UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MODÈLES HOMOGÈNES DE CAPTURE-RECAPTURE POUR POPULATIONS FERMÉES : ALGORITHMES D'ESTIMATION EN GRANDE DIMENSION

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

ARMELLE VENCESLAS MAFONE FOTSO

FÉVRIER 2024

REMERCIEMENTS

Je tiens avant tout exprimer ma sincère gratitude à mon directeur de recherche, Monsieur Mamadou Yauck pour sa participation à la réalisation de ce mémoire. Je le remercie pour sa disponibilité, sa patience, ses conseils, son soutien aussi bien financier que moral tout au long de ce travail de recherche.

Je remercie également les professeurs du département de Mathématique de l'Université du Québec à Montréal pour la qualité des enseignements reçus tout au long de ma formation.

J'exprime ma reconnaissance à toutes les personnes qui m'ont soutenues et encouragées lors de ce parcours. Je remercie ma mère, mes frères et soeurs pour leur soutien incondtionné pendant ma formation. Je remercie tous ceux qui de près ou de loin m'ont soutenue pendant ma formation.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

Les méthodes de capture-recapture sont des techniques d'échantillonnage utilisées pour estimer les caractéristiques démographiques des populations pour lesquelles il n'existe pas de base d'échantillonnage. Ce mémoire se propose d'élaborer, pour des populations dont la taille ne change pas sur une période donnée, ou populations dites fermées, des algorithmes d'estimation des paramètres des modèles de capture-recapture, ainsi que des formules explicites pour la variance des estimateurs, lorsque le nombre d'occasions de capture est grand. L'intérêt dans ce travail est porté sur les modèles dits homogènes, c'est-à-dire des modèles pour lesquels la probabilité de capture ne varie pas d'une unité à l'autre ; l'originalité du travail porte, en particulier, sur les modèles homogènes comportementaux. Il ressort de l'évaluation de la méthodologie proposée, à l'aide d'une étude de simulations, que les estimateurs de la taille de la population et des variances qui leur sont associées ont de bonnes propriétés sur de petits échantillons.

Mots clés : capture-recapture, population fermée, modèles log-linéaires, regréssion de Poisson, bootstrap.

INTRODUCTION

Les méthodes de capture-recapture sont des techniques d'échantillonnage utilisées pour résoudre les problèmes d'estimation des paramètres démographiques des populations (taille, survie, émigration, immigration). Cette méthode a été développée initialement en biologie pour l'étude des populations animales, mais elle a également été appliquée à d'autres domaines, tels que la démographie et les sciences sociales, les sciences médicales et, récemment, la technologie du téléphone mobile. En 1894, Petersen a proposé pour la première fois, une méthode d'analyse des données de capture-recapture à deux occasions, afin d'estimer la taille d'une population de poissons. Lincoln (1930) a proposé, pour des données à deux occasions de capture, le même estimateur pour la taille de la population d'oiseaux. Ledit estimateur fut par la suite dénommé l'etsimateur de Lincoln-Petersen, en référence aux deux auteurs. Chapman (1951) et Seber (1982) ont par la suite proposé des versions avec correction pour le biais de l'estimateur lorsque les fréquences de capture sont faibles.

Une expérience de capture-recapture peut être décrite comme suit : On capture un premier échantillon d'unités dans la population ; on marque chaque unité capturée, puis on les remet dans la population. Ensuite, en supposant que les unités déjà marquées se sont mélangées aux non marquées, on capture un deuxième échantillon d'unités ; on note la marque de celles qui ont déjà été capturées puis on marque celles qui sont capturées pour la première fois avant de les remettre dans la population. Enfin, on répète ce processus sur un nombre fini d'occasions de capture, puis on obtient l'historique des unités capturées au moins une fois. Cet historique consiste en une séquence de 0 (manqué) et 1 (marqué) résumée dans un tableau des historiques de capture individuels.

Les modèles de capture-recapture sont classées en deux catégories : les modèles de population fermée et les modèles de population ouverte. Une population fermée n'admet aucun ajout et aucune diminution : la taille de la population reste constante tout au long de l'étude. Une population ouverte est sujette à des ajouts (naissances, immigrations) et à des diminutions (morts, émigrations) au cours de l'expérience de capture-recapture, conduite sur un nombre fini de périodes de capture (Jolly 1965; Seber 1965). À chaque période de capture, les unités sont capturées, marquées et relâchées sur une seule occasion de capture. On considère généralement que les périodes de capture sont assez espacées dans le temps, rendant plausible l'hypothèse d'ouverture de la population.

En se limitant à deux occasions de capture, on obtient souvent des variances d'estimation biaisées. Aussi, une expérience à deux occasions de capture ne permet pas d'identifier les différences comportementales d'une unité à l'autre. Par conséquent, il devient important d'augmenter le nombre d'occasions de capture afin de mieux apprehender les paramètres du mécanisme de capture. Otis *et al*. (1978) ont élaboré les techniques d'estimation pour des expériences à plus de deux occasions de capture et identifient trois sources de variations pour la capture des unités : temporelle, comportementale ou hétérogène lorsque les chances de capture varient d'une unité à l'autre.

Dans une étude récente de Yauck *et al*. (2019), les méthodes de capture-recapture ont été appliquées à l'analyse des données d'activation des applications sur les téléphones intelligents. Les auteurs ont développé l'argumentaire selon lequel ces données d'activations, collectées en temps réel à travers le monde, peuvent être transformées en données de capture-recapture : les unités sont représentées par les propriétaires de téléphones intelligents et les occasions de capture correspondent aux jours de la semaine. Dans leur article, Yauck *et al*. (2019) ont suggéré que lorsque le nombre d'occasions de capture est grand, ce qui est le cas dans le contexte de l'écosystème mobile, les algorithmes d'estimation du nombre d'actvations d'applications mobiles, basée sur la maximisation de la vraisemblance pour des modèles de population ouverte, pourraient faire face à des défis computationnels : les programmes informatiques standard d'optimisation pourraient échouer. L'objectif de ce mémoire est de proposer, pour des modèles de populations fermées, des algorithmes d'estimation des paramètres, ainsi que des formules explicites pour la variance des estimateurs, lorsque le nombre d'occasions de capture est grand. L'accent sera mis sur les modèles dits homogènes, c'est-à-dire pour lesquels le mécanisme de capture ne change pas d'une unité à l'autre. La contribution spécifique de cette étude est l'extension des algorithmes d'estimation de paramètres démographiques des modèles de capture-recapture pour les populations fermées au modèle comportemental M_b. Dans ce mémoire, il est proposé un nouvel algorithme d'estimation des paramètres démographiques, ainsi que les formules explicites pour la variance des estimateurs ; d'un point de vue computationnel, cette approche est particulièrement utile lorsque le nombre d'occasions de capture est grand.

Ce travail est divisé en cinq chapitres. Le chapitre 1 présente une revue de la littérature sur les méthodes de capture-recapture pour une population fermée. Ce chapitre met l'accent sur la présentation des concepts pouvant aider à une meilleure compréhension des modèles de capture-recapture, notamment la formalisation des modèles de populations fermées et les hypothèses qui permettent la construction de ces modèles. Le chapitre 2 est consacrée à la présentation l'écosystème mobile, ainsi que le processus de transformation

de ces données en données de capture-recapture. Le chapitre 3 présente la methodologie déployée afin d'estimer la taille de la population à partir des données de capture-recapture dans le cas d'une population fermée. Ce chapitre propose les algorithmes d'estimation des paramètres des modèles homogènes de populations fermées et les formules de variances des estmateurs. Les chapitres 4 et 5 présentent respectivement les simulations permettant d'évaluer les estimateurs des paramètres et une étude de cas.

CHAPITRE 1

INTRODUCTION AUX MÉTHODES DE CAPTURE- RECAPTURE.

Ce chapitre est consacré à la présentation des concepts pouvant aider à une meilleure compréhension du document, notamment la formalisation des modèles de populations fermées et les hypothèses qui permettent la construction de ces modèles.

1.1 Définitions des concepts et notations

Considérons une expérience de capture-recapture sur une population fermée constituée de N unités. L'historique de capture d'une unité i est une suite de 0 (manqué) et 1 (capturé) consignée dans un vecteur $\boldsymbol{\omega}_i = (\omega_{i1}, \ldots, \omega_{i\ell})$, où ℓ représente le nombre d'occasions de capture, $\boldsymbol{\Omega}$ est l'ensemble des historiques de captures observables. Les historiques de captures sont souvent représentés dans une matrice $n \times \ell$:

$$\boldsymbol{H} = \begin{bmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1\ell} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2\ell} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{n1} & \omega_{n2} & \cdots & \omega_{n\ell} \end{bmatrix}.$$

Dès lors, on définit à partir des données observées, les statistiques suivantes :

- le nombre distinct d'unités de capturées au moins une fois pendant l'expérience est $n = \sum_{\omega=1}^{s} n_\omega$, où $s = 2^\ell - 1$; et n_ω est la fréquence associée à l'historique de capture observable $\boldsymbol{\omega}$.
- $u_j = \sum_{i=1}^{n} \prod_{k=1}^{j-1} (1 - \omega_{ik})\omega_{ij}$ est le nombre d'unités capturées pour la première fois à l'occasion j ;
- $n_j = \sum_{i=1}^{n} \omega_{ij}$ est le nombre d'unités capturées à l'occasion j ;
- f_j est le nombre d'unités capturées exactement j fois, $j = 1, 2, \ldots, \ell$;
- $C = \sum_\omega n_\omega \sum_{j=1}^{l} \omega_j$, est le nombre total de captures ; et ω_j est l'historique de capture observable à l'occasion de capture j.

La procédure d'estimation de la taille de la population N repose sur plusieurs paramètres :

- $P_\omega(\boldsymbol{\theta})$ est la probabilité de réalisation de l'historique $\boldsymbol{\omega}$, où $\boldsymbol{\theta}$ est un vecteur de paramètres de dimension au plus s ; $P_0(\boldsymbol{\theta}) = 1 - \sum_\omega P_\omega(\boldsymbol{\theta})$ est la fréquence du seul historique non observable $\boldsymbol{\omega} = (0, 0, \ldots 0)$;

— μ_ω est la fréquence prédite pour chaque historique de capture donnée par

$$\mu_\omega = NP_\omega(\boldsymbol{\theta}),\ \boldsymbol{\omega} \in \boldsymbol{\Omega}^0,$$

où $\boldsymbol{\Omega}^0 = \boldsymbol{\Omega} \cup \boldsymbol{\omega}^0$.

1.2 Procédure d'estimation

L'estimation de N se présente selon les trois approches suivantes : la vraisemblance multinomiale, la vraisemblance conditionnelle et la vraisemblance Poisson. Ces approches reposent sur l'hypothèse selon laquelle les unités sont capturées de manière indépendante.

1.2.1 Vraisemblance multinomiale

Ici, les fréquences observées des historiques de captures sont distribuées selon une loi multinomiale :

$$(N-n, n_\omega) \sim M(N, P_0(\boldsymbol{\theta}), P_\omega(\boldsymbol{\theta})). \tag{1.1}$$

On en déduit la vraisemblance multinomiale (Darroch, 1958) :

$$L_M(N, \boldsymbol{\theta}) = \frac{N!}{\prod_{\omega\in\Omega} n_\omega!(N-n)!} P_0^{N-n} \prod_{\omega\in\Omega} P_\omega(\boldsymbol{\theta})^{n_\omega}. \tag{1.2}$$

L'approche de maximisation directe de la vraisemblance complète (1.2) est dite non conditionnelle. Cette maximisation permet d'obtenir les estimateurs du maximum de vraisemblance de N et $\boldsymbol{\theta}$, notés $\hat{N}_U$ et $\hat{\boldsymbol{\theta}_U}$. Toutefois, cette méthode de maximisation directe semble difficile. Une solution serait d'utiliser une approche de maximisation dite conditionnelle, qui consiste à décomposer la vraisemblance multinomiale (1.2) selon la procédure discutée à la section suivante.

Exemple 1.1 *Considérons une expérience de capture-recapture avec deux occasions de captures ($\ell = 2$). L'ensemble des historiques de captures observables est donné par $\boldsymbol{\Omega} = \{(0,1),(1,0),(1,1)\}$. La probabilité de n'observer aucune capture est $P_0(\boldsymbol{\theta}) = 1 - P_{01}(\boldsymbol{\theta}) - P_{10}(\boldsymbol{\theta}) - P_{11}(\boldsymbol{\theta})$, avec $P_{01}(\boldsymbol{\theta})$, $P_{10}(\boldsymbol{\theta})$ et $P_{11}(\boldsymbol{\theta})$ les probabilités respectives d'observer les éléments de $\boldsymbol{\Omega}$. La vraisemblance multinomiale est donnée par :*

$$L_M(N, \boldsymbol{\theta}) = \frac{N!}{\prod_{\omega\in\Omega} n_\omega!(N-n)!} P_0^{N-n} P_{01}(\boldsymbol{\theta})^{n_{01}} P_{10}(\boldsymbol{\theta})^{n_{10}} P_{11}(\boldsymbol{\theta})^{n_{11}}.$$

En supposant que les probabilités de capture sont égales d'une occasion de capture à une autre, c'est á dire $P(\omega_j = 1) = P$, *pour tout* $j = 1, 2$, *on a :*

$$P_{01}(\boldsymbol{\theta}) = P_{01}(P) = (1-P)P,$$
$$P_{10}(\boldsymbol{\theta}) = P_{10}(P) = P(1-P),$$
$$P_{11}(\boldsymbol{\theta}) = P_{11}(P) = P^2.$$

Dans cet exemple $\Omega = P$ *et sa dimension est égale à* 1.

La vraisemblance devient donc :

$$L_M(N, \boldsymbol{\theta}) = L_M(N, P)\frac{N!}{n_{01}!n_{10}!n_{11}!(N-n)!}P^{n_{11}+n}(1-P)^{2N-n-n_{11}}.$$

1.2.2 Approche conditionnelle

Dans cette approche, la vraisemblance multinomiale peut s'écrire comme produit de deux termes :

$$L_M(N, \boldsymbol{\theta}) = L_1(N, P_0(\boldsymbol{\theta})) \times L_2(\boldsymbol{\theta}), \tag{1.3}$$

avec

$$L_1(N, P_0(\boldsymbol{\theta})) = \frac{N!}{n!(N-n)!}\{P_0(\boldsymbol{\theta})\}^{N-n}\{1-P_0(\boldsymbol{\theta})\}^n, \tag{1.4}$$

et

$$L_2(\boldsymbol{\theta}) = \frac{n!}{\prod_{\omega\in\Omega} n_\omega!}\prod_{\omega\in\Omega}\{Q_\omega(\boldsymbol{\theta})\}^{n_\omega}, \tag{1.5}$$

où $Q_\omega(\boldsymbol{\theta}) = P_\omega(\boldsymbol{\theta})/(1-P_0(\boldsymbol{\theta}))$ est la probabilité qu'une unité ait l'historique de capture $\boldsymbol{\omega}$ sachant qu'elle a été capturée au moins une fois au cours de l'expérience. La fonction L_1 représente la vraisemblance binomiale basée sur la probabilité de n et L_2 est la vraisemblance multinomiale du vecteur $(n_1, n_2, \ldots, n_{\ell-1})$ conditionnelle au nombre de captures n. La maximisation de la fonction de vraisemblance L_M peut se faire en deux étapes :

- Estimation de $\boldsymbol{\theta}$ en se servant de la vraisemblance conditionnelle L_2 , qui permet d'obtenir l'estimateur du maximum de vraisemblance $\hat{\boldsymbol{\theta}}_C$ de $\boldsymbol{\theta}$;
- Estimation de N à partir de la fonction L_1, qui permet d'obtenir l'estimateur du maximum de vrai-

semblance $\hat{N}_C$ de N. On a donc :

$$\hat{N}_C = [\frac{n}{1 - P_0(\hat{\boldsymbol{\theta}}_C)}], \tag{1.6}$$

où [ ] est la fonction partie entière.

Remarque 1.1 *Les estimateurs de N qui découlent des approches conditionnelle et non conditionnelle satisfont l'inégalité $\hat{N}_U < \hat{N}_C$. De plus, sous certaines conditions de régularité, ($\hat{N}_U$, $\hat{\boldsymbol{\theta}}_{\boldsymbol{U}}$) et ($\hat{N}_C$, $\hat{\boldsymbol{\theta}}_{\boldsymbol{C}}$) sont des estimateurs convergents de ($\hat{N}$, $\hat{\boldsymbol{\theta}}$) (Sanathanan, 1972).*

1.2.3 Vraisemblance Poisson

Les fréquences $\{n_\omega\}$ sont modélisées comme des variables suivant une distribution de Poisson :

$$n_\omega \sim \text{Poisson}(\mu_\omega = NP_\omega(\boldsymbol{\theta})). \tag{1.7}$$

La vraisemblance Poisson L_P s'écrit comme suit :

$$L_P(N, \boldsymbol{\theta}) = \prod_{\omega \in \Omega} \frac{e^{-\mu_\omega} \mu_\omega^{n_\omega}}{n_\omega!}. \tag{1.8}$$

Aussi, la fréquence prédite μ_ω s'écrit sous forme log-linéaire :

$$\log \mu_\omega = \gamma + \boldsymbol{X}_\omega^\top \boldsymbol{\beta}, \tag{1.9}$$

où γ est le logarithme de la fréquence prédite des unités non capturées, et $\boldsymbol{X}_\omega$ est le vecteur de dimension $d \times 1$ des variables explicatives du modèle, $\boldsymbol{\beta}$ le vecteur $d \times 1$ des paramètres du modèle. Ainsi, l'estimation des paramètres consignés dans le vecteur $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}^\top)^\top$ peut se faire en utilisant un modèle linéaire généralisé avec une distribution de Poisson et une fonction de lien. On remarque $\boldsymbol{X}_0$, la valeur de $\boldsymbol{X}$ de dimension $2^\ell - 1$, pour l'historique de capture non observable $\boldsymbol{\omega}^0 = (0, 0, \ldots, 0)$, est le vecteur d'éléments 0. La matrice des variables explicatives $\boldsymbol{X}$ du modèle est de dimension $(2^\ell - 1) \times d$; l'élément de la ligne ω correspond au vecteur $\boldsymbol{X}_\omega$. On définit le vecteur $\boldsymbol{y}$ des fréquences observées pour les $2^\ell - 1$ historiques de capture et μ son vecteur de fréquences prédites. les statistiques exhaustives minimales du modèle (1.9) peuvent être résumer dans le vecteur $(n, \boldsymbol{y}^\top \boldsymbol{X})^\top$. On définit une distribution de probabilité sur les 2^ℓ historiques de capture en posant μ_ω / N la probabilité de l'historique ω. On associe donc à la matrice $\boldsymbol{X}$ un vecteur aléatoire de dimension d; on pose pour ce vecteur aléatoire, $\boldsymbol{\mu}_{\boldsymbol{X}}$ et $\boldsymbol{\Sigma}$ comme le vecteur de dimension $d \times 1$ des espérances et la matrice de variances-covariances de dimension $d \times d$ respectivement.

Notons que les équations (1.8) et (1.9) définissent une régression de Poisson. L'estimation de la taille de la population N est donnée par :

$$\hat{N} = n + e^{\hat{\gamma}}, \tag{1.10}$$

avec $e^{\hat{\gamma}}$ étant l'estimateur du nombre d'unités non capturées, et $\hat{\gamma}$ représentant l'ordonnée à l'origine de la regression de Poisson.

Remarque 1.2 *La variance multinomiale de $\hat{N}$, $Var_M(\hat{N})$, peut être déduite de la variance sous le modèle log-linéaire suivant la relation*

$$Var_M(\hat{N}) \approx Var_P(\hat{N}) - N,$$

où Var_P représente la variance sous la distribution de Poisson (Yauck et Rivest, 2019).

Proposition 1.1 *Pour les modèles de la forme (1.9), la variance multinomiale asymptotique de l'estimateur du maximum de vraisemblance $\hat{N}$ est donnée par*

$$Var_M(\hat{N}) = \frac{N}{(1-P^*)^{-1} - 1 - \boldsymbol{\mu}_X^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_X}, \tag{1.11}$$

où $P^* = 1 - e^{\gamma}/N$ est la probabilité d'être capturée au moins une fois au cours de l'expérience, $\boldsymbol{\mu}_X$ le vecteur (de dimension $d \times 1$) des espérances de $\boldsymbol{X}$ et $\boldsymbol{\Sigma}$ la matrice de variance-covariance (de dimension $d \times d$) de $\boldsymbol{X}$.

Preuve 1.1 *La matrice d'information de Fisher pour le vecteur de paramètres du modèle log-linéaire $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}^\top)^\top$ est*

$$I_p(\boldsymbol{\theta}) = \sum_{\boldsymbol{\omega} \in \Omega} \mu_\omega \begin{pmatrix} 1 \\ \boldsymbol{X}_\omega \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{X}_\omega \end{pmatrix}^\top = N \left\{ \begin{pmatrix} -e^{\gamma}/N & 0 \\ 0 & \boldsymbol{\Sigma} \end{pmatrix} + \begin{pmatrix} 1 \\ \boldsymbol{\mu}_X \end{pmatrix} \begin{pmatrix} 1 \\ \boldsymbol{\mu}_X \end{pmatrix}^\top \right\}.$$

L'inversion de la matrice donne

$$I_p^{-1}(\boldsymbol{\theta}) = \frac{1}{N} \left\{ \begin{pmatrix} -N/e^{\gamma} & 0 \\ 0 & \boldsymbol{\Sigma}^{-1} \end{pmatrix} + \frac{\begin{pmatrix} -N/e^{\gamma} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_X \end{pmatrix} \begin{pmatrix} -N/e^{\gamma} \\ \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_X \end{pmatrix}^\top}{N/e^{\gamma} - 1 - \boldsymbol{\mu}_X^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_X} \right\}. \tag{1.12}$$

Une expansion de Taylor d'ordre 1 *de l'estimateur du maximum de vraisemblance pour* $\boldsymbol{\theta}$ *donne*

$$\hat{\boldsymbol{\theta}} \approx \boldsymbol{\theta} + I_p^{-1}(\boldsymbol{\theta}) S_p(\boldsymbol{\theta}),$$

où $S_p(\boldsymbol{\theta})$ *est la fonction de score pour la régression de Poisson :*

$$S_p(\boldsymbol{\theta}) = \left(n - NP^*,\ (\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{X}\right)^\top,$$

avec $P^* = 1 - e^\gamma / N$. *Une simple expansion de Taylor d'ordre* 1 *donne :*

$$\hat{N} - N \approx (n - NP^*) + e^\gamma (\hat{\gamma} - \gamma).$$

De (1.12), on a :

$$e^\gamma (\hat{\gamma} - \gamma) \approx \frac{\left(1 + \boldsymbol{\mu}_X^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_X,\ -\boldsymbol{\mu}_X^\top \boldsymbol{\Sigma}^{-1}\right)}{(1 - P^*)^{-1} - 1 - \boldsymbol{\mu}_X^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_X} \begin{pmatrix} n - NP^* \\ \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{\mu} \end{pmatrix}.$$

En substituant cette quantité dans l'expansion pour l'estimateur $\hat{N}$, *et en soustrayant* N *à la formule de la variance, on obtient* (1.11).

Proposition 1.2 *Une estimation de la variance de* (1.11), $v(\hat{N})$, *exprimée en fonction des paramètres log-linéaires, est donnée par*

$$v(\hat{N}) = e^{\hat{\gamma}} + e^{2\hat{\gamma}} v\left(\hat{\gamma}\right), \tag{1.13}$$

où $v\left(\hat{\gamma}\right)$ *représente une estimation de la variance de l'estimateur* $\hat{\gamma}$, *déduite par ajustement du modèle* (1.9) *aux données ; cette quantité est le premier élément de l'inverse de la matrice d'information de Fisher observée.*

Preuve 1.2 *On a*

$$\mathsf{Var}_P(\hat{N}) = \mathsf{Var}_P(n) + e^{2\gamma} \mathsf{Var}_P(\hat{\gamma}) + 2e^\gamma \mathsf{Cov}_P(n, \hat{\gamma}) = NP^* + e^{2\gamma} \mathsf{Var}_P(\hat{\gamma}) + 2e^\gamma \mathsf{Cov}_P(n, \hat{\gamma}) \tag{1.14}$$

car

$$\mathsf{Var}_P(n) = \mathbb{E}\{\mathsf{Var}_P(n|\tilde{N})\} + \mathsf{Var}_P\{\mathbb{E}(n|\tilde{N})\} = NP^*(1 - P^*) + NP^* = NP^*,$$

avec $\tilde{N}$ représentant une variable aléatoire telle que $\mathbb{E}(\tilde{N}) = N$. On montre par la suite que

$$\mathit{Cov}_P(n, \hat{\gamma}) = \frac{1}{e^\gamma} \frac{\left[1 + \boldsymbol{\mu}_X^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_X,\ -\boldsymbol{\mu}_X^\top \boldsymbol{\Sigma}^{-1}\right]}{(1-P^*)^{-1} - 1 - \boldsymbol{\mu}_X^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_X} \begin{pmatrix} NP^* \\ N\boldsymbol{\mu}_X \end{pmatrix} = \frac{N\left[P^* - \boldsymbol{\mu}_X^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_X (1-P^*)\right]}{e^\gamma \left[(1-P^*)^{-1} - 1 - \boldsymbol{\mu}_X^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_X\right]}.$$

On déduit donc $\mathit{Cov}_P(n, \hat{\gamma}) = N(1-P^*)/e^\gamma$. En substituant cette expression dans l'équation (1.14), il suit :

$$\mathit{Var}_P(\hat{N}) = N - e^\gamma + 2e^\gamma + e^{2\gamma} \mathit{Var}_P(\hat{\gamma}).$$

On a donc

$$\mathit{Var}_M(\hat{N}) = e^\gamma + 2e^\gamma + e^{2\gamma} \mathit{Var}_P(\hat{\gamma}).$$

On conclut la preuve en remplaçant les paramètres log-linéaires par leurs estimés.

1.3 Modélisation de la capture des unités

Soit P_{ij} (i= 1,2,...,N) la probabilité de capture de l'individu i, à l'occasion de capture j (j= 1,2,...,ℓ). Cette probabilité peut faire l'objet de variations temporelles (t) et comportementale (b pour behavior) dans le cadre de cette étude. Ainsi, de ces variations de la probabilité de capture, découlent des modèles M_0 (aucune variation), M_t, M_b.

1.3.1 Le modèle M_0

Dans le modèle M_0, on admet que la probabilité de capture ne varie ni d'une unité i à l'autre, et ni d'une occasion de capture j à l'autre : $P_{ij} = P$. En termes de paramétrisation log-linéaire, $d = 1$ et $_\mu = \sum_{j=1}{}^{\ell} \omega_j$ est le nombre de fois qu'une unité est capturée. Dans ce cas $\boldsymbol{\theta} = P$ et on a : $P_\omega(\boldsymbol{\theta}) = P^{\sum_{j=1}^{\ell} \omega_j} (1-P)^{\ell - \sum_{j=1}^{\ell} \omega_j}$ et $P_0(\boldsymbol{\theta}) = (1-P)^\ell$. Lorsque les fréquences prédites sont distribuées suivant une loi multinomiale, la vraisemblance multinomiale donnée à l'équation (1.2) devient :

$$L_M(N, P) = \frac{N!}{\prod_{\omega \in \Omega} n_\omega! (N-n)!} P^C (1-P)^{\ell N - C}, \tag{1.15}$$

où $C = \sum_{\omega \in \Omega} \sum_{j=1}^{\ell} n_\omega \omega_j$ est le nombre total de captures. Lorsque le nombre d'occasions de capture est supérieur à deux, la maximisation de cette fonction de vraisemblance devient complexe. C'est ainsi que Otis *et al.* (1978) ont construit des algorithmes de calcul de P et N.

La décomposition de la vraisemblance multinomiale (1.15), $L_M(N, P) = L_1(N, P) \times L_2(P)$, pour l'ap-

proche conditionnelle, aboutit à

$$L_1(N;P) = \frac{N!}{n!(N-n)!}\{(1-P)^\ell\}^{N-n}\{1-(1-P)^\ell\}^n \tag{1.16}$$

et

$$L_2(P) = \frac{n!}{\prod_{\omega\in\Omega} n_\omega!}\prod_{\omega\in\Omega}\{Q_\omega(P)\}^{n_\omega}, \tag{1.17}$$

avec

$$Q_\omega(P) = \frac{P^{\sum_{j=1}^\ell \omega_j}(1-P)^{\ell-\sum_{j=1}^\ell \omega_j}}{1-(1-P)^\ell}$$

représentant la probabilité qu'une unité ait l'historique de capture ω sachant qu'elle a été marquée au moins une fois au cours de l'expérience. La maximisation de la vraisemblance conditionnelle se fait comme suit. On estime P à l'aide de la fonction $L_2(P)$ pour obtenir $\hat{P}_C$. On maximise par la suite $L_1(N;\hat{P}_C)$ pour obtenir

$$\hat{N}_C = \left[\frac{n}{1-(1-\hat{P}_C)^\ell}\right]. \tag{1.18}$$

Dans le cas d'une regresssion de Poisson, on a $\boldsymbol{X_\omega} = \sum_{j=1}^\ell \omega_j$. Le modèle log-linéaire (1.9) s'ecrit :

$$\log\mu_\omega = \gamma + \sum_{j=1}^{\ell}\omega_j\beta, \tag{1.19}$$

avec $\gamma = \{\log N(1-P)^l\}$ et $\beta = \log\{P/1-P\}$. L'estimateur de la taille de population $\hat{N}$ est ainsi donné par :

$$\hat{N} = e^{\hat{\gamma}}(1+e^{\hat{\beta}})^\ell. \tag{1.20}$$

La variable $\boldsymbol{X_\omega} = \sum_{j=1}^\ell \omega_j$ suit une loi Binomiale de paramètres P et ℓ. On a ainsi $\boldsymbol{\mu_X} = \ell P$ et $\boldsymbol{\Sigma} = \ell P(1-P)$. En remplaçant $\boldsymbol{\mu_X}$ et $\boldsymbol{\Sigma}$ dans l'équation (1.11) la variance multinomiale asymptotique devient :

$$\text{Var}_M(\hat{N}) = \frac{N}{(1-P)^{-\ell} + (\ell-1) - (1-P)^{-1}}. \tag{1.21}$$

Exemple 1.2 *Pour une expérience à deux occasions de capture, on souhaite obtenir les expressions des estimateurs de N et P. La log-vraisemblance est donnée par :*

$$LL_P(\theta) = \sum_{\omega\in\boldsymbol{\Omega}}\{-\mu_\omega + n_\omega\log\mu_\omega - \log n_\omega\}$$

Le modéle log-linéaire M_0 est donné par :

$$\log \mu_\omega = \gamma + \sum_{j=1}^{\ell} \omega_j \beta.$$

avec $\gamma = \log\{N(1-P)^\ell\}$ et $\beta = \log\{P/(1-P)\}$. L'ensemble des historique observables est $\mathbf{\Omega} = \{(1,0),(0,1),(1,1)\}$. On a ainsi, pour chaque historique, les modèles suivants :

$\log \mu_{10} = \gamma + \beta$, $\log \mu_{01} = \gamma + \beta$, $\log \mu_{11} = \gamma + 2\beta$.

La log-vraisemblance devient donc :

$$\mathcal{L}L_P(\gamma, \beta) = -e^{\gamma+\beta}(2 + e^\beta) + n(\gamma + \beta) + n_{11}\beta - \log n_{10}! - \log n_{01}! - \log n_{11}!.$$

Les dérivées partielles de la log-vraisemblance par rapport à β et γ donnent :

$$\frac{\partial \mathcal{L}L_P(\gamma, \beta)}{\partial \gamma} = n - e^{\gamma+\beta}(2 + e^\beta).$$

$$\frac{\partial \mathcal{L}L_P(\gamma, \beta)}{\partial \beta} = -2e^{\gamma+\beta}(1 + e^\beta) + C.$$

En égalisant ces deux dérivées à zéro, on obtient :

$$n = NP(2 - P)$$

$$C = 2NP.$$

La resolution de ces deux équations permet d'obtenir les estimations suivantes de P et n :

$$\hat{N} = \frac{(n_1 + n_2)^2}{4n_{11}}.$$

$$\hat{P} = \frac{2n_{11}}{C}.$$

1.3.2 Le modèle M_t

Le modèle M_t, généralisant le modèle M_0 en incluant aux probabilités de capture une variable temporelle, suppose que les probabilités de capture varient dans le temps donc $P_{ij} = P_j$. Dans ce cas, $\boldsymbol{\theta} = (P_1, P_2, \ldots, P_\ell)$ et on a : $P_\omega(\boldsymbol{\theta}) = \prod_{j=1}^{\ell} P_j^{\omega_j}(1-P_j)^{1-\omega_j}$ et $P_0(\boldsymbol{\theta}) = \prod_{j=1}^{\ell}(1-P_j)$.

La vraisemblance multinomiale s'écrit alors :

$$L_M(N, P_j) = \frac{N!}{\prod_{\omega\in\Omega} n_\omega!(N-n)!} P_j^{n_j}(1-P_j)^{N-n_j}, \tag{1.22}$$

où $n_j = \sum_{J=1}^{\ell} \omega_j$ représente le nombre de captures à l'occasion de capture j. Une maximisation directe de cette vraisemblance permet d'obtenir l'estimateur du maximum de vraisemblance pour $(N, P_1, \ldots, P_\ell)$.

L'approche conditionnelle de décomposition de la vraisemblance (1.22), $L_M(N, \{P_j\}) = L_1(N, \{P_j\}) \times L_2(\{P_j\})$ aboutit à

$$L_1(N; P_1, \ldots, P_\ell) = \frac{N!}{n!(N-n)!} \left\{\prod_{j=1}^{\ell}(1-P_j)\right\}^{N-n} \left\{1 - \prod_{j=1}^{\ell}(1-P_j)\right\}^{n} \tag{1.23}$$

et

$$L_2(P_1, \ldots, P_\ell) = \frac{n!}{\prod_{\omega\in\Omega} n_\omega!} \prod_{\omega\in\Omega} \{Q_\omega(P_1, \ldots, P_\ell)\}^{n_\omega}, \tag{1.24}$$

avec

$$Q_\omega(P_1, \ldots, P_\ell) = \frac{\prod_{j=1}^{\ell} P_j^{\omega_j}(1-P_j)^{1-\omega_j}}{1 - \prod_{j=1}^{\ell}(1-P_j)}.$$

La maximisation se fait en deux étapes. On estime P_j, $j = 1, \ldots, \ell$, à partir de la fonction de vraisemblance $L_2(P_1, \ldots, P_\ell)$ pour obtenir $\hat{P}_{j,C}$, $j = 1, \ldots, \ell$. Ensuite, on maximise $L_1(N; \{\hat{P}_{j,C}\})$ pour aboutir à

$$\hat{N}_C = \left[\frac{n}{1 - \prod_{j=1}^{\ell}(1-\hat{P}_{j,C})}\right]. \tag{1.25}$$

Dans le cas d'une régression de Poisson, le modèle log-linéaire (1.9) devient :

$$\log \mu_\omega = \gamma + \sum_{j=1}^{\ell} \omega_j \beta_j, \tag{1.26}$$

avec $\gamma = \log\{N\prod_{j=1}^{\ell}(1-P_j)\}$ et $\beta_j = \log\{P_j/1-P_j\}$, $j = 1,2,...,\ell$. L'estimateur de la taille de population $\hat{N}$ est ainsi donné par :

$$\hat{N} = e^{\hat{\gamma}}\prod_{j=1}^{\ell}(1+e^{\hat{\beta}_j})^{\ell}. \tag{1.27}$$

Les variables $\omega_1,\ldots,\omega_\ell$ suivent des lois binomiales et sont indépendantes. On a ainsi $\boldsymbol{\mu_X} = (P_1,\ldots,P_l)^T$ et $\boldsymbol{\Sigma}$ une matrice diagonale d'éléments $P_j(1-P_j)$. En remplaçant μ_X et $\boldsymbol{\Sigma}$ dans l'équation (1.10), la variance asymptotique multinomiale s'écrit :

$$\text{Var}_M(\hat{N}) = \frac{N}{\prod_{j=1}^{\ell}(1-P_j)^{-}1+(\ell-1)\sum_{j=1}^{\ell}(1-P_j)}. \tag{1.28}$$

Lorsque $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_\ell = \beta$, alors le modèle M_t se reduit au modèle M_0. si $P_1 = P_2\cdots = P_\ell = P$, alors la variance mutinomiale asymptotique se reduit à celle du modèle M_0. On remarque donc que pour deux occasions de captures, le modèle M_t est une généralisation du modèle M_0.

Exemple 1.3 *Estimateur de Lincoln-Petersen.*

Considérons une expérience de capture-recapture à deux occasions de capture sur une population de N poissons. On s'intéresse l'estimation de la taille de population pour une expérience de capture-recapture menée sur deux jours consécutifs. Le processus de capture des unités est illustré à la figure 1.1.

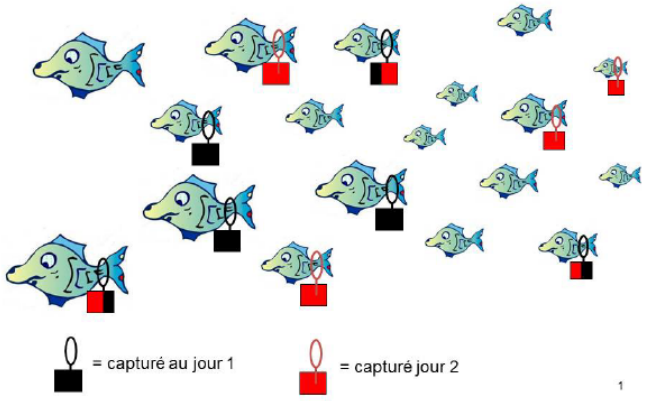

Figure 1.1 Représentation schématique d'une expérience de capture-recapture sur une population de poissons (Rivest et Baillargeon, 2013).

La marque noire représente la capture au premier jour et la marque rouge, la marque au deuxième jour. Les unités sont capturées aux jours 1 *et* 2 *avec des probabilités* P_1 *et* P_2 *respectivement ; on suppose que les captures sont indépendantes d'un jour à l'autre. L'ensemble des historiques de captures observables est donné par :* $\mathbf{\Omega} = \{(0,1),(1,0),(1,1)\}$*, avec pour fréquences respectives* n_{01},n_{10} *et* n_{11}*. Ces données sont présentées dans le tableau 1.1.*

Table 1.1 Données issues de l'expérience à deux occasions de captures

	Manqué jour 1	Capturé jour 1
Manqué jour 2	n_{00}	n_{10}
Capturé jour 2	n_{01}	n_{11}

Les fréquences prédites issues de cette expérience de capture-recapture sont ensuite consignées dans le tableau 1.2.

Table 1.2 Fréquences prédites découlant de l'expérience à deux occasions de captures

	Manqué jour 1	Capturé jour 1
Manqué jour 2	$\mu_{00} = N(1-P_1)(1-P_2)$	$\mu_{10} = NP_1(1-P_2)$
Capturé jour 2	$\mu_{01} = N(1-P_1)P_2$	$\mu_{11} = NP_1P_2$

Le nombre de poissons manqués est donné par $e^{\hat{\gamma}} = n_{10}n_{01}/n_{11}$*. En remplaçant cette valeur dans l'équation* (1.10)*, on obtient une estimation de la population de poissons donnée par* $\hat{N} = (n_{10}+n_{01}+n_{11}) + n_{10}n_{01}/n_{11}$*. Il suit donc que*

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

Cette estimation est appelé estimateur de Lincoln-Petersen (Lecren, 1965). La variance multinomiale est donnée par :

$$\mathit{Var}_M(\hat{N}) = \frac{N(1-P_1)(1-P_2)}{P_1P_2}.$$

Le modèle log-linéaire M_t *est donné par*

$$\log \mu_\omega = \gamma + \sum_{j=1}^{\ell} \omega_j \beta_j,$$

avec $\gamma = \log\{N\prod_{j=1}^{\ell}(1-P_j)\}$ *et* $\beta_j = \log\{P_j/1-P_j\}$, $j = 1, 2, ..., \ell$.

L'ensemble des historiques observables est $\mathbf{\Omega} = \{(1,0),(0,1),(1,1)\}$. *On a ainsi pour chaque historique les modèles suivants :*

$$\log \mu_{10} = \gamma + \beta_1.$$

$$\log \mu_{01} = \gamma + \beta_2.$$

$$\log \mu_{11} = \gamma + \beta_1 + \beta_2.$$

La log-vraisemblance dans ce cas devient

$$\mathcal{L}L_P(\gamma, \beta_1, \beta_2) = -e^{\gamma}(e^{\beta_1} + e^{\beta_2} + e^{\beta_1+\beta_2}) + n\gamma + n_1\beta_1 + n_2\beta_2 - \log n_{10}! - \log n_{01}! - \log n_{11}!.$$

Les dérivées partielles de la log-vraisemblance par rapport à β_1, β_2 *et* γ *donnent :*

$$\frac{\partial \mathcal{L}L_P(\gamma, \beta_1, \beta_2)}{\partial \gamma} = n - e^{\gamma}(e^{\beta_1} + e^{\beta_2} + e^{\beta_1+\beta_2}).$$

$$\frac{\partial \mathcal{L}L_P(\gamma, \beta_1, \beta_2)}{\partial \beta_1} = n_1 - e^{\gamma}(e^{\beta_1} + e^{\beta_1+\beta_2}).$$

$$\frac{\partial \mathcal{L}L_P(\gamma, \beta_1, \beta_2)}{\partial \beta_2} = n_2 - e^{\gamma}(e^{\beta_2} + e^{\beta_1+\beta_2}).$$

En égalisant les trois équations à zéro, on obtient : $n_1 = NP_1$, $n_2 = NP_2$ *et* $n = N1 - (1-P_1)(1-P_2)$. *Ainsi, on a les expressions des estimateurs suivants :*

$$\hat{N} = \frac{n_1 \times n_2}{n_{11}}.$$

$$\hat{P_1} = \frac{n_{11}}{n_2}.$$

$$\hat{P_2} = \frac{n_{11}}{n_1}.$$

1.3.3 Le modèle M_b

Dans ce modèle, l'hypothèse d'indépendance de capture des unités n'est plus considérée. Il est question de matérialiser le changement comportemental des unités après la première occasion de capture. On définit $P_{ij} = P$ la probabilité de capture pour la premiere occasion de capture et c la probabilité de recapture pour les unités capturées à la première occasion de capture. Ainsi, si $P > c$, la chance de revoir une unité après la première capture est petite et si $P < c$, on a plus de chance de revoir une unité après la première capture. Lorsque $P = c$, on revient au cas d'un modèle M_0. La fonction de vraisemblance multinomiale pour le modèle comportemental M_b s'écrit :

$$L_M(N, P, c) = \frac{N!}{(N-n)!} P^n (1-P)^{(\ell-1)N-n} c^n (1-c)^{N-n}, \tag{1.29}$$

La forme de la vraisemblance (1.29) suggère que l'estimation de c, qui représente un paramètre de nuisance, est indépendante de l'estimation des paramètres d'intérêt N et P. La variance multinomiale asymptotique pour l'estimateur du maximum de vraisemblance $\hat{N}$ est (Otis *et al.*, 1978)

$$\text{Var}_M(\hat{N}) = \frac{NP^*(1-P^*)}{\{1-(1-P)^\ell\}^2 - \ell^2 P^2 (1-P)^{\ell-1}}. \tag{1.30}$$

Lorsque $P = c$, la variance multinomiale de $\hat{N}$ pour le modèle comportemental M_b (1.30) se réduit à celle du modèle M_0 donnée à l'équation (1.21).

Pour l'approche conditionnelle, on définit, pour l'historique ω, t_ω comme l'occasion de la première capture, avec $t_\omega = 1, \ldots, \ell$. La décomposition $L_M(N, P, c) = L_1(N, p) \times L_2(P, c)$ donne

$$L_1(N, P) = \frac{N!}{n!(N-n)!} \{(1-P)^\ell\}^{N-n} \{1-(1-P)^\ell\}^n \tag{1.31}$$

et

$$L_2(P, c) = \frac{n!}{\prod_{\omega\in\Omega} n_\omega!} \prod_{\omega\in\Omega} \{Q_\omega(P)\}^{n_\omega}, \tag{1.32}$$

avec

$$Q_\omega(P) = \frac{c^{\sum_{j=t_\omega+1}^{\ell} \omega_j} P (1-P)^{t_\omega - 1} (1-c)^{\ell - t_\omega - \sum_{j=t_\omega+1}^{\ell} \omega_j}}{1-(1-P)^\ell}.$$

La maximisation se fait en deux étapes. D'abord, on estime le paramètre p sur la base de la fonction $L_2(p, c)$

pour déduire l'estimateur $\hat{p}_C$. Ensuite, on maximise $L_1(N; \hat{p}_C)$ avant de déduire l'estimateur pour N :

$$\hat{N}_C = \left[\frac{n}{1 - (1 - \hat{p}_C)^{\ell}} \right]. \tag{1.33}$$

Pour la régression de Poisson, le modèle des fréquences prédites des historiques μ_ω s'écrit :

$$\log \mu_\omega = \gamma + \boldsymbol{X}_{1,\omega}\beta_1 + \boldsymbol{X}_{2,\omega}\beta_2, \tag{1.34}$$

avec $\gamma = \log\{NP(1-c)^{(l-1)}\}$, $\beta_1 = \log\{1 - p/1 - c\}$, $\beta_2 = \log\{c/1 - c\}$. $\boldsymbol{X}_{1,\omega}$ représente le temps de la première capture moins 1 et suit une loi géométrique de paramètre P, $\boldsymbol{X}_{2,\omega}$ représente le nombre de captures à la suite de la première, ainsi la distribution conditionnelle $\boldsymbol{X}_{2,\omega}$ sachant que $\boldsymbol{X}_{1,\omega} = k$ suit une loi binomiale de paramètres $l - k - 1$ et c. Ainsi, pour ce modèle l'estimation de la taille de la population est donnée par :

$$\hat{N} = e^{\hat{\gamma}} \frac{(1 + e^{\hat{\beta}_2})^{\ell}}{1 + e^{\hat{\beta}_2} - e^{\hat{\beta}_1}}. \tag{1.35}$$

Lorsque $P = c$, alors $\beta_1 = 0$ et l'estimation de la taille de la population devient :

$$\hat{N} = e^{\hat{\gamma}} \frac{(1 + e^{\hat{\beta}})^{\ell}}{e^{\hat{\beta}}}. \tag{1.36}$$

Cet estimateur correspond à celui du modèle M_0. Le modèle M_b est aussi une généralisation du modèle M_0. Par ailleurs, il correspond à un modèle d'élimination, car après la première capture, l'historique résultante des autres captures ne permet pas d'estimer le paramètre de capture P.

CHAPITRE 2
ECOSYSTÈME MOBILE ET DONNÉES D'ACTIVATION

Dans ce chapitre, nous allons tout d'abord nous intéresser à la présentation de l'écosystème mobile. Ensuite, nous présenterons les données brutes, ainsi que le processus de transformation de ces données en données de capture-recapture. Nous finirons ce chapitre en nous focalisant sur la modélisation des données mobiles.

2.1 Présentation de l'écosystème mobile

L'écosystème mobile est un cadre dans lequel différents acteurs interagissent afin de tirer profit des activités des utilisateurs des appareils mobiles. Il est constitué des utilisateurs, éditeurs, annonceurs et plateformes d'approvisionnement.

Les utilisateurs disposent des téléphones intélligents dans lesquels sont installées des applications qui permettent de consulter leurs réseaux sociaux, accéder aux informations de la météo, passer des commandes. Une application mobile est un logiciel applicatif transportable et autonome, développé pour être installé sur un appareil électronique mobile. Elle est identifiée par un ou plusieurs programmes téléchargeables de façon gratuite ou payante depuis un magasin d'applications "Application Store ", permettant d'accéder à un contenu homogène et exécutable à partir du système d'exploitation du Smartphone.

L'éditeur est le propriétaire de l'application ; il permet à des tiers de payer pour placer leurs espaces publicitaires.Le coût de la publicité est principalement évalué en fonction du coût par clic pour les publicités qui dirigent vers des sites web, et du coût par impression pour les publicités qui mettent en avant des images, des vidéos ou d'autres types de visuels. Les annonceurs ont la possibilité de payer pour un nombre spécifique de clics ou d'impressions, à partir duquel la diffusion de la publicité est interrompue. Les éditeurs bien connus incluent Facebook, Twitter, Google, Linkedin et tripadvisor.

Les annonceurs qui sont généralement des entreprises, cherchent à promouvoir leurs produits et services en ligne. Par exemple, cela peut inclure des boutiques en ligne, des supermarchés, des sites d'actualités et d'informations, des banques, des compagnies d'assurances et des créateurs d'applications. Un exemple concret serait un annonceur qui souhaite cibler les amateurs de vélo, âgés de 18 à 45 ans, dans un rayon de 10 km, et utilisant des smartphones via Facebook. Dans ce cas, l'annonceur définit un objectif commercial,

tel que la vente de vélos ou la promotion de sa marque, identifie le public cible, crée la publicité, puis Facebook diffuse cette publicité aux utilisateurs correspondants.

La plateforme d'approvisionnement est un lieu de rencontre (virtuel) entre les éditeurs et les annonceurs ; elle fournit une infrastructure numérique (telle que des logiciels) qui aide à cibler les publics que les annonceurs souhaitent atteindre. Il fournit aux éditeurs des services de gestion d'espace publicitaire en sélectionnant et en plaçant des espaces publicitaires envoyés par les annonceurs. Les plates-formes d'approvisionnement populaires incluent Meta for Business, Meta Audience Network (publicité orienté app or in-app), Meta Pixel (orienté annonceur pour affiner le ciblage).

Le ciblage : les publicités s'affichent en fonction de vos activités dans les apps Meta (Pages suivis/likés, infos profils Facebook/IG telles que âge, employeur, formation, contenus créés, lieux visités et indiqués), activités chez d'autres entreprises (si vous partagez téléphone, e-mail pour newsletter, promos, coupons, qui peut être ajoutée à un fichier clientèle puis partagée avec Facebook via Meta Pixel, Facebook recoupe les données pour le ciblage), activités sur d'autres sites web et apps (qui peuvent envoyer les données à Facebook via Meta Pixel si vous consultez une page web, téléchargez une app, créé un panier en ligne ; Facebook peut utiliser ces infos pour vous montrer des pubs de vêtements, chaussures, etc.), votre localisation (endroit où vous vous connectez à Internet, activé Facebook/Instagram, ou indiqué sur vos profils pour ciblage basé sur votre présence à un endroit).

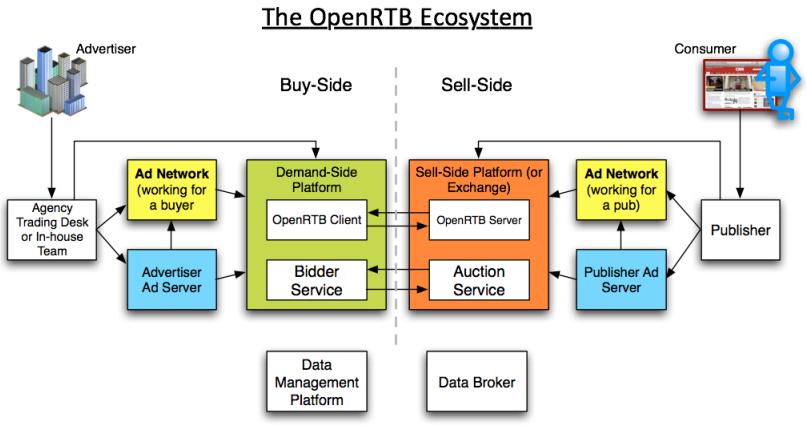

Figure 2.1 Représentation des interactions entre les acteurs de l'écosystème mobile (Gallo, 2015).

Le processus de vente et d'achat d'espaces publicitaires mobiles peut être formellement décrit comme suit. Lorsqu'un utilisateur mobile active l'application, il y a une vente enchère instantanée en temps réel entre les éditeurs pour avoir une chance de placer leurs annonces sur le smartphone activé. Au terme du processus complexe succinctement décrit par Gallo (2015), l'éditeur (acheteur) gagnant est sélectionné et son spot est placé sur le smartphone de l'utilisateur.

2.2 Transformation des données brutes en données de capture-recapture

Les informations sur les utilisateurs des appareils mobiles collectées lors des l'interactions entre édireurs et annonceurs, et sauvegardées dans un fichier sont les suivantes : ID (smartphone), localisation, date, type d'appareil, etc.

- *ID* : il s'agit de l'identifiant intra-app pour les smartphones fonctionnant sous le système iOs, Apple Inc, ou du numéro de série du téléphone pour les smartphones fonctionnant sous le système Android ;
- *Localisation* : fournit la latitude et la longitude de l'endroit visité par l'utilisateur grâce au système GPS ;
- *Date* : fournit la date et l'heure à laquelle la publicité a été lancée sur le téléphone mobile de l'utilisateur ;
- *Type d'appareil mobile* : il s'agit des informations sur le modèle du smartphone, son âge, la liste d'installation de l'application qui a été activée ainsi que le pouvoir de la batterie.

Ces données massives d'activations, collectées sur une base journalière à travers le monde, peuvent être acquises par les plateformes de marketing, qui entrevoient d'y extraire des informations utiles aux entreprises et entités financières. Ces informations sont souvent relatives à la fréquence et à l'étendue des mouvements de personnes qui se rendent dans des lieux publics de divertissement (e.g : stages, salles de cinéma ou parcs) ; visitent des magasins ou fréquentent des restaurants. L'utilité de telles informations réside dans l'intérêt grandissant des entreprises à mesurer, par exemple, la fréquence des visites dans leurs magasins de détails à la suite de campagnes publicitaires mobiles.

Table 2.1 Exemple de données d'activation issues du processus d'apurement

Numéro	ID	Date	Etat
1	abbb0fa-e0e3-408d-aa0f	$2014-07-10$	CA
2	abcb0fa-e0e3-408d-aa0f	$2015-09-07$	CA
3	abbd0fa-e0e3-408d-aa0f	$2015-06-27$	CA
4	abbb0fa-e0e3-408d-aa0f	$2015-06-29$	CA
⋮	⋮	⋮	⋮
3980	abbb0ha-e0e3-4888d-aa0f	$2014-07-17$	CA

Les données d'activations mobiles enregistrées sont ensuite transformées en données de capture-recapture suivants plusieurs critères : les activations provenant du même téléphone mobile et enregistrées à travers plusieurs applications sont liées grâce aux identifiants fournis lors de la transaction ; les activations frauduleuses sont isolées et supprimées à partir des techniques de detections de fraudes ; les données de géolocalisation érronées sont détectées et corrigées.

La transformation des données d'activation d'applications en données de capture-recapture peut se faire en suivant ces étapes :

— *Collecte des données* : la collecte les données relatives à l'activation des applications mobiles, inclut le nombre d'activations quotidiennes, les identifiants des utilisateurs et les informations temporelles précises. Ces données peuvent être obtenues en utilisant des fonctionnalités de suivi intégrées dans les applications ou en extrayant des informations à partir de sources telles que des journaux d'activité ;

— *Identification des unités* : Dans le contexte de la capture-recapture, les unités correspondent aux utilisateurs d'applications mobiles. Identifiez de manière unique chaque utilisateur à partir de leurs identifiants, par exemple, en utilisant des identifiants d'appareils uniques ou des identifiants publicitaires ;

— *Définition des occasions de capture* : Les occasions de capture correspondent aux jours ou aux périodes spécifiques pendant lesquelles les activations d'applications ont été enregistrées. Divisez les données en intervalles de temps appropriés pour définir les occasions de capture, comme les jours, les semaines ou les mois ;

— *Création de la matrice de capture-recapture* : Construisez une matrice de capture-recapture en uti-

lisant les unités (utilisateurs) en tant que lignes et les occasions de capture (jours) en tant que colonnes. Marquez les cellules de la matrice avec des 1 si l'unité a été capturée (activation enregistrée) à l'occasion correspondante, et avec des 0 sinon.

Table 2.2 Exemple de données d'activation issues du processus d'apurement transformées en données de capture-recapture

ID	jour 1	jour 2	...	jour 59
abbb0fa-e0e3-408d-aa0f	1	1	0	1
abcb0fa-e0e3-408d-aa0f	0	1	1	0
abbb0ha-e0e3-408d-aa0f	1	0	1	1
abbb0fa-e0e3-408d-aa0f	1	0	0	1
⋮	⋮	⋮	⋮	⋮
abbb0ha-e0e3-4888d-aa0f	0	1	1	0

Une fois transformées, ces données peuvent être modélisées comme étant obtenues à partir d'expériences de capture-recapture.

2.3 Modélisation des données mobiles d'activation

Considérons une population de N personnes envisageant des visites chez des concessionnaires d'automobiles au cours de la semaine. Posons l'hypothèse de non-variation de la taille de la population dans la semaine et supposons que chaque individu de la population possède un téléphone intelligent sur lequel est installée au moins une application mobile. Une personne de la population est enregistrée dans l'ensemble de données d'activation si trois conditions sont remplies : (i) la personne se rend effectivement dans l'entreprise au cours de la semaine (ii) elle active une application sur son appareil mobile lorsqu'elle se trouve sur place et (iii) cette activation est enregistrée par la plateforme du côté de l'offre qui fournit les données. Ainsi, la probabilité qu'un individu soit effectivement capturé dépend de la probabilité que cet imdividu se rende effectivement dans l'entreprise au cours de la semaine, la probabilité conditionnelle qu'il active une application sur son appareil mobile sachant qu'il a visité l'entreprise et la probabilité conditionnelle que cette activation soit enregistrée par la plateforme du côté de l'offre qui fournit les données conditionnellement à la visite et l'activation. Si ces trois probabilités conditionnelles ne varient pas durant la semaine, sont indépendantes et sont constantes d'un individu à l'autre, alors le nombre de captures pour un individu peut être modélisé par un modèle M_0.

Cependant, les probabilités conditionnelles peuvent varier dans le temps. Par exemple, il se peut qu'il y ait beaucoup plus d'achalandage durant les week-ends et moins durant la semaine. Il se peut également qu'il y ait une fréquentation beaucoup plus remarquée lors d'évènements spéciaux tels que les fêtes de fin d'années ou d'autres périodes de congés. On pourrait ainsi s'attendre à d'importantes variations temporelles pour les probabilités de capture des unités. Dans ce cas, sous l'hypothèse d'homogénéité inter-individuelle des probabilités conditionnelles, le nombre de captures intra-hebdomadaire peut être modélisé par un modèle M_t.

En outre, l'hypothèse de l'homogénéité inter-individuelle des probabilités conditionnelles est très peu plausible. En effet, la probabilité qu'une personne visite un concessionnaire peut dépendre de ses goûts et préférences pour certaines voitures et/ou d'autres facteurs. Aussi, la probabilité conditionnelle d'activer une application au cours de la visite chez le concessionnaire peut dépendre des facteurs socio-démographiques. Ainsi le nombre de captures au cours d'une semaine peut être modélisé par une distribution binomiale avec probabilité de succès la probabilité de revoir une personne après la première occasion de capture. Cette situation peut être modéliseé par un modèle M_b.

Le défi majeur pour l'ajustement des données de capture-recapture d'activations aux modèles de population fermée décrits plus haut concerne la dimensionnalité. En effet, lorsque le nombre d'occasions de capture est élevé, ce qui est très souvent le cas dans le contexte de la collecte de données d'activations, les programmes informatiques existants échouent dans la maximisation de la vraisemblance, mais également dans l'inversion de la matrice d'information de Fisher. Le problème est beaucoup plus sérieux lorsqu'il s'agit d'ajuster le modèle temporel M_t, dont le nombre de colonnes de la matrice de design est égal au nombre d'occasions de capture plus un. Au chapitre suivant, nous présenterons des algorithmes d'estimation des paramètres des modèles M_0, M_t et M_b basés sur des équations d'estimation obtenues à l'aide de l'approche des moments, ainsi que des formules explicites pour la variance des estimateur de N et p^*.

CHAPITRE 3
MODÈLES D'ACTIVATION POUR LES POPULATIONS FERMÉES

Ce chapitre présente la methodologie déployée afin d'estimer la taille de la population à partir de données d'activation d'applications mobiles dans le cas des populations fermées. Notre démarche de travail s'articule autour de l'algorithme d'estimation des paramètres dans le cas du modèle M_0, du modèle temporel M_t et du modèle comportemental M_b. La contribution spécifique dans cette méthodologie est le développement d'un nouvel algorithme pour le modèle M_b ; pour les modèles M_0 et M_t, les algorithmes développés peuvent être déduits des résultats de Yauck *et al.* (2019) dans le cas plus général des modèles pour populations ouvertes. Ledit algorithme se propose d'estimer des paramètres de population, ainsi que les formules explicites pour la variance des estimateurs, à partir des données d'activation d'applications mobiles dans le cas des populations fermées.

3.1 Données et modèles

Considérons une expérience de capture-recapture basée sur les enregistrements d'activations d'applications mobiles, réalisés à travers ℓ occasions de capture. Les données $\{n_\omega\}$ issues de cette expérience sont associées à l'historique de capture observable $\omega \in \mathbf{\Omega}$ tel que $|\mathbf{\Omega}| = 2^\ell - 1$. Nous souhaitons estimer la taille N de la population à partir des données observées et quantifier l'incertitude associée à cette estimation. Pour ce faire, nous supposons que les fréquences observées suivent des lois de Poisson indépendantes d'espérance $\{\mu_\omega\}$:

$$n_\omega \sim \text{Poisson}(\mu_\omega = NP_\omega(\boldsymbol{\theta})), \tag{3.1}$$

où $\boldsymbol{\theta} = (\gamma, \boldsymbol{\beta}^\top)^\top$ représente le vecteur des paramètres du modèle log-linéaire μ_ω :

$$\log \mu_\omega = \gamma + X_\omega^\top \boldsymbol{\beta}, \tag{3.2}$$

où $\boldsymbol{X}_\omega^\top$ est le vecteur de dimension $d \times 1$ des variables explicatives pour l'historique $\boldsymbol{\omega}$ dont la spécification dépend du modèle de population fermée M_0, M_t ou M_b. La vraisemblance Poisson L_P s'écrit comme suit :

$$L_P(N, \boldsymbol{\theta}) = \prod_{\omega \in \Omega} \frac{e^{-\mu_\omega} \mu_\omega^{n_\omega}}{n_\omega!}. \tag{3.3}$$

Notons que $\boldsymbol{X}$ représente la matrice de design du modèle de population fermée, contenant $2^\ell - 1$ lignes définies par les vecteurs $\boldsymbol{X}_\omega$ associés aux historiques observables $\boldsymbol{\omega}$ et $|\boldsymbol{\theta}| = d$; les fréquences observées associées aux historiques observables sont consignées dans le vecteur $\boldsymbol{y}$, et $\boldsymbol{\mu}$ représente le vecteur des fréquences prédites. En appliquant le logarithme à l'équation (3.3) et en remplaçant $\log \mu_\omega$ par son expression, on obtient la log-vraisemblance qui s'exprime comme suit :

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \sum_{\omega \in \Omega} [-\mu_\omega + n_\omega \log \mu_\omega] \\
&= \sum_{\omega \in \Omega} [-\mu_\omega + n_\omega (\gamma + \boldsymbol{X}_\omega^\top \boldsymbol{\beta})] \\
&= \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} n_{\omega_1}(\gamma + X_{\omega_1}\boldsymbol{\beta}) - \mu_{\omega_1} \\ n_{\omega_2}(\gamma + X_{\omega_2}\boldsymbol{\beta}) - \mu_{\omega_2} \\ \vdots \\ n_{\omega_{2^{l}-1}}(\gamma + X_{\omega_{2^{l}-1}}\boldsymbol{\beta}) - \mu_{\omega_{2^{\ell}-1}} \end{bmatrix} \\
&= \mathbf{1}^\top \begin{bmatrix} n_{\omega_1} \begin{bmatrix} 1 & X_{\omega_1} \end{bmatrix} \begin{bmatrix} \gamma \\ \beta \end{bmatrix} - \mu_{\omega_1} \\ n_{\omega_2} \begin{bmatrix} 1 & X_{\omega_2} \end{bmatrix} \begin{bmatrix} \gamma \\ \beta \end{bmatrix} - \mu_{\omega_2} \\ \vdots \\ n_{\omega_{2^{l}-1}} \begin{bmatrix} 1 & X_{\omega_{2^{\ell}-1}} \end{bmatrix} \begin{bmatrix} \gamma \\ \beta \end{bmatrix} - \mu_{\omega_{2^{\ell}-1}} \end{bmatrix},
\end{aligned}$$

où $\mathbf{1}^\top$ est le vecteur de dimension $1 \times s$, d'éléments 1. Quelques développements algébriques donnent

$$\mathcal{L}(\boldsymbol{\theta}) = \begin{bmatrix} n_{\omega_1} & n_{\omega_2} & \dots & n_{\omega_{2^{\ell}-1}} \end{bmatrix} \begin{bmatrix} 1 & X_{\omega_1} \\ 1 & X_{\omega_2} \\ \vdots & \\ 1 & X_{\omega_{2^{\ell}-1}} \end{bmatrix} \begin{bmatrix} \gamma \\ \beta \end{bmatrix} - \mathbf{1}^\top \begin{bmatrix} \mu_{\omega_1} \\ \mu_{\omega_2} \\ \vdots \\ \mu_{\omega_{2^{\ell}-1}} \end{bmatrix}.$$

La log-vraisemblance est ainsi donnée par

$$\mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\theta} - \mathbf{1}^\top \boldsymbol{\mu}. \tag{3.4}$$

Ainsi, en dérivant la log-vraisemblance par rapport à $\boldsymbol{\theta}$, on obtient ainsi la fonction de score-poisson de

dimension $1 \times s$. En effet,

$$\begin{aligned} \boldsymbol{S_P(\theta)} &= \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\boldsymbol{\partial \theta}} \\ &= \frac{\partial (\boldsymbol{y}^\top \boldsymbol{X\theta} - \mathbf{1}^\top \boldsymbol{\mu})}{\boldsymbol{\partial \theta}}. \end{aligned}$$

Or $\mu = e^{\boldsymbol{X\theta}}$; Il suit alors que,

$$\boldsymbol{S_P(\theta)} = \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{X}^\top e^{\boldsymbol{X\theta}}.$$

On obtient ainsi la fonction de score suivante :

$$\boldsymbol{S_P(\theta)} = \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{\mu}). \tag{3.5}$$

Dès lors, on égalise la fonction de score à 0 et on obtient les équations suivantes :

$$\sum_{\omega \in \Omega} \boldsymbol{X}_\omega n_\omega = \sum_{\omega \in \Omega} \boldsymbol{X}_\omega \mu_\omega, \tag{3.6}$$

et

$$\sum_{\omega \in \Omega} n_\omega = \sum_{\omega \in \Omega} \mu_\omega. \tag{3.7}$$

Dans la suite, la forme du vecteur $\boldsymbol{X}_\omega$ dépendra des hypothèses concernant les procédures de captures des unités d'une occasion à une autre, selon qu'on sera en présence d'un modèle M_0, M_t ou M_b.

3.2 Modélisation M_0

On considère dans cette section que le mécanisme de détection des unités suit le modèle homogène le plus simple, M_0. L'algorithme récursif d'estimation des paramètres du modèle et les formules explicites pour les variances asymptotiques des estimateurs sont présentés dans les paragraphes suivants.

3.2.1 Algorithme d'estimation des paramètres

Cet algorithme permet d'estimer $P = e^\beta/(1+e^\beta)$, la probabilité de capture des unités au cours de l'expérience et $N = n + e^\gamma$, la taille de la population. Ici, le vecteur des variables explicatives pour l'historique $\boldsymbol{\omega}$

est sous la forme $\boldsymbol{X}_{\boldsymbol{\omega}} = \sum_{j=1}^{\ell} \omega_j$. Les statistiques exhaustives du modèle (C, n) s'obtiennent en exploitant les équations (3.6) et (3.7). D'abord, en exploitant l'équation (3.6), on a :

$$\sum_{\omega \in \Omega} \boldsymbol{X}_{\omega} n_{\omega} = \sum_{\omega \in \Omega} \boldsymbol{X}_{\omega} \mu_{\omega}.$$

Or, nous avons

$$\sum_{\omega \in \Omega} \boldsymbol{X}_{\omega} n_{\omega} = \sum_{\omega \in \Omega} n_{\omega} \sum_{j=1}^{\ell} \omega_j = C.$$

De plus, on a

$$\sum_{\omega \in \Omega} \boldsymbol{X}_{\omega} \mu_{\omega} = \sum_{\omega \in \Omega} \mu_{\omega} \sum_{j=1}^{\ell} \omega_j = \sum_{\omega \in \Omega} NP_{\omega} \sum_{j=1}^{\ell} \omega_j = NP\ell.$$

En exploitant ensuite l'équation (3.7), on a :

$$\sum_{\omega \in \Omega} n_{\omega} = \sum_{\omega \in \Omega} \mu_{\omega}.$$

Or,

$$\sum_{\omega \in \Omega} n_{\omega} = n.$$

De plus,

$$\begin{aligned} \sum_{\omega \in \Omega} \mu_{\omega} &= \sum_{\omega \in \Omega} NP_{\omega} \\ &= \sum_{\omega \in \Omega^0} N[P_{\omega} - P_0] \\ &= N \left[\sum_{\omega \in \Omega^0} P_{\omega} - P_0 \right] \\ &= NP^*, \end{aligned}$$

avec $P^* = 1 - (1 - P)^{\ell}$. On obtient ainsi les équations suivantes :

$$C = NP\ell, \tag{3.8}$$

et

$$n = NP^*. \tag{3.9}$$

L'équation (3.8) peut s'écrire de la façon suivante :

$$P = \frac{C}{N\ell}. \tag{3.10}$$

La statistique exhaustive n est obtenue à partir de l'équation (3.9) donnée par :

$$n = NP^* = N\{1 - (1 - P)^{\ell}\}. \tag{3.11}$$

En remplaçant (3.10) dans (3.11), on obtient l'équation d'estimation pour N, $f_N^0(N) = 0$, où

$$f_N^0(N) = N \times \left\{1 - \left(1 - \frac{C}{N\ell}\right)^{\ell}\right\} - n. \tag{3.12}$$

L'algorithme d'estimation du modèle M_0 peut se faire en deux étapes :

- résoudre de l'équation (3.12) égale à 0, qui permet d'obtenir $\hat{N}$, l'estimateur de N ;
- remplacer N par $\hat{N}$ dans l'équation (3.10), pour obtenir $\hat{P}$, l'estimateur de P. Aussi, l'estimateur de P^* est donné par $\hat{P}^* = 1 - (1 - \hat{P})^{\ell}$.

3.2.2 Estimation des variances asymptotiques

Dans cette section, les formules explicites des variances asymptotiques des estimateurs $\hat{N}$ et $\hat{P}^*$ obtenues à partir des équations (3.10) et (3.12) sont proposées. Soit $p_2 = 1 - (1 - P)^{\ell} - \ell P(1 - P)^{\ell-1}$ la probabilité qu'un individu soit capturé au moins deux fois.

Proposition 3.1 *Dans le modèle de population fermée M_0, les variances asymptotiques de $\hat{N}$ et $\hat{P}^*$ sont données respectivement par :*

$$\mathit{Var}_M(\hat{N}) = \frac{N(1 - P^*)}{p_2} \tag{3.13}$$

et

$$\mathit{Var}(\hat{P}^*) = \frac{P^*(1 - P^*)p_1}{Np_2}, \tag{3.14}$$

où $p_1 = \ell P(1 - P)^{\ell-1}$ représente la probabilité qu'une unité soit capturée exactement une fois.

Preuve 3.1 *La démonstration des formules de variances asymptotiques sera basée sur le fait que le rapport*

entre une statistique et son espérance converge en probabilité vers 1 lorsque N tend vers l'infini : $C \approx N\ell P$, $n \approx NP^*$. *L'équation d'estimation* (3.12) *est fonction des statistiques exhaustives* C *et* n. *La variance asymptotique de* $\hat{N}$ *est obtenue par linéarisation. Puisque*

$$f_N^0(\hat{N}) - f_N^0(N) \approx (\hat{N} - N)\frac{df_N^0(\hat{N})}{dN}, \tag{3.15}$$

il suit que

$$Var_M(\hat{N}) = \left[Var\{f_N^0(N)\}/A^2\right] - N, \tag{3.16}$$

où $Var\{f_N^0(N)\}$ *est une approximation de la variance asymptotique de* $f_N^0(N)$, *vue comme une fonction des statistiques* C *et* n ; A *est la limite en probabilité de la dérivée de* $f_N^0(N)$ *par rapport à* N. *À partir des dérivations, on obtient l'expression de* A. *On a :*

$$\begin{aligned}
\frac{df_N^0(N)}{dN} &= 1 - \left(1 - \frac{C}{N\ell}\right)^{\ell} - N\left[\frac{C}{N^2}\left(1 - \frac{C}{N\ell}\right)^{\ell-1}\right] \\
&= 1 - (1-P)^{\ell} - \frac{C}{N}(1-P)^{\ell-1} \\
&= 1 - (1-P)^{\ell} - \ell P(1-P)^{\ell-1} \\
&= p_2.
\end{aligned}$$

Donc, $A = p_2$.

De plus, on a :

$$Var\{f_N^0(N)\} = \nabla f_N^0 \Sigma^0 (\nabla f_N^0)^T, \tag{3.17}$$

où Σ^0 *est une matrice de variance-covariance de* C *et* n *de dimension* 2×2, *et* ∇f_N^0 *est la limite du vecteur des dérivées partielles de* f_N^0 *par rapport à* C *et* n. *Pour calculer* Σ^0, *on définit* $\tilde{N}$ *comme une variable aléatoire qui suit une loi de Poisson de paramètre* N ; *elle représente le nombre d'individus de la population tout au long de l'expérience. Puisque* n *est une variable aléatoire suivant une loi de Poisson, sa variance est égale à son espérance :* $Var_p(n) = \mathbb{E}_p(n) = NP^*$. *Puisque* C *suit une loi Binomiale de paramètres* $\tilde{N} \times \ell$ *essais et de probabilité* P, *sa variance est calculée de la façon suivante :*

$$\begin{aligned}
Var_p(C) &= \mathbb{E}(\tilde{N})\ell P(1-p) + Var(\tilde{N})\ell^2 P^2 \\
&= NP\ell(\ell P + 1 - P).
\end{aligned}$$

Pour calculer la covariance entre C et n, on pose conditionnellement à $\tilde{N}$:

$$C = \sum_{i=1}^{\tilde{N}} X_i, \quad n = \sum_{i=1}^{\tilde{N}} \mathbf{1}_{(X_i>0)}.$$

La variable X_i suit une loi Binomiale avec ℓ essais et de probabilité P. On a donc :

$$\begin{aligned}
\mathsf{Cov}_p(C,n) &= \mathbb{E}\left\{\mathsf{Cov}(C,n|\tilde{N})\right\} + \mathsf{Cov}\left\{\mathbb{E}(C|\tilde{N}), \mathbb{E}(n|\tilde{N})\right\} \\
&= \mathbb{E}\left\{\mathsf{Cov}\left(\sum_{i=1}^{\tilde{N}} X_i, \mathbb{1}_{(X_i\leq 0)}\right)\right\} + \mathsf{Cov}\left(\tilde{N}\ell P, \tilde{N}P^*\right) \\
&= NP\ell(1-P^*) + N\ell PP^* \\
&= NP\ell.
\end{aligned}$$

Ainsi,

$$\Sigma^0 = N\begin{bmatrix} \ell P(\ell P + 1 - P) & \ell P \\ \ell P & P^* \end{bmatrix}. \tag{3.18}$$

Par ailleurs, on montre que

$$\nabla f_N^0 = p\lim_{N->\infty}\left(\frac{\partial f_N^0(C,n)}{\partial C}, \frac{\partial f_N^0(C,n)}{\partial n}\right) = ((1-P)^{\ell-1}, -1).$$

En remplaçant ces quantités dans les équations (3.16) *et* (3.17)*, on obtient la variance asymptotique de $\hat{N}$.*

En effet,

$$\begin{aligned}
Var\{f_N^0(N)\} &= \nabla f_N^0 \Sigma^0 \nabla (f_N^0)^T \\
&= N \begin{bmatrix} (1-P)^{\ell-1} & -1 \end{bmatrix} \begin{bmatrix} (\ell P + 1 - P) & \ell P \\ \ell P & P^* \end{bmatrix} \begin{bmatrix} (1-P)^{\ell-1} \\ -1 \end{bmatrix} \\
&= N \begin{bmatrix} \ell P (1-P)^{\ell-1} (\ell P + 1 - P) - \ell P & \ell P (1-P)^{\ell-1} - P^* \end{bmatrix} \begin{bmatrix} (1-P)^{\ell-1} \\ -1 \end{bmatrix} \\
&= N \begin{bmatrix} p_1 (\ell P + 1 - P) - \ell P & p_2 \end{bmatrix} \begin{bmatrix} (1-P)^{\ell-1} \\ -1 \end{bmatrix}, \quad avec \quad p_2 = P^* - p_1 \\
&= N \left[p_1 \ell P (1-P)^{\ell-1} + (1-P)(1-P)^{\ell-1} - \ell P (1-P)^{\ell-1} + p_2 \right] \\
&= N \left[p_1 \left\{ p_1 + (1-P)^{\ell-1} - 1 \right\} + p_2 \right] \\
&= N \left[p_1 \left(p_1 - P^* \right) + p_2 \right] \\
&= N \left[p_2 (1 - p_1) \right].
\end{aligned}$$

Ainsi,

$$\begin{aligned}
Var_M(\tilde{N}) &= \frac{N \left[p_2 (1-p_1) \right]}{{p_2}^2} - N \\
&= \frac{N \left[p_2 (1-p_1) \right]}{p_2} - N \\
&= N \frac{1 - p_1 - p_2}{p_2} \\
&= \frac{N(1-P^*)}{p_2}, \quad avec \quad P^* = p_2 + p_1.
\end{aligned}$$

Donc

$$Var_M(\hat{N}) = \frac{N(1-P^*)}{p_2}.$$

Considérons maintenant $\hat{P}^* = n/\hat{N}$. *On a par linéarisation*

$$Var(\hat{P}^*) = \nabla \hat{P}^* \Gamma^0 (\nabla \hat{P}^*)^T, \tag{3.19}$$

où Γ^0 *est est une matrice* 2×2 *de variance covariance de* $\hat{N}$ *et* n, $\nabla \hat{P}^*$ *est la limite du vecteur des dérivées partielles de* $\hat{P}^*$ *par rapport à* $\hat{N}$ *et* n. *On obtient ainsi :*

$$\nabla \hat{P}^* = p \lim_{N->\infty} \left(\frac{\partial \hat{P}^*(\hat{N}, n)}{\partial \hat{N}}, \frac{\partial \hat{P}^*(\hat{N}, n)}{\partial n} \right) = \frac{1}{N}(-P^*, 1).$$

Aussi, la linéarisation de $\hat{N}$ permet d'obtenir Cov$(n, \hat{N})$. En effet,

$$f_N^0(\hat{N}) - f_N^0(N) \approx (\hat{N} - N)\frac{df_N^0(\hat{N})}{dN}.$$

Alors,

$$(\hat{N} - N) \approx f_N^0(N)\left(-\frac{df_N^0(\hat{N})}{dN}\right)^{-1}.$$

Il vient que

$$\begin{aligned}
\text{Cov}(n, \hat{N}) &= \text{Cov}\left\{n, f_N^0(C, n)\left(-\frac{df_N^0(\hat{N})}{dN}\right)^{-1}\right\} \\
&= \text{Cov}\left\{n, f_N^0(C, n)(-A^{-1})\right\} \\
&= \text{Cov}\left\{n, \left[f_N^0(C, n) - f_N^0(\mathbb{E}(C), \mathbb{E}(n))\right](-A^{-1})\right\}.
\end{aligned}$$

$$f_N^0(C, n) - f_N^0(\mathbb{E}(C), \mathbb{E}(n)) \approx (C - \mathbb{E}(C))\frac{df_N^0(C, n)}{dC} + (n - \mathbb{E}(n))\frac{df_N^0(C, n)}{dn}.$$

Donc, nous avons

$$\begin{aligned}
\text{Cov}(n, \hat{N}) &= -A^{-1}\left\{\text{Cov}_p(n, C)\frac{df_N^0(C, n)}{dC} + \text{Var}_p(n)\frac{df_N^0(C, n)}{dn}\right\} \\
&= -A^{-1}\left\{NP\ell(1 - P)^{\ell-1} - NP^*\right\} \\
&= \frac{NP\ell(1 - P)^{\ell-1} - NP^*}{-(P^* - p_1)} \\
&= \frac{N(P^* - p_1)}{P^* - p_1} \\
&= N.
\end{aligned}$$

Ainsi, Cov$(n, \hat{N}) = N$. Donc, on a

$$\Gamma^0 = N\begin{bmatrix} \frac{1-p_1}{p_2} & 1 \\ 1 & P^* \end{bmatrix}.$$

En remplaçant ces quantités dans (3.19)*, on obtient la variance asymptotique de $\hat{P}^*$. On a :*

$$\begin{aligned}
\text{Var}(\hat{P}^*) &= \nabla\hat{P}^*\Gamma^0(\nabla\hat{P}^*)^T \\
&= \frac{1}{N}\begin{bmatrix} \frac{1-p_1}{p_2} & 1 \\ 1 & P^* \end{bmatrix}\begin{bmatrix} -P^* & 1 \end{bmatrix}\begin{bmatrix} -P^* \\ 1 \end{bmatrix}.
\end{aligned}$$

On a ainsi

$$\begin{aligned}
\textit{Var}(\hat{P}^*) &= \frac{1}{N}\begin{bmatrix}\frac{-P^*(1-p_1)}{p_2}+1 & 0\end{bmatrix}\begin{bmatrix}-P^*\\ 1\end{bmatrix}\\
&= \frac{1}{N}\left(\frac{(P^*)^2(1-p_1)}{p_2}-P^*\right)\\
&= \frac{P^*(P^*-p_2)-p_1(P^*)^2}{Np_2}\\
&= \frac{P^*p_1-p_1(P^*)^2}{Np_2}\\
&= \frac{P^*(1-P^*)p_1}{Np_2}.
\end{aligned}$$

Donc

$$\textit{Var}(\hat{P}^*) = \frac{P^*(1-P^*)p_1}{Np_2}.$$

Une autre méthode pour le calcul des variances asymptotiques des estimateurs $\hat{N}$ et $\hat{P}^*$ est un boostrap paramétrique. On souhaite obtenir des répétitions de bootstrap pour le vecteur des statistiques exhaustives n et C, qui se fait en deux étapes :

Etape 1 :

- Générer une réalisation de la taille de la population selon la distribution de Poisson : $\tilde{N} \sim Poisson(N)$; la valeur de N sera prise comme égale à son estimateur $\hat{N}$;
- Générer la statistique n selon une Binomiale $(\tilde{N}, P^*)$;
- Conditionnellement à n, on génère pour chaque individu i, $i = 1, 2, \ldots, n$ détecté au moins une fois au cours de l'expérience, $\boldsymbol{X}_i$ selon une loi binomiale et on calcule $C = \sum_{i=1}^{n} \boldsymbol{X}_i$.

Etape 2 :

- Egaliser les paramètres (N, P^*) à leurs estimés $(\hat{N}, \hat{P^*})$;
- Simuler L ensembles de statistiques exhaustives (n, C), puis estimer les paramètres N et P^* ;
- Calculer les variances asymptotiques en utilisant un boostrap paramétrique. La variance de l'estimateur de N est égale à la variance boostrap moins $\hat{N}$.

3.3 Modélisation M_t

On considère dans cette section que le processus de détection des captures des unités suit un modèle temporel M_t. L'algorithme récursif d'estimation des paramètres du modèle et les formules explicites pour les variances asymptotiques des estimateurs sont proposés dans les paragraphes suivants.

3.3.1 Algorithme d'estimation des paramètres

Cet algorithme permet d'estimer la probabilité de capture des unités à la j-ème occasion de capture, $P_j = e^{\beta_j}/(1+e^{\beta_j})$ et la taille de la population, $N = n+e^{\gamma}$. Le vecteur des variables explicatives pour l'historique $\boldsymbol{\omega}$ est $\boldsymbol{X}_{\boldsymbol{\omega}} = (\omega_1, \omega_2, \ldots, \omega_\ell)$. Les statistiques exhaustives du modèle sont donc :

$$\sum_{\omega\in\Omega} \boldsymbol{X}_\omega n_\omega = (n_1, n_2, \ldots, n_\ell)$$

et

$$\sum_{\omega\in\Omega} \boldsymbol{X}_\omega n_\omega = n.$$

La résolution de (3.4) égale à 0, conduit aux $\ell + 1$ équations d'estimation suivantes :

$$n_j = NP_j,\ j = 1, 2, \ldots, \ell \tag{3.20}$$

$$n = NP^*, \tag{3.21}$$

avec $P^* = 1 - \prod_{j=1}^{\ell}(1 - P_j)$. Les équations d'estimation (3.20) et (3.21) peuvent se reécrire de la façon suivante :

$$P_j = \frac{n_j}{N} \tag{3.22}$$

$$n = N\left\{1 - \prod_{j=1}^{\ell}(1 - P_j)\right\}. \tag{3.23}$$

Ainsi, l'équation d'estimation pour N, $f_N^t(N) = 0$, est obtenue en remplaçant (3.22) dans (3.23), où

$$f_N^t(N) = N \times \left\{1 - \prod_{j=1}^{\ell}(1 - P_j)\right\} - n. \tag{3.24}$$

L'algorithme d'estimation du modèle M_t se résume en deux étapes :

— Résoudre l'équation (3.24) égale à 0, qui permet d'obtenir $\hat{N}$, l'estimateur de N ;

— Remplacer N par $\hat{N}$ dans l'équation (3.22) pour obtenir $\hat{P}_j, j = 1, 2, \ldots, \ell$, l'estimateur de P_j. Donc l'estimateur de P^* est $\hat{P}^* = 1 - \prod_{j=1}^{\ell}(1 - \hat{P}_j)$.

3.3.2 Estimation des variances asymptotiques

Les formules explicites des variances asymptotiques des estimateurs $\hat{N}$ et $\hat{P}^*$ sont obtenues à partir de l'algorithme du modèle M_t. Soit p_2^* la probabilité qu'un individu soit capturé au moins deux fois. on a :

$$p_2^* = 1 - \prod_{j=1}^{\ell}(1 - P_j) - \sum_{j=1}^{\ell} P_j \prod_{s \neq j}(1 - P_s).$$

Proposition 3.2 *Dans le modèle de population fermée M_t, les variances asymptotiques de $\hat{N}$ et $\hat{P}^*$ sont données respectivement par :*

$$Var_M(\hat{N}) = \frac{N(1 - P^*)}{p_2^*} \tag{3.25}$$

et

$$Var(\hat{P}^*) = \frac{P^*(1 - P^*)p_1^*}{Np_2^*}, \tag{3.26}$$

où $p_1^* = \sum_{j=1}^{\ell} P_j \prod_{s \neq j}(1 - P_s)$ représente la probabilité qu'un individu soit capturé exactement une fois.

Preuve 3.2 *On utilise un même raisonnement que la démonstration de la Proposition 3.1. On a*

$$Var_M(\hat{N}) = \frac{1}{A^2} \nabla f_N^t \Sigma^t (\nabla f_N^t)^T - N, \tag{3.27}$$

avec A est la limite en probabilité de la dérivée de $f_N^t(N)$ par rapport à N. On montre pour le modèle M_t que $A = p_2^*$, Σ^t est la matrice $(\ell + 1) \times (\ell + 1)$ de variance-covariance de $\{n_j\}$ et n, et ∇f_N^t est la limite du vecteur des dérivées partielles de f_N^t par rapport à $\{n_j\}$ et n. On a :

$$Var_p(n_j) = NP_j,$$

$$Cov_p(n_j, n_k) = NP_jP_k, j \neq k,$$

$$Cov_p(n_j, n) = NP_j.$$

Il suit que

$$\Sigma^t = N \begin{bmatrix} P_1 & \dots & P_1P_\ell & P_1 \\ P_2P_1 & \dots & P_2P_\ell & P_2 \\ \dots & \dots & \dots & \dots \\ P_1 & \dots & P_\ell & P^* \end{bmatrix}. \tag{3.28}$$

Ensuite, on montre que

$$p\lim_{N->\infty} \frac{\partial f_N^t(\{n_j\}, n)}{\partial n_j} = \prod_{s\neq j}(1 - P_s).$$

On en deduit que

$$\nabla f_N^t = \left(\prod_{s\neq 1}(1 - P_s), \dots, \prod_{s\neq \ell}(1 - P_s), -1 \right).$$

En remplaçant ces quantités dans (3.27)*, on a*

$$\nabla f_N^t \Sigma^t (\nabla f_N^t) = N \begin{bmatrix} \prod_{s\neq 1}(1 - P_s) \\ \dots \\ \prod_{s\neq \ell}(1 - P_s) \\ -1 \end{bmatrix}^\top \begin{bmatrix} P_1 & \dots & P_1P_\ell & P_1 \\ P_2P_1 & \dots & P_2P_\ell & P_2 \\ \dots & \dots & \dots & \dots \\ P_1 & \dots & P_\ell & P^* \end{bmatrix} \begin{bmatrix} \prod_{s\neq 1}(1 - P_s) \\ \vdots \\ \prod_{s\neq \ell}(1 - P_s) \\ -1 \end{bmatrix}.$$

Le triple produit matriciel donne

$$\begin{aligned} \nabla f_N^t \Sigma^t (\nabla f_N^t)^\top &= N \left[-\sum_{j=1}^{\ell} P_j \prod_{s\neq j}(1 - P_s) - \sum_{j=1}^{\ell} P_j \prod_{s\neq j}(1 - P_s) \right] \\ &- N \left[\sum_{j=1}^{\ell} P_j \prod_{s\neq j}(1 - P_s) \prod_{s\neq j}(1 - P_s) + P^* + \left\{ \sum_{j=1}^{\ell} P_j \prod_{s\neq j}(1 - P_s) \right\}^2 \right] \\ &= N \left[1 - \prod_{j=0}^{\ell}(1 - P_j) - \sum_{j=1}^{\ell} P_j \prod_{s\neq j}(1 - P_s) \right] \left[1 - \sum_{j=1}^{\ell} P_j \prod_{s\neq j}(1 - P_s) \right] \\ &= N p_2^*(1 - p_1^*). \end{aligned}$$

Ainsi, nous avons

$$\begin{aligned} \mathit{Var}_M(\hat{N}) &= \frac{N p_2^*(1 - p_1^*)}{p_2^{*2}} - N \\ &= \frac{N(1 - p_1^* - p_2^*)}{p_2^*} \\ &= \frac{N(1 - P^*)}{p_2^*}. \end{aligned}$$

Donc, la variance asymptotique de $\hat{N}$ est

$$Var_M(\hat{N}) = \frac{N(1-P^*)}{p_2^*}.$$

Considérons maintenant $\hat{P}^* = n/\hat{N}$. On a par linéarisation

$$Var(\hat{P}^*) = \nabla\hat{P}^*\Gamma^t(\nabla\hat{P}^*)^\top, \tag{3.29}$$

où Γ^t est une matrice 2×2 de variance-covariance de $\hat{N}$ et n, $\nabla\hat{P}^*$ est la limite du vecteur des dérivées partielles de $\hat{P}^*$ par rapport à $\hat{N}$ et n. On obtient ainsi :

$$\nabla\hat{P}^* = p\lim_{N->\infty}\left(\frac{\partial\hat{P}^*(\hat{N},n)}{\partial\hat{N}}, \frac{\partial\hat{P}^*(\hat{N},n)}{\partial n}\right) = \frac{1}{N}(-P^*, 1).$$

La linéarisation de $\hat{N}$ permet d'obtenir Cov$(n, \hat{N}) = N$. Donc,

$$\Gamma^t = N\begin{bmatrix}\frac{1-p_1^*}{p_2^*} & 1\\ 1 & P^*\end{bmatrix}.$$

En remplaçant ces quantités dans (3.29)*, on obtient la variance asymptotique de $\hat{P}^*$.*

Une autre technique pour le calcul des variances asymptotiques des estimateurs $\hat{N}$ et $\hat{P}^*$ est un boostrap paramétrique. On souhaite obtenir des répétitions bootstrap pour le vecteur des statistiques exhaustives n_j et n, dont le principe est le suivant :

Etape 1 :

- Générer une réalisation de la taille de la population selon la distribution de Poisson : $\tilde{N} \sim Poisson(N)$; la valeur de N sera prise comme égale à son estimateur $\hat{N}$;
- Générer la statistique n selon une Binomiale $(\tilde{N}, P^*)$;
- Générer la statistique n_j selon des lois Binomiales $(\tilde{N}, P_j)$;

Etape 2 :

- Egaliser les paramètres (N, P^*) à leurs estimés $(\hat{N}, \hat{P^*})$;
- Simuler L ensembles de statistiques exhaustives (n_j, n), puis estimer les paramètres N et P^* ;

— Calculer les variances asymptotiques en utilisant un boostrap paramétrique. La variance de l'estimateur de N est égale à la variance boostrap moins $\hat{N}$.

3.4 Modélisation M_b

On considère dans cette section que le processus de détection des captures des unités suit un modèle comportemental M_b. L'algorithme récursif d'estimation des paramètres du modèle et les formules explicites pour les variances asymptotiques des estimateurs sont proposés dans les paragraphes suivants.

3.4.1 Algorithme d'estimation des paramètres

Considérons le modèle log-linéaire comportemental où P est la probabilité de la première capture et c est la probabilité de recapture moins un. On définit pour l'historique de capture ω, $X_{1,\omega}$ représente le temps de la première capture moins 1 et $X_{2,\omega}$ le nombre de captures à la suite de la première capture. Le modèle des fréquences prédites des historiques μ_ω s'écrit :

$$\log \mu_\omega = \gamma + X_{1,\omega}\beta_1 + X_{2,\omega}\beta_2, \tag{3.30}$$

avec $\gamma = \log\{NP(1-c)^{l-1}\}$, $\beta_1 = \log\{(1-P)/(1-c)\}$, $\beta_2 = \log\{c/(1-c)\}$. On souhaite déterminer l'estimation de la taille de la population

$$N = e^{\gamma}\frac{(1+e^{\beta_2})^{\ell}}{1+e^{\beta_2}-e^{\beta_1}}$$

et l'estimateur de la probabilité d'au moins une capture P^*. Les statistiques exhaustives de N et P^* sont données par n et C_1, où $C_1 = \sum_{\omega\in\Omega} n_\omega X_{1,\omega} = \sum_{i=1}^{n} X_{1,i}$; pour chaque individu i, $X_{1,i}$ représente le temps de la première capture moins 1. Premièrement, on calcule l'espérance de la variable C_1 comme suit :

$$\begin{aligned}
\mathbb{E}(C_1) &= \mathbb{E}\left(\sum_{i=1}^{n} X_{1,i}\right) \\
&= \mathbb{E}\left\{\mathbb{E}\left(\sum_{i=1}^{n} X_{1,i}|n\right)\right\} \\
&= \mathbb{E}\left\{\sum_{i=1}^{n}\mathbb{E}\left(X_{1,i}|n\right)\right\} \\
&= \mathbb{E}\left(\sum_{j=0}^{\ell-1}\sum_{i=1}^{n} jP(X_{1,i}=j|n)\right).
\end{aligned}$$

Après plusieurs développements, on obtient :

$$\begin{aligned}
\mathbb{E}(C_1) &= \sum_{j=0}^{\ell-1} jNP(1-P)^j \\
&= NP(1-P) \times \frac{d}{dP}\left\{\frac{1-(1-P)^\ell}{P}\right\} \\
&= NP(1-P) \times \frac{1-(1-P)^\ell - \ell P(1-P)^{\ell-1}}{P^2} \\
&= \frac{N(1-P)(P^* - p_1)}{P} \\
&= \frac{Np_2(1-P)}{P}.
\end{aligned}$$

Donc l'espérance de C_1 est :

$$\mathbb{E}(C_1) = \frac{Np_2(1-P)}{P}. \tag{3.31}$$

En résolvant l'équation de la fonction score égale zéro, on obtient :

$$C_1 = \frac{Np_2(1-P)}{P}. \tag{3.32}$$

De plus, l'équation qui entraine la statistique n est $n = N\{1-(1-P)^\ell\}$. Donc obtient

$$N = \frac{n}{1-(1-P)^\ell} \tag{3.33}$$

En remplçant N par son expression dans l'équation (3.32), on obtient l'équation d'estimation pour P, $f_P^b(P) = 0$, où

$$f_P^b(P) = \frac{np_2(1-P)}{P\{1-(1-P)^\ell\}} - C_1. \tag{3.34}$$

L'algorithme d'estimation du modèle M_b se résume en deux étapes :

- Resoudre l'équation d'estmation pour P (3.34) égale à 0, qui permet d'obtenir $\hat{N}$, l'estimateur de N ;
- Remplacer N par $\hat{N}$ dans l'équation (3.34) pour obtenir $\hat{P}$, l'estimateur de P. Donc l'estimateur de N est $\hat{N} = n/1-(1-\hat{P})^\ell$.

3.4.2 Estimation des variances asymptotiques

Pour le modèle M_b, les formules explicites des variances asymptotiques de $\hat{N}$ et $\hat{P^*}$, obtenues à partir des équations (3.31) et (3.32) seront proposées. Considérons ici que le rapport entre une statistique et son espérance converge en probabilité vers 1 lorsque N tend vers l'infini : $C_1 \approx \frac{Np_2(1-P)}{P}$, $n \approx NP^*$.

L'équation d'estimation (3.34) est fonction des statistiques exhaustives C_1 et n. La variance asymptotique de $\hat{N}$ est obtenue par linéarisation. Puisque

$$f_P^b(\hat{P}) - f_P^b(P) \approx (\hat{P} - P)\frac{df_P^b(P)}{dP}, \tag{3.35}$$

il suit que

$$\text{Var}(\hat{P}) = \left[\{\text{Var} f_P^b(P)\}\right] / A^2, \tag{3.36}$$

où Var$f_P^b(P)$ est une approximation de la variance asymptotique de $f_P^b(P)$, vue comme une fonction des statistiques C_1 et n, A est la limite en probabilité de la dérivée de $f_P^b(P)$ par rapport à P. À partir des dérivations, on obtient l'expression de A. Soit $p_2^{'} = \frac{(\ell-1)p_1}{(1-P)}$. On a :

$$\begin{aligned}
\frac{df_P^b(P)}{dP} &= \frac{[Np_2^{'}(1-P) - Np_2]P - Np_2(1-P)}{P^2} \\
&= \frac{Np_2^{'}(1-P)P - Np_2}{P^2} \\
&= \frac{Np_1P(\ell-1) - Np_2}{P^2}.
\end{aligned}$$

Donc,

$$A = \frac{Np_1P(\ell-1) - Np_2}{P^2}.$$

De plus, on a :

$$\text{Var}\{f_P^b(P)\} = \nabla f_P^b \Phi^b (\nabla f_P^b)^T, \tag{3.37}$$

où Φ^b est une matrice de variance-covariance de C_1 et n de dimension 2×2, et ∇f_P^b est la limite du vecteur des dérivées partielles de $f_P^b(P)$ par rapport à C_1 et n. Aussi, n est une variable aléatoire suivant une loi de Poisson, sa variance est égale à son espérance : $\text{Var}_P(n) = \mathbb{E}_P(n) = NP^*$. La variance de C_1 est calculée

de la façon suivante :

$$
\begin{aligned}
\mathsf{Var}(C_1) &= \mathsf{Var}(\sum_{i=1}^{n} \boldsymbol{X}_{1,i}) \\
&= \mathbb{E}\left\{\sum_{i=1}^{n} \mathsf{Var}(X_{1,i}|n)\right\} + \mathsf{Var}\left\{\sum_{i=1}^{n} \mathbb{E}(X_{1,i}|n)\right\} \\
&= \mathbb{E}\left\{\sum_{i=1}^{n} \left((\mathbb{E}(X_{1,i}^2|n) - \mathbb{E}^2(X_{1,i}|n)\right)\right\} + \mathsf{Var}\left\{\sum_{i=0}^{n}\sum_{j=0}^{\ell-1} \frac{jNP(1-P)^j}{P^*}\right\} \\
&= \mathbb{E}\left\{\sum_{i=1}^{n}\left(\sum_{j=0}^{\ell-1} \frac{j^2NP(1-P)^j}{P^*}\right)\right\} - \mathbb{E}\left\{\sum_{i=1}^{n}\left(\sum_{j=0}^{\ell-1} \frac{jNP(1-P)^j}{P^*}\right)^2\right\} \\
&+ \mathsf{Var}\left\{n\sum_{j=0}^{\ell-1} \frac{jNP(1-P)^j}{P^*}\right\}; \\
&= \mathbb{E}\left\{n\left(\sum_{j=0}^{\ell-1} \frac{j^2NP(1-P)^j}{P^*}\right)\right\} - \mathbb{E}\left\{n\left(\sum_{j=0}^{\ell-1} \frac{jNP(1-P)^j)^2}{P^*}\right)\right\} \\
&+ \left(\sum_{j=0}^{\ell-1} \frac{jNP(1-P)^j}{P^*}\right)^2 \mathsf{Var}(n) \\
&= NP^*\sum_{j=0}^{\ell-1} \frac{j^2NP(1-P)^j}{P^*} - NP*\left(\sum_{j=0}^{\ell-1} \frac{jNP(1-P)^j}{P^*}\right)^2 \\
&+ NP*\left(\sum_{j=0}^{\ell-1} \frac{jNP(1-P)^j}{P^*}\right)^2 \\
&= NP^*\sum_{j=0}^{\ell-1} \frac{j^2NP(1-P)^j}{P^*} \\
&= N\sum_{j=0}^{\ell-1} j^2P(1-P)^j \\
&= NP(1-P)^2\sum_{j=0}^{\ell-1} j^2P(1-P)^{j-2} + \frac{Np_2(1-P)}{P} \\
&= NP(1-P)^2\frac{d}{dP}\left\{\left(\frac{1-(1-P)^\ell - \ell P(1-P)^{\ell-1}}{P^2}\right)\right\} + \frac{Np_2(1-P)}{P} \\
&= NP(1-P)^2 \times \frac{p_1P(\ell-1) - p_2(1-P)}{P^3(1-P)} + \frac{Np_2(1-P)}{P} \\
&= N(1-P)^2 \times \frac{p_1P(\ell-1) - p_2(1-P)}{P^2(1-P)} + \frac{Np_2(1-P)}{P} \\
&= \frac{N(1-P)}{P^2} \times [p_1P(\ell-1) - p_2(1-P)] + \frac{Np_2(1-P)}{P}.
\end{aligned}
$$

Ainsi,

$$\mathsf{Var}(C_1) = \frac{N(1-P)}{P^2} \times [p_1 P(\ell - 1) - p_2(1-P)] + \frac{N p_2 (1-P)}{P}.$$

Pour calculer la covariance entre C_1 et n, on définit $C_1 = \sum_{i=1}^{n} X_{1,i}$, où la variable $X_{1,i}$ est le temps de la première capture moins un pour l'individu i, et suit une loi géométrique de paramètre P. On a donc :

$$\begin{aligned}
\mathsf{Cov}_P(C_1, n) &= \mathbb{E}(C_1 \times n) - \mathbb{E}(C_1) \times \mathbb{E}(n) \\
&= \mathbb{E}\left\{\mathbb{E}(C_1 \times n | n)\right\} - \mathbb{E}(C-1) \times \mathbb{E}(n) \\
&= \mathbb{E}\left\{ n^2 \sum_{j=0}^{\ell-1} \frac{jP(1-P)^j}{P^*} \right\} - NP^* \times \frac{N p_2 (1-P)}{P} \\
&= \left\{ \sum_{j=0}^{\ell-1} \frac{jP(1-P)^j}{P^*} \right\} \mathbb{E}(n^2) - NP^* \times \frac{N p_2 (1-P)}{P} \\
&= \left((NP^*)^2 + NP^*\right) \times \sum_{j=0}^{\ell-1} \frac{jP(1-P)^j}{P^*} - NP^* \times \frac{N p_2 (1-P)}{P} \\
&= \left(N^{2*} + N\right) \times \frac{p_2(1-P)}{P} - NP^* \times \frac{N p_2 (1-P)}{P} \\
&= \frac{N p_2 (1-P)}{P} \\
&= \frac{N p_2 (1-P)}{P}.
\end{aligned}$$

Donc

$$\mathsf{Cov}_P(C_1, n) = \frac{N p_2 (1-P)}{P}.$$

Ainsi, on a

$$\Phi^b = N \begin{bmatrix} \frac{1-P}{P^2}\left[(p_1(1-P)(\ell-1)\right] + \frac{p_2(1-P)}{P} & \frac{p_2(1-P)}{P} \\ \frac{p_2(1-P)}{P} & P^*. \end{bmatrix}.$$

Par ailleurs, on montre que

$$\nabla f_P^b = p \lim_{N->\infty} \left(\frac{\partial f_P^b(C_1, n)}{\partial n}, \frac{\partial f_P^b(C_1, n)}{\partial c_1} \right) = \left(\frac{p_2(1-P)}{PP*}, -1 \right).$$

En remplaçant ces quantités dans les équations (3.36) et (3.37), on obtient la variance asymptotique de $\hat{P}$:

$$\begin{aligned}\mathsf{Var}\{\nabla f_P^b(P)\} &= \frac{Np_2^2(1-P)^3}{P^4P^{*2}}[p_1P(\ell-1)-2p_2(1-P)] \\ &+ \frac{Np_2^3(1-P)^3}{P^3P^{*2}} - \frac{2Np_2^2(1-P)^2}{P^2P^*} + NP^*.\end{aligned}$$

Il vient donc que la variance asymptotique de $\hat{P}$ est :

$$\mathsf{Var}(\hat{P}) = \frac{p_2^2(1-P)^3}{NP^{*2}(p_1P(\ell-1)-p_2)^2} + \frac{p_2^3(1-P)^3P}{NP^{*2}(p_1P(\ell-1)-p_2)^2} - \frac{[2p_2(1-P)^2+P^{*2}P^2]P^2}{NP^*(p_1P(\ell-1)-p_2)^2}.$$

Considérons maintenant $\hat{N} = n/\hat{P}^*$. On a par linéarisation

$$\mathsf{Var}(\hat{N}) = \nabla\hat{N}\Delta^b(\nabla\hat{N})^T - N, \tag{3.38}$$

où Δ^b est est une matrice 2×2 de variance covariance de $\hat{P}$ et n, $\nabla\hat{N}$ est la limite du vecteur des dérivées partielles de $\hat{N}$ par rapport à $\hat{P}$ et n. On obtient ainsi :

$$\nabla\hat{N} = p\lim_{N->\infty}\left(\frac{\partial\hat{N}(\hat{P},n)}{\partial\hat{P}}, \frac{\partial\hat{N}(\hat{P},n)}{\partial n}\right) = \frac{1}{P^*}\left(\frac{-N\ell(1-P)^{\ell-1}}{P^*}, 1\right).$$

Aussi, la linéarisation de $\hat{N}$, permet d'obtenir $\mathsf{Cov}(n,\hat{P})$. En effet,

$$f_P^b(\hat{P}) - f_P^b(P) \approx (\hat{P}-P)\frac{df_P^b(P)}{dP}.$$

Alors, nous avons

$$(\hat{P}-P) \approx f_P^b(P)\left(-\frac{df_P^b(P)}{dP}\right)^{-1}.$$

Il vient que

$$\begin{aligned}\mathsf{Cov}(n,\hat{P}) &= \mathsf{Cov}\left\{n, f_P^b(P)\left(-\frac{df_P^b(P)}{dP}\right)^{-1}\right\} \\ &= \mathsf{Cov}\left\{n, f_P^b(P)(-A^{-1})\right\} \\ &= \mathsf{Cov}\left\{n, \left[f_P^b(P) - f_P^b(\mathbb{E}(C_1),\mathbb{E}(n))\right](-A^{-1})\right\}.\end{aligned}$$

Or

$$f_P^b(P) - f_P^b(\mathbb{E}(C_1),\mathbb{E}(n)) \approx (C_1-\mathbb{E}(C_1))\frac{df_P^b(P)}{dC_1} + (n-\mathbb{E}(n))\frac{df_P^b(P)}{dn}.$$

Donc, nous avons

$$\begin{aligned}\mathsf{Cov}(n,\hat{P}) &= -A^{-1}\left\{\mathsf{Cov}_p(n,C_1)\frac{df_P^b(P)}{dC_1}+\mathsf{Var}_p(n)\frac{df_P^b(P)}{dn}\right\}\\ &= -A^{-1}\left\{-\frac{Np_2(1-P)}{P}+\frac{p_2(1-P)}{PP^*}\times NP^*\right\}\\ &= 0.\end{aligned}$$

Ainsi, on peut écrire

$$\mathsf{Cov}(n,\hat{P})=0.$$

Donc, on a

$$\Delta^b=\begin{bmatrix}\mathsf{Var(P)} & 0\\ 0 & NP^*\end{bmatrix}.$$

En remplaçant ces quantités dans (3.38), on obtient la variance asymptotique de $\hat{N}$, donnée par

$$\begin{aligned}\mathsf{Var}(\hat{N}) &= \frac{Np_1^2p_2^2(1-P)^3}{P^{*4}P^2(p_1P(\ell-1)-p_2)^2}+\frac{Np_1^2p_2^3(1-P)^3}{P^{*4}P(p_1P(\ell-1)-p_2)^2}\\ &- \frac{Np_1^2}{P^{*3}}\times\frac{2p_2(1-P)^2+P^{*2}P^2}{(p_1P(\ell-1)-p_2)^2}+\frac{N-NP^*}{P^*}.\end{aligned}$$

Une méthode alternative pour le calcul des variances asymptotiques des estimateurs $\hat{N}$ et $\hat{P}^*$ est un boostrap paramétrique. On souhaite obtenir des répétitions bootstrap pour le vecteur des statistiques exhaustives n et C_1, qui se fait en deux étapes :

Etape 1 :

- Générer une réalisation de la taille de la population selon la distribution de Poisson : $\tilde{N}\sim Poisson(N)$; la valeur de N sera prise comme égale à son estimateur $\hat{N}$;
- Générer la statistique n selon une Binomiale $(\tilde{N},P^*)$;
- Conditionnellement à n, on génère pour chaque individu i, $i=1,2,\ldots,n$ détecté au moins une fois au cours de l'expérience, $\boldsymbol{X}_{1,i}$ selon une loi Géométrique tronquée sur $\{1,2,\ldots,\ell-1\}$ et on calcule $C_1=\sum_{i=1}^n \boldsymbol{X}_{1,i}$.

Etape 2 :

— Egaliser les paramètres (N, P^*) à leurs estimés $(\hat{N}, \hat{P^*})$;

— Simuler L ensembles de statistiques exhaustives (n, C_1), puis estimer les paramètres N et P^* ;

— Calculer les variances asymptotiques en utilisant un boostrap paramétrique. La variance de l'estimateur de N est égale à la variance boostrap moins $\hat{N}$.

Les deux prochains chapitres seront consacrés à l'évaluation de la performance de la méthode d'estimation et l'application des algorithmes d'estimation des paramètres présentés ci-dessus.

CHAPITRE 4

ETUDE DE SIMUALTIONS

Ce chapitre présente le processus de simulation ainsi que les résultats de la simulation qui permettent d'évaluer pour les modèles homogènes M_0 et M_t le biais et la précision des estimations obtenues à partir de la méthodologie proposée.

4.1 Données et métriques d'évaluations

La simulation vise à évaluer, pour les modèles homogènes M_0 et M_t, le biais et la précision des estimations obtenues à partir de la méthodologie présentée au chapitre 3 (trois). Les données de simulations sont générées à partir des paramètres suivants : la probabilité d'au moins une capture P^*, la taille de la population N et le nombre d'occasions de capture ℓ. Ces paramètres sont fixés à $P^* = (0.2, 0.3)$, $N = (1000, 5000)$ et $\ell = (5, 10, 20)$. On fixe $B = 1000$ le nombre de répétitions pour chaque combinaison des paramètres.

La qualité des estimations est évaluée à l'aide du biais relatif, de l'erreur quadratique moyenne relative, du biais relatif de la variance, du taux de couverture de l'intervalle de confiance et de la longueur de l'intervalle de confiance, définis comme suit.

- Le biais relatif (BR) : il évalue l'écart entre l'estimation d'un paramètre et la vraie valeur de ce paramètre. Il est calculé comme suit :

$$RB(\hat{N}) = \frac{\sum_{i=1}^{B}(\hat{N}_i - N)/N}{B},$$

  où N représente la vraie valeur de la taille de population ; $\hat{N}_i$, l'estimation de la taille de population pour chaque répétition i et B est le nombre total de répétitions.
- L'erreur quadratique moyenne relative(EQMR) : elle évalue l'erreur moyenne relative entre l'estimation d'un paramètre et la vraie valeur de ce paramètre. Elle est caculée en divisant l'erreur quadratique moyenne (EQM) par la vraie valeur du paramètre.

$$EQM(\hat{N}) = \left(\frac{\sum_{i=1}^{B}(\hat{N}_i - N)^2}{B}\right)^{1/2},$$

$$EQMR(\hat{N}) = \frac{EQM(\hat{N})}{N}.$$

— Le biais relatif de la variance : il évalue l'écart entre l'estimation de la variance d'un échantillon et la vraie valeur de la variance de la population. Il se calcule comme suit :

$$BRV(\hat{N}) = \frac{\sum_{i=1}^{B}(Var(\hat{N}_i) - Var(N))/B}{Var(N)}$$

— Le taux de couverture de l'intervalle de confiance (95%Cov) : il évalue la précision de l'intervalle de confiance. Il représente la proportion de fois où l'intervalle de confiance contient la vraie valeur du paramètre, lorsque l'estimation de ce paramètre est répétée plusieurs fois sur des échantillons différents.

— La longueur relative de l'intervalle de confiance (RLCI) : elle repésente l'étendue de l'intervalle de confiance d'un paramètre. Elle est donnée par la différence relative entre la borne supérieure et la borne inférieure de l'intervalle de confiance. Elle se calcule de la manière suivante :

$$RLCI(\hat{N}) = \frac{N_{sup} - N_{inf}}{N},$$

où N_{sup} est la borne supérieure et N_{inf} la borne inférieure de l'intervalle de confiance.

4.2 Résultats des simulations

Dans cette section, l'évaluation du biais et de la precision des estimations se fera pour le cas des modèles M_0 et M_t.

4.2.1 Résultats de la simulation dans le cas du modèle M_0.

Le tableau 4.1 présente les résultats pour l'estimation de la taille de la population à l'aide du modèle M_0.

Lorsque la taille de la population est $N = 1000$ et la probabilité d'au moins une capture est $P^* = 0.2$, le biais relatif, l'erreur quadratique moyenne relative, le taux de couverture de l'intervalle de confiance et la longueur de l'intervalle de confiance diminuent lorsque le nombre d'occasions de capture augmente ; tandis que le biais relatif de la variance de l'estimation de la taille de population augmente avec le nombre d'occasions de capture.

En effet, lorsque le nombre d'occasions de capture vaut respectivement $5, 10, 20$, on a un biais relatif de $0.05, 0.04, 0.03$, respectivement. Dans chacun des cas, bien que le biais relatif soit positif, il est faible. Ainsi, l'estimation de la taille de population se rapproche de la vraie valeur de la taille de la population à mesure que le nombre d'occasions de capture augmente. L'erreur quadratique moyenne relative est respectivement égale à $0.27, 0.25, 0.23$ lorsque le nombre d'occasions de capture vaut respectivement $5, 10, 20$. L'erreur quadratique est relativement faible. Ainsi, elle semble indiquer une estimation précise, avec une faible erreur par rapport à la variabilité réelle. Quant au taux de couverture de l'intervalle de confiance à un niveau de confiance de 95%, lorsque le nombre d'occasions de capture vaut respectivement $5, 10, 20$, il vaut respectivement $0.95, 0.94, 0.94$. Le taux de couverture est élevé et proche du niveau de confiance. Cela pourrait indiquer une bonne précision de l'intervalle de confiance et une estimation fiable. Aussi, la longueur de l'intervalle de confiance est respectivement $0.94, 0.88, 0.83$ pour respectivement $5, 10, 20$ occasions de capture. Ainsi, ces résultats confirment une précision de l'estimation de la taille de population. Par ailleurs, le biais relatif de la variance de l'estimation de la taille de la population vaut respectivement $-0.07, -0.07, 0.01$ lorsque le nombre d'occasions de capture vaut $5, 10, 20$. Cela signifie que le biais relatif se rapproche de zéro et indiquerait une bonne précision de l'estimateur et une absence de biais.

Lorsque la probabilité est $P^* = 0.3$, le biais relatif, le taux de couverture de l'intervalle de confiance et la longueur de l'intervalle de confiance augmentent avec le nombre d'occasions de capture, tandis que le biais relatif de la variance des estimations et la longueur de l'intervalle de confiance diminuent lorsque le nombre d'occasions de capture augmente. On constate que le biais relatif est positif et faible ; ce qui indiquerait que l'estimation de la taille de population se rapproche de la vraie la valeur de la taille de la population. Aussi, le taux de couverture de l'intervalle de confiance à un niveau de confiance de 95%, est élevé et proche du niveau nominal. Cela indique une bonne précision de l'intervalle de confiance et une estimation fiable. Par ailleurs, la longueur de l'intervalle de confiance est faible (valeur proche de 0.50). Cela veut dire que l'intervalle de confiance est étroit et indique une bonne précision de l'estimation de la taille de population. Le biais relatif de la variance de l'estimation de la taille de la population est proche de zéro, ce qui indique une absence de biais pour l'estimation de la précision associée aux estimateurs.

Ainsi, nous constatons que pour une taille de population $N = 1000$, la qualité de l'estimation de la taille de la population augmente avec la probabilité de capture.

Lorsque la taille de la population est fixée à $N = 5000$, et la probabilité de capture $P^* = 0.2$, Le biais

relatif de la variance et la longueur de l'intervalle de confiance diminuent lorsque le nombre d'occasions de capture augmente. Tandis que le biais relatif de l'estimation de la taille de la population, l'erreur quadratique moyenne relative et la couverture de l'intervalle de confiance avec un niveau de confiance de 95% ne varient pas lorsque le nombre d'occasions de capture augmente. En effet, le biais relatif de la variance est positif lorsque le nombre d'occasions de capture est $\ell = 5\ (0.05)$ et négatif pour lorsque le nombre d'occasions de capture vaut 10 et 20, soit respectivement -0.03 et -0.04. Toutefois, ce biais est proche de zéro. Ceci indiquerait une bonne précision de l'estimateur et une absence de biais. Aussi la longueur de l'intervalle de confiance vaut respectivement $0.38, 0.36,$ et 0.35 lorsque le nombre d'occasions de capture est égal à $5, 10, 20$ respectivement. On a ainsi un intervalle de confiance étroit avec un taux de couverture de 0.96 pour chaque cas. Cela témoigne de la précision de l'estimation de la taille de la population.

Lorsque la probabilité d'au moins une capture P^* vaut 0.3, de façon générale, les indicateurs d'appréciation de l'estimation de la taille de la population, tels que le biais relatif, l'erreur quadratique moyenne relative, le biais relatif de la variance et la longueur de l'intervalle de confiance sont faibles et le taux de couverture de l'intervalle de confiance est proche du niveau de confiance, soit 95%. En effet, lorsque le nombre d'occasions de capture est égale à 5 et 10 respectivement, le biais relatif vaut 0.005 et 0.004 lorsque le nombre d'occasions de capture vaut 20 ; cela indique que l'estimation de la taille de la population est précise. La longueur de l'intervalle de confiance est respectivement de $0.23, 0.22$ et 0.21 lorsque le nombre d'occasions de capture vaut $5, 10$ et 20 ; ceci indique que l'intervalle de confiance est étroit. Le taux de couverture de l'intervalle de confiance étant proche du niveau de confiance de 95%, permet d'affirmer que l'estimation de la taille de la population est précise.

Par ailleurs, lorsqu'on observe les indicateurs de performance de l'estimation de la taille de la population, on constate que le biais relatif, l'erreur quadratique moyenne relative et la longueur de l'intervalle de confiance lorsque la taille de la population est égale 5000, sont plus faibles que lorsqu'elle est égale à 1000. Ainsi, cela signifie que la précision de l'estimation de la taille de la population devient meilleure à mesure que la taille de la population augmente.

Table 4.1 Résultats de la simulation pour l'estimation de N dans le cadre du modèle du M_0

N	p^*	ℓ	$BR(\hat{N})$	$EQMR$	$BR(v(\hat{N}))$	95% Cov.	$RLCI(\hat{N})$
1000	0.2	5	0.05	0.27	-0.07	0.95	0.94
		10	0.04	0.25	-0.07	0.94	0.88
		20	0.03	0.23	0.01	0.94	0.83
	0.3	5	0.01	0.13	0.09	0.95	0.52
		10	0.02	0.13	0.04	0.96	0.51
		20	0.02	0.13	0.01	0.96	0.49
5000	0.2	5	0.01	0.1	0.05	0.96	0.38
		10	0.01	0.1	-0.03	0.96	0.36
		20	0.01	0.1	-0.04	0.96	0.35
	0.3	5	0.005	0.06	0.05	0.96	0.23
		10	0.005	0.05	0.06	0.95	0.22
		20	0.004	0.05	0.14	0.95	0.21

4.2.2 Résultats de la simulation dans le cas du modèle M_t.

Les résultats de l'étude de simulations pour le modèles M_t montrent de façon générale, un biais important lorsque la probabilité de capture vaut $P^* = 0.2$. Ce resultat peut s'expliquer par le fait que les p_j du vecteur de probabilités de capture sont inférieures à 0.1 (Otis *et al.* (1978)).

Lorsque la probabilité est $P^* = 0.3$, le biais relatif, l'erreur quadratique moyenne relative et la longueur de l'intervalle de confiance diminuent avec le nombre d'occasions de capture, tandis que le biais relatif de la variance des estimations et le taux de couverture de l'intervalle de confiance augmentent lorsque le nombre d'occasions de capture augmente. On constate que le biais relatif est positif et faible ; ce qui indique que l'estimation de la taille de population se rapproche de la vraie la valeur de la taille de la population.

De plus, lorsqu'on observe ces indicateurs de performance de l'estimation de la taille de la population, on constate que le biais relatif, l'erreur quadratique moyenne relative et la longueur de l'intervalle de confiance lorsque la taille de la population est égale 5000, sont plus faibles que lorsqu'elle est égale à 1000. Ainsi, cela signifie que l'estimation de la taille de la population est meilleure lorsque la taille de la population augmente. Les résultats de l'étude de simulations pour le cas du modèle M_t sont présentés dans le tableau 4.2.

Table 4.2 Résultats de la simulation pour l'estimation de N dans le cadre du modèle du M_t

N	p^*	ℓ	$BR(\hat{N})$	$EQMR$	$BR(v(\hat{N}))$	95% Cov.	$RLCI(\hat{N})$
1000	0.2	5	0.28	1.20	-0.404	0.71	1.79
		10	0.34	1.18	-0.51	0.76	1.65
		20	0.42	1.38	-0.51	0.74	1.81
	0.3	5	0.014	0.13	-0.67	0.78	0.28
		10	0.011	0.11	-0.66	0.78	0.25
		20	0.005	0.10	-0.64	0.79	0.24
5000	0.2	5	0.027	0.30	-0.81	0.56	0.45
		10	0.034	0.30	-0.84	0.54	0.40
		20	0.045	0.27	-0.84	0.56	0.39
	0.3	5	0.003	0.05	-0.66	0.82	0.12
		10	0.002	0.04	-0.60	0.87	0.11
		20	0.002	0.04	-0.60	0.89	0.10

CHAPITRE 5
ÉTUDE DE CAS : ACTIVATIONS D'APPLICATIONS MOBILES

Dans ce chapitre, nous allons nous intéresser tout d'abord à la description des données utilisées dans le cadre de cette étude. Nous finirons ce chapitre en ajustant les modèles M_0, M_t et M_b sur ce jeu données afin d'estimer la taille de la clientèle ayant visité les onze (11) concessionnaires automobiles au cours des 533 jours d'expérience.

5.1 Description des activations quotidiennes d'applications chez les concessionnaires.

L'échantillon de données sur lequel se focalise notre analyse est recueilli par Ninth Decimal, une plateforme de marketing basée en Californie, concernant les activations quotidiennes d'applications chez onze (11) concessionnaires automobiles d'une grande marque dans une zone métropolitaine américaine. Un historique de capture désigne une liste de jours où un appareil a été activé.

Le tableau 5.1 présente la répartition du nombre de clients ayant visité les concessionnaires selon la fréquence d'activation des applications. La répartition du nombre de clients au cours des 533 jours de collecte permet de constater qu'un total de 9332 personnes ont activé une application au moins une fois. Parmi ces individus, $77,05\%$ ont activé des applications une seule fois, 10.90% ont activé les applications deux fois, 3.70% ont activé les applications trois fois et seulement 5.14% ont activé les applications plus de six fois. Cela signifie qu'une fois un individu enregistré, la fréquence des visites diminue progressivement au courant des semaines qui suivent le premier enregistrement.

Table 5.1 Répartition du nombre de clients selon la fréquence d'activation d'applications.

Fréquence d'activation	Nombre d'activations	Pourcentage (%)
1	7190	77.05
2	1017	10.90
3	345	3.70
4	183	1.96
5	117	1.25
6 et plus	480	5.14
Total	9332	100.00

Lorsqu'on s'intéresse au nombre d'activations d'applications par semaine, on constate qu'en moyenne 184 personnes visitent les concessionnaires automobiles chaque semaine. Le plus grand nombre d'enregistrements est observé à la 70ième semaine avec 509 enregistrements, tandis que le plus faible nombre d'enregistrements est de 50, observé à la 21ième semaine. Le graphique suivant nous permet de visualiser l'évolution du nombre d'activation d'applications au cours des semaines. On observe que le nombre d'activations varie d'une semaine à l'autre. Donc, il semble pertinent d'avoir des probabilités de détection qui varient d'un jour à l'autre.

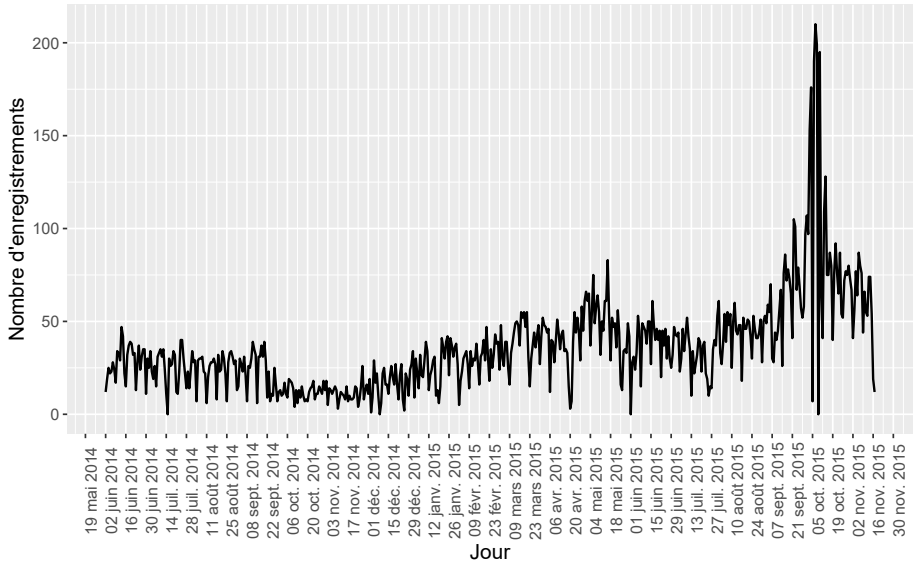

Figure 5.1 Evolution du nombre d'activations au cours des semaines.

Lorsqu'on s'interesse au nombre d'activations d'applications en fonction des jours de la semaine, on constate que le nombre total d'activations au cours de l'expérience est $C = 19075$. En effet, le nombre d'activation d'applications est plus grand les vendredis avec 3272 appareils activés, suivi des mardis et mercredis avec respectivement 3043 et 2923 appareils activés ; tandis que le nombre d'activations d'applications est plus faible les samedis et dimanches avec respectivement 2800 et 1384 activations. Ces resultats montrent que les clients visitent plus les concessionnaires automobiles les jours ouvrables qu'en fin de semaine. Le nombre d'activations d'applications en fonction des jours de la semaine est présenté sur le graphique 5.1.

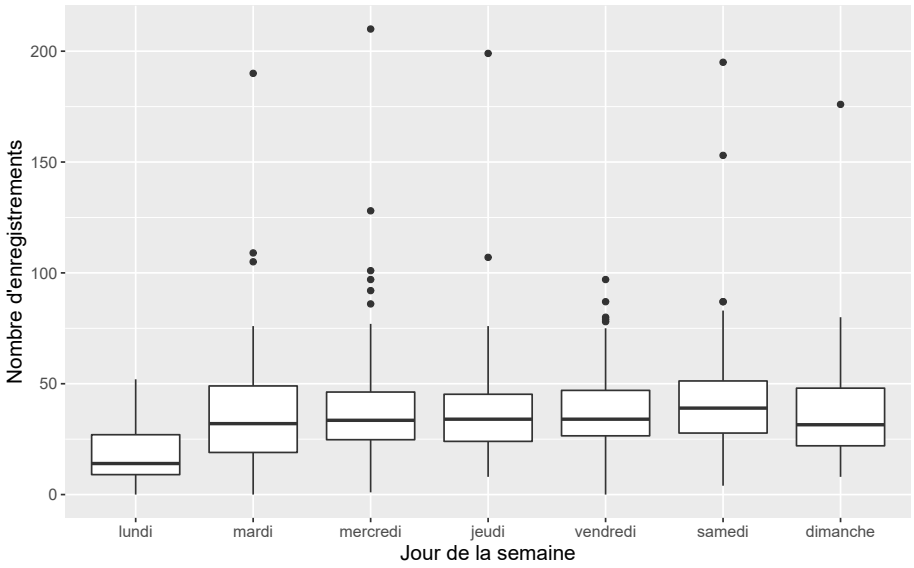

Figure 5.2 répartition du nombre de clients selon les jours de la semaine.

Afin de mieux appréhender l'estimation du nombre de clients chez les concessionnaire automobiles, nous allons ajuster les modèles M_0, M_t et M_b au jeu de données.

5.2 Estimations du nombre d'activations d'applications par les modèles M_0, M_t et M_b.

Le tableau 5.2 présente les résultats d'estimation des paramètres des modèles M_0, M_t et M_b. Pour le modèle M_0, le nombre de clients chez les onze (11) concessionnaires au cours des 533 jours est estimé à $\hat{N} = 11532$, avec un intervalle de confiance de 95% de $[11401; 11663]$. La probabilité qu'une application soit activée sur un téléphone intelligent au cours de l'expérience est $\hat{P} = 0.003$. Donc, sur 1000 personnes arrivées chez les concessionnaires, 3 personnes en moyenne ont activé une application sur leurs téléphones intelligents.

Par ailleurs, l'analyse du modèle M_0 montre que la probabilité pour qu'un appareil soit activé au moins une fois pendant les 533 jours est $\hat{P}^* = 0.81$, avec un intervalle de confiance de 95% de $[0.803; 0.815]$. Cela signifie que sur 100 personnes ayant visitées les concessionnaires dans la zone métropolitaine américaine, environ 81 personnes ont activé une aplication sur leurs smartphones. Aussi, il est à noter que 32 personnes sur 100 qui se sont rendues chez les concessionnaires ont activé une application exactement une fois, soit une probabilité $\hat{p}_1$. Tandis que 49 sur 100 personnes ont activé une application chez les concessionnaires

deux fois, soit une probabilité $\hat{p}_2 = 0.49$.

L'évolution du nombre de d'activations d'applications au cours des semaines montrait que le nombre d'activations est différent d'une semaine à l'autre. Donc, les probabilités de détection pourraient vraisemblablement varier d'un jour à l'autre. Il semble pertinent d'ajuster le modèle temporel M_t à ce jeu de données.

Le modèle M_t montre que le nombre de clients chez les onze concessionnaires au cours des 533 jours est estimé à $\hat{N} = 11526$, avec un intervalle de confiance au niveau de confiance de 95% de $[11395; 11657]$. Le vecteur de probabilités de détection d'activation d'applications au cours de 533 jours d'expérience montre que les probabilités de détection varient entre 0 et 0.018. Donc, sur 100 personnes arrivées chez les concessionnaires au cours de l'expérience,on a entre 0 et 18 personnes qui ont activé une application sur leurs téléphones.

De plus, l'analyse du modèle temporel M_t montre que la probabilité pour qu'un appareil soit activé au moins une fois pendant les 533 jours est $\hat{P}^* = 0.81$, avec un intervalle de confiance au niveau 95% de $[0.804; 0.815]$. Cela signifie que sur 100 personnes ayant visitées les concessionnaires dans la zone métropolitaine américaine, environ 81 personnes ont activé une application sur leurs téléphones. Aussi, il est à noter que 32 personnes sur 100 qui se sont rendues chez les concessionnaires ont activé une application exactement une fois, soit une probabilité $p_1 = 0.32$. Tandis que 49 sur 100 personnes ont activé une application chez les concessionnaires deux fois, soit une probabilité de $\hat{p}_2 = 0.49$.

La répartition du nombre de clients ayant visités les concessionnaires selon la fréquence d'activation d'applications au cours des 533 montre que les clients ont tendance, après la première visite, à ne plus se rendre chez les concessionnaires. Cela pourrait signifier un changement de comportement des clients au cours de l'expérience. Il semble donc judicieux d'ajuster le modèle comportemental M_b à notre jeu de données.

L'analyse du modèle comportemental M_b montre que le nombre de clients chez les onze concessionnaires au cours des 533 jours est estimé à $\hat{N} = 11001$, avec un intervalle de confiance de 95% de $[10123; 11880]$. La probabilité de détection de la première activation d'applications au cours de 533 jours d'expérience est 0.003. Donc, sur 1000 personnes arrivées chez les concessionnaires au cours de l'expérience, on a 3 personnes qui ont activé une application sur leurs téléphones pour la première fois.

Par ailleurs, une analyse minitieuse du modèle temporel M_b montre que la probabilité pour qu'un appareil

soit activé au moins une fois pendant les 533 jours est $\hat{P}^* = 0.85$, avec un intervalle de confiance au niveau de confiance 95% de $[0.79; 0.91]$. Cela signifie que sur 100 personnes ayant visitées les concessionnaires dans la zone métropolitaine américaine, environ 85 personnes ont activé une aplication. Aussi, il ressort que 29 personnes sur 100 qui se sont rendues chez les concessionnaires ont activé une application exactement une fois, soit une probabilité $p_1 = 0.29$. Tandis que 56 sur 100 personnes ont activé une application deux fois, soit une probabilité $\hat{p}_2 = 0.56$.

Ainsi, l'analyse des modèles M_0, M_t et M_b montre que la probabilité pour qu'un appareil soit activé au moins une fois est plus grande pour modèle comportemental M_b ($\hat{P}^* = 0.85$). Tandis que pour les modèles M_0 et M_t, sa valeur est de $\hat{P}^* = 0.81$. L'erreur-type de l'estimation de la probabilité pour qu'un appareil soit activé au moins une fois est dix fois plus grande pour le modèle M_b par rapport aux modèles M_0 et M_t. De plus l'erreur type de l'estimation du nombre de clients ayant visités les concessionnaires est plus grande pour modèle comportemental M_b que pour les modèles M_0 et M_t. De plus, les analyses des modèles M_0 et M_t montrent des résutats assez similaires.

Il est important de noter qu'un travail de sélection de modèles pour les données observées n'a pas été effectué pour deux raisons. La première concerne la fonction de vraisemblance qui n'est pas utilisée dans le contexte d'une estimation par la méthode des moments, et du fait du défi de la dimensionnalité pour les programmes statistiques standard d'optimisation numérique. Par ailleurs, les modèles dits hétérogènes, qui supposent des variations dans les probabilités de capture individuelles, n'ont pas été considérées dans le cadre de ce mémoire. Il serait donc prudent, dans le contexte de cette analyse, de se limiter seulement à la comparaison des estimations des paramètres pour les trois modèles M_0, M_t et M_b.

Table 5.2 Résultats de l'estimation des paramètres des modèles M_0, M_t, M_b

Paramètres	M_0	M_t	M_b
$\hat{N}$	11532	11526	11001
$IC(\hat{N})$	[11401 ; 11663]	[11395 ; 11657]	[10123 ; 11880]
P^*	0.81	0.81	0.85
$IC(\hat{P}^*)$	[0.803 ; 0.815]	[0.804 ; 0.815]	[0.79 ; 0.91]
$se(\hat{N})$	66.82	66.70	448.17
$se(\hat{P}^*)$	0.003	0.003	0.032

CONCLUSION

Les méthodes de capture-recapture sont des techniques d'échantillonnage utilisées pour l'estimation des paramètres démographiques des populations difficiles à rejoindre. Ces paramètres incluent la taille de la population, la survie, les naissances et l'émigration. Initialement développées et popularisées dans les domaines tels que la biologie et l'écologie, les méthodes de capture-recapture ont trouvé expansion dans les sciences médicales, sociales et plus récemment dans la technologie du téléphone intelligent (Yauck *et al.*, 2019). Pour ce plus récent champ d'application, il ressort un défi méthodologique du fait du nombre élevé d'occasions de capture, rendant les outils numériques existants de maximisation de la vraisemblance inadaptés pour la dimensionnalité des jeux de données.

Ce mémoire a présenté, pour les modèles de population fermée homogènes M_0 (aucune variation dans la probabilité de capture d'une unité), M_t (variation temporelle) et M_b (variation comportementale), des algorithmes d'estimation des paramètres, ainsi que des formules explicites pour les variances des estimateurs, lorsque le nombre d'occasions de capture est grand. L'originalité du travail concerne l'algorithme d'estimation des paramètres pour le modèle comportemental M_b ainsi que la formule explicite pour les variances des estimateurs qui lui sont associés. Exploitant les propriétés intéressantes de la régression de Poisson, notamment la flexibilité de la fonction de score, des équations d'estimation ont été proposées pour la taille de la population N et la probabilité d'au moins une capture p^*. Ensuite, en utilisant des expansions de Taylor d'ordre 1, des formules explicites pour les variances des estimateurs $\hat{N}$ et $\hat{p}^*$ ont été proposées.

Une étude de simulation a permis d'évaluer la performance de la méthodologie proposée à partir de cinq indicateurs de performance : le biais relatif, l'erreur quadratique moyenne relative, le biais relatif de la variance, le taux de couverture de l'intervalle de confiance et la longueur de l'intervalle de confiance. Les résultats des simulations montrent que, dans le cas du modèle M_0, plus la taille de la population et la probabilité de capture augmentent, meilleure est la performance des estimateurs. Dans le cas du modèle temporel M_t, un biais important est noté lorsque les probabilités de capture sont faibles ; ce résultat confirme la remarque faite par Otis *et al.* (1978). Par contre, lorsque la probabilité de capture augmente, le biais diminue et le taux de couverture de l'intervalle de confiance se rapproche du niveau de confiance nominal de 95%.

Les algorithmes d'estimation des paramètres ont été appliquées sur des données de recueillies par Ninth

Decimal, une plateforme de marketing basée en Californie, concernant les activations quotidiennes d'applications chez des concessionnaires automobiles d'une grande marque, dans une zone métropolitaine américaine. De plus les algorithmes d'estimation pour les modèles M_0 et M_t ont été proposés en utilisant une approche basée sur les modèles de populations fermées. Leurs résultats peuvent être retrouvés être retrouvés sur la base de populations ouvertes dans l'article de Yauck *et al.* (2019)

Dans le contexte de la téchnologie mobile, les modèles homogènes pour des populations fermées supposent que les probabilités d'activations d'un téléphone intelligent sont identiques d'un individu à l'autre. Ainsi, les modèles proposés dans ce mémoire ne prennent pas en compte les différences individuelles qui pourraient affecter la probabilité d'activation d'une application mobile. Il serait donc intéressant d'étendre cette étude aux modèles dits hétérogènes afin de mieux modéliser les comportements généralement complexes des utilisateurs de téléphones intelligents.

RÉFÉRENCES

Chapman, D. G. (1951). Some properties of the hypergeometric distribution with applications to zoological sample censuses. *Univ. Cal. Publ. Stat.*, *7*, 131–160.

Darroch, J. N. (1958). The multiple recapture census I : Estimation of a closed population. *Biometrika*, *45*, 343–359.

Gallo, M. (2015). Openrtb api specification. Récupéré de `http://www.iab.com/wp-content/uploads/2016/01/OpenRTB-API-Specification-Version-2-4-DRAFT.pdf`

Jolly, G. M. (1965). Explicit estimates from capture-recapture data with both death and immigration-stochastic model. *Biometrika*, *52*, 225–247.

Lincoln, F. C. (1930). Calculating waterfowl abundance on the basis of banding returns. *United States Department of Agriculture Circular*, *118*, 1–4.

Otis, D. L., Burnham, K. P., White, G. C. et Anderson, D. R. (1978). *Statistical Inference from Capture Data on Closed Animal Populations*, volume 62 de *Wildlife Monographs*. Wildlife Society.

Rivest, L.-P. et Baillargeon, S. (2013). Capture-recapture methods for estimating the size of a population : Dealing with variable capture probabilities. *Statistics in Action*, *54*, 289–303.

Sanathanan, L. (1972). Estimating the size of a multinomial population. *Ann. Math. Stat.*, *43*, 142–152.

Seber, G. A. F. (1965). A note on the multiple-recapture census. *Biometrika*, *52*, 249–259.

Seber, G. A. F. (1982). *The Estimation of Animal Abundance and Related Parameters* (2nd éd.). New York : Macmillan.

Yauck, M. et Rivest, L.-P. (2019). On the estimation of population sizes in capture–recapture experiments. *Journal of Multivariate Analysis*, *173*, 512–524.

Yauck, M., Rivest, L.-P. et Rothman, G. (2019). Capture-recapture methods for data on the activation of applications on mobile phones. *Journal of the American Statistical Association*, *114*(525), 105–114.