_{12}, \lambda_{13}, \lambda_{23}) = \prod_{\omega \in \Omega} \frac{\mu_\omega^{n_\omega} e^{-\mu_\omega}}{n_\omega!} .$$

Les estimateurs du maximum de vraisemblance sont obtenus en maximisant cette fonction. On peut, par exemple, montrer que pour que la fonction L soit maximale, il faut que

$$e^{\hat{\lambda}} = \frac{n_{111} \times n_{100} \times n_{001} \times n_{010}}{n_{101} \times n_{110} \times n_{011}} .$$

5 Modèles pour des populations ouvertes

Lorsque les occasions de capture sont espacées dans le temps, des naissances et des morts sont possibles entre deux périodes d'observation. Si des captures et des recaptures ne sont réalisées que sur le site de reproduction d'une espèce à la période des amours, les occasions de capture sont les années et la taille de la population change d'une année à l'autre. Une telle population est dite ouverte. À chaque année il y a des naissances, dont le nombre est noté B à la figure 2 ; il y a aussi des décès. À la figure 2, ϕ dénote la probabilité de survie entre deux occasions de capture.

Les données sont similaires à celles de la section 3. S'il y a trois années d'observations, il y a sept historiques de capture observables. Évidemment, un historique particulier ne permet pas de déterminer exactement le destin d'un animal. Par exemple, l'historique $(0, 1, 0)$ peut être associé à un animal présent dans la population lors des trois années de l'étude mais qui a été manqué aux années 1 et 3. Il peut également représenter un animal né après la première année, capturé à la deuxième et décédé à la troisième. L'objectif de cette section est donc d'utiliser l'estimateur de Petersen dans ce contexte et d'évaluer la qualité des estimations qu'il permet de calculer.

Pour la suite, on s'intéresse plus particulièrement à deux paramètres : ϕ_1, la probabilité de

Historique			μ
ω_1	ω_2	ω_3	
1	1	1	$M_2 p_2 r$
1	1	0	$M_2 p_2 (1-r)$
1	0	1	$M_2 (1-p_2) r$
1	0	0	$M_2\{(1-\phi_1)/\phi_1 + (1-p_2)(1-r)\}$
0	1	1	$U_2 p_2 r$
0	1	0	$U_2 p_2 (1-r)$

Tableau 8 – Fréquences prédites μ_ω pour un modèle de population ouverte

survie entre les années 1 et 2, ainsi que N_2, la taille de la population à l'année 2. Pour estimer ces paramètres, la fréquence n_{001} n'est pas informative. On n'utilisera donc que les fréquences des six autres historiques de capture. Le tableau 8 donne les fréquences prédites selon le modèle de la figure 2 en fonction des paramètres suivants :

a) ϕ_1, p_2 tel que définis à la figure 2 ;
b) $r = \phi_2 p_3$ est la probabilité de survivre de l'année 2 à 3 et d'être capturé à l'année 3 ;
c) $M_2 = N_1 p_1 \phi_1$, l'espérance du nombre d'animaux marqués juste avant les captures à l'année 2 ;
d) $U_2 = N_1(1-p_1)\phi_1 + B_1$, l'espérance du nombre d'animaux non marqués dans la population avant les captures à l'année 2.

Pour estimer $N_2 = M_2 + U_2$, on peut utiliser l'estimateur de Petersen basé sur les deux « listes » suivantes : (1) présents dans la population au temps 2 et (2) capturés au temps 1 et 3. Ces animaux sont forcément présents dans la population au temps 2 ; on va donc les utiliser pour estimer p_2, la probabilité de capture au temps 2. L'indépendance des captures d'une occasion à l'autre, évoquée à la figure 2, fait en sorte que cette estimation de p_2 est non biaisée.

	Manqué liste 2	Capturé liste 2
Manqué liste 1	$n_0 = ?$	n_{101}
Capturé liste 1	$n_{011} + n_{110} + n_{010}$	n_{111}

Tableau 9 – Fréquence des manqués et des capturés pour estimer N_2 avec l'estimateur de Petersen

L'estimateur de Petersen obtenu à l'aide des données du Tableau 9 est

$$\hat{N}_2^P = n_{+1+} + n_{101} + \frac{n_{101} \times (n_{011} + n_{110} + n_{010})}{n_{111}} = n_{+1+} \times \left(1 + \frac{n_{101}}{n_{111}}\right),$$

où n_{+1+} est la somme des fréquences pour les quatre historiques qui ont une capture au temps 2. Le paramètre ϕ_1 peut, quant à lui, être estimé par S_1, le nombre de survivants à l'année 2 parmi les n_{1++} animaux relâchés marqués à l'année 1, divisé par n_{1++}. Bien que S_1 soit inconnu, il peut être estimé à l'aide de l'estimateur de Petersen basé sur les deux listes suivantes : (1) capturé marqué à l'année 2 et (2) capturé à l'année 3 et marqué à l'année 1. On a donc

$$\hat{\phi}_1^P = \frac{1}{n_{1++}} \left(n_{111} + n_{110} + n_{101} + \frac{n_{110} \times n_{101}}{n_{111}}\right).$$

Ces estimateurs sont convergents au sens de Fisher, c'est-à-dire que si, dans les expressions pour $\hat{N}_2^P$ et $\hat{\phi}_1^P$, on remplace les fréquences n_ω par leurs valeurs prédites données au tableau 8, on trouve les paramètres N_2 et ϕ_1. Démontrons ce résultat pour $\hat{\phi}_1^P$. La valeur prédite pour n_{1++} est $\mu_{1++} = M_2/\phi_1$ selon le tableau 8. De même, la valeur prédite pour $n_{111} + n_{110} + n_{101} + n_{110}n_{101}/n_{111}$ est

$$\begin{aligned}\mu_{111} + \mu_{110} + \mu_{101} + \frac{\mu_{110}\mu_{101}}{\mu_{111}} \\ = M_2 p_2 r + M_2 p_2 (1-r) + M_2 (1-p_2) r + \frac{M_2 p_2 (1-r) M_2 (1-p_2) r}{M_2 p_2 r} = M_2.\end{aligned}$$

Ainsi $\hat{\phi}_1^P$ est convergent au sens de Fisher. On peut montrer de manière semblable que $\hat{N}_2^P$ est convergent au sens de Fisher.

Les estimateurs $\hat{N}_2^P$ et $\hat{\phi}_1^P$ font-ils la meilleure utilisation possible de l'information disponible ? Existe-il d'autres estimateurs de ces paramètres qui soient convergents au sens de Fisher et plus précis que $\hat{N}_2^P$ et λ_1^P ? On peut présumer que $\hat{\phi}_1^P$ n'est pas le meilleur estimateur. En effet, au tableau 8, le rapport de cotes des capturés à l'année 2 vaut 1. Ainsi

$$\frac{\mu_{110}}{\mu_{111}} = \frac{\mu_{010}}{\mu_{011}} = \frac{\mu_{110} + \mu_{010}}{\mu_{111} + \mu_{011}}.$$

En remplaçant n_{110}/n_{111} par $(n_{110}+n_{010})/(n_{111}+n_{011})$ dans la formule pour $\hat{\phi}_1^P$, on obtient un estimateur convergent au sens de Fisher. Il est donné par

$$\hat{\phi}_1 = \frac{1}{n_{1++}}\left\{n_{111}+n_{110}+n_{101}+\frac{n_{101}\times(n_{110}+n_{010})}{n_{111}+n_{011}}\right\}.$$

Il est plus précis que $\hat{\phi}_1^P$ car il utilise cinq des fréquences du tableau 8, alors que $\hat{\phi}_1^P$ n'en utilise que trois. Pour déterminer à coup sûr les meilleurs estimateurs de N_2 et ϕ_1, il faut maximiser la vraisemblance de Poisson pour les cinq paramètres de ce modèle. C'est ce qui est fait à la section suivante.

6 Maximisation de la vraisemblance de Poisson pour un modèle de populations ouvertes

La vraisemblance de Poisson construite à partir des fréquences des six historiques de capture du tableau 8 est donnée par

$$L(\phi_1,p_2,r,M_2,U_2) = \prod_{6\ \omega}\frac{\mu_\omega^{n_\omega}e^{-\mu_\omega}}{n_\omega!}\,.$$

À une constante de proportionnalité près, elle est égale à

$$(1-p_2)^{n_{101}}M_2^{n_{1++}}p_2^{n_{+1+}}r^{n_{++1}}(1-r)^{n_{+10}}U_2^{n_{01+}}\left\{\frac{1-\phi_1}{\phi_1}+(1-p_2)(1-r)\right\}^{n_{100}}\exp\left(-\frac{M_2}{\phi_1}-U_2p_2\right),$$

où $n_{++1}=n_{101}+n_{011}+n_{111}$. Soit $\ell(\phi_1,p_2,r,M_2,U_2)$ la log-vraisemblance, c'est-à-dire le logarithme de cette expression. Ses dérivées partielles par rapport à M_2 et U_2 sont données par

$$\frac{\partial}{\partial M_2}\,\ell(\phi_1,p_2,r,M_2,U_2)=\frac{n_{1++}}{M_2}-\frac{1}{\phi_1}\,,\quad \frac{\partial}{\partial U_2}\,\ell(\phi_1,p_2,r,M_2,U_2)=\frac{n_{01+}}{U_2}-p_2.$$

Donc, pour (r,p_2,ϕ_1) fixé, la vraisemblance est maximale à $M_2=\phi_1 n_{1++}$ et $U_2=n_{01+}/p_2$. De plus le maximum de ℓ est, à une constante près, égal à

$$\begin{aligned}\ell_2(\phi_1,p_2,r) = {}& n_{1++}\ln(\phi_1)+n_{11+}\ln(p_2)+n_{101}\ln(1-p_2)+n_{++1}\ln(r)\\ &+n_{+10}\ln(1-r)+n_{100}\ln\left\{\frac{1-\phi_1}{\phi_1}+(1-p_2)(1-r)\right\}.\end{aligned}$$

En posant les dérivées partielles de $\ell_2(\phi_1, p_2, r)$ par rapport à (r, p_2, ϕ_1) égales à 0, on obtient, au prix de quelques manipulations, les équations suivantes :

$$\begin{aligned} 1 - \phi_1 + \phi_1(1-p_2)(1-r) &= \frac{n_{100}}{n_{1++}}\,, \\ \frac{n_{11+}}{p_2} - \phi_1(1-p_2)(1-r)n_{1++} &= n_{11+} + n_{101}, \\ \frac{n_{++1}}{r} - \phi_1(1-p_2)(1-r)n_{1++} &= n_{++1} + n_{+10}. \end{aligned}$$

Ces trois équations admettent une solution explicite, à savoir

$$\begin{aligned} \hat{p}_2 &= \frac{n_{+11} \times n_{11+}}{n_{101}n_{+1+} + n_{+11}n_{11+}}\,, \\ \hat{\phi}_1 &= \frac{1}{n_{1++}} \left\{ n_{111} + n_{110} + n_{101} + \frac{n_{101} \times (n_{110} + n_{010})}{n_{111} + n_{011}} \right\}, \\ \hat{r} &= \frac{n_{++1} \times n_{+11}}{n_{+1+} \times (n_{+11} + n_{101})}\,. \end{aligned}$$

On note que l'estimateur de ϕ_1 est celui suggéré un peu plus tôt, alors que l'estimateur de N_2 est donné par $\hat{N}_2 = \hat{M}_2 + \hat{U}_2 = \hat{\phi}_1 n_{1++} + n_{01+}/\hat{p}_2$. On obtient ainsi

$$\hat{N}_2 = n_{+1+} \left\{ 1 + n_{101} \times \frac{n_{1+1}}{(n_{111} + n_{110}) \times (n_{111} + n_{011})} \right\}.$$

Puisque cet estimateur a été construit en maximisant la vraisemblance de Poisson, il est plus précis que l'estimateur naïf $\hat{N}_2^P$ déduit du modèle de Petersen. Ce sont les estimateurs standard pour les modèles de populations ouvertes. Ces estimateurs sont présentés, dans un cadre un peu plus général, au chapitre 4 du livre de Pollock et coll. [6].

7 Discussion

La première utilisation connue de l'estimateur de Petersen remonte au mathématicien français Laplace en 1786. Pour estimer la taille de la population française, Laplace a multiplié n_{1+}, le nombre de naissances pour le pays tout entier, par l'inverse du rapport du nombre de naissances sur la population totale dans un certain nombre de paroisses. Lecren (1965)[4] traite de l'histoire de ces modèles en écologie.

C'est le statisticien canadien Stephen Fienberg qui a été le premier à suggérer l'utilisation de modèles log-linéaires pour traiter les données de capture-recapture dans les années 1970. Les modèles pour populations ouvertes ont été découverts simultanément (et indépendamment) par les statisticiens George Jolly et George Seber dans les années 1960. Des logiciels spécialisés sont disponibles pour ajuster ces modèles dits de Jolly–Seber ; voir Baillargeon et Rivest (2007)[1]. Les

utilisateurs peuvent utiliser les logiciels sans nécessairement connaître la théorie mathématique sous-jacente.

Aujourd'hui les modèles de capture-recapture sont toujours un sujet de recherche actif en statistique. Pour les modèles de populations fermées, le traitement de l'hétérogénéité des probabilités de capture fait l'objet de nombreux travaux. Dans certaines expériences, on mesure sur chaque unité capturée des variables associées aux probabilités de capture, et des modèles statistiques qui tiennent compte de ces variables dans l'estimation de N sont maintenant disponibles.

Dans les populations ouvertes, des variantes de l'expérience de la figure 2 sont utilisées pour estimer les probabilités de survie. Chaque année, des animaux marqués sont relâchés et sont suivis au cours des années suivantes. Les modèles pour traiter les données de recapture de ces animaux permettent d'estimer des taux de survie. Les tailles de population ne sont pas estimables à partir de ces données car la sélection des animaux marqués est souvent déterministe ; on marque par exemple tous les animaux présents sur un certain site une journée donnée. Des généralisations de ces modèles qui permettent à la survie de changer selon "l'état" d'un animal sont aujourd'hui disponibles. Ce paramètre peut par exemple varier selon l'âge de l'animal, jeune versus adulte, ou bien selon son statut reproducteur.

Références

[1] Baillargeon, S. and Rivest, L.-P. (2007). The Rcapture package : Loglinear models for capture-recapture in R. *Journal of Statistical Software*,19(3) `http://www.jstatsoft.org/`

[2] Bezerra, K., Gurgel, R., Llozue, C. et Castaneda, D.N. (2011). Estimating the number of street children and adolescents in two cities of Brazil using capture–recapture. *Journal of Paediatrics and Child Health*, 47, 524–529.

[3] Genest, C. et Rivest, L.-P. (2008). Comment compter les cerfs ? *Accromath*, 3 (2), 2–7.

[4] Lecren, E.D. (1965). A note on the history of mark-recapture population estimates. *Journal of Animal Ecology*, 34, 453–454.

[5] Murphy, J. (2009). Estimating the World Trade Center tower population on September 11, 2001 : A capture-recapture approach. *American Journal of Public Health*, 99, 65–67.

[6] Pollock, K.H., Nichols, J.D., Brownie C. et Hines, J.E. (1990). *Statistical Inference for Capture Recapture Experiments*. Wildlife Monograph N° 107. A publication of the Wildlife Society.

[7] Rivest, L.-P. (2011). A lower bound model for multiple record systems estimation with heterogeneous catchability. *The International Journal of Biostatistics*, 7 (1), Article n° 23.

[8] Tauvo, M. (2010). *Child Beggars in Dakar. Understanding Children Work*. UNICEF World Bank.